doi:10.3772/j.issn.2095-915x.2019.01.009

基于依存分析的军事领域英文实体关系抽取研究

开放科学（资源服务）标识码（OSID）

李芊芊 张克亮

战略支援部队信息工程大学 洛阳 471003

摘要：在海量信息的分析、处理中，实体关系抽取技术发挥着至关重要的作用。由于英文实体关系抽取存在信息量少和关系表述内聚性低的问题，针对英语军事文本，提出一种改进的基于依存分析的英文实体关系抽取方法。首先根据依存句法分析识别句子中的谓词，然后提取该谓词支配的所有依存关系，接着对上述依存关系进行二级扩展，得到完整的主语、宾语、介词短语，同时对状语和补语从句进行二级扩展，得到从句中的谓词并再次进行二级扩展，最后根据三元组的构成模式输出关系三元组。在英语军事领域语料上的实验表明，提出的方法具有较高的准确率和召回率，与斯坦福 coreNLP3.6.0 关系抽取方法相比，F 值提高了 57.71%。

关键词：关系抽取；依存分析；英文实体关系抽取；军事文本

中图分类号：TP319 G35

Entity Relation Extraction Based on Dependency Parsing in Military Field

LI Qianqian ZHANG Keliang

China PLA Information Engineering University at Luoyang, Henan 471003, China

Abstract: Entity relation extraction plays an important role in the analysis and processing of mass information. To reduce incoherent and uninformative extractions, an improved method of English entity relation extraction based on dependency parsing is proposed. The predicate of the sentence is firstly recognized according to the results of dependency parsing. Then, all dependency relations of the predicate are further extended to output complete subject, object and prepositional phrase.

基金项目：国家自然科学基金项目“多语种言语识别项目”(11590771)。

作者简介：李芊芊（1995-），研究生，研究方向：语言信息处理、知识工程，E-mail：619075219@qq.com；张克亮 (1964-)，博士，教授，博士生导师，研究方向：计算语言学、语言信息处理、知识工程等。

Meanwhile, the adverbial clause and complement clause is further extended to find and extend the predicate in the clause. Finally, the relation triples are output according to the pattern of triples. Experiments on the military corpus show the proposed method has a higher precision rate and recall rate, and the F measure is 57.71% higher than Stanford coreNLP3.6.0 method.

Keywords: Relation extraction; dependency parsing; English entity relation extraction; military texts

概述

大数据时代，海量的信息对人们知识获取提出了挑战，如何从海量信息中快速准确获取我们需要的知识成了一个重要的研究课题，从大规模、非结构化文本中自动抽取知识，也一直是人工智能的主要目标[1]。在海量信息的分析、处理和生成中，信息抽取技术（Information Extraction）发挥着至关重要的作用。信息抽取是从一段文本中抽取特定的信息，形成结构化的数据并存入一个数据库供用户查询和使用。信息抽取包含三个关键技术：实体抽取、关系抽取和事件抽取。其中，实体抽取是关系抽取和事件抽取的基础，是从文本中识别出人名、地名、机构名、日期、数额等实体信息。

关系抽取（Relation Extraction）是信息抽取的关键技术之一。Alexander Schutz 等人[2]认为关系抽取是自动识别由一对概念和联系这对概念的关系构成的相关三元组。关系抽取技术突破了传统的必须经过人工阅读、理解的方式来获得语义关系的限制，取而代之的是语义关系的自动查找和抽取[3]。

本文提出一种面向英语军事文本的基于依存分析的英文实体关系抽取方法。由于英文军事文本具有平均句长较长、状语从句和补语从句频率较高等特点，本研究不采用 coreNLP3.6.0 方法中切割语句、简化语句的思路，提出了一种针对英文长文本的“二级扩展”关系抽取方法。本方法首先利用斯坦福大学研发的 coreNLP 工具进行依存分析（Dependency Parsing），提取谓词支配的所有依存关系，并对主语、宾语和介词短语进行二级扩展，得到完整的主语、宾语和介词短语，同时对状语和补语从句进行二级扩展，得到从句中的谓词并再次进行二级扩展，最后按照三元组的构成模式输出关系三元组。在英语军事领域语料上的实验表明，提出的方法具有较高的准确率和召回率，与斯坦福 coreNLP3.6.0 关系抽取方法相比，F 值提高了 57.92%。

1 相关工作

关系抽取的关键技术有：基于模式匹配[4]的关系抽取、基于词典驱动[5]的关系抽取、基于机器学习的关系抽取。其中，基于机器学习的关系抽取方法包括基于特征向量[6]或核函数[7-8]的有监督机器学习、半监督[9]机器学习、无监督[10]机器学习、基于远程监督[11-12]的机器学习、基于深度学习[13-14]的机器学习和开放式关系抽取技术。

其中，开放式关系抽取旨在从开放的网络文本中自动发现非限定类型的语义关系实例集

doi:10.3772/j.issn.2095-915x.2019.01.009

合。针对英文的开放式关系抽取研究成果已较为丰富。Sekine[15] 尝试了按需抽取的思路，通过自动构造简单模板，完成非限定关系的关系抽取任务，其工作表明在开放文本环境下，利用浅层模式匹配的方法具有直观优势。Banko[16] 提出一种无监督的方法，用命名实体之间的动词作为关系短语，并实现了 TextRunner 信息抽取系统，该方法对于关系抽取的研究路线产生重要影响。

Fader[17] 提出三元组抽取的两个问题：（1）抽取的三元组信息含量少，忽略了关键信息，如“Faust made a deal with the devil.” 中抽取的信息为 (Faust, made, a deal)，而不是 (Faust, made a deal with, the devil)；（2）三元组的关系表述内聚性低，如“The Mark 14 was central to the torpedo scandal of the fleet.”中抽取的关系“was central torpedo”内聚性低，难以作为具有普适性的关系。

与 TextRunner 系统先识别实体再识别关系描述词不同，Fader 采用先识别关系描述词再识别实体的方法研发了 ReVerb 系统，由于词汇信息和句法信息对抽取关系表述进行限制，ReVerb 系统的性能优于 TextRunner 系统。ReVerb 系统的关系描述词需要满足 V|VP|VW*P 的句法模式，其中 V 包含动词和副词，W 包括名词、形容词、副词、代词和冠词，P 包括介词、小品词和标记词。识别出关系描述词后，ReVerb 系统抽取关系表述左右两边最近的两个名词短语作为论元，并组成三元组。ReVerb 系统由于采用了先识别关系描述词再识别实体的方法，因而关系表述词的内聚性较高，但是实体的识别过程中存在实体识别不准确的问题。

Mausam[18] 利用自举方法对 WOE 系统 [19] 进行提升，开发了 Ollie 系统。Ollie 系统利用自举方法在依存分析的基础上建立了覆盖范围较广的词汇化模式，尽管它的大部分模式是有效的，但是 Ollie 系统依旧存在关系表述内聚性低的问题，在不同领域文本中的适用性较差。

为了识别隐含关系并减少错误的关系抽取，Vo 等 [20] 在 ClausIE 系统 [21] 的基础上进一步改进基于从句框架的方法，它充分利用从句结构和从句类型抽取关系。对于每个从句，根据从句内部的语法组成结构来判断从句是否可以抽取为三元组关系。此外，Vo 等提出一种自训练算法来抽取指定类型的关系。实验表明该方法可以获得较高的准确率，提高了三元组的信息含量，而且抽取的关系表述也更为灵活，内聚性较高。

在 2013 TAC KBP 的填空任务中，斯坦福 coreNLP3.6.0 关系抽取系统 [22] 是表现最好的开放式关系抽取系统。该方法不再利用大量的规则从文本中抽取含有关键信息的三元组，而是先处理语句从中抽取内聚性高的从句（entailed clause），并利用简化后的从句生成关系三元组。

斯坦福 coreNLP3.6.0 关系抽取方法主要包括三个步骤：

（1）首先从训练语料中学习到一个分类器，将一个句子切割成较短的语句，切割出的从句需要句法合理、语义完整，而且继承原语句的限定约束。Gabor 等将该问题进一步简化为搜索问题，在依存句法分析的基础上，利用生成（Yield）、递归（Recurse）和停止（Stop）三种操作遍历依存句法树，并利用噪音语料训练从句生成模型，最后利用一些特征如依存关系

类型、依存弧父节点的依存关系类型、父节点的词性等特征训练多项式逻辑回归分类器；

（2）然后利用自然逻辑将上述切割出的语句尽可能地简化，同时保留语句中最必需的成分。自然逻辑理论为语句中的词汇变化提供了理论基础，将依存关系弧分为两类，一种可以删除依存弧的“状态元”（被支配者），一种不能删除依存弧的“状态元”，如“cute rabbit”可以简化为“rabbit”，“Jack runs”不能简化为“runs”；

（3）最后利用 14 条手工构建的规则将简化后的语句组合成关系三元组。

斯坦福 coreNLP3.6.0 关系抽取方法参见以下示例（表 1）：

表 1　斯坦福 **coreNLP3.6.0** 关系抽取方法示例

例句：Born in a small town, she took the midnight train going anywhere.		
	步骤 1：从左向右，生成从句	
步骤 2：从上向下，简化从句	she took the midnight train going anywhere	she took the midnight train
	Born in a small town, she took the midnight train	she took midnight train
	Born in a town, she took the midnight train	
步骤 3：生成三元组：（she, took, midnight train）		

通过对斯坦福 coreNLP3.6.0 关系抽取方法的实验发现，该方法的信息含量较为全面，但是关系表述内聚性不够高，可读性不高。

针对上述问题，同时为了解决军事英文文本中的关系抽取问题，本研究借鉴斯坦福 coreNLP3.6.0 关系抽取思路，并进行相应改进，设计了基于依存分析的军事领域英文实体关系抽取方法。

2　基于依存分析的实体关系抽取思路

2.1　依存分析

依存语法 (Dependency Grammar) 又称从属关系语法、配价语法，最早是 1959 年法国语言学家特斯尼耶尔 (Lucien Tesniére) 在《结构句法基础》(Element de Syntaxe Structurale) 中提出的。特斯尼耶尔未对依存语法下一个正面定义，目前比较通用的定义是周国光给配价语法下的定义：它主要研究以谓词为中心而构句时由深层语义结构映现为表层句法结构的状况及条件，谓词与体词之间的同现关系，并据此划分谓词的词类 [23]。

依存语法理论最基本的概念是“关联”、“结合”和“转位”[24]。该理论认为，关联、结合和转位是概括一切结构句法现象的三大核心。特斯尼耶尔使用图示 (stemma) 来表示依存关系，直接受动词支配的名词词组形成“行动元”(actant)，直接受动词支配的副词词组形成“状态元”(circonstant)。“行动元”，即某种方式或者某种名称的事或物，它是通过非常简单的名称或者消极的方式来参与过程，即传统语法中的主语、宾语、间接宾语（补语）。“状态元”则相当于传统语法中的补语。

doi:10.3772/j.issn.2095-915x.2019.01.009

1960 年，美国语言学家海斯 (Hays) 提出“依存分析法”。1970 年，美国语言学家罗宾孙 (Robinson) 在《依存结构和转换规则》中提出四条公理，为依存语法的形式化描述及在计算语言学中的应用奠定了基础，这四条公理是：（1）一个句子只有一个成分是独立的；（2）其他成分直接依存于某一成分；（3）任何一个成分都不能依存于两个或两个以上的成分；（4）如果 A 成分直接依存于 B 成分，而 C 成分在句子中位于 A 和 B 之间，那么 C 或者直接依存于 A，或者直接依存于 B，或者直接依存于 A 和 B 之间的某一成分。

英文中，句子成分间相互依存和被依存的现象普遍存在。本文利用斯坦福大学研发的 coreNLP3.6.0[25] 依存分析模块进行依存关系分析。该模块主要分析的依存关系有 50 多种，包括 acomp（动词的形容词补语）、advcl（动词的状语从句）、amod（名词短语的形容修饰语）、nsubj（动词的名词性主语）、dobj（直接宾语）、nmod（复合复词修饰）等。

例 1：*Airmen from the 36th Maintenance Group participated in a 12 hourlong munitions loading exercise July 13, 2016 at Andersen Air Force Base, Guam.*

coreNLP3.6.0 对例 1 进行依存分析，分析部分结果参见以下示例（表 2）：

表 2 例 1 的依存分析结果

序号	依存关系	序号	依存关系
1	root(ROOT-0, participated-7)	7	nmod:from(Airmen-1, Group-6)
2	nsubj(participated-7, Airmen-1)	8	case(munitions-12, in-8)
3	case(Group-6, from-2)	9	det(munitions-12, a-9)
4	det(Group-6, the-3)	10	nummod(munitions-12, 12-10)
5	amod(Group-6, 36th-4)	11	amod(munitions-12, hourlong-11)
6	compound(Group-6, Maintenance-5)		

根据斯坦福 coreNLP 的分析结果，可以得出该句的支配动词为“participated”。每条依存关系分为三个部分，依存关系、支配词和被支配词，如“nsubj(participated-7, Airmen-1)”中，“nsubj”表示名词性主语关系，“participated-7”表示支配词，“Airmen-1”表示被支配词，“7”和“1”表示该单词在句中的序号位置。

2.2 基于依存分析的关系抽取

基于依存分析的关系抽取流程图如下，（1）首先根据依存句法分析识别句子中的谓词即支配词；（2）然后提取该谓词支配的所有依存关系，包括主语、宾语、介词短语、状语和补语等；（3）对主语、宾语和介词短语进行二级扩展，得到完整的主语、宾语和介词短语；（4）对状语和补语从句进行二级扩展，得到从句中的谓词，返回步骤（2）；（5）按照三元组的构成模式输出关系三元组。图 1 分别对各个步骤进行介绍：

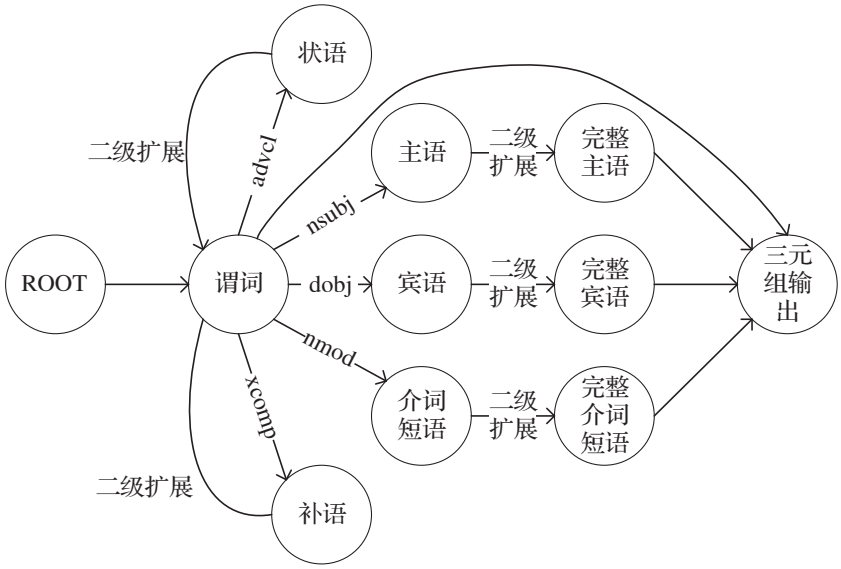

图 **1** 基于依存分析的关系抽取流程图

2.2.1 谓词的主要依存关系

根据斯坦福大学依存关系 [26] 得出谓词相关的依存关系解释参见表 3：

表 **3** 谓词相关的依存关系解释

依存关系	解释	例句
advcl：as/while/to/after/by/in/if	修饰动词的状语从句	The 36th Wing concluded exercise Beverly Palm, earning a grade of… advcl(concluded, earning)
advmod	副词修饰	Airmen and sailors worked together. advmod(worked, together)
aux	助动词	They are taking part in a challenging exercise. aux(taking, are)
auxpass	过去式的助动词	Kennedy has been killed. auxpass(killed, been)
ccomp	补语从句	He says that you like to siwm. ccomp(says, like)
compound:prt	复合	A Lancer takes off during Exercise Cope North 2017 compound:prt(takes,off)
dobj	直接宾语	…teach more than 30 Airmen and Soldiers… dobj(teach, Soldiers)
nsubj	名词性主语	instructors travelled to Guam to teach Airmen… nsubj(travelled, instructors)
nsubjpass	被动名词主语	Andersen was tested on their ability to receive… nsubjpass(tested, Andersen)
nmod:tmod	时间修饰语	Airmen and sailors worked together Aug.12. nmod:tmod(worked, Aug.)
nmod:in/from/through/to/on/at/during	介词短语	Airmen are taking part in the final exercise. nmod:in(take, exercise) Airmen delivered 237,000 pounds of fuel to B-52 Stratofortresses from the Squadron. nmod:from(delivered, Squadron)
xcomp	开放从句补语	Lancers are scheduled to deploy to Andersen. xcomp(scheduled, deploy)
xsubj	控制主语（开放从句动词→实际控制对象）	Tom likes to eat fish. xsubj(eat, Tom)

doi:10.3772/j.issn.2095-915x.2019.01.009

针对关系抽取任务，上表中的绝大部分依存关系是关系抽取任务中值得提取的部分：nsubj 作为依存关系的主语，dobj 作为直接宾语，nmod 主要表示介词短语与谓词之间的关系，包括介词 at、on、to、poss、during、in、of、from、with、for 等。

上表中一些依存关系对于关系抽取的意义不大：aux 作为助动词，advmod 表示副词修饰谓词的依存关系，nsubjpass 和 auxpass 表示被动语句中的依存关系，不具备关系抽取的主语。

此外，还有一些需要进一步处理的依存关系：xcomp 为开放补语从句（open clausal complement），表示谓词和从句中表示语义的词之间的关系；advcl 为修饰动词的状语从句表示谓词和从句中主要词的依存关系，包括 as、while、after 等引导的状语从句；compound:prt 表示与谓词构成的紧密成分（词组等），如 compound:prt(takes, off)、compound:prt(takes, on) 等。

2.2.2 从句的谓词抽取

基于依存分析的关系抽取任务的关系表达主要依靠输入语句的主句谓词，这种抽取方式的准确性较高，避免了 Fader 提出的内聚性低的关系表述问题，但是仅仅依靠谓词表达关系限制了关系抽取的召回率，因而需要从其他角度挖掘除主句谓词外的关系表达，提高召回率。

主要从如下状语从句、现在分词作状语、开放式补语从句三个方面进行分析：

（1）状语从句（advcl: 先行词）

状语从句包括时间状语、原因状语、地点状语、目的状语、结果状语、条件状语等，是提取关系三元组中不可缺少的一部分。

例 2：*A B-52 Stratofortress aircrew provides surveillance while U.S. Navy Sailors launch a harpoon missile.*

例 2 中 while 引导的时间状语从句可以提取三元组 (U.S. Navy Sailors, launch, a harpoon missile)。状语从句中的动词和主句谓词的依存关系表示为：advcl: 状语从句先行词 (主句谓词，从句动词)，其中状语从句先行词包括 while、as、after、if、by 等。

状语从句中谓词的主语和主句中的主语并无关联性和一致性，因而在本研究中假定：如果依存关系中有状语从句的谓词对应的主语关系（如例 2 中的 nsubj(launch, Sailors)），则以该主语为状语从句谓词对应的主语；如果依存关系中没有状语从句的谓词对应的主语关系，则以主句主语作为状语从句的谓词对应的主语。

（2）现在分词作状语（advcl）

现在分词和主句谓词的依存关系表示为：advcl(现在分词，主句谓词)。现在分词作状语的情况与状语从句类似，其区别在于现在分词作状语的逻辑主语与主语主语一致。

例 3：*The 36th Wing concluded exercise Beverly Palm* 13-7 *last week, earning a grade of Effective.*

例 3 中“earning”的主语和“concluded”的主语一致，皆为“The 36th Wing”。该情况下的三元组提取模板如下(主句主语，现在分词，现在分词对应的宾语或状语），在例 3 中为 (The 36th Wing, earning, a grade of Effective)。

（3）开放式补语从句 xcomp（open clausal complement）

开放式补语从句借用了词汇功能语法的概

念，即没有主语的补语从句或谓语，其依存关系表示为 xcomp(主句谓词，补语从句谓词)。该从句的主语通常由更高一层的从句主语或宾语决定，依存关系表示为 nsubj:xsubj(补语从句谓词，补语从句谓词的主语)。

例 4：*Andersen Air Force Base is scheduled to participate in Citadel Pacific.*

例 4 的依存关系有 root(ROOT, scheduled)、xcomp(scheduled, participate)、nsubj:xsubj(participate, Base)，从这些依存关系可以分析该句的谓词是“scheduled”，开放式状语从句谓词是“participate”，补语从句谓词的主语是“Base”。

2.2.3 论元和从句的二级扩展

在依存关系中可以作为论元的有 nsubj(名词性主语)、xsubj(开放式补语从句的实际主语)、dobj(直接宾语) 和 nmod(介词短语)，因为依存关系体现的是两个单词之间的关联关系，需要对单词进行二级扩展，才能得到所需的关系三元组。

（1）nsubj 名词性主语、dobj 直接宾语扩展方法

例 5：*Team Andersen Airmen flew to the land to demonstrate the capability and flexibility.*

对例 5 进行依存分析得到表 4：

例 5 中的名词性主语依存关系为 nsubj(flew, Airmen)，需要对单词“Airmen”进行二级扩展，找到它支配的其他单词，如依存关系 compound(Airmen-3, Team-1) 和 compound(Airmen-3, Andersen-2)，将这些支配按照它们在句中的顺序拼接起来，得到“Team Andersen Airmen”。同理得出直接宾语 dobj(demonstrate, capability) 的二级扩展结果为“the capability and flexibility”。

表 **4** 例 **5** 的依存分析结果

序号	依存关系	序号	依存关系
1	root(ROOT-0, flew-4)	9	mark(demonstrate-9, to-8)
2	compound(Airmen-3, Team-1)	10	xcomp(flew-4, demonstrate-9)
3	compound(Airmen-3, Andersen-2)	11	det(capability-11, the-10)
4	nsubj(flew-4, Airmen-3)	12	dobj(demonstrate-9, capability-11)
5	nsubj:xsubj(demonstrate-15, Airmen-3)	13	cc(capability-11, and-12)
6	nmod:to(flew-4, land-7)	14	dobj(demonstrate-9, flexibility-13)
7	case(land-7, to-5)	15	conj:and(capability-11, flexibility-13)
8	det(land-7, the-6)		

（2）nmod 介词短语扩展方法

nmod 介词短语的扩展方法与主语宾语的扩展方法相同，区别在于 nmod 介词短语的种类较多，包括 in、at、from、through、to 等，将介词信息附加在谓词上可以得到更加精确的三元组关系。如例 5 将依存关系“nmod:to(flew-4, land-7)”中的介词“to”附加在谓词上得到关系“flew to”，最终得到关系三元组为 (Team

doi:10.3772/j.issn.2095-915x.2019.01.009

Andersen Airmen, flew to, the land)。

（3）advcl 状语从句和 xcomp 开放式补语从句扩展方法

在识别状语从句 advcl 和开放式补语从句 xcomp 的过程中，同样需要应用二级扩展方法。根据例 5 中的依存关系得出该句中的开放式补语从句谓词为“demonstrate”，对从句进行二级扩展，根据“nsubj:xsubj(demonstrate-15, Airmen-3)”得出从句主语为“Airmen”，对“Airmen”继续进行二级扩展，得出“Team Andersen Airmen”，最终生成的三元组为 (Team Andersen Airmen, demonstrate, the capability and flexibility)。

2.2.4 关系三元组构成

根据依存分析，三元组构成有以下几种模式：

（1）名词性主语 nsubj+ 谓词 + 直接宾语 dobj；

例 6：*Four B-2 Spirit pilots assigned to the 13th Expeditionary Bomb Squadron completed a 24-hour training mission after flying a 10,000-mile roundtrip flight to Alaska as part of the Continuous Bomber Presence here.*

例 6 的依存分析结果参见表 5：

表 **5** 例 **6** 的部分依存分析结果

依存关系	依存关系
root(ROOT-0, completed-12)	det(mission-16, a-13)
nummod(pilots-4, Four-1)	amod(mission-16, 24-hour-14)
compound(pilots-4, B-2-2)	compound(mission-16, training-15)
compound(pilots-4, Spirit-3)	dobj(completed-12, mission-16)
nsubj(completed-12, pilots-4)	det(mission-16, a-13)

根据该模式，抽取出的关系三元组为（Four B-2 Spirit pilots，completed，a 24-hour training mission）。

（2）名词性主语 nsubj+ 谓词 +nmod 介词短语；

例 7：*B-52 aircrews from the 20th Expeditionary Bomb Squadron flew to Hawaii last week for an opportunity to sharpen their war fighting skills in the latest Koa Lightning exercise.*

例 7 的依存分析结果参见表 6：

表 **6** 例 **7** 的部分依存分析结果

依存关系
root(ROOT-0, flew-9)
compound(aircrews-2, B-52-1)
nsubj(flew-9, aircrews-2)
nmod:to(flew-9, Hawaii-11)

根据该模式，抽取出的关系三元组为（B-52 aircrews，flew to，Hawaii）。

（3）状语从句主语 nsubj+ 状语从句谓词 + 状语从句宾语 dobj；

例 8：*A B-52 Stratofortress aircrew provides surveillance while U.S. Navy Sailors launch a harpoon missile.*

例 8 的依存分析结果参见表 7：

表 **7** 例 **8** 的部分依存分析结果

依存关系	依存关系
root(ROOT-0, provides-5)	nsubj(launch-11, Sailors-10)
mark(launch-11, while-7)	advcl:while(provides-5, launch-11)
compound(Sailors-10, U.S.-8)	det(missile-14, a-12)
compound(Sailors-10, Navy-9)	

根据该模式，抽取出的关系三元组为（U.S. Navy Sailors，launch，a harpoon missile）。

（4）状语从句主语 nsubj+ 状语从句谓词 + 状语从句 nmod 介词短语；

该模式与模式 3 相似，不再举例解释。

（5）主句主语 nsubj+ 现在分词（作状语）advcl+ 状语从句的宾语 dobj；

例 9：*The 36th Wing concluded exercise Beverly Palm* 13-7 *last week, earning a grade of Effective.*

例 9 的依存分析结果参见表 8：

表 **8** 例 **9** 的部分依存分析结果

依存关系	依存关系
root(ROOT-0, concluded-4)	det(grade-14, a-13)
det(Wing-3, The-1)	dobj(earning-12, grade-14)
amod(Wing-3, 36th-2)	case(Effective-17, of-15)
nsubj(concluded-4, Wing-3)	nmod:of(grade-14, Effective-17)
advcl(concluded-4, earning-12)	

根据该模式，抽取出的关系三元组为（The 36th Wing，earning，a grade of Effective）。

（6）主句主语 nsubj+ 现在分词（作状语）advcl+ 状语从句的 nmod 介词短语；

该模式与模式 6 相似，不再举例解释。

（7）开放式补语从句的实际主语 xsubj+ 开放式补语从句谓词 + 开放式补语从句的直接宾语 dobj；

例 10：*Team Andersen Airmen flew to the land ‘down under’ this week to demonstrate the capability and flexibility of the B-52 Stratofortress bomber to our Australian partners.*

例 10 的依存分析结果参见表 9：

表 **9** 例 **10** 的部分依存分析结果

依存关系	依存关系
root(ROOT-0, flew-4)	mark(demonstrate-15, to-14)
compound(Airmen-3, Team-1)	xcomp(flew-4, demonstrate-15)
compound(Airmen-3, Andersen-2)	det(capability-17, the-16)
nsubj(flew-4, Airmen-3)	dobj(demonstrate-15, capability-17)
nsubj:xsubj(demonstrate-15, Airmen-3)	cc(capability-17, and-18)
conj:and(capability-17, flexibility-19)	dobj(demonstrate-15, flexibility-19)

根据该模式，抽取出的关系三元组为（Team Andersen Airmen，demonstrate，the capability and flexibility）。

（8）开放式补语从句的实际主语 xsubj+ 开放式补语从句谓词 + 开放式补语从句的 nmod 介词短语；

该模式与模式 7 相似，不再举例解释。

（9）固定表达形成的三元组。

上述模式部分解决了句子中的多谓词问题，但是在实际的关系抽取任务中，往往会出现多个名词性主语、多个 nmod 介词短语、多个宾语的情况，本研究的解决策略是固定谓词，其他成分进行全排列。如从例句中提取主语有“Andersen Airmen”和“Parachuters”，谓词为“finished”，宾语为“Exercise Cope North 2017”，对上述论元进行全排列得到三元组 (Andersen Airmen, finished, Exercise Cope North 2017) 和三元组 (Parachuters, finished, Exercise Cope North 2017)。

doi:10.3772/j.issn.2095-915x.2019.01.009

3 实验与结果分析

3.1 实验设计

虽然关系抽取任务中存在标准数据集，如 SemEval-2010 Task 8，但是该数据集主要针对 Cause-Effect、Instrument-Agency、Member-Collection、Message-Topic、Product-Producer、Content-Container、Entity-Origin、Entity-Destination、Component-Whole 等关系，这些关系表述并不符合军事领域的需求。开放式关系抽取中的 TAC KBP 任务没有针对军事领域的语料，coreNLP3.6.0 方法在 2013 TAC KBP 中表现优异，却在军事领域语料中表现一般，标准数据集的评测结果不能较好地反映在军事领域中的实际应用效果。因而，针对军事领域的英文实体关系抽取任务，需要构建军事领域语料进行测试。

为了测试本文实体关系抽取方法在军事领域文本中的效果，为提高实验数据的代表性，本研究从美国陆军官网、美国海军官网、美国陆战队官网、美国安德森空军基地官网、美国嘉手纳空军基地官网等军事网站爬取了 6100 篇英文文本，共 144112 句，并随机抽取其中的 1000 句作为实验数据。经研究分析，该类报道通常在文章的导语部分介绍军事活动的主要内容、参与单位、活动时间和活动地点，如“Airmen from the 36th Maintenance Group participated in a 12 hourlong munitions loading exercise July 13, 2016 at Andersen Air Force Base, Guam.”。文章的其他部分通常描述军事活动的细节和军事人员的感受。因而，为提高抽取效果，从每篇报道中抽取导语构成美国军事基地军事报道语料库，作为本实验的实验语料。

基于依存分析的英文实体关系抽取首先借助斯坦福大学的 coreNLP 工具对军事领域文本进行依存分析，coreNLP 版本选择 3.6.0，然后采取本研究提出的关系抽取方法进行关系抽取，并与斯坦福 coreNLP3.6.0 关系抽取方法进行比较分析。

由于自建的军事领域语料库不是标准数据集，关系抽取结果没有指定的标准的答案，判断抽取出的关系表述正确与否采用专家判断加上抽样验证的方法。本文判断关系表述正误的标准为：（1）关系表述包含的信息量的多少；（2）关系表述内聚性的高低。如果一个关系表述包含原语句中的部分信息（部分但不矛盾），而且关系表述内聚性较高、具有可懂性，那么这个关系表述是正确的。为了确保关系表述的准确性，在专家判断后，会抽取一部分判断结果由另一位专家进行抽样验证。

在实验中，采用准确率 P、召回率 R 和 F 值进行评价，实验结果分别从上述三个角度进行评价，计算公式如下：

$$P=\frac{\text{抽取出的正确关系表述数量}}{\text{抽取出的关系表述数量}} \quad (1)$$

$$R=\frac{\text{抽取出的正确关系表述数量}}{\text{文档集合中实际含有的关系表述数量}} \quad (2)$$

$$F=\frac{P\times R\times 2}{P+R} \quad (3)$$

3.2 实验及结果分析

实验结果参见表 10：

表 **10** 对比实验结果

测试方法	语句数	P 准确率	R 召回率	F 值
基于依存分析的关系抽取	1000	1771/2143=0.826411	1771/2235=0.792393	0.809044
斯坦福 coreNLP3.6.0 关系抽取	1000	1222/8304=0.147157	1222/2235=0.546756	0.231899

实验结果说明，在没有引入领域知识、只依靠依存分析的前提下，本文的关系抽取方法比 coreNLP 方法，准确率提高了 67.92%，召回率提高了 24.56%，F 值提高了 57.71%。

这主要是因为美国军事基地军事报道语料库中的平均句长为 29.9，coreNLP 的关系抽取方法的语句切分粒度过细，生成的备选三元组过多，在英文长句的关系抽取中不占优势，较大地影响了系统的准确度，而且 coreNLP 方法的关系表述内聚性较低，可懂性较低。

例 11：*Engineers from the Royal Australian Air Force, Republic of Singapore Air Force, Republic of Korea Air Force and the Japan Air Self-Defense Force began the first multilateral partner-nation Silver Flag exercise Feb. 13, at Andersen Air Force Base, Guam.*

下面分别利用 coreNLP 和本文依存方法对例 11 的处理结果进行比照说明，关系抽取结果分别为表 5 和表 6，参见表 11、12：

表 **11** **coreNLP** 对例 **11** 的关系抽取结果

论元 1	关系	论元 2
10.0engineer	begin	partner-nation silver flag exercise
10.0engineer	begin	multilateral partner-nation silver flag exercise
10.0engineer	begin	first multilateral parter-nation silver flag exercise
10.0Air Force	be of	Republic of Korea
10.0engineer	begin partner-nation silver flag exercise at	Andersen Air Force Base
10.0Royal Australian Air Force	republic of	Singapore Air Force
10.0engineer	begin partner-nation silver flag exercise at	Guam
10.0engineer	begin partner-nation silver flag exercise at_time	Feb. 13
10.0engineer	begin	first partner-nation silver flag exercise

表 **12** 本文依存方法对例 **11** 的关系抽取结果

论元 1	关系	论元 2
Engineers from the Royal Australian Air Force Republic of Singapore Air Force Republic of Korea Air Force and the Japan Air Self-Defense Force	begin	the first multilateral partner-nation Silver Flag exercise
Engineers from the Royal Australian Air Force Republic of Singapore Air Force Republic of Korea Air Force and the Japan Air Self-Defense Force	begin at	Feb. 13
Engineers from the Royal Australian Air Force Republic of Singapore Air Force Republic of Korea Air Force and the Japan Air Self-Defense Force	begin at	Andersen Air Force Base Guam

doi:10.3772/j.issn.2095-915x.2019.01.009

针对例 11，coreNLP 关系抽取的准确率为 1/9=11.11%，召回率为 1/3=33.33%；本文依存方法的准确率为 3/3=100%，召回率为 3/3=100%。从上例可以发现，coreNLP 方法的关系表述内聚性低，关系三元组的信息量较低。

尽管基于依存分析的关系抽取方法比斯坦福 coreNLP3.6.0 三元组抽取的性能指标有了大幅度的提升，但仍然存在着如下问题：

（1）涉及动词词组的谓词提取不完整。如“Intense training took place April 18-21 at Andersen Air Force Base.”中提取的谓词为 took，而应该为 took place；

（2）依存关系分析错误。如“Paratroopers assigned to 3rd Battalion, conduct M67 fragmentation grenades live-fire training at Kraft Range Dec. 12, 2017.”中提取的谓词为 assigned，而应该为 conduct；

（3）描述关系的谓词种类多。100 句演习活动报道中使用了 50 多种谓词，如“participate”、“conduct”、“complete”等，需要根据领域知识进一步规范；

（4）忽略了部分固定表达形成的领域关系。如利用固定表达“assigned to”可以迅速抽取三元组 (personnel, assigned to, unit)。

针对上述问题提出如下解决方法：问题（1）的解决相对容易，因为在美国军事基地军事报道语料库中涉及的动词词组类型较少，可以罗列如“take place、take off、take part in”等。问题（2）的解决需要依靠依存句法分析准确率的提高。问题（3）的解决只依靠谓词来定义三元组关系是不够的，需要进一步引入领域知识对关系进行规范，正如刘知远所说，利用依存语法识别表达语义关系的短语，来抽取实体间关系的方法简单有效。这种方法抽取的实体间关系丰富而且自由，一般是一个以动词为核心的短语。但是这种方法的关系语义无法规范化，同一种关系可能会有多种不同的表示，如何对这些自动发现的关系进行聚类规约是一个挑战性的问题 [27]。问题（4）的解决可以利用固定表达提取领域关系。在军事领域一些固定表达可以弥补依存关系的不足，提高关系抽取的效率，此外还可以利用固定表达和领域知识抽取关系，如已知安德森空军基地的下属单位“374 AMXS”，因而可以从“Senior Airman Austin Endsley, 374 AMXS crew chief.”抽取三元组 (Austin Endsley, assigned to, 374 AMXS)。

4 结束语

本文提出了一种面向英语军事文本的基于依存分析的英文实体关系抽取方法。实验结果说明，本方法能够在保证三元组信息量和关系表述内聚性的前提下从英文自由文本中抽取出实体关系，在军事领域语料的测试下，其准确率和召回率显著高于斯坦福 coreNLP 三元组抽取方法。

下一步将进一步引入领域知识对关系表示进行规范，并充分利用固定表述提高本方法在军事领域文本中的关系抽取效果。

参考文献

[1] Poon H, Domingos P. Unsupervised ontology induction from text[C]. Proceedings of the 48th annual meeting of the Association for Computational

Linguistics. Association for Computational Linguistics, 2010: 296-305.

[2] Schutz A, Buitelaar P. RelExt:a tool for relation extraction from text in ontology extension[C]. International Conference on the Semantic Web. Springer-Verlag, 2005: 593-606.

[3] 刘绍毓 , 李弼程 , 郭志刚 , 等 . 实体关系抽取研究综述 [J]. 信息工程大学学报 , 2016, 17(5): 541-547.

[4] Appelt D E, Hobbs J R, Bear J, et al. SRI International FASTUS system:MUC-6 test results and analysis[M]. DBLP, 1995.

[5] Aone C, Ramos-Santacruz M. REES:A Large-Scale Relation and Event Extraction System[J]. Proceedings of Anlpnaacl, 2000.

[6] 车万翔 , 刘挺 , 李生 . 实体关系自动抽取 [J]. 中文信息学报 , 2005, 19(2): 1-6.

[7] Bunescu R C, Mooney R J. A shortest path dependency kernel for relation extraction[C]. Conference on Human Language Technology and Empirical Methods in Natural Language Processing. Association for Computational Linguistics. 2005: 724-731.

[8] Panyam N C, Verspoor K, Cohn T, et al. Exploiting graph kernels for high performance biomedical relation extraction[J]. Journal of Biomedical Semantics, 2018, 9(1): 7.

[9] Etzioni O, Cafarella M, Downey D, et al. Unsupervised named-entity extraction from the Web: An experimental study[J]. Artificial Intelligence, 2005, 165(1): 91-134.

[10] Hasegawa T, Sekine S, Grishman R. Discovering relations among named entities from large corpora[C]. Meeting on Association for Computational Linguistics. Association for Computational Linguistics. 2004: 415.

[11] Hoffmann R, Zhang C, Ling X, et al. Knowledge-based weak supervision for information extraction of overlapping relations[C]. Meeting of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2011: 541-550.

[12] Qu J, Ouyang D, Hua W, et al. Distant supervision for neural relation extraction integrated with word attention and property features[J]. Neural Networks, 2018, 100.

[13] Zeng D, Liu K, Chen Y, et al. Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks[C]. Conference on Empirical Methods in Natural Language Processing. 2015: 1753-1762.

[14] Zheng S, Hao Y, Lu D, et al. Joint Entity and Relation Extraction Based on A Hybrid Neural Network[J]. Neurocomputing, 2017, 257(1): 1-8.

[15] Sekine S. On-Demand Information Extraction[C]. ACL 2006,International Conference on Computational Linguistics and Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Sydney, Australia, 17-21 July. DBLP, 2006.

[16] Banko M. Open Information Extraction for the Web [D]. Seattle: University of Washington, 2009.

[17] Fader A, Soderland S, Etzioni O. Identifying relations for open information extraction[C]. Proceedings of the conference on empirical methods in natural language processing. Association for Computational Linguistics, 2011: 1535-1545.

[18] Mausam, Schmitz M, Bart R, et al. Open language learning for information extraction[C]. Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Association for Computational Linguistics, 2012.

[19] Wu F, Weld D S. Open information extraction using Wikipedia[C]. 2010 Proceedings of ACL. 2010: 118-127.

[20] Vo D T, Bagheri E. Self-training on refined clause patterns for relation extraction[J]. Information Processing & Management, 2017: S0306457316303259.

[21] Corro L D, Gemulla R. ClausIE: clause-based

doi:10.3772/j.issn.2095-915x.2019.01.009

open information extraction[C]. 2013 International Conference on World Wide Web. 2013: 355-366.

[22] Angeli G, Premkumar M J J, Manning C D. Leveraging linguistic structure for open domain information extraction[C]. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. 2015: 344-354.

[23] 周国光 . 汉语配价语法论略 [J]. 南京师大学报 (社科版), 1994(4): 103-106.

[24] Tesniere, Lucien. Elements de syntax structural[M]. Paris: Klincksieck. 1976, 11-13.

[25] Manning C D, Surdeanu M, Bauer J, et al. The Stanford CoreNLP Natural Language Processing Toolkit[C]. Meeting of the Association for Computational Linguistics: System Demonstrations. 2014.

[26] Marnee M C D, Manning C D. Stanford typed dependencies manual[EB/OL]. (2008-09-01) [2018-01-13]. https://nlp.stanford.edu/software/dependencies_manual.pdf.

[27] 刘知远 , 崔安颀 . 大数据智能 [M]. 北京 : 电子工业出版社 , 2016.