doi: 10.3969/j.issn.1006-1576.2011.05.008

作战文书关键信息抽取方法

李畅[1]，王永良[2]，冯晓洁[3]，聂峰[1]
（1. 炮兵指挥学院 联合作战教研室，河北 宣化 075100；2. 炮兵指挥学院 军事运筹教研室，河北 宣化 075100；
3. 总参通信训练基地，河北 宣化 075100）

摘要：为提高特定领域作战文书关键信息抽取的召回率和精确率，提出一种作战文书关键信息的抽取方法。分析作战文书的记述特征，构建领域本体以明确抽取内容，而后利用领域词典标记文书文本，并采用信息抽取模式抽取关键信息。实验结果表明，该方法在实际项目中可以很好地完成抽取任务，能促进 C^4I 系统建设和作战模拟的发展。

关键词：作战文书；信息抽取；领域词典；领域本体；抽取模式
中图分类号：TP391　　**文献标志码**：A

Key Information Extraction from Operational Document

Li Chang[1], Wang Yongliang[2], Feng Xiaojie[3], Nie Feng[1]
(1. Combined-Operations Staff Room, Artillery Command Academy of PLA, Xuanhua 075100, China;
2. Staff Room of Military Operation Research, Artillery Command Academy of PLA, Xuanhua 075100, China;
3. General Staff Communications Training Base, Xuanhua 075100, China)

Abstract: To improve the recall rate and precision rate of key information extraction from operational document, an approach is presented. First, it analyzes the character of operational document and builds domain ontology to ascertain what should be extracted. Then, it tags the document using domain lexicon. At last, it extracts key information employing extraction pattern. The experiment results show that the approach can accelerate the development of C^4I and can be applied in practice.

Keywords: operational document; information extraction; domain lexicon; domain ontology; extraction pattern

0　引言

作战文书是军队各级机关在作战和其他军事行动中形成的具有法定效力和规范体式的各种电报、文件等信息载体的统称，是军用文书的重要组成部分。作战文书中的关键信息主要包括敌我双方的作战任务、兵力部署、武器装备编制、作战阶段的划分以及敌我态势演变等。实现计算机正确抽取作战文书的关键信息，是实现 C^4I 系统作战态势自动标绘的关键技术，是作战模拟中实现对计算机生成的虚拟兵力（computer generate force, CGF）实施常规指挥的关键，对促进 C^4I 系统建设和作战模拟的发展具有重要的现实意义。

从已公开发表的文献来看，目前针对作战文书关键信息抽取的研究并不多，在设计实现完善的信息抽取（information extraction）系统方面还处在探索阶段[1-2]。目前与信息抽取相关的理论研究主要集中在领域知识表达[3-5]、领域词典建立[6]，以及抽取模式的获得[7-8]等方面。而在实用系统研究中，于琨等[9]利用双层文本分类的方法实现了中文简历信息的自动抽取；周明建等[10]利用本体论实现了 Web 信息抽取；梁晗等[11]基于框架理论对灾难事件进行抽取。通过分析发现，这些针对特定领域的文本信息抽取系统存在的主要问题是使用单一的理论，信息抽取的召回率和精确率普遍不高。

因此，针对所要抽取的领域文本，对其特征进行深入分析，并广泛融合各种理论研究适合于抽取需求的方法，实现特定领域信息抽取系统的实用化。

1　作战文书特征分析

作战文书的种类多样，包括命令、指示、计划、通报等，其特点是具有高度的权威性、严格的准确性以及严密的规范性。作战文书对于其组成要素、汉字使用甚至印刷字号都有特殊而严格的规定。以作战文书中的“行军命令”为例，分析作战文书的记述特征。

特征 1　记述部队名称通常用番号，如“陆军第××集团军步兵××师装甲团”，也可以使用规范化

收稿日期：2011-01-25；修回日期：2011-03-07
基金项目：炮兵指挥学院科研创新基金项目（PY2010B0012）
作者简介：李畅（1980—），男，内蒙古人，硕士，讲师，从事作战模拟、数据挖掘研究。

的简称，如“××军××师装甲团”；

特征 2　记述地名，以作战文书所用地图上的名称为准，地名后面应当注明地图坐标，如“张庄（××，××）”。文书中的同一地名，前面已注明地图坐标的，后面可以省略。

特征 3　作战文书的标题表述方式较为固定，可以作为自身的属性区别于其他类作战文书。如标题“向×××地区机动/行军”表明该文书为“行军命令”。

特征 4　关键信息按照主题可以分为若干类。“行军命令”中较为常见的有:“敌情”、“行军部署”、“组织指挥”、“疏散地域区分”、“相关保障内容”、“行军要求”、“情况处置”等。这些信息在行文位置上彼此独立，一般以一个自然段表示一类信息。

特征 5　关键信息的记述方式相对固定。如“行军命令”中记述关键信息“调整点”的方式为“调整点在赵川（××，××）”，基本不会使用其他的记述方式。

特征 1、2 表明，作战文书关键信息的抽取方法不涉及未登录实体名称及地名的识别，但应建立对应的领域词典。特征 3、4 表明，作战文书具有多层分类结构，可利用类间层次将各类信息组织成树状结构，如图 1。特征 5 表明，待抽取的关键信息可以使用语义分析的方法进行抽取。

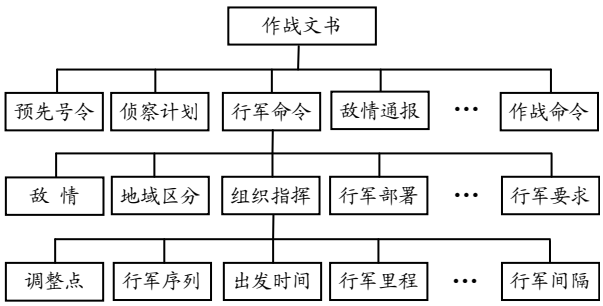

图 1　作战文书各类信息结构图

2　信息抽取模型

2.1　构建领域词典

与通用词汇不同，军事术语能更准确地表达作战文书中的信息，并且通常是一个不可拆分的信息单元。如“疏散隐蔽地域”有其特定的军事含义，若将其切分成几个单独的词：“疏散”+“隐蔽”+“地域”,则这些单独的词所表达的信息与短语自身信息截然不同。

笔者规定如下构词规则，对自然语言适当加以限制,从而显著降低复杂性和减少机器处理的困难。

1) 基本词。凡是《军语》中收录的领域词语，认为是基本词，统一标注为 M。

2) “兵种”+“编制名词”。如,“炮兵师”、“装甲团”。特别地，如果发现有连续的编制名词出现，或者是后面加括号“欠+编制名词”，都认为是一个编制词元，不可分割。如,“××军××师（欠三团）”,标记为 B。

3) 时间词。如,“××日××时××分××秒”，标记为 T。

4) 数量词。如,“30 km”、“20 min”，标记为 N。

5) 地点词。如,“（××，××）”，以坐标形式出现的格式，均认为是一个地点名词。特别地，如果地点名词后面还有方位词+数量词，或者在地点名词前还有数字地图地名字段中的词，则认为这是一个地点词元，不可分割。如,“张庄（××，××）以东 200 m”，标记为 L。

2.2　领域知识表示

作战文书种类众多，所包含的信息关系复杂，如何组织领域知识、提高软件的复用程度是一项重要而迫切的研究内容。利用作战文书领域 Ontology[12-14]作为信息抽取的外部知识。

2.2.1　领域 Ontology 定义

作战文书领域 Ontology 定义是对所要抽取信息的一个描述，声明了在信息抽取过程中所关心的信息项。处理作战文书的根本目标是要抽取出感兴趣的信息，这些信息需要被明确指出，才能被抽取过程所了解。“本体定义”正是完成了这个功能，在“本体定义”中定义了信息抽取的目标及信息类型，这样，在后续的抽取过程中，只需处理已经定义了的信息项，对于与目标无关的信息，抽取过程可以不做具体的分析。

定义 1　领域 Ontology 领域 Ontology 利用二元组(C, R)来表示，其中 C 表示领域中的概念，R 表示概念之间的关联，R 通过三元组$(C1, C2, r)$来表示，其中，$C1 \in C, C2 \in C$，r 表示 $C1$ 和 $C2$ 之间的关系名称。

定义 2　概念 C 属性集合 R 概念 C 属性集合表示为 $R=\{r\}$，集合 R 中的每个属性 r 通过二元组(Name[$C.r$],Type[$C.r$])来表示，其中 Name[$C. r$]表示概念 C 的属性 r 的名称集合，Type[$C. r$]表示概念 C 的属性 r 的信息类型集合。

2.2.2 领域 Ontology 表示

领域 Ontology 的具体表现形式是彼此联系的，带有不同属性的概念之间的知识框架树。利用 XML 来描述这种知识框架树，具体形式如例 1。

例 1 作战文书领域 Ontology 描述片断（以行军命令为例）

<Ontology> //作战文书领域本体（“行军命令”）
<Concept> //行军部署
<Attribute> //Key-Information
<Name> </ Name> //属性名称(“车间距”)
<Type> </ Type>
//属性的信息类型（“数量类型”）
</ Attribute>
<Attribute> // Key-Information
<Name> </ Name>
//属性名称(“小休息地点”)
<Type> </ Type>
//属性的信息类型（“地点类型”）
</ Attribute>
……//属性描述部分完成属性信息的描述
</ Concept>
……//概念描述部分，通过概念中属性来描述概念信息
</ Ontology>

2.3 信息抽取模式

通过对大量作战文书的分析发现，大多数信息只需在相邻文本中找到匹配的信息类型即可，只有少部分的信息需要进一步处理，下面分别加以讨论。

2.3.1 “关键词+信息类型”模式

不同于基于语法分析的信息抽取，“关键词+信息类型”模式使用的是信息类型分析的方法。这种方法的优点在于利用信息类型分析指导信息抽取，抽取过程中只考虑信息类型的识别，而不进行语句的语法分析，从而大大简化了抽取过程。其抽取过程如例 2。

例 2 待抽取的输入文本为：“出发点为兴仁堡（××，××），团先头××日××时××分准时通过出发线。调整点在赵庄（××，××）。行军途中在关底（××，××）小休息一次，时间 30 min。”

由行军命令本体定义可知，“出发点”、“出发线”、“调整点”、“小休息时间、地点”为抽取内容，其信息类型如表 1。L 为地点类型、T 为时间类型、N 为数量类型。

表 1 “行军命令”信息类型表（部分）

关键信息	信息类型 1	信息类型 2
出发点	L	
出发线	L	T
调整点	L	
小休息	N	L

从表 1 中可以看出，“出发线”有 2 个信息类型属性，其一为地点类型，其二为时间类型。这是由于在作战文书中，“出发线”虽然本意是指地点，但是大多数情况下都比较关心越过出发线的时间。

在抽取过程中，首先识别出要抽取的关键信息，如“出发点”。通过本体定义，可知其信息类型为地点类型 L，因此，以标点符号为分界点，找到被分词算法标记为 L 的相邻信息元“兴仁堡（××，××）”，即完成关键信息“出发点”的信息抽取。

2.3.2 事件类信息抽取模式

上节所介绍的方法虽然可以抽取作战文书中的大部分属性类信息，但是对于描述事件类的信息抽取则无能为力[15]，如对“行军序列”的抽取。因此，必须建立相应抽取模式。

事件类信息抽取算法描述如下。

1) 针对每一个句子，查看组成它的词语是否在“触发词—事件类别”对照表中，如果存在这样的词 w，则认为这个句子是一个候选事件。

2) 选取候选事件的词法、上下文、词典信息等三类语言学特征，采用二元分类的方法判断候选事件是否是真正的待抽取事件。

3) 待抽取事件类别的确定，相应地就获得了该类事件的模板，即获得了要抽取的元素标签，可将事件元素识别任务转换为对文本中每个候选元素进行类别标签识别的分类任务。

例如，从文书文本“团决定编成 5 个梯队，成一路纵队，按照一营、群直、二营、三营、后装分队的序列。”抽取事件类信息。首先由“编成”、“序列”等关键词，判断该句可能为“行军序列”的候选事件；然后由该事件的三类语言学特征判断其确为“行军序列”事件；最后，将文本中的每个候选元素按照事先定义的模板进行分类，即得到事件类关键信息。

3 实验与结果分析

3.1 语料介绍及评价方法

测试语料包括“行军命令”、“敌情通报”等 5

种共 268 篇作战文书。经统计，5 种文书共涉及各类关键信息 669 项。测试语料统计情况如表 2。

表 2　实验语料统计情况

文书种类	文书数量	L 型	N 型	T 型	M 型	事件 E 型
行军命令	69	37	16	22	41	25
敌情通报	52	33	19	17	38	19
作战命令	57	41	22	20	35	17
预先号令	41	38	17	19	39	23
侦察计划	49	36	20	17	43	15

实验采用传统的召回率和精确率作为基本测试指标，使用宏平均（macro-average）定义平均精确率和平均召回率（N 为类别总数，Precision_i 为第 i 类的精确率，Recall_i 为第 i 类的召回率。

$$\text{平均召回率}=\frac{\sum_{i=1}^{N}\text{Recall}}{N};$$

$$\text{平均精确率}=\frac{\sum_{i=1}^{N}\text{Precision}_i}{N}$$

使用精确匹配标准[16]判断抽取出的信息是否准确。在各型信息抽取中，若抽取出的信息与标准答案存在 90%以上的交集，且不相交部分低于 10%，认为抽取成功。

3.2　实验结果与分析

表 3 是实验结果的统计数据。该结果表明该方法在抽取作战文书关键信息中获得了 93.1%以上的平均召回率和 88.3%以上的平均精确率，可以满足作战文书关键信息抽取的实际工程需要。

表 3　实验结果统计数据

测试指标	L 型	N 型	T 型	M 型	事件 E 型
平均召回率/%	93.7	94.1	96.2	93.1	72.1
平均精确率/%	88.3	89.7	89.1	89.6	54.2

较高的平均召回率和平均精确率主要得益于作战文书行文的规范化及人工建立的高质量抽取模式。但实验结果表明事件 E 型关键信息的平均召回率和平均精确率较低。比如在实验中，“疏散地域区分”的平均召回率为 72.1%，平均精确率仅为 54.2%。究其原因在于，虽然作战文书相对具有用语规范、句式简单等特点，但基于知识工程人工制定的抽取模式依然不能很好地解决中文自然语言特有的灵活性与多变性。因此，将在今后的工作中，引入基于机器学习技术自动获取抽取模式的方法，以提高这类信息抽取的召回率和精确率。

4　结论

作战文书关键信息抽取方法主要采取使用领域 Ontology 表达知识、构建领域词典、编制模式抽取规则等步骤，避免了进行深层句法分析，降低了系统实现的难度，在作战文书关键信息抽取的应用中表现出一定成效。下一步，将针对实验中暴露出来的问题，进一步完善领域词典、完善模式抽取规则，并尝试使用浅层语法分析，进一步提高系统性能。

参考文献：

[1] 樊延平, 马亚龙, 袁野. 军事想定数据挖掘技术研究[J]. 系统仿真学报, 2006, 18(8): 172-174.

[2] 李跃进, 赵晶林, 鸿飞. 基于 Internet 的军事演习信息抽取系统[J]. 计算机工程与应用, 2006, 42(14): 214-218.

[3] 陈刚, 陆汝钤, 金芝. 基于领域知识重用的虚拟领域本体构造[J]. 软件学报, 2003, 14(3): 350-355.

[4] Zelenko D, Aone C, Riehardella A. Kemel methods form reslation extraction[J]. Journal of Machine Learning Research, 2003(3): 1080-1106.

[5] Fayyad U M, Piatetsky-Shapiro G. Advances in Knowledge Discovery and Data Mining [M]. Cambridge, MA, MIT Press, 1996.

[6] 陈文亮, 朱靖波, 朱慕华. 基于领域词典的文本特征表示[J]. 计算机研究与发展, 2005, 42(12): 2155-2160.

[7] 于江德, 王立新, 樊孝忠. 基于自扩展的信息抽取模式自动获取[J]. 小型微型计算机系统, 2009, 30(5): 891-894.

[8] 李伟, 黄颖. 基于 HtmlParser 的网页信息提取[J]. 兵工自动化, 2007, 26(7): 41-42.

[9] 于琨, 管刚, 周明. 基于双层级联文本分类的简历信息抽取[J]. 中文信息学报, 2006, 20: 59-66.

[10] 周明建, 高济, 李飞. 基于本体论的 Web 信息抽取[J]. 计算机辅助设计与图形学学报, 2004, 16(4): 535-541.

[11] 梁晗, 陈群秀, 吴平博. 基于事件框架的信息抽取系统[J]. 中文信息学报, 2006, 20(2): 40-46.

[12] 高军, 王腾蛟, 杨冬青, 唐世渭. 基于 Ontology 的 Web 内容二阶段半自动抽取方法[J]. 计算机学报, 2004, 27(3): 310-317.

[13] Agichtein E. Extracting relations from large yext collections[D]. Columbia University, 2005.

[14] 马静, 吴一占, 刘思峰. 基于领域本体的信息抽取模式生成与系统实现[J]. 情报学报, 2008, 27(2): 193-198.

[15] 周法国, 王映龙, 杨炳儒. 非结构化信息抽取关键技术研究探讨[J]. 计算机工程与应用, 2009, 45(14): 1-6.

[16] A. Fum, N. Kuslmerick. Multi-level Boundary Classification for Information Extraction [A]. ECML-2004[C]. 2004.