文章编号: 1672-1497(2012)06-0075-05

基于本体的受限领域文本信息提取方法

孙 剑, 周深根, 陈超敏
(装甲兵工程学院装备指挥与管理系, 北京 100072)

摘 要: 针对指挥信息化模拟训练系统中自由格式报文转换问题, 根据仿真驱动信息的需求, 讨论了对自由格式报文进行信息提取的实现方法。在受限领域(军事领域)中引入本体概念, 构建了领域本体模型及语法规则, 提出了基于本体的信息提取方法, 采用可扩展标记语言(eXtensible Markup Language, XML)和 VC++.NET 技术相结合, 实现对文书的信息提取, 为进一步的数据转换和应用奠定了基础。

关键词: 自由文本; 本体; 信息提取
中图分类号: E072; TP391.43 文献标志码: A

Information Extraction Method of Limited Area Based On Ontology

SUN Jian, ZHOU Shen-gen, CHEN Chao-min
(Department of Equipment Command and Administration, Academy of Armored Force Engineering, Beijing 100072, China)

Abstract: Aiming at data conversion for non-formatted message in simulated training system of command informatization, this paper discusses an information extraction method from non-formatted message according to the demand of simulation-driven information. The paper builds the domain ontology model and the syntax rules after the ontology concept is introduced into the limited area (military area) and proposes the information extraction method based on the ontology. With the technical support of eXtensible Markup Language (XML) Schema and VC++.NET, the method achieves information extraction in practice. It lays foundation for data conversion in the actual simulation work.

Key words: non-formatted message; ontology; information extraction

文本信息提取是建立在分词基础上的一种对信息的有效提取。汉语自动分词是计算机自动识别文本中词的边界的过程, 是把没有词语分割标记的句子自动切分成有一定语义的词串[1]。信息提取一般是在特定的领域中进行, 其过程涉及 2 个方面: 一是待处理的文本集以及用户感兴趣的信息特性; 二是系统过滤文本集并以一定格式输出匹配的信息。

为解决指挥信息系统和作战仿真系统互操作的问题, 本文对军事领域这一受限领域中的文本进行分析研究, 探索信息提取的方法。在对指挥信息系统与作战仿真系统交互中文本的要素提取中, 待处理文本是由指挥信息系统发出的带有军事目的的文本, 用户(这里指的是作战仿真系统中的虚拟兵力)感兴趣的信息特性是文本能产生驱动的信息, 而过滤后的文本集的输出格式是作战仿真系统中的数据格式。

1 受限领域信息提取常用方法

由于指挥信息系统中存在的自由格式报文以及一些文书的格式, 作战仿真系统不易识别, 两者之间的数据结构存在较大的差异。造成这种语义异构的因素主要有: 不同的信息源采用多种术语表示同一概念; 同一术语在不同的信息源中表达不同的含义。虽然各信息源的数据格式差异较大, 但各信息源中的概念之间却存在着各种联系, 为信息提取提供一定的可能性。

1.1 信息提取的基本含义

信息提取是指从一段文本中提取指定的一类信

收稿日期: 2012-07-12
作者简介: 孙 剑(1986-), 男, 硕士研究生。

息(如事件、事实)并将其形成结构化的数据。根据MUC(Message Understanding Conferences)的定义,信息提取主要包含5个典型的提取阶段。

1) 命名实体的提取:提取文本中相关的命名实体,包括人名、地名、机构等的识别。

2) 实体关系提取:提取命名实体之间的关系(即事实),一般为二元关系。

3) 事件结构提取:提取文本中的事件信息,一般是一种多元语义关系。如:对于"进攻"这一事件,就有时间、地点、执行单位、执行方式等要素需要提取,这些要素共同形成"进攻"这一类事件的信息结构。

4) 共指信息发现:发现文本中代词和名词共指的信息。

5) 共同事件合并:合并提取出的相同事件。

1.2 军用文书信息提取基本方法

军用文书作为自然语言理解的一个受限领域,虽然句子简练、用词精确、格式方法固定、内容相对确定且行文规范,但在计算机智能理解方面仍然缺乏有效的方法。目前,对军用文书的处理有以下2种基本方法[3]。

1) 模板方式。通过对军用文书进行分析,将其制成一系列的模板,固定不变的参数设置为常量,需要变动的参数设置为变量,通过与所需分析文书的匹配来确定文书的内容。这种方式操作相对简单,但过于僵化,适应性差,精确度不高。

2) 自然语言处理方式。采用常规的分词、词性标注、短语分析方法,最终将军用文书转化成一棵语法树,通过与已有规则库的匹配,得到最终结果。这种方式与模板方式相比,精度大大提高了,但需要非常庞大的规则库,且没有考虑特定领域的语义表达。

对军用文书这一专业受限领域进行计算机理解,目的在于获取指定信息,分析过程通常可以是面向结果的、浅层的或部分的语言分析。因此,在进行词法和语法分析中直接采用语义标注获取指定信息的词汇、短语块语言结构,而不需要清楚每一语句的完整句法结构树。借助上下位关系可定义词的语义类别,如用上位词表明其下位词的语义类别。语义标注通过给词加上语义类别标签,将原文的内容抽象到一个较高的层次。本文在考察了以上2种方式之后,通过对指挥信息系统与作战仿真系统间交互数据特点的分析,引入本体技术,并借鉴自然语言处理方式,从语义层次上对军用文本进行有效的信息提取。

2 基于本体的信息提取方法

语法分析是基于词性标注(即依据汉语词语的词性对字符串进行切分并标注的方法)进行的,由于词性标注的高度抽象化,使得语义信息在向上传递中遗失部分语义,而语义标注一般需要借助概念层次来完成。

特定领域的信息抽取任务与通用的自然语言理解任务不同。对于通用的自然语言理解来说,系统必须对输入句子的所有意义(包括隐含意义)进行识别。一般来说,理解分为2步:1)通过句法分析将输入的句子映射到一个句法结构中,如句法树;2)通过句法到语义的转换分析,实现将句法结构映射到意义表达。而对特定领域的信息抽取来说,完全句法分析和深入的语义解释是没有必要的,输入的文本只需映射到有限数目的事件分类即可。此外,需要抽取的信息的类型也是预先定义好的,因而在相关的句子中,需要被解释的只是一些携带相关信息的短语单元。

2.1 信息的组成方式

信息一般包含语法、语义和语用等信息。语法信息是一种形式信息,在信息空间中只涉及到形式符号本身;语义信息进一步涉及到了形式符号的内在含义;语用信息则涉及到符号的用法及使用环境的因素。

在军事文书中,特定军事力量的部署及行动的表述,一般都有特定的一个或一组固定动词与之相对应。因此,在对军事文书进行自动识别和理解时,可以通过固定的动词对其语句进行种类划分,获取包含人们所关心的内容的语句,然后只对这一部分语句进行相关切分,这样既减少了冗余信息对切分过程的干扰,又可以提高全文理解的速度。

2.2 本体在信息提取中的应用

本体是领域的术语及其关系的清晰形式化规范,即对研究领域的概念、概念的特征和属性,以及属性约束进行明确的形式化描述[3]。本体作为领域内部不同主体之间进行交流的一种语义基础,在语义互操作中体现了相当大的优势,因此,在本文的设计中引入本体的概念,利用本体的抽象能力实现语义标注切分并提取有效语素,为切分后的实体映射构造条件。

本体既准确地描述了概念涵义及概念之间的内

在关联,又可通过逻辑推理获取概念之间蕴含的关系,具有很强的表达概念语义和获取知识的能力,因此可用来解决语义异构的问题[4]。本文军事本体中只包含领域知识内的概念及概念之间的关系,不包含推理知识。在一般面向特定领域本体应用中,词典作为文本与本体的接口,从词映射到概念,再利用概念所含有的约束进行推理,本体构成搜索空间,概念间的约束知识作为启发式信息指导搜索。

2.3　基于本体的信息提取主要思路

本文针对一体化指挥信息系统与作战仿真系统之间的数据异构现象,利用本体的高度抽象能力,在军事领域本体库的支撑下,通过基础分词、语义分类标注、浅层句法分析、信息结构化,将来自一体化指挥信息系统中的作战文书及自由格式的报文中包含的信息提取出来,为实现一体化指挥信息系统与作战仿真系统间的交互提供结构化的信息。文本要素提取思路如图 1 所示。

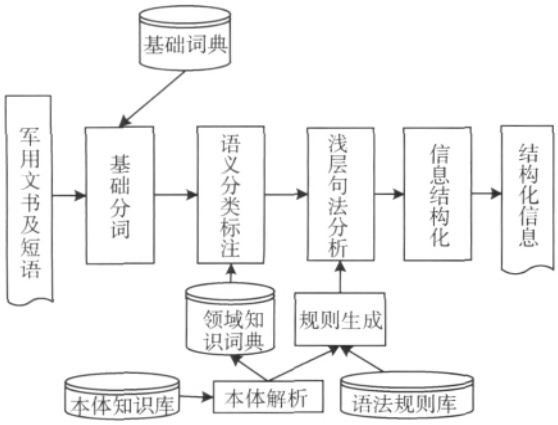

图 1　文本要素提取思路

基础分词模块主要是利用中科院 ICTCLAS 分词系统[5]作为基础,对来自一体化指挥信息系统中的军用文书及短语报文进行初步的分词,获取最基本的分词。语义分类标注的主要功能是对基础分词模块输出的分词进行语义标注,利用本体知识库中的领域本体概念构造用户词典,从语义层次上对基本分词进行解析并标注。浅层句法分析主要利用语法规则对标注的字串进行分析,进一步定位各要素的位置及功能,并提取有用信息及信息语义标签。信息结构化是将提取出来的信息及信息语义标签转换到结构化的数据结构中,并存入数据库以备用。

从上文的思路可以看出: 基于本体的信息提取与一般的分词的区别在于,在本体的支撑下,引入领域内的相关词汇及词汇之间的相互关系,在此基础上进行语义标注并分析提取领域内所关心的信息。因此,基于本体的信息提取的核心是对底层知识库的设计。

3　知识库设计

知识库是存储指挥信息系统和仿真系统间数据转换所需知识的文档或数据库文件,主要包括领域本体库和语法规则库 2 部分。

3.1　领域本体设计

在基于本体的文本信息提取中,本体主要的作用相当于一个语义词典,能够为分词提供基本的语义信息和语义标注。按照汉语语义完整性表达的前提,在一个完整语义的字串中一般包括以下成分: 实施者、实施对象、实施手段等[6]。基于本体文本信息提取的词典中关键词的作用就是,将这个有复杂成分的字串切分成单一的成分串。

在指挥信息系统中,信息流动呈现复杂多样的形态,而在作战过程中,人员关心的是与作战有关的态势信息以及能够影响作战过程的指挥信息,一般来讲,对于响应指挥员的模拟兵力关心的是与行为有关的信息载体——任务。任务是由一个或者多个动作组成的,具有明确意图的目的行为过程,动作必须由具体的执行实体来完成,并使实体的状态发生变化[7]。战场中实体的行为是以任务为驱动、以动作为中心的一组要素的组合,因此,战场中的实体行为可抽象为以下信息模型:

行为 : = 动作(实体,时间,位置,方式,状态) 。

从以上的实体行为模型不难看出: 在建立军事领域的作战本体时,主要关心的是 6 个方面,即动作、实体、时间、位置、方式和状态[8-9]。在建立本体时,主要参照以上实体的行为模型来分别构建本体。以动作本体为例,文中将动作分为火力、机动和动作转换 3 类分别进行构建,如图 2 所示。

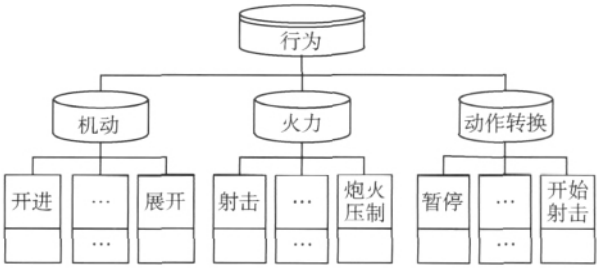

图 2　动作本体

3.2 语法规则库设计

语法规则的主要功能是对切分并标注好的字符串进行句法分析，获取具有相应功能的信息实体。如：在前文的分析中，获得了“位置”概念的词，但是并不能确定该“位置”词在整个句子中的功能角色，需要对整个句子进行句法分析，以便确定该词是属于“目标位置”还是“路线”等。根据军用文书书写规范简洁的特点，对其进行浅层句法分析时，比一般自由文本简单。本文采用基于规则的浅层句法分析，利用编写好的规则指导句法分析，获取功能信息。

军用文书中的语法较为单一，除了主、谓、宾的主体结构外，其他的成分大多以介词结构的形式出现，而在对军用文书进行解析时，主要是需要解析出介词结构中包含的信息要素中的语义成分。因此，这里主要介绍规则中介词结构规则的构建与解析。文书中用到的介词主要有“对”、“向”、“沿”、“于”、“至”、“在”6类，每个介词后跟的成分都有相应的语法功能，如“对”后一般跟的是表示目标的信息成分。通过分析归类，本文建立了相应介词结构规则并形式化，如图3所示。

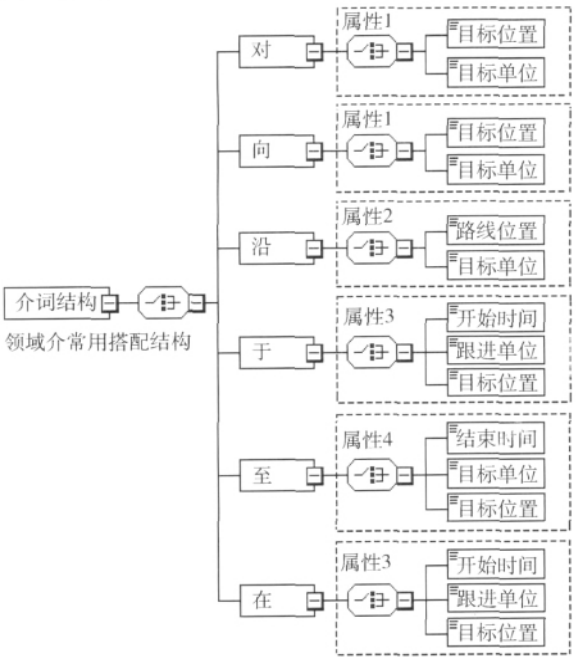

图3 介词结构规则

4 基于本体的信息提取算法设计

基于本体的信息提取算法的基本思想是：首先，从领域本体库中提取领域相关的实体，构建领域相关的专业词典，结合现有的分词系统将用户关心的常见实体词汇划分出来，利用本体的概念抽象能力对实体进行合并处理；其次，利用规则库中的语法规则对分词结果进行分析，划分各实体的语用功能；最后，得到用户关心的具有语义和语用标注的信息要素。其算法流程如图4所示。

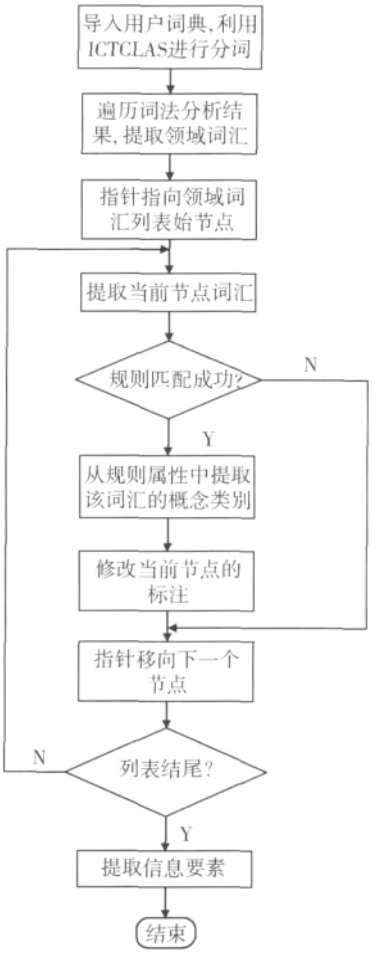

图4 信息提取算法

5 试验及结果分析

以VC++为基础开发平台，以可扩展标记语言(eXtensible Markup Languge，XML) Schema为规则载体，对本文中提出的算法构建计算机模型，并对军事文书中的信息进行提取。输入字符串为“第4机步团沿1、2号道路向伊和陶勒盖(94，11)、呼特格乌拉(91，16)、巴音杜信敖包(90，20)方向攻击前进，抢占呼特格乌拉(91，16)、巴彦超格图(90，17)、巴音杜信敖包(90，20)地域有利地形后，向1212高地(89，25)方向实施主要攻击，配合纵深攻击群、左翼攻击群粉碎敌反击。”

系统运行后抽取结果如表 1 所示。

表 1　试验结果

行为类型	执行单位	路线	目标位置
攻击前进	第 4 机步团	1、2 号道路	呼特格乌拉(91,16)、巴音杜信敖包(90,20)
抢占	第 4 机步团		呼特格乌拉(91,16)、巴彦超格图(90,17)、巴音杜信敖包(90,20)
攻击	第 4 机步团		1212 高地(89,25)

从表 1 可以看出: 对于“攻击前进”这一命令抽取中的“目标位置”并不完整,原文中应为“伊和陶勒盖(94,11)、呼特格乌拉(91,16)、巴音杜信敖包(90,20)”3 个目标点。通过检查在本体中没有收录“伊和陶勒盖”这一实例,发现在信息抽取过程中出现了缺漏现象,在完善本体之后再对上面的实例进行试验,得到了完整的抽取结果,如表 2 所示。

表 2　完善本体后的试验结果

行为类型	执行单位	路线	目标位置
攻击前进	第 4 机步团	1、2 号道路	伊和陶勒盖(94,11)、呼特格乌拉(91,16)、巴音杜信敖包(90,20)
抢占	第 4 机步团		呼特格乌拉(91,16)、巴彦超格图(90,17)、巴音杜信敖包(90,20)
攻击	第 4 机步团		1212 高地(89,25)

试验中,对一些典型的军用文书中的短语指令(如“机动、集火射击、集结、展开、武装侦察”等)进行了测试,在确保本体完整性的前提下,基本上能够将仿真实体需要的驱动信息从文本中提取出来,并对信息进行结构化处理。试验结果表明: 文中提出的基于本体的受限领域文本要素提取的方法是可行的,对军用文书及短语中的要素提取能够满足系统驱动的要求。

6　结论

本体代表一种用机器可以理解的语言和逻辑建立的对信息资源的结构化描述规范,目前广泛用于异构系统的交互及异构数据的集成研究中。文中在解决指挥信息系统与作战仿真系统间异构数据的交互中,引入本体概念,目标是捕获军事领域的知识,将其抽象并构建相应的概念集,在不同层次的形式化模式上明确了概念之间的关系。同时对军用文书的书写规范进行了研究,构建了介词规则,指导浅层句法分析,有效地解决了文本要素提取问题,为进一步的映射转换提供结构化的数据准备。同时在试验中得出: 对文本信息提取的完整性取决于领域本体的完整性,本体中包含该领域的实体概念越完整,在信息提取中提取出的要素就越完整。

参考文献:

[1] 孙茂松,邹嘉彦. 汉语自动分词研究评述[J]. 当代语言学,2001,3(1):22-32.

[2] 顾晓明. 一个基于本体的作战文书理解系统设计实现[D]. 南京: 东南大学,2006.

[3] 高茂庭,王正欧. Ontology 及其应用[J]. 计算机应用,2003,23:31-33.

[4] 甘健侯,姜跃,夏幼明. 本体方法及其应用[M]. 北京: 科学出版社,2011.

[5] 刘群,张华平,俞鸿魁,等. 基于层叠隐马模型的汉语词法分析[J]. 计算机研究与发展,2004,41(8):1421-1429.

[6] 段永威,秦峰. 异构数据源数据转换工具的设计与实现[J]. 现代图书情报技术,2004(4):59-62.

[7] 胡晓峰,司光亚,吴琳,等. 战争模拟引论: 上[M]. 北京: 国防大学出版社,2004:125.

[8] 岳磊,马亚平,徐俊强,等. 面向语义的作战命令形式化描述及本体构建[J]. 指挥控制与仿真,2012,34(1): 11-14.

[9] 程恺,车军辉,张宏军,等. 作战任务的形式化描述及其过程表示方法[J]. 指挥控制与仿真,2012,34(1):15-19.

(责任编辑: 尚彩娟)