从网页目录到轻量级本体——自然语言处理及形式化 *

鞠 奇, 杨凤杰
(吉林大学 计算机科学与技术学院 智能信息处理教研室, 长春 130012)

摘 要: 为了解决以自然语言表示节点标签的分类树很难通过自动软件 agents 来进行自动推理的问题,通过词性标志、词义辨析、连接词辨析和受约束的自然语言定义及转换等步骤,将分类树中每一个节点对应的自然语言标签转换成了机器能够识别的逻辑表达式,从而使整个分类树转换成了一个轻量级本体,它适合应用在数据整合的语义匹配、文档分类和语义搜索等方面的自动推理,从而促进了本体知识的自动化推理,为以后文本自动检索奠定基础。

关键词: 分类; 描述逻辑公式; 轻量级本体; 词性标志; 词义消歧; 等位词消歧; 受限自然语言
中图分类号: TP301.2 **文献标志码**: A **文章编号**: 1001-3695(2010)04-1352-05
doi:10.3969/j.issn.1001-3695.2010.04.039

From Web directories to lightweight ontology: natural language processing and formalization

JU Qi, YANG Feng-jie
(*Intelligent Information Processing Lab*, *College of Computer Science & Technology*, *Jilin University*, *Changchun* 130012, *China*)

Abstract: In order to solve the problem that classifications were very hard to be reasoned about by automated software agents and represent annotations of little use for semantic Web applications since their labels or nodes were written in natural language, this paper introduced an approach to transform a hierarchical directory into lightweight ontology by a series of steps, including part-of-speech tagging, word sense disambiguation, coordination disambiguation, and new controlled natural language definition and conversion, which then helped formalize the natural language labels into simple description logic formulae and provided the significant basis for further ontology reasoning and document retrieval.

Key words: classification; description logic formulae; lightweight ontology; part-of-speech tagging; word sense disambiguation; coordination disambiguation; controlled natural language

0 引言

随着网络资源越来越丰富繁杂,为了组织和操作这些数据,基于训练集和属性值的分类法成为一种不可或缺的方法。早期的分类法主要应用在图书管理系统方面,现在则扩展到了网页、图片等各种数字资源等方面。分类法用自然语言标签来描述资源的内容,然而当进行自动化操作时,由于自然语言标签不能实现推理分类而具有很大的局限性。对于分类和其他自然语言处理的相关领域,自动翻译过程变得非常必要。一般而言,一个轻量级本体是由一组概念在 IS-A 关系下组成的层次。例如,数据字典、产品分类和主题图等经常被看做轻量级本体[1]。从这种观点来看,网页目录也经常被叫做轻量级本体[2]。与轻量级本体相对的是形式知识本体,它通常使用形式化逻辑来描述约束、关系和其他应用到概念上的规则[3]。

现阶段的自然语言处理主要集中于四大方向:语言学、数据处理、人工智能和认知科学、语言工程。对于处理英文方面,比较成功的系统有加拿大蒙特利尔大学 TAUM—METE 机器翻译系统和 Babel Fish 机器翻译系统;对于处理中文方面,中国科学院计算所数字化研究室研制的汉英机器翻译系统处于比较先进的地位,它们都是从词法分析、语法分析、机器翻译等入手提高机器对自然语言的识别。

本文首次引进轻量级本体这一概念使得机器在理解自然语言没有歧义的情况具备“思考”的能力,利用其高速的计算能力来理解自然语言并与人类进行沟通。具体来说,就是致力于把 DMOZ 目录转换到一个轻量级本体,通过一系列过程转换,词义歧义和等位词歧义问题将很大程度上得到改进,同时将整个目录用机器识别的形式化逻辑公式来表示,从而实现自动推理[4]。图 1 体现了本文的中心思想:通过对网页自然语言标签标志、词义歧义消除、等位词歧义消除和形式化过程,最终将自然语言标签转换为逻辑公式的形式,以实现机器的自动推理。

本文描述了自然语言处理中的词义的歧义消除,基于 NLP 工具给出了等位词歧义消除精确度的概念,并对比了原始过程

收稿日期: 2009-08-17; **修回日期**: 2009-09-22 **基金项目**: 国家自然科学基金资助项目(60773097); EASTWEB 欧洲合作项目(111084)
作者简介:鞠奇(1983-),男,湖北云梦人,硕士,主要研究方向为本体及语义网(qi.ju02@gmail.com); 杨凤杰(1964-),女,吉林长春人,副教授,硕士,主要研究方向为人工智能、智能信息处理、离散数学.

精确度与加入启发式规则后过程的精确度,讨论了一种与分类形式化和推理服务有联系的受限自然语言的定义及应用,综合以上几个过程列出了一些试验性能评估。

1　词义歧义消除

资源丰富的网页目录呈现的是以自然语言为标签的分类,标签中的每一个词都有多种含义,实现网页目录本体化的第一步就是满足词义的单一性。本文借助 WordNet 词汇表。

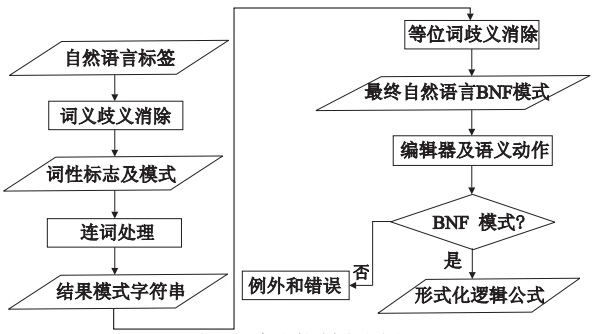

图1　本文控制流程图

1.1　WordNet

WordNet 是一个免费的英语词汇数据库,其设计灵感源自于与人类词汇记忆相关的心理语言学理论[5]。在英语中分词被组织成同义词集,用于分别代表每个基本词汇概念。在 WordNet 中,对于每个分词的不同词义,都可以是某个同义词集的一部分,而该分词的不同词义使得其可以属于多个不同的同义词集。WordNet 同时表示了同义词集之间的语义关系和分词之间的词汇关系[6];Li 等人[7]指出用其作为算法的源信息时,若只考虑取其第一词义则正确率为 57%,而考虑第一二词义时可达到 67%;Mihalcea 等人[8]指出当 WordNet 与其他资源进行组合和交互核对时效果更好,如若算法允许在信度比较低时不给出结果的话,正确率将提高到 92%[9]。当使用一个小而典型的分词集合来确定上下文时,Natase 等人[10]指出,若允许算法不给出结果的话可获得 82% 的平均准确度。

1.2　方法

对于任何给定的分类法,词义歧义消除的最终目标就是为每个分词找出其概念的含义。标签中被称为概念分词的绝对分词是本文考虑的对象,如在 WordNet 中作为形容词和名词的分词[11],方法如下:

a)深度优先遍历目标分类。深度优先比宽度优先在共享内存和执行时间上更具优势。

b)遍历目标分类中每个节点时过滤出所有的概念分词,因为几乎所有自然语言标签中的分词都是名词或形容词。

c)检测当前节点的每个概念分词是否一词多义。若不是则跳过以下步骤,此分词惟一的词义即是最终结果。

d)为每个分词标志其词性,并保留一致性的词义,同时删除这个分词其他词性的词义。

e)仅保留同一标签中有上下位关系的名词分词词义,其上下位关系可以通过 WordNet 的上下位层次检测[11]。

f)结合根节点到当前标签节点的词义环境,尽可能地通过相似度演算算法和其他方法得出分词间的相似度,保留特定相似度的词义。

g)对于未经 a) ~ f) 处理的分词取其第一词义。

此时每一个自然语言标签中的分词只存在一种词义,即后面提到的名词概念或形容词概念,这些单一词义为等位词消岐过程提供了方便。

2　等位词歧义消除

等位词指的是句子或短语组成部分有相同语法形式的一种句法关系;从分类树的角度来看,等位词歧义是指对于一个复杂的短语或语句会导致多种分类树结构,从而产生句法歧义。在英语中,这类歧义是一类潜在的普遍问题,然而等位词歧义消除是自然语言处理中最具挑战性和普遍性的难题之一。具体来说,在本文中针对连词 and、or 的等位词歧义来自对两种不同模式的不同理解:

a)A and B C, 可分解为 A and(B C)或(A and B)C 两种方式。在前者中 A 是一个短语,B 作为 C 的一个修饰符;对于后者,A、B 同时作为 C 的修饰符。

b)A B and C,可分解为(A B)and C 或 A(B and C)两种。在第一种情况下 A 作为 B 的一个修饰符,C 为一个独立的短语;在第二种情况下,A 作为 B、C 的共同修饰符。

2.1　数据集

本文关注 AND 连接词短语,它在从生物医学文摘的语言库提取的连接词中占 87.07%,同时 OR 连接词占10.34%[12],它们共占英国国家语料库中的分词的 3%。本文研究 DMOZ 分类树中带有 AND 的四分词标签,据统计,其占所有四分词标签的 55.73%,且出现的平均次数为 2.59。

2.2　方法

本文侧重于内部带有 AND 连接词的两种标签模式:A and B C, A and B C。其中:A、B 和 C 是 POS 标记。OpenNLP 工具可用来处理等位词歧义消除,它支持一组基于 Java 的用来执行基于句子检测、分词化和 POS 标记的句子语块化过程的 NLP 工具。语块标记由语块类型名组成,如名词短语为 I-NP,动词短语为 I-VP。大多数语块两类句块标签,即代表第一个分词的 B-CHUNK 和其他分词的 I-CHUNK,它们之间的分隔符是“O”(outside)。例如,内容为“swimming pools and spas”的标签将被语块为“swimming_B-NP|pools_I-NP|and_O|spas_B-NP|”。

借助于 WordNet 可知,相似性是基于 POS 标记匹配、分词匹配和基于 WordNet 的语义类似的估量来衡量的。由此通过连续计算本文中连接词左右分词的相似程度可知任意两个分词之间的相似程度。

由上本文对于两类典型的等位词消歧得到一些启发式规则如下:

1)A and B C 类型

a)若 A 为 NNS,则 A and(B C);

b)若 A 为 NN,B 为 JJ, 则 A and (B C);

c)若 A、B 为 NN,分别计算 A、B 和 B、C 的相似性;

d)若 A 为 JJ,B 为 JJ, 则(A and B)C;

e)若 A 为 JJ,B 为 NN, 则 A and(B C);

f)若 A、B 为 VBG, 则(A and B)C。

2)A B and C 类型

a)若 B 为 NNS, 则(A B)and C;

b)若 A 为 JJ,B 和 C 都为 NN 或者 VBG, 则 A(B and C)。

根据 2.2 节的方法,本文抽取 1 288 个典型的四分词,进行了测试,结果如表 1 ~3 所示。

表 1　等位词歧义消除测试结果

模式类别	A and B C		
	Cor	Total	Acc
改进前	556	904	61.50%
改进后	735	904	81.31%

表 2　等位词歧义消除测试结果

模式类别	A B and C		
	Cor	Total	Acc
改进前	219	385	56.88%
改进后	271	385	70.39%

表 3　等位词歧义消除测试结果

模式类别	全部(包括 A and B C 和 A B and C)		
	Cor	Total	Acc
改进前	775	1289	60.12%
改进后	1006	1289	78.04%

其中:Cor 为正确标签模式个数;Total 为全部标签模式个数;Acc 为准确率。由表 1 可知,单纯地借用 OpenNLP 工具对数据进行等位词歧义消除测试,模式 A and B C 的准确率为 61.50%,比模式 A B and C 高出约 5 个百分点,对选取数据总的准确率为 60.12%;当采用启发式规则 1)和 2)后,可以看出相对应的结果分别提高了 20%、16%、18% 左右,在很大程度上改进了等位词歧义消除的准确率,为下文的模式匹配和形式化转换提供了基础。目前这方面做得比较好的有 Kurohashi 等人[13],他们提出了用日语来分析等位词结构的方法,准确率为 81.3%,但在选取数据、使用语言方面存在一些局限性。其中 Goldberg 采用无监督学习方式决定歧义连接词短语的分块[14],得到准确率为 72%;还有就是 Resnik,他强调了在解决等位词歧义时的语义相似性,准确率为 71.2%。本文的结果稍微优于这些,可能是与研究的范围、测试数据合集不一样,也可能是方法原理的不同。

3　受限自然语言

经过词义歧义消除和等位词歧义消除之后,最初的自然语言标签转换为了一个没有任何歧义、具备学习能力的机器所能识别的自然语言,通过本章所述,处理这些自然语言成为机器能够完全接受的表达形式,如逻辑表达式等。

受限自然语言被定义为在语法、词典和类型方面具有明确约束的自然语言子集。这些约束通常具备规则的形式,以减少完备自然语言的歧义和复杂性[15,16]。目前,最成功的受限自然语言可能是 ASD Simplified Technical English[17],它主要是为了提高航空飞行器维护文档对非专业读者的可读性。但是,语言的可读性仍需要与机器可处理性相结合才能使语言在问答式中变得灵活可用。一些相关的受限语言,如 Attempto Controlled English[18]、Common Logic Controlled English[19] 和 Processable English[20],以方便人类和机器合作工作的方式结合,并平衡了可读性与可操作性,而事实上它们等价于不可判定的一阶谓词逻辑的一部分,并且已经作为知识表示语言广泛应用于各领域。

实质上,虽然形式化语言理论更广泛地定义了规则和语法,本文中仅考虑与上下文无关的形式语法。Backus-Naur Form (BNF)作为形式化元句法用来表示与上下文无关的语法已经普遍地应用在语言的描述方面,所以本文也采取 BNF 来形式化自然语言。

3.1　数据集

对 DMOZ 目录分析可知,除了很少非英语标签的分支路径,如 Top/World、Top/Adult/World、Top/Kids 和 Teens/International 等,DMOZ 目录共包含 491 512 个路径。通过 OpenNLP 工具得到普通名词模式和专有名词模式共 6 620 种,它们的主要区别在于专有名词模式的实例都是专有名词,此类名词由于它们的固定性和特殊性不属于本文的考虑范围。为了发掘出所有的普通名词模式,首先,为每一个标签模式中的模式找出其对应的所有实例,并通过其所在的原始 DMOZ 目录上下文来检测这些实例是否是专有名词,若结果为真,则这些模式就是专有名字模式;反之,就是普通名词模式。这一过程通过在以下启发式规则引导下通过手工标志来完成:

a)在 6 620 种标签模式中,4 459 种模式仅代表一个目录标签实例,此外,对应实例数小于 5 的模式一共有 5 749 种,占所有标签数的 86.84%。通过观察,几乎所有这些模式的实例都是专有名词,并且这些模式都不符合普通模式类型,如 CD/RP/IN/NN/、CD/NNS/VB/、SYM/VBG/CD/、JJ/IN/JJ/DT/NN/NN/ 等。

b)一些专有名词模式包含一些典型词性标志,如很多电影名都包含 CD、DT 等,因此这些专有名词模式也应排除在外。

c)一些在 DMOZ 中的人名有一个相对统一的形式:last name/、/first name/middle name, 其中 first and middle names 能够被缩写,几乎所有像 NN/、/NN、NNS/、/NN、NN/、/CC、NN/、/NNS 等的模式都是用来表示人名的;带有逗号的普通名词模式一般会出现两个逗号或者逗号后面紧跟一个连接词标志 CC,如 NN/、/NN/CC/NN/、NN/、/NN/、/CC/NN/。

d)团体组织名称,尤其是一些专科院校、高等院校和研究机构,都有一个比较固定的模式,如 NN/NN/IN/JJ/NN/、JJ/NN/IN/NN/NN/ 等都包含一个介词标志 IN。

e)在专有名词短语中占有很大比例的地点名词直接跟随在目录路径中自然语言 localities 之后,通过此方式能找到大部分地点专有名词。

通过大量的手工标注工作来区分普通名词模式和专有名词模式,最后得到了 174 种普通标签模式,它们代表了 82 277 个实例,占所有 DMOZ 目录总实例数的 65.30%。

3.2　自然语言标签 BNF 定义

通过对这 174 种普通标签模式的分析可以看出,几乎所有的标签模式都包含一个或多个名词短语或者形容词短语,它们通过一些连接词连接。考虑到尽可能多的情况,本文总结出 DMOZ 中自然语言标签模式的 BNF 定义如下:

a)NL_Label ::= Phrase {Conn Phrase}

b) Phrase ::= Adjectives [NounPhrase] | NounPhrase

c) Adjectives ::= Adjective {Adjective}

d) NounPhrase ::= Noun {Noun}

e) Conn ::= ConjunctionConn | PrepositionConn

f) Noun ::= NN | NNS | VBG

g) Adjective ::= JJ | VBN

h) ConjunctionConn ::= , | CC

i) PrepositionConn ::= IN

以上定义的上下文无关的形式化语法由以下四个部分组成:

a) 开始符: NL_Label;

b) 非终结符: NL_Label、Phrase、Connective、ConjunctionConn、PrepositionConn、NounPhrase、AdjectivePhrase、Adjective、Noun;

c) 终结符:NN、NNS、VBG、JJ、VBN、CC、,、IN;

d) 规则集:以上 DMOZ 中自然语言标签模式的 BNF 定义 a) ~i)。

以上各符号解释参照表4、5。

表4　符号解释

符号	解释	符号	解释
::=	"定义为"	{ }	重复 0 次或多次
[ ]	可选项	( )	短语优先级

表5　BNF 缩写的元符号解释

缩写形式	含义	缩写形式	含义
NN	单数名词或不可数名词	VBN	动词过去式
NNS	名词复数	CC	等位连词
VBG	动名词,现在分词	,	逗号
JJ	形容词	IN	介词,从属连词

其中:CLP 为普通标签模式;DMOZ-I 为实例总数。

进一步,检验该 BNF 定义的所包含的普通标签模式覆盖率和其代表的 DMOZ 实例数如表 6 所示。本文定义的自然语言标签 BNF 的模式覆盖率还是很高的,同时也覆盖了65.04%的所有标签,排除 2.1 节数据集中所示的一些情况,这个覆盖率还是比较理想的。

表6　自然语言 BNF 定义覆盖率统计

比较项	总数量	覆盖数	比例/%
CLP	174	166	95.40
DMOZ-I	125 338	81 531	65.05

3.3　形式化 BNF 定义

自然语言标签的 BNF 定义后,为了实现机器自动化,首先建立了一个对应的机器能够识别处理的形式化语言,它使用逻辑语言来表示标签模式,一个满足 3.2 节定义的自然语言标签,通过转换后生成的形式化逻辑公式也必定满足以下的形式化逻辑定义:

a) FL_Label ::= Formula {, Formula}

b) Formula ::= AtomicConcept | ComplexConcept

c) AtomicConcept ::= T | ⊥ | NounConcept | AdjectiveConcept

d) ComplexConcept ::= Formula OR Formula | Formula AND Formula

其中:初始符为 FL_Label;非终结符为 FL_Label、Formula、AtomicConcept、ComplexConcept;终结符为 T、⊥、AND、OR、NounConcept、AdjectiveConcept;规则集为 a) ~d)。

3.4　语义动作

一般地,语义动作的形式是:表达式[动作]。其中,动作实际上是当解析器解析时被运行的一些 Java 代码;同时,语义动作也可以被附在解析器内部任何层次的表达式上。一个动作是一个 C/C++或 Java 的功能函数或当表达式匹配成功时被调用的函数对象。为了实现本文中自然语言标签到形式化逻辑公式的过度,在 3.2 节中原 BNF 定义的相关位置添加语义动作如下:

a) #0 NL_Label ::= Phrase {Conn Phrase} #1

b) Phrase ::= Adjectives [NounPhrase] | NounPhrase #2

c) Adjectives ::= Adjective {Adjective} #3

d) NounPhrase ::= Noun {Noun} #4

e) Conn ::= ConjunctionConn | PrepositionConn #5

f) Noun ::= NN | NNS | VBG #6

g) Adjective ::= JJ | VBN #7

h) ConjunctionConn ::= , | CC #8

i) PrepositionConn ::= IN #9#10

部分语义动作内容如下:

#1:{String a="", b, d, c="";} { c = Phrase ( ) {a = c ;} (b = Conn ( ) (c = Conn ( ))? d = Phrase ( ) {a = a + " " + b + " " + d ;}) *";" {return a ;} }

#2:{String a="", b="", c="";} {a = Adjectives ( ) (b = NounPhrase ( ) {{c = " " + "AND" + " ";}})? { a = a + c + b ;} { return a;} |a = NounPhrase ( ) {return a ;}}

#3:{String a="", b;} {b = Adjective( ) {a = b;} (b = Adjective( ) {a = a + " " + "AND" + " " + b;}) * { return a ;}}

其中,JavaCC 是当前最流行的一种解析产生器,它是一个能够读入语法定义并将其转换成为 Java 源代码以用来识别该语法的匹配,而且,JavaCC 可以编译带有语义动作的标准 BNF 模式,并产生相应的 Java 源代码,用来检测目标模式是否符合 3.2节中 BNF 定义,并在检测的过程中根据语义动作产生相应的目标输出。

4　实验与测试

当给定的一个自然语言标签首先满足了自然语言模式 BNF 定义检测后,会通过其中的语义动作生成逻辑表达形式,这种逻辑形式接着会通过形式化定义的 BNF 检验,如果满足,就是本文最终需要转换的格式。

4.1　局部测试

对于经过等位词歧义消除过程后的标签模式,运用 JavaCC 解析产生的 Java 源代码对其进行测试并产生结果,如下:

Input1: NN1 CC_and NN2 NNS3

Ouput1: {NounConcept1, NounConcept2 AND NounConcept3}

Input2: (NN1 CC_and NN2) NNS3

Ouput2: {NounConcept1 AND NounConcept3, NounConcept1 AND NounConcept3}

Input3: NN1 CC_and JJ NNS2

Ouput3: {NounConcept1, AdjectiveConcept AND NounConcept2}

Input4: NN CC_or NNS

Ouput4: {NounConcept OR NounConcept}

Input5: NN1 CC_, NNS2 CC_, JJ NN3

Ouput5: {NounConcept1, NounConcept2, AdjectiveConcept AND NounConcept3}

Input6: NN1 (NN2 CC_and NNS3)

Ouput6: {NounConcept1 AND NounConcept2, NounConcept1 AND

NounConcept3 }

每一个 NounConcept 和 AdjectiveConcept 都表示了一个概念。各种符号解释请参照表 3。

4.2 流程测试

这一过程是在等位词歧义消除的基础上对表 1 的整个流程图的实现,具体分为以下三种情况:

a) Correct: 给定自然语言标签完全符合表 3 中 BNF 定义,并产生出相应的逻辑表达式。

Input natural label :(bank and warehouse) guarder
Disambiguation pattern :(NN CC_and NN) NN
Processed pattern: NN NN CC NN NN
Drafted output: NounConcept AND NounConcept OR NounConcept AND NounConcept
Final output: { NounConcept AND NounConcept, NounConcept AND NounConcept }

b) Error:标志模式中出现了没有定义在表 3 中的符号,如专有名词"United States"。

其对应的标志模式 NNP NNPS 中的符号 NNP(专有名词符号)并未出现在 2.2 节中,即该标志模式在此被定义成错误的模式。源码输出如下:

Input natural label: United States
Disambiguation pattern: NNP NNPS
Processed pattern: NNP NNPS
Error.
Lexical error at line 1, column 3. Encountered: "P" (80), after: ""

c) Exception:给定自然语言的标志模式不符合表 3 中 BNF 定义,亦不能匹配上述 BNF 文法,但是,源码会给出不匹配的具体原因,以便作进一步修改。与 b)不同的是,该标志模式中的所有符号都出现在表 5 节中,如:

Input natural label: music related
Disambiguation pattern: NN VBN
Processed pattern: NN VBN
Exception.
Encountered "VBN" at line 1, column 4.
Was expecting one of:
"NN" …
"NNS" …
"VBG" …
"JJ" …
"VBN" …
";" …

从以上局部测试和流程测试可以看出,本文所描述的系统能够准确地识别并处理网页目录的自然语言标签,并转换为机器能够理解的逻辑表达式,进而实现机器的自动化推理,为网页搜索提供了智能支持。

5 结束语

处理并完善 DMOZ 目录中的命名实体识别,保证后续转换过程的精确度;同时寻求一些优化技术以提高词义消歧精度及实现高效的等位词歧义消除过程,其中包括对长修饰符的等位词分析;最终实现并结合各部分,组成一个自动化操作网页目录自然语言系统,从而实现从网页自然语言目录到轻量级本体的过渡。

参考文献:

[1] ZHU H W, MADNICK S. A lightweight ontology approach to scalable interoperability[C]//Proc of CISL. 2006.

[2] USCHOLD M, GRUNINGER M. Ontologies and semantics for seamless connectivity [J]. ACM SIGMOD Record, 2004,33(4):58-64.

[3] GUARINO N. Formal ontology and information systems[C]// Proc of Formal Ontologies in Information Systems (FOIS '98). Trento: IOS Press, 1998:3-15.

[4] NATALYA F. Semantic integration: a survey of ontology-based approaches[J]. ACM SIGMOD Record, 2004,33(4): 65-70.

[5] FELLBAUM C. WordNet: an electronic lexical database [M]. [S. l.]: MIT Press, 1998.

[6] ALBERTO J C, ALEJANDRO V, MARCO C. Using WordNet for word sense disambiguation to support concept map construction[C]// Proc of the 10th International Symposium on String Processing and Information Retrieval. 2003.

[7] LI X, SZPAKOWICS S, MATWIN S. A WordNet-based algorithm for word sense disambiguation [C]//Proc of IJCAI' 95. 1995: 1368-1374.

[8] MIHALCEA R, MOLDOVAN D. A method for word sense disambiguation of unrestricted text[C]// Proc of ACL'99. 1999:152-158.

[9] MIHALCEA R, MOLDOVAN D. An iterative approach to word sense disambiguation[C]//Proc of Flairs 2000. 2000:219-223.

[10] NASTASE V, SZPAKOWICS S. Word sense disambiguation in Roget's Thesaurus using WordNet[C]//Proc of NAACL WordNet and Other Lexical Resources Workshop. 2001:17-22.

[11] GIUNCHIGLIA F, MARCHESE M, ILYA Z I. Encoding classifications into lightweight ontologies[C]//Lecture Notes in Computer Science, vol 4380. 2007:57-81.

[12] NENADIC G, SPASIC I, ANANIADOU S. Mining biomedical abstracts: what is in a term[M]. [S. l.]: Springer-Verlag, 2005: 797-806.

[13] KUROHASHI S, NAGAO M. Dynamic programming method for analyzing conjunctive structures in Japanese[C]//Proc of the 14th Conference on Computational Linguistics. 1992:170-176.

[14] PHILIP R. Semantic similarity in taxonomy: an information-based measure and its application to problems of ambiguity in natural language[J]. Journal of Artificial Intelligence Research, 1999,11: 23-29.

[15] HUIJSEN W O. Controlled language: an introduction[C]//Proc of CLAW. 1998:1-15.

[16] O'BRIEN S. Controlling controlled English: an analysis of several controlled language rule sets[C]//Proc of EAMT-CLAW'03. [S. l.]: Dublin City University, 2003:105-114.

[17] Specification ASD-STE100, ASD simplified technical documentation in the international aerospace maintenance language[S]. 2005.

[18] FUCHS N E, KALJURAND K, SCHNEIDER G. Attempted controlled english meets the challenges of knowledge representation, reasoning, interoperability and user interfaces[C]//Proc of the 19th International Florida Artificial Intelligence Research Society Conference. 2006:664-669.

[19] JOHN F S. Common logic controlled english draft[EB/OL]. (2004-02-04). http://www.jfsowa.com/clce/specs.htm.

[20] SCHWITTER R. English as a formal specification language[C]// Proc of the 13th International Workshop on Database and Expert Systems Applications. 2002:228-232.