概念空间中上下位关系的意义识别研究

刘 磊[1)] 曹存根[2)] 张春霞[3)] 田国刚[2)]

[1)](北京工业大学应用数理学院 北京 100124)

[2)](中国科学院计算技术研究所 北京 100190)

[3)](北京理工大学软件学院 北京 100081)

摘 要 针对上下位关系在分类层级结构建立阶段遇到的多义性问题,给出一种概念空间中上下位关系意义识别的方法.单个概念的意义识别问题被转换为概念空间中上下位关系的意义识别.首先利用并列语境解决语境稀疏问题,获取上下位关系意义的语境.然后利用《同义词词林》对每个语境进行词义修正,以三种特征计算特征词权重,构建"关系-词"的高维向量空间,然后通过潜在语义分析降维,获取上下位关系意义的潜在语义,最后组平均聚类后得到关系的意义划分.在实验中,给出了聚类阈值自动调整函数,分析了词林和潜在语义分析的作用,实验结果证实了方法的有效性.

关键词 知识获取;上下位关系;潜在语义分析;关系获取;概念空间;意义聚类

中图法分类号 TP301 **DOI 号**: 10.3724/SP.J.1016.2009.01651

Sense Recognition Research of Hyponymy Based on Concept Space

LIU Lei[1)] CAO Cun-Gen[2)] ZHANG Chun-Xia[3)] TIAN Guo-Gang[2)]

[1)](*College of Applied Sciences, Beijing University of Technology, Beijing* 100124)

[2)](*Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190)

[3)](*School of Computer Software, Beijing Institute of Technology, Beijing* 100081)

Abstract For the polysemy of hyponymy in the phase of building taxonomic hierarchy, this paper presents a method of sense recognition of hyponymy based on concept space. The problem of sense recognition of single concept is transformed into recognition of hyponymy in concept space. Firstly, the contexts of hyponymy are acquired iteratively using coordinate relation patterns. Secondly CiLin and the weight of feature words are used to construct a hyponymy-word vector space. Then LSA is used to reduce the dimension of the vector space. In the final phase, the senses of hyponymy can be recognized using average-group clustering. The relation of decreasing degree of similarity and threshold of clustering, and the effect of CiLin and LSA in experiment are analyzed. Experimental results show that the method is adequate of partitioning the senses the hyponymy.

Keywords knowledge acquisition; hyponymy relation; latent semantic analysis; relation acquisition; concept space; sense clustering

1 引 言

上下位关系是一种基本的语义关系,常用于本体、知识库、词典的构建和验证.给定概念 c_1 和 c_2,若 c_2 的外延包含 c_1 的外延,则认为 c_1 和 c_2 具有上下位关系,称 c_2 为 c_1 的上位概念(hypernym),c_1 为 c_2 的下位概念(hyponym),记作 $ISA(c_1, c_2)$.判断 $ISA(c_1, c_2)$

收稿日期:2007-03-19;最终修改稿收到日期:2009-06-03.本课题得到国家自然科学基金(60573064,60705022,60773059)、国家"八六三"高技术研究发展计划项目基金(2007AA01Z325)和北京工业大学博士启动基金(X0006014200803)资助.**刘 磊**,男,1979 年生,博士,讲师,研究方向为知识获取、本体学习.E-mail: liuliu_leilei@bjut.edu.cn.**曹存根**,男,1964 年生,研究员,博士生导师,研究领域为知识获取与共享、文本挖掘.**张春霞**,女,1974 年生,博士,讲师,主要研究方向为知识获取、文本挖掘.**田国刚**,男,1977 年生,博士,主要研究方向为知识获取、属性获取.

是否成立的简单方法是看句子:"c_1是一种/类/个c_2"是否可以接受.例如:

ISA(中国,国家),即中国是一个国家.

ISA(生物酶,催化剂),即生物酶是一种催化剂.

上下位关系获取作为文本知识获取(Knowledge Acquisition from Text,KAT)中一个基本而又关键的问题,其获取方法可以分为两大类:一类是基于模式的方法,主要利用语言学和自然语言处理技术,通过词法分析和语法分析获取上下位关系模式,然后利用模式匹配获取上下位关系;另一类是基于统计的方法,主要基于语料库和统计语言模型,通过聚类计算概念间关联度来获取上下位关系[1-2].

所获取的上下位关系可以进一步用于建立分类层次(taxonomic hierarchy)结构.但构成上下位关系的概念的二义性问题,增加了层次结构建立的难度.实际上,概念本身并没有二义性,它能唯一地、准确地指向现实世界中的实体或对象.但在文本中,构成上下位关系的概念是由词表示的,这里称为概念词.当上下位关系从文本中提取后,其概念词已脱离了所在的语境,其概念词的意义就不易确定.因此所谓概念二义性,就是由于一个概念词可以表示多个概念引起的.例如概念词"木马"至少可以表示3种概念:

(1)木马是一种玩具.

(2)木马是一种运动器械.

(3)木马是一种病毒.

在相关研究中,Caraballo等利用连接词和同位语获取名词,通过上下文中名词的连接关系或同位关系构造名词特征向量,利用聚类得到名词间上下位关系,此外,他还利用Hearst的方法对层次结构的内部结点进行了标记[1-2].Cimiano等将句法模式、启发式规则、Web等多种方法混合在一起建立上下位关系层次结构.但他们的度量方法都没有考虑到多义词的影响[3].Rydin等以领域无关文本作为语料获取上下位关系,在建立上下位关系的分类层次结构时,利用启发式规则部分解决了多义词的影响,但启发式规则的覆盖范围比较有限[4].Grefenstette等利用WordNet验证上下位关系,同时考虑了多义词的影响,认为上位概念词的多义性可以通过WordNet中已知下位概念的语境来识别,但是对上位和下位概念词同为多义词的情况没有进一步讨论[5].后来,Kenji提出利用聚类算法自动发现多义词的意义,并以Web文档作为语料库减少数据稀疏,但没有从整体上考虑分类层次结构的建立问题[6].

我们认为借鉴自然语言处理中的词义消歧和识别研究[7],是解决概念二义性问题的重要途径.研究人员主要采用以各类词典为基础的规则方法或以大规模语料库为基础的统计方法进行词义消歧和识别.McRoy采用了一种混合方法进行词义消歧,她以词的内聚性为基础,考虑了多种可用的资源(如词典、句法标记、选择性约束等)[8].Schutze提出一种无监督的词义识别方法.此方法在多义词两种主要意义的识别中得到了较好结果.但是文中只假设一个语境是一组词的列表,而实际上语境还包含更丰富的语法、语义信息[9].

目前汉语词义消歧已有大量的研究工作[10].如刘群等提出了基于《知网》的词汇语义相似度计算方法[11],王惠提出了一种基于语法、语义知识库的汉语词义消歧策略,认为词的不同意义会在句法或词汇搭配层面上表现出不同的组合特征[12].鲁松等提出基于向量空间模型的词义消歧无导学习方法,通过计算词语权重,基于k-NN计算相似度实现词义消歧[13].

本文给出了一种在概念空间中进行上下位关系意义识别方法.本文方法是基于这样一种假设:概念词的意义选取可以由所构成上下位关系共同决定.方法的主要特点体现在以下几个方面:

(1)在给定的概念空间中,将单个概念词的意义识别问题,转化为上下位概念对的意义识别.

(2)在上下位关系的语境向量表示中,综合考虑特征词的3种权重:$Tfidf$值、词频对数似然比、词距离.

尽量保证获取更多的语言信息.并且使用启发式并列关系模式来减少语境的稀疏问题.

(3)利用《同义词词林》减少特征选取中同义特征项的影响.利用潜在语义分析将传统向量空间模型的高维词汇空间映射为低维潜在语义空间,减少了向量的噪声、冗余、歧义问题.

(4)在层次聚类中,通过分析类与类之间相似性的下降程度,给出相似度阈值的调整函数.

(5)方法具有无监督性,不需要预先标注训练集合.

本文第2节给出概念空间的基本定义和构造算法,并通过分析概念空间结构,给出上下位关系意义识别的思路;第3节详细说明上下位关系意义识别的整体算法框架;第4节通过实验分析对方法进行了评价;最后一节对本文提出方法的优缺点进行全面的总结和讨论.

2　概念空间的基本定义和构造

本文所要讨论的上下位关系的意义识别和分析都是在概念空间中进行的,为此我们下面引入概念空间的定义.

定义 1.　概念空间是一个 5 元组 $\Omega=(C,R,M,\eta,\gamma)$,其中

(1) 集合 C 称为 Ω 的概念词集合,其中的每一个概念词称为 Ω 的一个概念结点(或称结点);

(2) $R\subseteq C\times C$ 称为 Ω 的直接上下位关系集合.对 $c_1,c_2\in C$,且 $r=(c_1,c_2)\in R$,我们称 c_1 为 c_2 的直接下位,c_2 为 c_1 的直接上位.(c_1,c_2) 称为 Ω 的一条有向边,其中分别称 c_1 和 c_2 为该边的起始结点和终止结点.

(3) M 为 Ω 的意义集合.M 中的每一个元素 m 为概念词所表示的一种意义.

(4) 映射 $\eta:C\to 2^M$ 为 C 到 M 幂集上的函数,对任意 $c\in C$,$\eta(c)=\{m_1,\cdots,m_k\}$ 表示概念词 c 具有 $\{m_1,\cdots,m_k\}$ 种意义.在本文中,我们假设 $\eta(c)\neq\varnothing$,也即每个概念词至少有一个意义.

(5) 映射 $\gamma:R\to M\times M$ 为 R 到 $M\times M$ 上的函数,满足对任意 $c_1,c_2\in C$,若 $(c_1,c_2)\in R$,则 $\gamma((c_1,c_2))=(m_i,m_j)$,其中 $m_i\in\eta(c_1)$,$m_j\in\eta(c_2)$.γ 给出了概念词 c_1 和 c_2 只有在 c_1 的意义为 m_i、c_2 的意义为 m_j 时,(c_1,c_2) 才构成直接上下位关系,我们记作 $(c_1,c_2)|(m_i,m_j)$.在下文中,我们将 $\gamma((c_1,c_2))$ 简记为 $\gamma(c_1,c_2)$.

在上述定义中,C 和 R 构成的是概念词空间,M 中的每个元素实际表示了一个概念,通过映射 η,建立了概念词和概念的对应关系,通过映射 γ,使得 R 中概念词级别的上下位关系映射为概念级别的上下位关系.

定义 2.　在 Ω 中 R 具有保意义的传递性(sense-preserving transitivity),对任意 $c_1,c_2,\cdots,c_n\in C(n>2)$,如果 $(c_1,c_2)|(m_1,m_2),(c_2,c_3)|(m_2,m_3),\cdots,(c_{n-1},c_n)|(m_{n-1},m_n),m_1,m_2,\cdots,m_n\in M$ 成立,可以推出 c_1 在意义为 m_1、c_n 在意义为 m_n 时,c_1 和 c_n 上下位关系也成立,其中 c_1 为 c_n 的下位,c_n 为 c_1 的上位,记作 $(c_1,c_n)'|(m_1,m_n)$.

概念空间的构造算法如下.

算法 1.　概念空间构造算法.

输入:待分析上下位关系集合 $ISA=\{(c_1,c_2),(c_3,c_4),\cdots,(c_{n-1},c_n)\}$

输出:概念空间 Ω

1. 初始化概念空间 $\Omega=(C,R,M,\eta,\gamma)$ 为空,赋变量初值 $i=1$;

2. 若 $ISA\neq\varnothing$,读入一个关系 (c_1,c_2),否则转步 4;

(1) 若 $c_1\notin C$,$c_2\notin C$,则 $C=C\cup\{c_1,c_2\}$;$R=R\cup\{(c_1,c_2)\}$;$M=M\cup\{m_i\}\cup\{m_{i+1}\}$;$\eta(c_1)=\{m_i\}$;$\eta(c_2)=\{m_{i+1}\}$;$\gamma(c_1,c_2)=(m_i,m_{i+1})$;$i=i+2$;

(2) 若 $c_1\notin C$,$c_2\in C$,则 $C=C\cup\{c_1\}$;$R=R\cup\{(c_1,c_2)\}$;$M=M\cup\{m_i\}\cup\{m_{i+1}\}$;$\eta(c_1)=\{m_i\}$;$\eta(c_2)=\eta(c_2)\cup\{m_{i+1}\}$;$\gamma(c_1,c_2)=(m_i,m_{i+1})$;$i=i+2$;

(3) 若 $c_1\in C$,$c_2\notin C$,则 $C=C\cup\{c_2\}$;$R=R\cup\{(c_1,c_2)\}$;$M=M\cup\{m_i\}\cup\{m_{i+1}\}$;$\eta(c_2)=\{m_i\}$;$\eta(c_1)=\eta(c_1)\cup\{m_{i+1}\}$;$\gamma(c_1,c_2)=(m_{i+1},m_i)$;$i=i+2$;

(4) 若 $c_1\in C$,$c_2\in C$,$(c_1,c_2)\notin R$,则 $R=R\cup\{(c_1,c_2)\}$;$M=M\cup\{m_i\}\cup\{m_{i+1}\}$;$\eta(c_1)=\eta(c_1)\cup\{m_i\}$;$\eta(c_2)=\eta(c_2)\cup\{m_{i+1}\}$;$\gamma(c_1,c_2)=(m_i,m_{i+1})$;$i=i+2$;

(5) 若 $(c_1,c_2)\in R$,则不作处理;

3. $ISA=ISA-\{(c_1,c_2)\}$,转步 2;

4. 输出概念空间 Ω.

为了便于分析,表 1 中给出了一组含有多义概念词的上下位关系,首先以表 1 中关系为输入,由算法 1 建立概念空间,见图 1.图中含有 9 个概念词(以圆圈表示)、8 个上下位关系(以箭头表示)、16 个意义 $\{m_1,\cdots,m_{16}\}$、9 个概念词的意义映射(以大括号标出)、8 个关系的意义映射(在箭头上标出).从图可以看出,“病毒”、“木马”、“蠕虫”、“程序”都是多义概念词.正是由于这些概念词的多义性问题,无法保证关系传递性成立,例如已知(流感,病毒)$|(m_9,m_{10})$ 和(病毒,程序)$|(m_8,m_7)$ 成立,不能确定(流感,程序)$'|(m_9,m_7)$ 是否成立,因为不知道是否 $m_{10}=m_8$;同样已知(蠕虫,病毒)$|(m_{11},m_{12})$ 和(病毒,程序)$|(m_8,m_7)$ 成立,也不能确定(蠕虫,程序)$'|(m_{11},m_7)$ 是否成立.因此只有识别出概念词在不同关系中所取意义哪些是相同的(如“病毒”的意义 $m_3=m_8=m_{12}$,$m_6=m_{10}$),哪些是不同的(如“木马”的意义 $m_1\neq m_4$,“蠕虫”的意义 $m_{11}\neq m_{14}$),才能根据保意义的传递性特征建立正确的上下位关系层次结构.

表 1　一组上下位关系

ISA(木马,玩具)
ISA(木马,病毒)
ISA(病毒,生物)
ISA(病毒,程序)
ISA(流感,病毒)
ISA(蠕虫,病毒)
ISA(蠕虫,动物)
ISA(审判,程序)

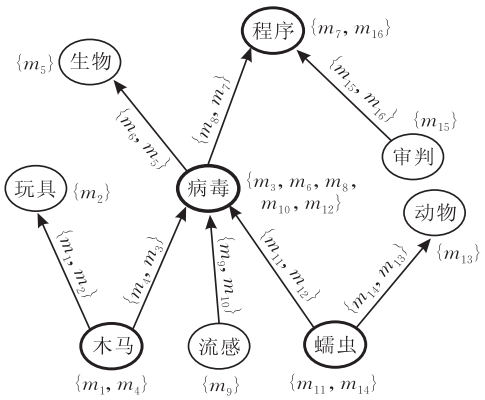

图 1 概念空间示例

对于一个关系$(c,c')|(m,m')$，我们认为概念词 c 和 c' 的意义选取是由所构成上下位关系共同决定的，即 c 决定了 c' 的意义选取，同时 c' 也决定了 c 的意义选取. 这样单个概念词的意义识别问题，可以转化为上下位概念对的意义识别. 以图 1 中“病毒”的意义 m_3 和 m_8 识别为例，判断 m_3 和 m_8 是否相等可以转化为判断上下位关系意义(m_4,m_3)和(m_8,m_7)是否相等. 在已知 m_3 和 m_8 所对应的概念词相同(都是“病毒”)的前提下，若能判断$(m_4,m_3)=(m_8,m_7)$，则认为 $m_3=m_8$，意义合并得到(木马，病毒)|(m_4,m_3)，(病毒，程序)|(m_3,m_7)，根据保意义的传递性，则推出(木马，程序)′在关系意义(m_4,m_7)下成立.

上下位关系意义的识别需要考虑 3 个问题：(1)选取哪些上下位关系进行意义识别？(2)上下位关系的意义如何表示？(3)如何进行上下位关系意义识别？

上下位关系意义识别的选取以概念空间中的点为中心，由定义 3 给出.

定义 3. 在 Ω 中，对 $c\in C$，所有关联于结点 c 的边称为 c 的意义识别集，用 $\rho_h(c)$ 表示，边的数目记作$|\rho_h(c)|$，c 称为 $\rho_h(c)$ 的核心概念词.

如 ρ_h(病毒)=｛(木马，病毒)，(病毒，生物)，(病毒，程序)，(流感，病毒)，(蠕虫，病毒)｝.

对于第二个问题，由于上下位关系的意义是由其上位概念词和下位概念词的意义决定的，而概念词的意义一般与其所在的语境密切相关，因此可以假设对于一个关系$(c,c')|(m,m')$，(m,m')是由 c 和 c' 共现的语境特征决定的，可以利用 c 和 c' 的共现语境构造(m,m')的特征向量，作为(m,m')的一种语义表示.

而根据 Miller 等提出的语境假设(如果两个词的语境相似，则两个词的语义相关)[14]，可以假设如果两个关系的语境相似，则两个关系的意义相似. 这样，关系的意义识别可以看作是表示上下位关系意义的语境的聚类过程：即上下位关系的意义识别问题，就是对每个 $\rho_h(c)$ 进行上下位关系意义聚类的过程. 如 ρ_h(病毒)就应该被聚为两类｛(木马，病毒)，(病毒，程序)，(蠕虫，病毒)｝和｛(病毒，生物)，(流感，病毒)｝，进而推出 $m_3=m_8=m_{12}$，$m_6=m_{10}$. 根据保意义的传递性得到 ISA(蠕虫，程序)和 ISA(流感，生物)，如图 2 所示.

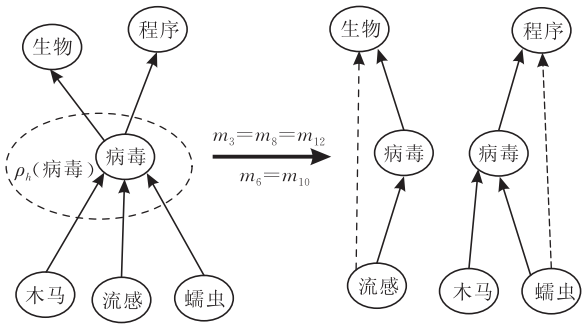

图 2 “病毒”的意义识别

3 上下位关系意义识别方法

在上下位关系意义识别中，我们借鉴了词义消歧中的方法，算法的整体框架如图 3 所示. 首先，从概念空间中提取意义识别集，然后获取意义识别集中每个上下位关系的语境，从语境中选取合适的语

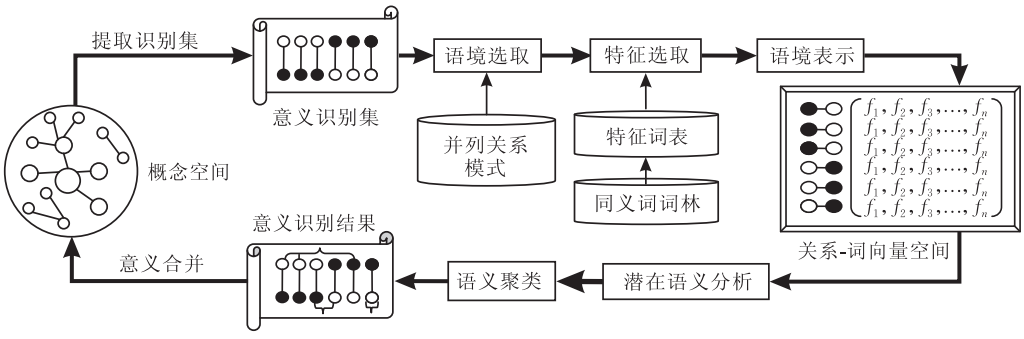

图 3 上下位关系意义识别过程

境特征,构造“关系-词”向量矩阵,再利用潜在语义分析技术,通过奇异值分解构造“关系-语义”向量矩阵,然后利用层次聚类将语义相似的关系聚为一类,认为每个类都表示上下位关系的一种特定意义,最终利用意义识别结果对概念空间的意义进行合并.

在整个流程中,还采用了一些辅助策略,包括利用启发式并列关系模式减少语境获取中的数据稀疏问题,利用《同义词词林》(以下简称为《词林》)减少特征选取中同义特征项的影响.下面分小节介绍意义识别集 $\rho_h(c)$ 的识别算法,并以 ρ_h(病毒)中 5 个关系为例说明.

3.1　语境选取

$\rho_h(c)$ 中每个上下位关系的语境选取与通常单个词的语境选取略有不同,在语境选取时需要考虑上位概念词和下位概念词的距离问题.设 $(c_1,c_2)\in\rho_h(c)$,(c_1,c_2) 的语境来源于语料中同时蕴含 c_1 和 c_2 的所有文档,且满足条件:(1) c_1 和 c_2 的位置距离 $d_1<\alpha$,(2)每个概念词前后限定的距离 $d_2<\beta$,其中 α,β 为整数阈值,语境集合记作 $CT(c_1,c_2)=\{ct_1,ct_2,\cdots,ct_n\}$,其中 ct_i 称为 (c_1,c_2) 的第 i 个语境项.这里,距离用出现的特征词数目表示,关于特征词将在 3.2 节详细介绍.

语境选取要求上、下位概念词必须同时出现在一篇文档中,这可能会导致语境稀疏问题,因此我们使用启发式并列关系模式增加关系的语境数目.首先给出并列关系模式的定义.

定义 4.　称一个模式为并列关系模式,如果从满足模式的句子中能提取具有并列关系的概念词.若两个概念词 c_1,c_2 具有并列关系,记作 $CR(c_1,c_2)$.

下面列举了 3 种并列关系模式的 BNF 表示,其中〈?C1〉和〈?C2〉表示模式变量.注意,我们这里使用了一种非规范的表示——〈$t_1|t_2$〉,按照 BNF〈$t_1|t_2$〉::=$t_1|t_2$,表示“或”关系.给定一个匹配并列关系模式的句子,若〈?C1〉中存在概念词 c_1,〈?C2〉中存在概念词 c_2,则认为 $CR(c_1,c_2)$ 成立.

模式 1.　〈?C1〉{〈、|或者|或是|或|及|和|与〉〈?C2〉}*〈是|为|指|即〉.

模式 2.　〈?C1〉{〈、|或者|或是|或|及|和|与〉〈?C2〉}*〈各|之|这〉〈种|类|些|样|流〉〈的〉.

模式 3.　〈如|象|包括|分|包含|囊括|涵盖|有|是|指|即〉〈?C1〉{〈、|或者|或是|或|及|和|与〉〈?C2〉}*〈等〉.

下面的例句匹配了模式 3,能够得到并列关系 CR(水稻,玉米,红薯,烟叶).

农作物主要/有/{{水稻}$_{c_1}$}$_{〈?C1〉}$、{{玉米}$_{c_2}$}$_{〈?C2〉}$、{{红薯}$_{c_3}$}$_{〈?C3〉}$、{{烟叶}$_{c_4}$}$_{〈?C4〉}$/等/.

利用并列关系模式,$\rho_h(c)$ 中每个关系的增量语境都可以按照算法 2 获取.

算法 2.　利用并列关系模式增量获取语境算法.

输入:语料库 G,并列关系模式集 CRP,上下位关系 (c,c')

输出:(c,c') 的增量语境

1. 初始化下位概念 c 的并列概念词集合 Q 为 $\{c\}$,并为 c 增加未处理标记;

2. 从 Q 中取一个未处理过的概念 c_i,在 G 中搜索同时满足下列条件的句子 s,组成句子集合 S:(1) c_i,c' 同时在 s 中出现;(2) s 满足并列关系模式 $crp\in CRP$;(3) c_i 是并列概念词之一;

3. 从每一个 $s\in S$ 中获取 c_i 的并列概念词 w,若 $w\notin Q$,则将 w 加入到 Q 中,且 w 标记为未处理;

4. 若 Q 中存在未处理过的概念,返回步 2,否则以 $Q=\{c,c_1,\cdots,c_k\}$ 中的概念词为下位,构成上下位关系集合 $R=\{(c,c'),(c_1,c'),\cdots,(c_k,c')\}$;

5. 对 R 中的每一个关系 r,从 G 中获取 r 的语境 $CT(r)$.

6. (c,c') 的增量语境 $CTU(c,c')=\bigcup CT(r),r\in R$.

另外引入语境合并规则:对 $(c_1,c),(c_2,c)\in\rho_h(c)$,若 $CR(c_1,c_2)$ 成立,则认为 (c_1,c) 和 (c_2,c) 的意义相同,直接将它们的语境合并,即

$$CTU(c_1,c)=CTU(c_2,c)=CTU(c_1,c)\bigcup CTU(c_2,c).$$

以 ρ_h(病毒)中(蠕虫,病毒)为例,已知如下例句:

例句 1.　哪个杀毒软件对蠕虫、欢乐时光等病毒最有效.

例句 2.　连接指定站点,下载其它木马、蠕虫、后门等病毒.

则可以通过算法 2 从例句中首先获取“蠕虫”的并列概念词“欢乐时光”、“木马”、“后门”,再将并列概念词与“病毒”所构成上下位关系的语境作为增量语境加入到(蠕虫,病毒)的语境中.而且由于(蠕虫,病毒),(木马,病毒)$\in\rho_h$(病毒),CR(蠕虫,木马)成立,根据语境合并规则得到

$$\begin{aligned}CTU(\text{蠕虫,病毒})&=CTU(\text{木马,病毒})\\&=CTU(\text{蠕虫,病毒})\bigcup CTU(\text{木马,病毒}).\end{aligned}$$

3.2　特征词表

语境中的特征信息以特征词表示.特征词 w 是句子中的符合下列条件的词:(1) w 是实体词(即词性为名词、动词、形容词等);(2)语料库 G 中含有词条 w 的文档数 $Df_w>\alpha$;(3) G 中 w 的倒排文档频度 $Idf_w>\beta$;其中 $\alpha,\beta\in R$ 是指定的阈值,Idf_w 的定义见式(1),其中,$|G|$ 为 G 中的文档总数.

$$Idf_w=\log\left(\frac{|G|}{Df_w}\right)\tag{1}$$

特征词表的构造步骤为:首先以一个候选特征词集合(如分词词表)为输入,选取满足特征词定义的词构成候选特征词表,记作 $FWL=\{w_1,w_2,w_3,\cdots,w_n\}$,然后引入《词林》[15-16]将互为同义词的特征词合并,以减少同义词对语境的语义度量的影响.《词林》是现代比较常用的词义分类词典,收词近7万,描述了一个由宽泛概念到具体词义的5层语义分类体系.我们取第5层作为同义特征词合并的依据.这样特征词表变为 $FWL=\{Syn(w_1),Syn(w_2),\cdots,Syn(w_k)\}$,其中 $Syn(w_i)$ 为特征词 w_i 的同义词集合.例如:

Syn(树林)={树林,丛林,林,林海,森林,树丛,树林子},

Syn(风流人物)={风流人物,风云人物,名家,名流,名人,名士,名宿,头面人物,闻人,知名人士}.

3.3 特征向量空间

$\rho_h(c)$ 中每个关系的语境特征向量以特征词为基础构造,其特征向量的每一维为特征词的权重,权重计算考虑了3个因素:$Tfidf$ 值、词频对数似然比、词距离.

(1) $Tfidf$ 值.用于体现特征词在整个语料中对文档的重要性.$Tfidf_{w,d}$ 的计算见式(2),其中 $Wef_{w,d}$ 是特征词 w 在一篇文档 $d\in G$ 中出现的次数;Idf_w 是 w 在 G 中的 Idf 值;$Tfidf_{w,d}$ 是 w 对于文档 d 的 $Tfidf$ 值.

$$Tfidf_{w,d}=Wef_{w,d}\times Idf_w \tag{2}$$

(2)词频对数似然比.用于体现特征词在语境中对文档的重要性.假设与上下位关系意义相关的特征词,其在语境中出现频率要比在整个语料库中要高,则特征词 w 对于关系 r 的词频对数似然比为式(3)[17].

$$Logw(w,r)=\min\{1,|CTU(r)|/10\}\times\log\left(\frac{Pr(w|r)}{Pr(w)}+1\right) \tag{3}$$

这里 $Pr(w)$ 为 w 在整个语料库中出现的频率,$Pr(w|r)$ 是 w 在 r 的语境中出现的频率,$|CTU(r)|$ 是 r 的语境数目.使用 $\min\{1,|CTU(r)|/10\}$ 是为了避免 r 的语境稀疏所引起的错误估计.

(3)词距离.在语境中,假设特征词 w 与关系 $r(c,c')$ 中概念词的距离越近,重要性越高,则距离计算见式(4).这里 $dis(w,c)$ 指 w 与 c 间隔的特征词个数.

$$Disw(w,r)=10/\max\{10,\min\{dis(w,c),dis(w,c')\}\} \tag{4}$$

设关系 $r(c,c')$,语境 $CTU(c,c')=\{ct_1,ct_2,ct_3,\cdots,ct_m\}$,特征词列表 $FWL=\{Syn(w_1),Syn(w_2),\cdots,Syn(w_k)\}$,则 r 的语境特征向量 $\boldsymbol{ConVec}(c,c')$ 按式(5)~(7)计算:其中式(5)计算特征词在 r 的语境中的特征权重值,由于3个权重因素的计算公式没有归一化,所以采用三者的乘积作为权重值;式(6)计算特征词的同义词集合在 r 的语境中的特征权重值,这里通过计算同义词集合中每个同义词的权重值之和得到;式(7)计算关系 r 的语境特征向量,向量的每一维都是一个同义词集合在语境中的特征权重值.

$$Unionw(w_i,r)=Tfidf(w_i)\times Logw(w_i,r)\times Disw(w_i,r) \tag{5}$$

$$Unionw(Syn(w_i),r)=Unionw(w_{i_1},r)+Unionw(w_{i_2},r)+\cdots+Unionw(w_{i_n},r) \tag{6}$$

$$\boldsymbol{ConVec}(r)=(Unionw(Syn(w_1),r),Unionw(Syn(w_2),r),\cdots,Unionw(Syn(w_k),r)) \tag{7}$$

$\rho_h(c)$ 中每个关系都可以通过上述公式得到语境向量,这样以 $\rho_h(c)$ 中的关系为行,以特征词为列,就构成了一个高维"关系-词"空间特征矩阵 $\boldsymbol{A}_{m\times n}$,其中 m 为关系数,n 为特征项数,$a_{ij}\in\boldsymbol{A}_{m\times n}$ 是关系 i 对应特征项 j 的特征权重值.若矩阵某一列的数值皆为0,则去掉此列.语境向量空间示例见表2,为便于显示,ρ_h(病毒)的5个关系用行显示,同义特征词集用列显示.

表2 语境向量空间示例

		特征权重值				
		(病毒,生物)	(病毒,程序)	(木马,病毒)	(蠕虫,病毒)	(流感,病毒)
1	Syn(人类)	5.286264	0.042735	0	0.065161	2.708873
2	Syn(报警)	0	0.940581	0.550891	0.009923	0
3	Syn(病人)	11.357896	0.002186	0	0	11.175187
4	Syn(磁盘)	0	1.378750	0.528703	0.005281	0
5	Syn(机体)	7.577363	0.047664	0	0	0.966117
6	Syn(微生物)	4.306056	0	0	0.026196	8.290844
…	…	…	…	…	…	…

3.4　潜在语义分析

语境向量空间矩阵 $\boldsymbol{A}_{m\times n}$将进行潜在语义分析.潜在语义分析(Latent Semantic Analysis,LSA)作为传统向量空间模型的进一步改进,通过统计方法提取并量化潜在语义结构,将传统向量空间模型的高维词汇空间映射为低维潜在语义空间,减少了向量的噪声、冗余、歧义问题[18].其中奇异值分解(Singular Value Decomposition,SVD)是目前常用的 LSA 空间构造方法.它利用矩阵的奇异值分解降秩技术,提取 k 个最大的奇异值及其对应的奇异矢量构成新矩阵来近似表示原矩阵.

$\boldsymbol{A}_{m\times n}$的潜在语义分析过程为:首先利用 SVD 将 $\boldsymbol{A}_{m\times n}$中隐含的语义关系分解成线性独立的向量,得到 3 个满秩矩阵:$\boldsymbol{A}_{m\times n}=\boldsymbol{U}_{m\times q}\boldsymbol{S}_{q\times q}(\boldsymbol{V}_{n\times q})^{\mathrm{T}}$,其中 $\boldsymbol{S}_{q\times q}$为奇异值对角矩阵,$\boldsymbol{S}_{q\times q}=\mathrm{diagram}\{\lambda_1,\lambda_2,\cdots,\lambda_q\}$,$q=\min(m,n)$,$\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_q\geq 0$,且当 $1\leq i\leq r$ 时,$\lambda_i>0$,当 $i>r$ 时,$\lambda_i=0$,这里 $r=\mathrm{rank}(\boldsymbol{A}_{m\times n})$;$(\boldsymbol{U}_{m\times r})^{\mathrm{T}}\boldsymbol{U}_{m\times r}=I$,$(\boldsymbol{V}_{n\times r})^{\mathrm{T}}\boldsymbol{V}_{n\times r}=I$,$\boldsymbol{U}_{m\times r}$和 $\boldsymbol{V}_{n\times r}$的列分别称为 $\boldsymbol{A}_{m\times n}$的左奇异值向量和右奇异值向量.然后选取 $\boldsymbol{S}_{q\times q}$前 k 个最大的奇异值,其余的设置为 0,通过删除 $\boldsymbol{S}_{q\times q}$相应的行和列获得新的对角矩阵 $\boldsymbol{S}_{k\times k}$,同时保留 $\boldsymbol{U}_{m\times q}$和 $\boldsymbol{V}_{n\times q}$中相应的行和列,获得矩阵 $\boldsymbol{U}_{m\times k}$和 $\boldsymbol{V}_{n\times k}$,删减矩阵相乘后得到接近原始矩阵 $\boldsymbol{A}_{m\times n}$的近似矩阵 $\boldsymbol{A}'_{m\times n}=\boldsymbol{U}_{m\times k}\boldsymbol{S}_{k\times k}(\boldsymbol{V}_{n\times k})^{\mathrm{T}}$.

$\boldsymbol{A}'_{m\times n}$包含了 $\boldsymbol{A}_{m\times n}$的主要特征(即关系与特征项的语义关系),这样词级别的语境特征空间 $\boldsymbol{A}_{m\times n}$经过 LSA 后就转换为语义级别的向量空间 $\boldsymbol{A}'_{m\times n}$.语义向量空间示例由表 2 经过潜在语义分析($k=2$)得到,如表 3 所示.

表 3　语义向量空间示例

		特征值				
		(病毒,生物)	(病毒,程序)	(木马,病毒)	(蠕虫,病毒)	(流感,病毒)
1	*Syn*(人类)	4.7982	−0.1550	−0.4627	−0.5002	3.2647
2	*Syn*(报警)	−0.1912	0.0641	0.1849	0.2051	0.2174
3	*Syn*(病人)	11.6487	0.1098	0.2750	0.3405	10.8441
4	*Syn*(磁盘)	−0.2178	0.0731	0.2109	0.2338	0.2481
5	*Syn*(机体)	6.1931	−0.4786	−1.3981	−1.5362	2.5421
6	*Syn*(微生物)	5.3939	0.3892	1.1004	1.2396	7.0523
…	…	…	…	…	…	…

3.5　语义聚类

最后对 $\boldsymbol{A}'_{m\times n}$采用凝聚法聚类,以 cosine 距离作为相似性度量方法,以组平均策略(group-average)作为计算策略.假设两个关系 r_1,r_2分别具有语境特征向量 $\boldsymbol{F}_1$和 $\boldsymbol{F}_2$,则 cosine 距离表示见式(8).组平均策略是指两个类的相似度是类中成员的平均相似度.设两个类为 $CL_1=\{r_{a_1},r_{a_2},\cdots,r_{a_m}\}$,$CL_2=\{r_{b_1},r_{b_2},\cdots,r_{b_n}\}$,则类间相似度计算见式(9).在聚类最开始阶段,每个关系各自是一类,首先将最相似的两类合并成一个新类,然后迭代合并最相似的两个类,直到要合并的类的相似性小于规定的相似度阈值为止.聚为一类的关系将看作是意义相同的,会在概念空间中进行意义合并.

$$Sim(r_1,r_2)=\cos(\boldsymbol{F}_1,\boldsymbol{F}_2)=\sum_{f_1\in F_1,f_2\in F_2}f_1f_2\Big/\sqrt{\sum_{f_1\in F_1}f_1^2\sum_{f_2\in F_2}f_2^2}\tag{8}$$

$$SIM(CL_1,CL_2)=\frac{\sum_{i=1}^{m}\sum_{j=1}^{n}sim(r_{ai},r_{bj})}{m\times n}\tag{9}$$

我们对 $\boldsymbol{A}_{m\times n}$和 $\boldsymbol{A}'_{m\times n}$分别进行相似性计算,见表 4.表 4 中括号内数据为 $\boldsymbol{A}_{m\times n}$的相似性结果,从表中可以看出通过 LSA,*Sim*((蠕虫,病毒),(病毒,程序))从 0.584 提高到 0.697,*Sim*((病毒,程序),(病毒,生物))从 0.019 降低到−0.018.这说明 LSA 使得意义相同关系的相似性提高,反之降低.最后通过凝聚法聚类,ρ_h(病毒)关系被聚为两类(相似度阈值=0.4),如图 4 所示.

表 4　相似性比较

	相似度				
	(病毒,生物)	(病毒,程序)	(木马,病毒)	(蠕虫,病毒)	(流感,病毒)
(病毒,生物)		−0.018(0.019)	−0.046(0.000)	−0.037(0.007)	0.855(0.830)
(病毒,程序)	**−0.018(0.019)**		0.790(0.720)	0.697(0.584)	0.130(0.150)
(木马,病毒)	−0.046(0.000)	0.790(0.720)		0.983(0.972)	0.205(0.282)
(蠕虫,病毒)	−0.037(0.007)	**0.697(0.584)**	0.983(0.972)		0.220(0.265)
(流感,病毒)	0.855(0.830)	0.130(0.150)	0.205(0.282)	0.220(0.265)	

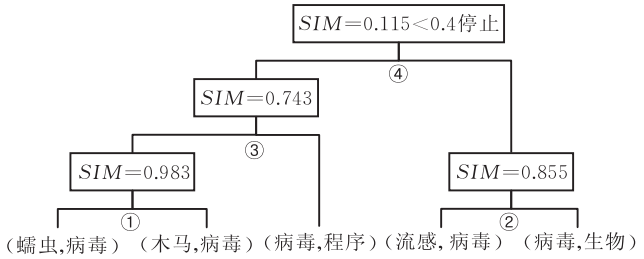

图 4 ρ_h(病毒)意义聚类过程

4 实验及评价

4.1 实验数据和结果

(1)特征词表构造过程. 首先以《计算所分词词典》(116736 词)作为候选特征词,随机选取约 180 万篇 Web 网页作为训练语料,根据特征词的选取条件($Df_w>1$;$Idf_w>4.0$)得到 85816 个特征词,然后利用《词林》合并同义特征词,最终得到 64661 个同义特征词集合,其示例如表 5 所示. 由于在《词林》中出现的部分词(6062 个)具有多个词义,因此一个特征词可能同时属于多个同义特征词集合. 如表 5 中"先生"同时属于 Syn(老师)和 Syn(大夫)两个同义特征词集合.

表 5 部分同义特征词集合

特征词集合	同义特征词列表
Syn(老师)	老师|导师|讲师|讲席|教师|教书匠|教员|师长|先生|园丁
Syn(大夫)	大夫|郎中|先生|医|医生|医师
Syn(编制)	编制|建制|体制
Syn(便宜)	便宜|低廉|公道|贱|克己|廉|廉价

(2)聚类的评价指标. 准确率(Precision)、召回率(Recall)、F 分值(F-Measure),其计算见式(10). 由于只有聚为一类的关系才看作是能意义合并的,因此更需要综合考虑准确率和召回率,即 F 分值的取值. 被聚类关系是否正确由人工评价.

准确率(P)=聚类正确的关系总数/被聚类的关系总数,

召回率(R)=聚类正确的关系总数/关系总数,

F 分值(F)=$2\times P\times R/(P+R)$ (10)

(3)上下位关系意义识别过程. 首先利用基于模式的方法获取一组上下位关系[15,18],根据算法 1 生成概念空间. 我们从中选取了以"病毒""程序""木马"等为核心概念的 12 个意义识别集进行意义识别,以特征词词典和 4G 的 Web 测试语料为输入,经过本文算法(参数取值:语境窗口 $d_1<50$,$d_2<50$;LSA 中 k 取 10)得到识别结果见表 6. 其中第 3 列括号内为经人工评价正确的关系数目,第 7 列 $MAX(F)$表示聚类的最大 F 分值. 需要说明的是,聚类时不知道核心概念的真正意义有多少种,因此聚类可以看作是一种无导的词义识别过程. 由于聚类阈值不能直接指定,我们输出了相似度阈值范围在[0,1]、阈值步长为 0.01 的所有结果,希望从实验结果中找到阈值估计方法.

表 6 实验数据和结果

序号	核心概念	关系数目	部分关系实例	实例语境数	实例语境向量非零维数	$MAX(F)$	$MAX(F)$时相似度阈值	核心概念的多义性解释
1	病毒	67(56)	(病毒,生物) (病毒,程序) (爱虫,病毒) (流感,病毒)	49 877 207 46	2926 8176 3372 557	0.831	0.61～0.64	计算机病毒 微生物
2	程序	97(92)	(进程,程序) (审判,程序) (触发器,程序)	274 145 26	5180 3395 327	0.726	0.20～0.37	计算机软件 事件次序
3	木马	12(11)	(木马,病毒) (木马,玩具) (特洛伊,木马)	195 23 178	3450 173 3257	0.791	0.73～0.74	玩具 计算机病毒 体育器械
4	语言	156(138)	(汉语,语言) (梵语,语言) (java,语言)	205 9 85	6224 785 2576	0.812	0.81～0.83	计算机语言 人类语言
5	容器	37(31)	(罐,容器) (锅,容器) (按钮,容器)	35 25 124	455 402 3655	0.764	0.35～0.38	计算机术语 真实容器
6	传奇	35(29)	(传奇,网络游戏) (诸葛亮,传奇)	48 6	795 153	0.731	0.66～0.70	一款游戏 人生、事件的传奇
7	蠕虫	52(47)	(蠕虫,病毒) (蠕虫,动物)	670 56	5103 3122	0.823	0.65～0.69	计算机病毒 一种生物
8	大夫	14(12)	(大夫,古代官名) (白求恩,大夫)	34 26	766 541	0.782	0.31～0.35	古代官名 医生

（续　表）

序号	核心概念	关系数目	部分关系实例	实例语境数	实例语境向量非零维数	$MAX(F)$	$MAX(F)$时相似度阈值	核心概念的多义性解释
9	分子	19(16)	(水分子,分子) (分子,数学概念) (知识分子,分子)	52 23 18	2421 385 392	0.802	0.41～0.47	数学上分子 化学上分子 归属整体的个体
10	变态	21(18)	(虐杀动物,变态) (变态,动物性本能)	16 25	296 375	0.733	0.27～0.32	生物不同形态变化 人心理不正常
11	龙骨	10(8)	(龙骨,药材) (龙骨,船构件)	19 6	302 134	0.754	0.26～0.34	中药材 船的主要部件
12	媒体	42(38)	(视听,媒体) (人民日报,媒体)	112 562	3755 4522	0.773	0.79～0.83	媒介,手段 专指新闻媒体

4.2　数据分析

4.2.1　聚类相似度阈值的调整

上述实验中每组取 $MAX(F)$时,所取聚类相似度阈值皆不相同,因此将此算法真正用于概念空间中所有上下位关系意义聚类时,不可能直接指定一个固定的相似度阈值.通过分析聚类结果,我们发现算法在迭代合并最相似的两个类 CL_1 和 CL_2 时,如果这两个类实际是代表了两种关系意义时,其相似性会迅速降低.例如:

CL_1＝{(丁型肝炎病毒,病毒),
(口蹄疫,病毒),…,(噬菌体,病毒)},
CL_2＝{(网页恶意代码,病毒),
(病毒,程序),…,(爱虫,病毒)},
$SIM(CL_{11},CL_{12})=0.824552$,
CL_{11}和 CL_{12}为 CL_1的子类,
$SIM(CL_{21},CL_{22})=0.676623$,
CL_{21},CL_{22}为 CL_2的子类,
$SIM(CL_1,CL_2)=0.335438$.

因此我们通过相似度的下降程度动态调整相似度聚类阈值.每次迭代的相似度阈值为

$$\min\{SIM(CL_{11},CL_{12}),SIM(CL_{21},CL_{22})\}/2.$$

通过此公式计算其它意义识别集(如以“语言”、“容器”、“传奇”为核心概念),发现可以近似得到 $MAX(F)$.

4.2.2　《词林》和 LSA 的作用

在聚类中我们采用《词林》和 LSA 改进关系语境,这里以 12 个意义识别集的平均准确率为例,说明《词林》和 LSA 对聚类结果的影响.我们分别采用三种方法进行聚类:(M1)语境不经过词林和 LSA 处理;(M2)语境只经过词林处理;(M3)语境经过词林和 LSA 处理.聚类结果的准确率数据比较见图 5,可以看出在阈值[0.09,0.70],M2 和 M3 的准确率明显比 M1 高,表明了词林和 LSA 的改进作用,但在阈值 0.70 之后,M1、M2 和 M3 的准确率接近,这说明词林和 LSA 对语境相似度高的关系没有明显改进.在阈值[0.21,0.37],M3 的准确率明显比 M2 高,但在阈值 0.37 后,M2 和 M3 的准确率接近,这表明 LSA 对于原来相似度低的关系的聚类有明显改善.但对本身语境相似度高的关系影响不大.另外,通过分析关系两两相似性结果,发现 LSA 和词林会使部分原本意义不同的关系相似性降低,意义相同的关系相似性升高,这将有利于上下位关系意义识别.

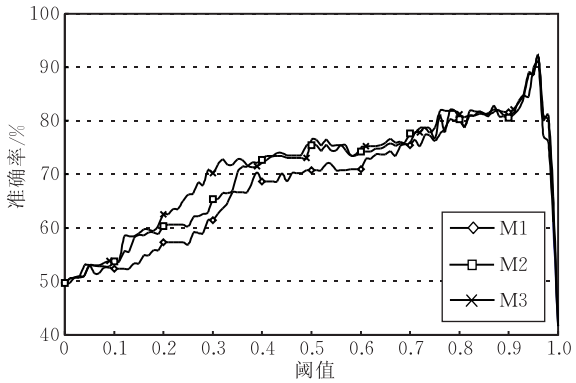

图 5　语境处理方法准确率比较

4.2.3　与其它方法的比较

我们将本文方法分别与 Rydin 和 Kenji 所提出的方法做了对比.Rydin 是利用启发式规则解决多义词问题,对多义概念词 c,则意义合并规则为:(1)利用模式从同一句子中获取 c 的下位集合,其意义相同.(2)对从不同句子中获取的 c 的两个下位集合分别为$\{c_1,c_2,\cdots,c_n\}$和$\{c'_1,c'_2,\cdots,c'_m\}$,若两个集合交集不为空,则其意义相同[4].Kenji 提出利用聚类算法自动发现多义词的意义,对多义概念词 c 的任意两个上位为 c_1和 c_2,首先以 c_1和 c_2的所有下位的并集,构建其对应的特征向量 $\boldsymbol{f}_1$和 $\boldsymbol{f}_2$,并利用 cosine 公式计算相似性,然后用单链策略对 c 所有上位层次聚类,在指定阈值停止.

我们以“病毒”“程序”“木马”等为核心概念的

12 个意义识别集，4G 的 Web 测试语料为输入，分别用 Rydin、Kenji 和本文方法进行意义识别. 由于 Rydin 方法只支持多义词为上位，Kenji 方法只支持多义词为下位的情况，因此对于其不能处理的关系没有输入，其实验结果见表 7.

表 7 不同方法比较

方法	平均准确率	平均召回率	平均 F 分值
Rydin	0.942	0.235	0.376
Kenji	0.854	0.542	0.663
本文方法	0.802	0.724	0.761

通过比较可以看出，Rydin 方法虽然准确率很高(0.942)，但召回率很低(0.235)，其原因是其意义合并的规则非常强，但符合规则的下位毕竟太少. Kenji 方法需要获取多义概念词的每个上位的下位集合信息，虽然规则较强，但存在下位稀疏的问题. 而本文提出的方法总体上具有更好 F 分值(0.761)且改进幅度较大，并且同时支持多义概念词为上位或着为下位的情况. 另外在上下位关系意义识别的过程中，也将一些错误上下位关系识别出来. 这些错误关系一般是一种比喻，例如(病毒，毒瘤)、(流行，病毒)，这是由于这些错误关系的语境文本与其它正确关系的文本差别很大而没有被聚类.

5 结论和讨论

目前上下位关系获取的研究大多数都是基于概念词级别的. 但必须区分出这些概念词的实际含义，将其转换为概念级别上下位关系，才能用于扩充和构建本体和知识库. 对于概念词的多义性问题，一般会借助已有的语义词典，如 WordNet、MindNet、HowNet 等来解决，但是这些词典很少收录专业领域方面的词汇. 当上下位关系获取的来源为领域无关的自由文本时(如 Web 文本)，许多概念词在词典找不到对应的词义解释. 因此需要更多的方法识别概念词在不同上下位关系中的意义.

本文在上下位关系的意义识别研究中，首先给出了上下位关系的概念空间结构，提出概念词的意义选取可以由所构成上下位关系共同决定，将单个概念词的意义识别问题，转化为上下位概念对的意义识别. 然后根据 Miller 等提出的语境假设，综合利用了向量空间模型、《词林》、LSA、层次聚类等方法和资源，给出了一种无导的上下位关系意义识别方法. 而且根据算法已经给出的上下位关系意义的向量表示，当有新的关系加入概念空间时，可以用类似信息检索方式将新的关系意义归到意义最接近的关系意义中.

《词林》虽然存在词量不足、粒度过大等问题[13]，但对减少特征选取中同义特征项的影响起到了一定的作用；而 LSA 将高维词汇空间映射为低维潜在语义空间，减少了向量的噪声. 实验证明本文提出的思想和方法是可行的. 但一些问题仍然存在，需要进一步解决.

(1) 语境选取错误. 通过分析一些关系的语境发现，由于分词、未登录词识别等错误，使得语境不是关系的真正语境.

(2) 语境稀疏问题. 虽然采用并列关系模式使语境数目增加了近 50%，但仍然存在稀疏问题，可以考虑利用搜索引擎 API 在线搜索数据作为数据来源.

(3) 语义信息丢失. 虽然利用 Tfidf 值、词频对数似然比、词距离三个特征词权重表示语境信息，但仍然丢失许多语义信息. 如果将词性、语序，甚至概念词在语境中的属性考虑进来，可能效果会更好.

(4) 普通概念问题. 由于普通上位概念的下位常常属于多个领域，算法对其的意义聚类效果不明显，如普通概念“问题”、“方法”、“现象”等.

参 考 文 献

[1] Hearst Marti A. Automatic acquisition of hyponyms from large text corpora//Proceedings of the 14th International Conference on Computational Linguistics. Nantes, France, 1992: 539-545

[2] Caraballo Sharon A. Automatic construction of a hypernym-labeled noun hierarchy from text//Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics. Maryland, USA, 1999: 120-126

[3] Cimiano P, Pivk A, Schmidt Thieme L, Staab S. Learning taxonomic relations from heterogeneous evidence//Proceedings of the ECAI-2004 Workshop on Ontology Learning and Population. Valencia, Spain, 2004: 1-6

[4] Rydin S. Building a hyponymy lexicon with hierarchical structure//Proceedings of the ACL-02 Workshop on Unsupervised Lexical Acquisition. Philadelphia, USA, 2002: 26-33

[5] Grefenstette Gregory, Hearst M A. A method for refining automatically-discovered lexical relations: Combining weak techniques for stronger results//Proceedings of the AAAI Workshop on Statistically-Based Natural Language Programming Techniques. Stanford, USA, 1992: 64-72

[6] Kenji Miura, Yoshimasa Tsuruoka, Junichi Tsujii. Automatic acquisition of concept relations from Web documents

with sense clustering//Proceedings of the International Joint Conference on Natural Language Processing. Hainan, China, 2004: 37-40

[7] Purandare Amruta, Pedersen Ted. SenseClusters — Finding clusters that represent word senses//Proceedings of the 19th National Conference on Artificial Intelligence (AAAI04). California, USA, 2004: 1030-1031

[8] McRoy Susan W. Using multiple knowledge sources for word sense discrimination. Association for Computational Linguistics, 1992, 18(1): 1-30

[9] Schutze H. Automatic word sense discrimination. Association for Computational Linguistics, 1998, 24(1): 97-123

[10] Lu Zhi-Mao, Liu Ting, Li Sheng. The research progress of statistical word sense disambiguation. Acta Electronica Sinica, 2006, 34(2): 333-343(in Chinese)
(卢志茂, 刘挺, 李生. 统计词义消歧的研究进展. 电子学报, 2006, 34(2): 333-343)

[11] Liu Qun, Li Su-Jia. Word similarity computing based on how-net//Proceedings of the CLSW3. Taipei, China, 2002: 1-18(in Chinese)
(刘群, 李素建. 基于《知网》的词汇语义相似度计算//第 3 届汉语词汇语义学研讨会. 台北, 中国, 2002: 1-18)

[12] Wang Hui. A study of chinese word sense disambiguation in machine translation based on grammatical & semantic knowledge-bases//Proceedings of the 7th Chinese Joint Conference on AI. Guilin, China, 2002: 75-82(in Chinese)
(王惠. 机器翻译中基于语法、语义知识库的汉语词义消歧策略//第 7 届中国人工智能联合学术会议. 桂林, 中国, 2002: 75-82)

[13] Lu Song, Bai Shuo, Huang Xiong. An unsuptervised approach to word sense disambiguation based on sense-words in vector space model. Journal of Software, 2002, 13(6): 1082-1089(in Chinese)
(鲁松, 白硕, 黄雄. 基于向量空间模型中义项词语的无导词义消歧. 软件学报, 2002, 13(6): 1082-1089)

[14] Miller G, Charles W. Contextual correlates of semantic similarity. Language and Cognitive Processes, 1991, 6(1): 1-28

[15] Mei Jia-Ju, Zhu Yi-Ming, Gao Yun-Qi, Yin H X. Tongyici Cilin (Dictionary of Synonymous Words). Shanghai: Shanghai Lexicographical Publishing House, 1983(in Chinese)
(梅家驹, 竺一鸣, 高蕴琦. 同义词词林. 上海: 上海辞书出版社, 1983)

[16] Chen Keh-Jiann, You Jia-Ming. A study on word similarity using context vector models. Computational Linguistics and Chinese Language Processing, 2002, 7(2): 37-58

[17] Karov Yael, Edelman Shimon. Similarity-based word sense disambiguation. Computational Linguistics, 1998, 24(1): 41-59

[18] Landauer Thomas K, Foltz Peter W, Laham Darrell. Introduction to latent semantic analysis. Discourse Processes, 1998, 25(1): 259-284

[19] Liu Lei, Cao Cungen, Wang Haitao. Acquiring hyponymy relations from large chinese corpus. WSEAS Transactions on Business and Economics, 2005, 4(2): 125-132

[20] Tian Guogang, Cao Cungen, Liu Lei, Wang Haitao. MFC: A method of co-referent relation acquisition from large-scale Chinese corpora//Proceedings of the ICNC'06-FSKD'06. Xi'an, China, 2006: 1256-1261

LIU Lei, born in 1979, Ph. D., lecturer. His major research interests include knowledge acquisition and ontology learning.

CAO Cun-Gen, born in 1964, Ph. D., professor, Ph. D. supervisor. His major research interests include knowledge acquisition and text mining.

ZHANG Chun-Xia, born in 1974, Ph. D., lecturer. Her major research interests include domain knowledge acquisition and text mining.

TIAN Guo-Gang, born in 1977, Ph. D.. His major research interests include domain attribute acquisition and text mining.

Background

KAT(Knowledge Acquisition from Text) is an important approach of large-scale knowledge acquisition at present. Especially, automatic acquisition of concepts and semantic relations from text has received much attention in the last ten years. The authors' research interests include the acquisition of concept, relation and attribute in KAT. Relation acquisition focuses on three basic semantic relations (hyponymy, whole-part relation, and co-referent relations). The authors have designed and developed a prototype system of acquiring knowledge from semi-structure text, and use it to acquire specialize knowledge from large-scale Chinese corpus. Research on hyponymy acquisition and verification is a basic and crucial problem in KAT. Hyponymy relations play a crucial role in various NLP(Natural Language Processing) systems, such as systems for information extraction, information retrieval, and dialog systems. Hyponymy relations are also important in accuracy verification of ontologies, knowledge bases and lexicons. The problem of polysemy hinders the building taxonomic hierarchy using hyponymy seriously. Furthermore, it hinders knowledge whole verification and knowledge database.