

文章编号:1003 - 0077(2005)01 - 0063 - 08

知网与同义词词林的信息融合研究

梅立军,周 强,臧 路,陈祖舜

(智能技术与系统国家重点实验室 清华大学 计算机系,北京 100084)

摘要:本文主要探讨了将知网(HowNet)和同义词词林进行信息融合的方法。我们针对知网对词的概念描述和同义词词林对词的语义分类的特点,提出了一种词典信息融合的方法:首先为词林的每个词集确定一个与知网中 DEF 类似的概念描述,在此基础上对两部词典中同时收录且均只有一个义项的词语进行双向意义联结,最后根据分类算法对两部词典中同时收录非单一义项的词语进行双向意义联结。实验表明,本文提出的处理策略达到了 93% 的信息融合正确率,融合后形成的新词典兼有词林的分类学信息和知网的描述信息。

关键词: 计算机应用; 中文信息处理; 词典信息融合; 知网; 同义词词林; 分类

中图分类号: TP391 **文献标识码:** A

Merge Information in HowNet and TongYiCi CiLin

MEI Li-jun, ZHOU Qiang, ZANGLu, CHEN Zr shun

(State Key Laboratory of Intelligent Technology and Systems, Dept. of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

Abstract: In this paper, we study the problem of merging information in HowNet and a Chinese thesaurus — TongYiCi CiLin. In order to integrate both the conception descriptions of words in HowNet and the semantic categories of words in TongYiCi CiLin, we propose several useful merging strategies: Firstly, we establish a DEF description for each SynSet in TongYiCi CiLin, which is similar with the word sense definition in HowNet. Then, we make bidirectional link for the words which have only one sense in both dictionaries. Finally we make bidirectional link for other words with multiple senses by using a classification algorithm based on salient frequency and vector distance of two sense descriptions. Experimental result shows that these merging strategies are effective and the merging accuracy is about 93%. The merged results form a new dictionary, which not only has semantic category of TongYi CiLin, but also has conception description of HowNet.

Key words: computer application; Chinese information processing; information merging; HowNet; TongYiCi CiLin; classification

1 引言

语义资源的开发和应用是自然语言处理的基础问题。近几年来,国内外研究人员通过人工总结或人机辅助处理,开发出许多大规模的语义计算资源,在英语方面,有 WordNet^[1]、FrameNet^[2]、MindNet^[3]等;在汉语方面,有知网 HowNet^[4]、同义词词林^[5]等。针对不同的应用需

收稿日期:2004 - 05 - 17

基金项目:国家自然科学基金(601330008)、国家 937 基金(G1998030507) 国家高技术研究发展 836 计划(2001AA114040)

作者简介:梅立军(1983 →),男,硕士研究生,主要研究方向为词汇语义学。

求,如何有效地选择或整合现有的语义资源,已成为语义计算的一个新的研究热点。

目前的研究大体可以分为以下两种不同的语义资源整合方式:

一种是在本体论(Ontology)层面的整合。典型研究工作是 Pease (2002) 等人提出的 SUMO (Suggested Upper Merged Ontology)^[6]。他们通过对现有不同抽象本体类和具体本体类描述进行有机整合,形成了包含 1000 多个项(terms)和约 3700 个断言(statements)的本体概念描述体系。并通过与 WordNet 的映射处理^[7],将 WordNet 中约 50,000 多个同义词集(SynSet)与相应 SUMO 类建立连接,形成一个更适合自然语言理解应用的语义计算资源。

另一种是在词义描述层面上的融合。在英语方面,Yarowsky (1992)探索了将 Roget 义类辞典中的语义类与 COBUILD 中的词语义项之间的意义联结方法^[8]。在汉语方面,姬东鸿等(1998)进行了《现代汉语词典》中的词语义项与《同义词词林》中的义类集合之间的意义联结研究,通过为同义词集中每个词语标上适当的现汉词语义项,初步实现了基于义项的同义关系描述^[9]。杨尔弘等(1999)通过开发人机交互的标注环境,进一步实现了对现汉约 4 万词语的词林义类标注^[10]。Carpuat 等(2002)使用基于语料库统计和双语词典匹配方法探索了英语 WordNet 的 SynSet 描述与中文 HowNet 的概念定义(DEF)的连接方法。

本文初步探索了知网和词林信息融合的方法。利用知网的 DEF 对同义词集中的不同词语进行标注,通过义原特征统计和分布向量计算进行自动排歧,实现了知网的 DEF 与词林的同义词集之间的双向意义联结:即为知网的每个 DEF 标注相应词林的义类代码,为词林的同义词集的每个词语标注相应的知网 DEF 描述。初步形成一个新的词语语义描述词典:对每个词语义项,可以给出相对应的知网 DEF 描述和词林义类代码。从而为深入研究两部词典构造背后的语义范畴和本体论体系的差异打下了很好的基础。

在接下来的章节中,首先介绍知网与同义词词林的语义信息描述特点,并分析在两者的信息融合过程中可能存在的问题,然后结合这些问题提出我们的词典融合算法,最后通过实验结果进行初步的融合算法性能评价。

2 词典融合问题描述

在词义描述方面,知网和词林各有其特点。知网通过不同义原及其关系描述符形成的 DEF 对词语意义进行描述。知网的设计者从汉语 6 千多个常用汉字的义项中经人工总结提炼出约 1500 个义原,对汉语的 62364 个词语义项进行了描述^[4]。而词林则通过同义词群(SynSet)描述词义。词林的设计者按照同义原则将现代汉语的 63895 个词语义项组织成 3925 个同义词群,其上进一步划分为 1428 个小类,94 个中类和 12 个大类。这样,针对每个词语义项,形成了唯一的 8 位义类编码^[5]。

我们认为,同义、近义关系是词汇语义学研究的一个重要内容。英文的 WordNet^[7]就是在此基础上进行的一个大型语言工程项目。而词林中由语言学家人工总结的 3925 个词群以及相应的 1428 个小类,为我们研究汉语词语的同义、近义组合关系提供了很好的素材。但目前的词林的信息组织方式是面向人的,并没有明确指出不同同义词群中各个词语的哪个义项在哪些方面形成同义、近义关系。而在这些方面,知网中提供的基于义原的 DEF 描述式可以提供很好的补充和帮助。因此,我们希望通过词林与知网的信息融合研究,为词林的同义词群中的每个词语标注义项对应的知网 DEF 描述式,并进一步提取形成针对不同同义词群的义原组

目前我们使用的是知网 2000 网络版,有关内容可参阅文献[4]。

合描述,以便更精确地揭示这些词语形成同义、近义关系的内在心理基础。

词林与知网的信息融合研究也是我们目前正在进行的基于情境的词汇语义描述研究^[12]的一部分,其基本研究思路为:词语是概念的符号体现。概念产生于特定的认知图示。概念,或标识它的词语的义项,只有在产生它的特定图示中才能描述、定义清楚。概念的使用则是在使用环境中对照、还原、引用产生它的图示的过程。我们用情境作为认知图示的数学模型,把情境理论作为词汇语义学和建基其上的自然语言语义学的统一理论框架。为此,需研究以下基本问题:构建情境表示图式和用情境定义、描述概念,建立情境代数以刻画情境间的关系、变换与运算,实现用代数演算体现概念思维,建立情境网以实现图示的结构、概念的组织方式等。

目前的研究重点是情境描述体系的开发和基于情境的语义词典的构成与组织问题,对此我们有如下考虑:词语的含义是多个层次的,词林、知网和现汉从不同侧面对词语的某个层次的语义进行了描述,而我们希望通过三部词典的融合将词语的多个层次的语义都提取出来。由于前人已经做过词林和现汉的融合,如果我们再将词林和知网进行融合,就可以将这三部词典的意义描述信息交叉关联起来,从而为情境描述体系和基于情境的语义词典的开发提供一个很好的基础语义资源支撑平台。

由于知网和词林编撰时期不同,应用目标也存在着差异,所以两部词典中收录词条差别较大。初步统计表明:知网收录了 50222 个词条,共 62364 条义项记录;而词林收录了 52256 个词条,共 63895 条义项记录。二者同时收录的词条只有 30926 个,词林中有 21330 个词条不在知网中出现,知网中有 19296 个词条不在词林中出现。在这 30926 个同时出现的词语中,知网和词林对相应词语义项的选择也存在着较大差异,差异如表 1 所示:

表 1 知网和词林对应词语义项分布统计表

知网义项数 \ 词林义项数	1	2	3	4	5	6	7	8	> 8
1	21562	2388	317	51	6	9	1	0	1
2	2659	1551	380	102	33	9	3	0	0
3	310	357	221	111	40	17	7	2	1
4	32	67	89	97	46	30	14	6	3
5	5	21	26	41	34	19	22	7	4
6	0	4	8	12	17	22	11	8	15
7	0	1	4	3	10	3	5	11	4
8	1	0	0	1	3	11	5	3	13
> 8	0	0	1	0	2	6	4	6	31

根据对这两部词典词语重合的实际情况,我们将资源融合分成两步:

第一步,对两部词典中同时收录的词进行双向意义联结,对这 30926 个词语的主要义项,可以给出相对应的知网 DEF 描述和词林义类代码,初步形成一个新的词语语义描述词典;

第二步,在第一步工作结果的基础上,为所有知网中出现的但在第一步处理后仍然没有找到恰当的分类信息的义项进行分类。

由于篇幅限制,本文将只针对资源融合的第一步工作进行详细介绍。

3 词典融合的双向意义联结方法

针对两部词典中同时出现的词语进行双向意义联结的基本步骤如下:

(1) 为词林 SynSet 定义 DEF 描述

对词林的每个 SynSet (8 位编码),我们根据知网信息为其定义一个与知网中词语义项定义类似的 DEF 描述。当确定知网中的词语义项在词林中的分类时,可根据词语义项的 DEF 和词林中对应 SynSet 的 DEF 的相似程度来确定该词语义项的词林分类。在此过程中,还可以完成对 SynSet 中每个词语的知网义项排歧,从而初步实现“词林-知网”的意义联结。

(2) 词林和知网中单一义项词的特殊处理

对两部词典同时收录的词条,如果其在知网和词林中的义项数目都仅有一个,在这种情况下,这两部词典的义项含义相同的概率非常大,因为收录的唯一义项应当是主要义项,而主要义项一般来说都是相同的。考虑到这两部词典的编著者不同,所以在进行单一义项词的意义联结过程中,还是需要进行特殊处理,不能简单的直接进行意义联结。

(3) 针对知网 DEF 义项的分类算法

如果我们为知网中的每个 DEF 确定一个词林分类,就等于为知网添加了词林的分类信息,从而实现了“知网-词林”的初步融合。基本的分类算法有两种:频度统计和特征向量计算。我们将两者的处理优势进行结合,形成了综合处理算法。

3.1 定义词林 SynSet 的 DEF 描述

知网中的每个 DEF 都是通过一个义原(特征)的集合来定义的,其中第一个义原描述了该概念的基本属性,我们称之为**主特征**;其它义原从不同方面描述概念,我们称之为**次特征**。如果我们为词林的一个 SynSet 定义一个 DEF 的话,那么这个 DEF 必定是由能够代表该 SynSet 所有词语义项的共同含义的义原集合组成。我们使用义原的特征向量来表示它,其基本内容包括: { 义原, 该义原在 SynSet 中的突显度 }。具体计算方法如下:

第一步:将 SynSet 中所有可以在知网中找到对应义项(可能有多个对应义项)的词,统计知网中这些词所有 DEF 中出现过的义原,将这些不同的义原作为特征向量的基;

第二步:统计得到每个义原出现的次数,再除以 SynSet 的有效词语总数(扣除不能贡献 DEF 的词语),由此可以得到每个义原的突显度;(除以 SynSet 中有效词的个数的目的是为了减少词林 SynSet 中含有的词的多少对于分类时的影响。)

第三步:在前两步的基础上,我们根据 DEF 中的主次特征的划分,将义原分为主特征基和次特征基,再根据义原的突显度,可以得到主特征向量和次特征向量。

如果某个 SynSet A 中的有效词语总数小于 4,则根据这几个有效词来确定 A 的特征向量可能就不会太准确,所以在这种情况下我们还需要计算在词林分类层次中 A 的上一级 SynSet A' (6 位编码)的义原突显度和特征向量,用来作为 A 的特征向量的补充。

为了便于向量距离计算,我们需要把知网义项的 DEF 和计算出的词林 SynSet 的 DEF 的特征向量进行归一化处理。知网中共有特征(义原)1538 个,归一化后特征向量基就是由这 1538 个义原组成。

在处理过程中,对在知网中可找到多个对应义项的词,会将所有知网中的义项都考虑进来,这样无疑会对 SynSet 的 DEF 集合定义产生噪音,本文并没有对这个 DEF 集合进行语义排歧,因为 SynSet 的 DEF 定义是我们为知网中义项确定词林分类信息的整体算法的一部分,虽然 SynSet 的 DEF 定义本身会存在一定的集合噪声,并没有通过排歧算法来消除,但在整个算法中,通过了夹角余弦的计算来排除了集合噪声对确定词林分类信息的影响。

3.2 词林和知网中单一义项词的处理

知网和词林中的单一义项词语,如满足下面两个条件中任一个,就可直接进行意义联结。

条件 1:在词林中该词对应的 SynSet 的 DEF 描述向量(不计算该词)含有该词在知网中的

DEF 中的所有义原(特征),并且二者的重合度大于某个阈值 ;

说明:重合度 = $\frac{\text{SynSet 的 DEF 和知网义项 DEF 中相同义原的突显度累加和}}{\text{SynSet 中有效词数目}}$

条件 2:在词林中该词对应的 SynSet 的 DEF 描述向量(不计算该词)含有该词在知网中的 DEF 中的部分义原(特征),并且二者的重合度大于某个阈值。

我们的处理目标是选择合适的阈值使直接意义联结的单义项词语的分布密度达到最大。

3.3 分类算法

分类算法主要有以下两种方法:

(1) 特征频度统计的分类算法

特征频度统计算法主要分两步:第一步:对知网中一个待确定分类词的义项,在词林中找出所有包含该词的 SynSet,对每个 SynSet 根据其相应 DEF 进行特征频度统计,找出主特征重合度最大的那个 SynSet,根据该 SynSet 的主特征值按比例确定阈值,保留主特征计数大于阈值的 SynSet 分类;第二步:统计第一步处理留下的 SynSet 分类 DEF 的次特征重合度,取次特征重合度最大的 SynSet 作为最终确定的分类。如果这两步所有 SynSet 的 DEF 主特征或次特征都小于某一阈值,则认为该词无法分类。

(2) 特征向量计算的分类算法

通过比较知网中待分类的词的 DEF 和词林中含有待分类词的 SynSet 分类的特征向量间的距离也可以确定分类。向量间的距离可以用向量间的夹角的余弦值来表示,余弦值越大,向量间距越小,词条义项和语义类的相关度越高。计算出待确定分类义项的特征向量和多个待选分类特征向量间的夹角余弦后,取与待确定分类特征向量间交角余弦最大的分类作为分类结果。统计表明,如出现以下两种情况,则分类结果错误较多,可视为无法对待该词语义项有效分类:

设:待选分类特征向量(主次特征向量和)与待分类特征向量间的夹角余弦为: $\cos(A, C)$

待选分类特征向量(次特征向量)与待分类特征向量间的夹角余弦为: $\cos(A1, C)$

情况 1:如果 $\cos(A, C) < \cos(A1, C)$ 且 $\cos(A, C) < \theta$ (为阈值),说明待选分类主特征向量与待分类特征向量差别较大,分类结果很可能错误(经过实验,确定阈值为 0.3)。

情况 2:如果 $\cos(A, C) < \theta$ (为阈值),说明待分类词和目标分类的相似度很低,分类结果很可能有错误(经过实验,确定阈值为 0.15)。

特征频度的分类方法考虑了一个词林 SynSet 中所有义原突显度对确定知网义项分类的积极影响。但有时虽然词的词性是相同的,但词义有细微的差别,仅凭频度统计就不足以判断出正确分类。特征向量计算方法考虑了 SynSet 的 DEF 中无用义原对于判定知网义项分类的错误影响,可判断那些明显错误的分类。但其作为判断分类的唯一根据也会产生错误,因为余弦值反映的是 SynSet 中无用义原和有用义原突显度的分布情况,是一个相对的值。

单独采用特征频度统计或特征向量计算来分类都存在问题,因而需将两种方法结合起来,取长补短。特征频度统计可以得到很高的正确率,因此先进行特征频度统计,在此基础上再利用特征向量计算加以修正。主要分两步:第一步,通过特征频度统计方法为知网中可以进行分类的义项计算出一个初步的词林分类;第二步,使用计算夹角的方法对第一步分类结果进行校验,找出其中不合适的分类,并尝试用夹角计算的结果来代替原分类方法。如仍然无法满足要求,则算作无法分类。

4 实验结果分析

4.1 SynSet 的 DEF 描述

利用上面介绍的方法,我们可得到如下的一个词林 SynSet 的 DEF 描述:

Ec040101:SynSet = {洁白, 涅白, 凝脂, 纯白, 皓, 皓皓, 皎洁, 白, 粉白,}

主特征向量: {0.85(aValue| 属性值), 0.1(RainSnow| 雨雪), 0.1(food| 食品), 0.1(shape| 物形), 0.05(Tool| 用具), 0.05(decorate| 装饰), 0.05(material| 材料), 0.05(metal| 金属)}

次特征向量: {0.7(color| 颜色), 0.6(white| 白), 0.1(brightness| 明暗), 0.1(bright| 明), 0.05(MakeUp| 化妆), 0.05(apply| 涂抹), 0.05(clothing| 衣物), 0.05(original| 原)}

对这个 SynSet 进行人工分析,可发现它的 DEF 描述中突显度比较高的几个义原(aValue| 属性值), (color| 颜色), (white| 白), 能够很好的反映该 SynSet 的含义,可作为确定分类的依据。

4.2 单一义项词的意义联结分析

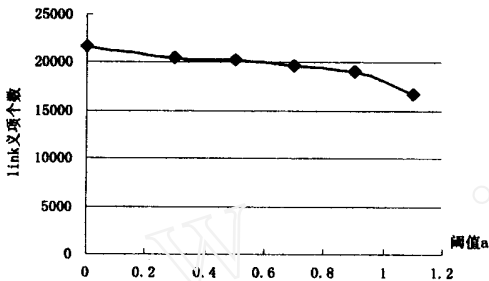


图1 单一义项 link 阈值效果曲线

经过实验,我们发现对于一个给定的阈值, 阈值取 $1/2$ 时,双向意义联结的错误率比较低,因此在阈值的选取过程中,取 $= 1/2$,这样可以根据阈值的选取来确定这两个阈值。图1共记录了(0, 0), (0.3, 0.15), (0.5, 0.25), (0.7, 0.35), (0.9, 0.45), (1.1, 0.55)这六组(,)阈值情况下的意义联结效果。考虑到收录词语的唯一义项应当是主要义项,因此这些单一义项词大部分应该是可以进行双向意义联结的,所以最后确定阈值为:

$= 0.5$, $= 0.25$ 。在此阈值下,可以对 94% 的单一义项词(20303 个知网义项)进行双向意义联结。对其中 300 个意义联结结果的抽样分析表明,正确分类数为 276,正确率达到 92%。

经过分析,分类错误的原因在于这两部词典仍然存在着理解上的分歧。例如词语‘宛’,在知网中定义为“DEF = aValue| 属性值,form| 形状,curved| 弯”,而词林中的 SynSet 则定义为{ 犹,犹如,似乎,似,宛,宛然,宛如,}。

4.3 自动分类效果分析

目前对自动分类的效果评价是:从分类结果中随机抽取约 20% 的样本,人工进行分类,如果自动分类结果与人工分类的结果一致,算作正确分类,否则作为错误分类。

说明:第一步频度统计分类正确率 = $\frac{\text{样本中的正确分类数}}{\text{抽取的样本总数}}$

第二步特征向量分类正确率 = $\frac{\text{样本中确属错误分类的数目}}{\text{找出的可能错误分类样本数}}$

特征向量重新分类正确率 = $\frac{\text{样本中重新分类正确的数目}}{\text{重新分类的样本数}}$

(1)在分类算法的第一步特征频度统计中,我们共为 8798 个知网义项确定了词林分类。经过对 400 个确定的分类统计表明,正确率达到 94%。

第一步融合正确例子:为知网中“白”义项(编号 000724)确定词林分类 Fc040119:

白,000724,DEF = look| 看,manner = unsatisfied| 不满 Fc040119:SynSet = {白眼,白}

第一步融合错误例子:将知网“关口”义项(编号 01696)分到词林 Ca040301 类中:

关口,016196,DEF = part| 部件,%place| 地方,important| 主,# military| 军

Ca040301:SynSet = {契机, 节骨眼, 当口儿, 关键, 关口, 关, 关头, 转捩点, 转折点, 转机}

错误原因:用频度统计的方法时,词林 Ca040301 类的 SynSet 中的“关头”、“转折点”和“转捩点”都对次特征“important|主”的突显度产生了错误影响。

(2)在分类算法的第二步特征向量计算中,我们共从第一步的结果中找出 1051 个可能的分类错误,并为其中 8 个知网义项重新划分了词林分类。统计表明,特征向量分类的正确率达到约 95%,重新分类的正确率为 87.5%。

修改成功例子:根据词林 Ca040301 中“转折点”的知网义项 DEF = {time|时间, @change|变, important|主}中“time|时间”和“change|变”这两个对确定分类无用的特征来减弱“important|主”对确定分类的错误的积极作用,将“关口”(知网编号 01696)重新分到词林 SynSet Cb200101:雄关,关,关隘,关口,边关。

4.4 词典融合总体效果

经过处理,我们共对 29101 个知网义项进行了分类,其中通过单义项词直接联结的有 20303 个,通过特征分类方法进行分类的有 8798 个,有 12781 个义项无法有效分类。对随机抽样得到的 400 个结果进行统计,意义联结正确率达到约 93%,说明本文的融合策略比较有效。通过特征向量计算进行校正的词不多,因为特征频度统计本身精度(94%)已比较高,特征向量计算主要是找出不恰当的分类。出现错误的原因主要有如下两点:

1)词林中缺乏与知网中某一义项相对应的分类,可能将义项错分到某一相近的分类中,而不是判断为不能分类。如知网编号 24711 的“慷慨”(DEF = aValue|属性值, behavior|举止, strong|强),被分到词林词集 Ee350201(侠义, 先人后己, 舍己为人...),但实际应分到词集 Ic050201(激昂, 激昂慷慨, 激昂, 慷慨激昂, 拍案而起...),但该词集并没有“慷慨”这个词。

2)某些知网义项在词林中有合适分类,但计算出的相关度低于阈值,因而被视为无法有效分类。例如知网中编号 37073 的“散发”(DEF = spread|撒),按频度统计被分到词集 Jd070202(发, 泛, 散发, 散),但特征向量计算认为是错误分类,因为其夹角余弦值小于阈值。

5 结语

在这篇论文中,我们详细介绍了将知网和同义词词林进行融合的方法,并对融合的结果和正确率进行了分析和统计。融合结果统计表明:对于两部词典中同时收录的词,大部分都可以根据本文提出的融合策略来很好地进行双向意义联结,还有一小部分词不能根据本文提出的融合策略确定分类信息,主要因为两部词典的编写者的出发点存在差异。

词典的融合是一项很复杂的工作,本文只给出了一个初步的融合策略,提供了一种词典融合的思路,还有很多工作有待完善,例如:单一义项词意义联结中的阈值确定,词典融合的结果评价标准的制定(目前仍是人工评价),两部词典中不同时收录的词语的融合,利用知网的义原描述信息对现有同义词群进行意义排序和重组,以及深入分析两部词典在义项选择中的差异情况等,我们将在后续论文中对这些问题的处理进行更深入的研究。

参 考 文 献:

- [1] Miller, George A., & Fellbaum C. . Semantic Network of English[A]. In :Beth Levin and Steven Pinker (Eds.) Lexical & Conceptual Semantics[C]. Elsevier Science Publishers, B. V. ,Amsterdam, the Netherlands, 1991.
- [2] Baker, Collin F. ,Fillmore and et al. The Berkeley FrameNet project [A]. In : Proceedings of the COLING- ACL 98 [C]. Montreal, Canada : 1998, 86 - 90.
- [3] Richardson S. D. ,Dolan W. B. and Vandervende L. . MindNet : acquiring and structuring semantic information from

- text[A]. In: Proc. of COLING- ACL '98[C]. 1998,1098 - 1102.
- [4] 董振东. 语义关系的表达和知识系统的建造[J]. 语言文字应用,1998,(3):76 - 82.
- [5] 梅家驹,等. 同义词词林[M]. 上海:上海辞书出版社出版,1983.
- [6] Pease ,A. ,Niles ,I. ,and Li J. The Suggested Upper Merged Ontology : A Large Ontology for the Semantic Web and its Applications [A]. In : Working Notes of the AAAI- 2002 Workshop on Ontologies and the Semantic Web[C]. Ed monton ,Canada : 2002.
- [7] Niles ,I. and Pease ,A. Linking Lexicons and Ontologies : Mapping WordNet to the Suggested Upper Merged Ontology [A]. In : Proceedings of the 2003 International Conference on Information and Knowledge Engineering[C]. Las Vegas ,Nevada : 2003.
- [8] D. Yarowsky. Word Sense Disambiguation Using Statistical Models of Roget 's Categories Trained on Large Corpora [A]. In : Proc. Of COLING '92[C]. Nantas ,France : 1992 ,454 - 460.
- [9] Ji Donghong ,Gong junping ,Huang Changning. Combining a Chinese Thesaurus with a Chinese Dictionary [A]. In : Proc. of COLING- ACL 98[C]. 1998 ,600 - 606.
- [10] 杨尔弘,黄昌宁,李涓子.《现代汉语词典》的义类标注[A]. 黄昌宁,董振东主编. 见:计算语言学文集 [C]. 北京:清华大学出版社,1999,167 - 173.
- [11] Marine Carpuat ,Grace Ngai ,Pascale Fung and et al. Creating a Bilingual Ontology : A Corpus - Based Approach for Aligning WordNet and HowNet [A]. In : Proceedings of the 1st Global WordNet Conference[C],2002.
- [12] 陈祖舜,周强,赵强. 情境 ——组织/ 存放词汇语义知识的恰当框架[J]. Computational Linguistics and Chinese Language Processing ,2002 7(2): 1 - 36.

(上接第 62 页)

- [6] Atallah ,M.J. ,C.J. McDonough ,V. Raskin ,and S. Nirenburg. Natural Language Processing for Information Assurance and Security :An Overview and Implementations [C]. In :M. Shaeffer(ed.) ,NSPW '00 :Proceedings of Workshop on New Paradigms in Information Security ,Cork ,Ireland ,September 2000. New York :ACM Press ,51 - 65.
- [7] 张玉洁,山本和英. 汉语语句的自动改写[J]. 中文信息学报,2003(6):31 - 38.
- [8] Mikhail J. ,Atallah ,Samuel S. ,Wagstaff ,Jr. . Watermarking with Quadratic Residues[C]. Working Paper ,Department of Computer Science ,Purdue University ,1996.
- [9] Mikhail J. Atallah ,Victor Raskin ,Mchael Crogan ,Christian Hempelmann ,Florian Kerschbaum ,Dina Mohamed and Sanket Naik. Natural Language Watermarking :Design ,Analysis ,and a Proof - of - Concept Implementation[C]. Information Hiding 2001 :185 - 199
- [10] Mikhail J. Atallah ,Craig J. McDonough ,Victor Raskin. Natural Language for Information Assurance and Security :An Overview and Implementation[C]. Published in : M. Schaefer(ed.) ,Proceedings. New Security Paradigm Workshop. September 18th - 22nd ,Ballycotton ,County ,Cork Ireland. New York :ACM Press ,2000 .pp. 51 - 65.
- [11] Nirenburg ,S. ,and V. Raskin 2003. Ontological Semantics. Cambridge [M] , MA :MIT Press(forthcoming) . Pre - publication draft ,<http://crl.nmsu.edu/Staff/Pages/Technical/sergei/book/index-book.html>.