文章编号：1000-8608(2010)04-0603-06

基于义原关系的多策略汉语词义消歧方法

车 超[1]， 金 博[1]， 滕弘飞*[1,2]， 屈福政[2]

（1.大连理工大学 计算机科学与技术学院，辽宁 大连 116024；
2.大连理工大学 机械工程学院，辽宁 大连 116024）

摘要：词义消歧是自然语言处理领域的重点和难点问题.提出了一种基于知网中义原关系的多策略词义消歧方法.该方法利用知网中义原间最基本和最重要的部件-整体和属性-宿主关系进行词义消歧，并辅以基于值-属性关系、中文信息结构和语义相关度的消歧方法.在SENSEVAL-3汉语词义消歧任务测试文本上的实验表明，该方法与官方结果相比，具有较好的计算性能.

关键词：知网；词义消歧；义原关系；语义相关度

中图分类号：TP391 **文献标志码**：A

0 引 言

词义消歧是根据一个多义词所在的上下文语境来确定其在该语境中的意义.多义词分布的普遍性和无规律性决定了词义消歧是自然语言处理领域研究的难点和热点之一，同时词义消歧又在自然语言处理的很多领域有重要的应用，诸如机器翻译、信息检索、自然语言内容语义分析、语法分析、语音识别和文语转换等[1].

词义消歧研究的关键问题是如何获取消歧知识源，目前获得知识源主要有词典资源或人工标注的语料库两种途径.由于大规模人工标注的语料库需要耗费大量的手工劳动并且存在数据稀疏问题，很多学者使用语义词典作为词义消歧的知识源.值得注意的是，在汉语词义消歧中，知网是一个常用的语义词典.基于知网的语义消歧方法有很多，如杨尔弘等[2]提出一种基于义原同现频率的词义消歧方法，统计义原同时出现的频率，建立互信息评价函数，对多义词进行消歧，试验证明消歧效果良好；闫蓉等[3]利用词语之间的优先组合关系，计算各实词概念与歧义词概念之间的相似度，以判断歧义词词义；余晓峰等[4]以基于知网的词汇语义相似度计算为基础，通过相似度的大小来判断歧义词的词义.此外，Wang[5]利用知网中义原同现、汉语信息结构和属性-属性对进行消歧.除Wang利用了义原间的值-属性关系外，利用义原间关系消歧的方法不多见.

本文利用知网中义原间最基本和最重要的属性-宿主和部件-整体关系进行词义消歧，并加入对形容词消歧作用明显的基于值-属性关系的消歧方法和通用性较强的基于信息结构和语义相关度的词义消歧方法作为补充，提出一种多策略相结合的语义消歧方法.

1 知网简介

知网是一个以汉语和英语的词语所代表的概念为描述对象，以揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库[6].在知网中，词语的每个概念被表示为义项.每个义项由一个或多个义原表示.义原是知网中最基本的、不易于再分割的最小单位，知网通过对约6 000个汉字进行考察和分析抽取了800多个义原，并总结了如部分、主体、客体、从属、时空、材料等若干种义原间的语义关系.

收稿日期：2008-05-17； **修回日期**：2010-04-27.
基金项目："八六三"国家高技术研究发展计划资助项目(2006AA04Z109)；国家自然科学基金资助项目(60674078).
作者简介：车 超(1981-)，男，博士生；滕弘飞*(1936-)，男，教授，博士生导师，E-mail：tenghf@dlut.edu.cn；屈福政(1957-)，男，教授，博士生导师.

2 词义消歧

本文的词义消歧步骤如图 1 所示. 首先对歧义词根据词性进行类别消歧. 如果类别消歧不能消除歧义,则分析歧义词及上下文中存在什么样的义原关系. 如果存在部件-整体关系,则进行基于部件-整体关系的消歧;如果存在属性-宿主关系,则进行基于属性-宿主关系的消歧;如果存在值-属性关系,则进行基于值-属性关系的消歧. 如果上述方法还不能消除歧义,则进行基于中文信息结构和语义相关度的消歧. 在本文所使用的 6 种消歧策略中,起主要作用的是基于 3 种义原关系的消歧策略. 类别消歧对歧义词进行简单的消歧,将词性不对的词义删除;而基于中文信息结构和语义相关度的消歧起到一个补充的作用,对不能用 3 种义原关系消歧的词语进行消歧. 下面按处理的顺序,分别介绍这几种消歧方法.

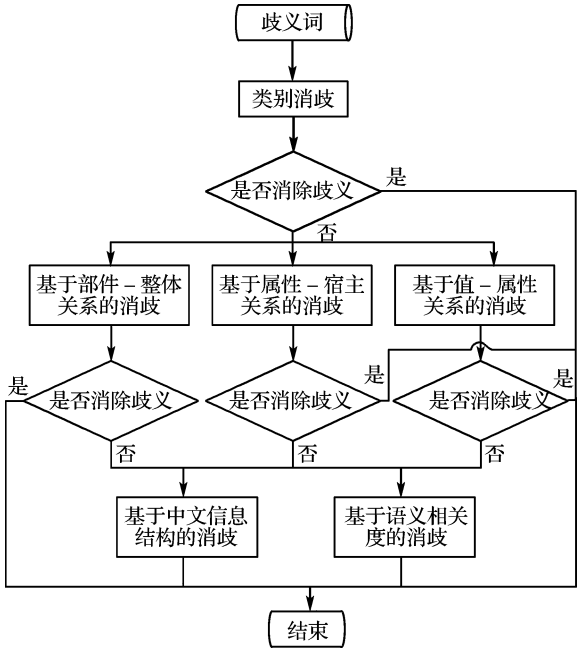

图 1 本文方法词义消歧过程示意图

Fig. 1 The disambiguation process of the method in this paper

2.1 类别歧义的消歧

类别歧义是指由于词性而引起的歧义,即一个词在不同词性下有不同的意思,因而通过词性就可区别词义. 对于这类歧义,只要根据词性来判断词义即可. 可以对输入的文本进行分词、词性标注,然后根据该词在知网中的概念对应的词性信息,确定该词在相应上下文中的意思. 据统计,根据词性标记,大约能消除 1/5 的歧义. 剩下的大部分多义词,还要通过其他方法处理.

2.2 基于部件-整体关系的词义消歧

知网哲学认为,每一个事物都可能是另外一个事物的部件,同时每一个事物也可能是另外一个事物的整体. 利用此关系可有效消除具有部件 - 整体关系的多义词的歧义. 其具体方法如下:歧义词 W^d 具有义项 $C_1^d, C_2^d, \cdots, C_n^d$,且部件义原 $S_i^d \in C_i^d (1 \leqslant i \leqslant n)$,如果歧义词 W^d 所在的上下文中存在这样的一个名词 W^n,其某个义项的第一义原 $S_1^n \in S_i^d$ 或 $S_1^n = S_i^d$,则义项 C_i^d 为歧义词 W^d 的概念. 其中 $S_1^n \in S_i^d$ 表示在义原层次树中,S_1^n 是 S_i^d 的孩子节点,且 S_1^n 和 S_i^d 之间路径长度 $d \leqslant 2$,以下同.

现举例分析,“他打碎了杯子把”. 分词并标注词性后得到“他/pron 打/v 碎/adj 了/stru 杯子/n 把 /n”. 其中“/”后的字母表示词语的词性,其具体含义与知网中的词性标注含义相同,以下同,不再赘述. 这里对“把”进行消歧. 在知网中,“把”有 13 个义项,经过类别消歧,剩下词性为 n 的 3 个义项如表 1 所示. 其中 C_2 中有部件义原“%implement|器具”. 上下文中“杯子”的第一义原“tool|用具”在义原树中属于“%implement|器具”的孩子节点,所以“把”的概念为义项 C_2.

表 1 “把”和“杯子”在知网中的义项

Tab. 1 The concepts of the "把" and "杯子" defined in HowNet

词	义项序号	词性	义项
把	C_1	n	part|部件,%LandVehicle|车,*drive|驾驭,hand|手
	C_2	n	part|部件,%implement|器具,*hold|拿,*OpenShut|开关,hand|手
	C_3	n	part|部件,%plant|植物,body|身
杯子	C_1	n	tool|用具,cubic|体,@put|放置,#drinks|饮品

上述规则是用表示整体的词确定表示部件词的语义,反之亦然,规则如下:歧义词 W^d 具有义项 $C_1^d, C_2^d, \cdots, C_n^d$,其中 $C_i^d (1 \leqslant i \leqslant n)$ 的第一义原为 S_1^d,歧义词 W^d 所在的上下文中存在一个名词 W^n 含有部件义原 S_j^n,且 $S_1^d \in S_j^n$ 或 $S_1^d = S_j^n$. 则选择义项 C_i^d 作为 W^d 的概念.

2.3　基于属性-宿主关系的词义消歧

在知网的哲学中，任何一个事物都包含着多种属性，事物之间的异或同是由属性决定的，没有了属性就没有了事物. 属性和它的宿主之间的关系是固定的，即有什么样的宿主就有什么样的属性，反之亦然. 可以利用这种固定关系进行词义消歧，其具体过程如下：

歧义词 W^d 具有义项 $C_1^d, C_2^d, \cdots, C_n^d$，属性义原 $S_i^d \in C_i^d (1 \leqslant i \leqslant n)$，如果歧义词 W^d 的上下文中存在这样的一个名词 W^n，其某个义项的第一义原 $S_1^n \in S_i^d$ 或 $S_1^n = S_i^d$，则义项 C_i^d 为歧义词 W^d 的概念. 其中 $S_1^n \in S_i^d$ 表示在义原层次树中，S_1^n 是 S_i^d 的孩子节点，且 S_1^n 和 S_i^d 之间路径长度 $d \leqslant 2$.

现举例说明上述规则的应用，句子“小伙子是干生意的材料”，经分词标注词性后得到“小伙子/n 是/v 干/v 生意/n 的/stru 材料/n”，其中“/”之后标注的是词的词性. 这里对“材料”消歧，“材料”在知网中有 3 个义项，如表 2 所示. 由于都是名词，无法通过类别消歧. 上下文中的“小伙子”含有义原“human|人”，而“材料”的义项 C_1 中有属性义原“&human|人”，所以选择 C_1 为“材料”的概念.

表 2　“材料”和“小伙子”在知网中的义项

Tab. 2　The concepts of "材料" and "小伙子" defined in HowNet

词	义项序号	词性	义项
材料	C_1	n	attribute|属性，quality|质量，&human|人
	C_2	n	information|信息
	C_3	n	material|材料，generic|统称
小伙子	C_1	n	human|人，young|幼，male|男

上面是利用宿主词来消除属性词的歧义，也可以利用属性词来消除宿主词的歧义，规则如下：歧义词 W^d 具有义项 $C_1^d, C_2^d, \cdots, C_n^d$，其中 $C_i^d (1 \leqslant i \leqslant n)$ 的第一义原为 S_1^d，歧义词 W^d 所在的上下文中存在一个名词 W^n 含有属性义原 S_j^n，且 $S_1^d \in S_j^n$ 或 $S_1^d = S_j^n$. 则选择义项 C_i^d 作为 W^d 的概念.

2.4　基于值-属性关系的词义消歧

在知网中，值-属性关系指形容词或副词的属性值义原与所修饰的名词中的属性义原具有一一对应的关系. 也就是说名词有什么样的属性，才能使用含有什么样属性值的形容词或副词来修饰；反之亦然. 因此，可以利用形容词或副词和名词之间这种一一对应关系进行词义消歧. 本文基于值-属性关系的语义消歧方法借鉴了 Wang[5] 的思想，并进行了改进，改进体现在用本方法消歧时不需要事先建立属性-义原对应表，其具体方法如下.

歧义词 W^d 具有义项 $C_1^d, C_2^d, \cdots, C_n^d$，且属性值义原 $S_i^d \in C_i^d (1 \leqslant i \leqslant n)$，如果歧义词 W^d 所在的上下文中存在这样的一个名词 W^n，W^n 含有一个表示属性的义项 C_i^n，若 C_i^n 第二义原 $S_2^n = S_i^d$，则 C_i^d 为歧义词 W^d 的概念. 现举例分析，“他个子很矮”. 分词并标注词性后得到“他/pron 个子/n 很/adv 矮/adj”. 其中“/”后的字母表示词性. 对“矮”进行消歧，“矮”在知网中有 2 个义项，如表 3 所示. 所有义项的词性均为 adj，无法通过类别消歧. 上下文中“个子”的属性义项中第二义原是“height|高度”，而“矮”的义项 C_1 中有“height|高度”义原，所以选择 C_1 为“矮”的概念.

表 3　“矮”和“个子”在知网中的义项

Tab. 3　The concepts of "矮" and "个子" defined in HowNet

词	义项序号	词性	义项
矮	C_1	adj	aValue|属性值，height|高度，low|矮
	C_2	adj	aValue|属性值，rank|等级，LowRank|低等
个子	C_1	n	attribute|属性，height|高度，&human|人

上述方法是利用表示属性的词来消除表示属性值的词的歧义，反之亦然，规则如下：歧义词 W^d 具有多个义项：$C_1^d, C_2^d, \cdots, C_n^d$，表示属性的 $C_i^d (1 \leqslant i \leqslant n)$ 的第二义原为 S_{i2}^d，歧义词 W^d 所在的上下文中存在一个表示属性值的形容词或副词 W^a，若属性值义项的第二义原 $S_2^n = S_{i2}^d$，则义项 C_i^d 为歧义词的概念.

2.5　基于中文信息结构的词义消歧

上述 3 种消歧策略只能对具有特定义原关系的部分多义词使用，不具备通用性. 所以当上述 3 种策略使用完之后，还有一部分多义词无法消歧，需要通用性较好的方法来进行余下的工作.

信息结构(message structure)是由两个或两个以上的字、词或短语构成，句法和语义合理，并传达了特定信息的结构. 中文信息结构是中文中句法和语义合理的一个语言片段. 知网的中文信息结构抽取器通过构建一系列的语义规则来识别信息结构，本文借助这些语义规则来消歧. 这些规

则构建的主要目的是为了识别信息结构，所以对于词义消歧稍显不足.为了方便消歧，本文增加了一些语义规则，如下文中句法结构“V←N”包含的一些语义规则.

具体的消歧方法是：先找到歧义词及其上下文对应的语法结构，再在该语法结构下查找对应的语义规则消歧.下面以对“穿”消歧的过程为例说明.例句“他穿拖鞋就去上课”，经分词标注词性后为“他/pron 穿/v 拖鞋/n 就/adv 去/v 上课/v”，其中“/”之后标注的是词的词性.“穿”在知网中有3个义项，如表4所示.该句中“穿-拖鞋”对应句法结构“V←N”，该语法结构对应的规则如下：

[CN.def]==｛PutOn|穿戴｝ &&[R1.def.class] == ｛clothing|衣物｝；

[CN.def]==｛stab|扎｝ &&[R1.def]==｛Vdirection|动趋，upper|上｝；

[CN.def]==｛ start|开始｝ &&[R1.def] == ｛ fact|事情，function|活动｝；…

其中[CN.def]表示当前扫描节点的概念定义；[R1.def] 和[R1.def.class]分别表示当前扫描节点的下一个节点的概念定义和概念定义的类.“穿”的下一个节点的概念类为“clothing|衣物”，推测“穿”的概念应为“PutOn|穿戴”，所以其概念为 C_1.

表4 “穿”和“拖鞋”在知网中的义项

Tab.4 The concepts of ″穿″ and ″拖鞋″ defined in HowNet

词	义项序号	词性	义项
穿	C_1	v	PutOn|穿戴
	C_2	v	cross|越过
	C_3	v	stab|扎
拖鞋	C_1	n	clothing|衣物，#foot|脚

2.6 基于语义相关度的词义消歧

知网中的义原根据上下位关系构成了一个树状的义原层次体系，可以通过计算义原树中的语义距离来得到义项间的语义相关度，根据语义相关度进行语义消歧.具体方法是：计算歧义词的各个义项与上下文的语义相关度，取语义相关度最大的义项作为歧义词的概念.

本文的语义相关度计算方法借鉴了刘群等[7]的词语相似度计算方法，但又与其不同.其不同主要表现在义原的权重赋值上.

词语相似度计算中，约定由于第一独立义原反映了一个概念最主要的特征，将其权值定义得比较大[7].本文认为在词义消歧中，每个概念的第一独立义原不一定对词语相关度的贡献最大.特殊情况下，第一义原对消歧甚至都不起作用.比如多义词“运动”的3个义项为：“fact|事情，AlterLocation|变空间位置”，“fact|事情，exercise|锻练，sport|体育”，“fact|事情，function|活动，politics|政”.3个义项的第一义原都是“fact|事情”，此义原对消歧没有任何贡献，所以本文不给第一义原赋予更多的权重，一般情况下约定所有义原的权重是相等的.下面具体介绍语义相关度的计算方法.

由于知网中所有的概念都用义原表示，先介绍义原间语义距离的计算方法.假设义原 S_1 和 S_2 在义原层次体系中的路径距离为 d，则两个义原之间的语义距离为[7]

$$sim(S_1,S_2)=\frac{\alpha}{d+\alpha} \tag{1}$$

其中 α 是一个可调节的参数，本文取 $\alpha=1.6$.

假设义项 C_1 含有义原 $S_{11},S_{12},\cdots,S_{1n}$，义项 C_2 含有义原 $S_{21},S_{22},\cdots,S_{2m}$，义项 C_1 和 C_2 的相关度为

$$R(C_1,C_2)=\sum_{i=1}^{n}\max\{sim(S_{1i},S_{2j})\}/n;\quad j=1,2,\cdots,m \tag{2}$$

假设词 W 含有义项 $C_1,C_2,\cdots,C_n$，则义项 C 与词 W 的相关度为

$$R(C,W)=\max\{R(C,C_i)\};\ i=1,2,\cdots,n$$

3 数据实验

3.1 实验过程及结果

本文实验是在SENSEVAL-3汉语语义消歧任务[8]的测试文本上进行的.SENSEVAL是由ACL(Association for Computational Linguistics)举办的权威性评测会议，旨在针对不同语言、不同词汇的文本语义自动分析系统(包含词义消歧)进行评测.SENSEVAL-3汉语词义消歧任务的语料采用知网2000中的义项进行标注，而本文方法又是基于知网的，因此本文方法采用其作为测试语料.SENSEVAL-3汉语词义消歧任务为20个歧义词提供了比例为2∶1的测试集和训练集.本文实验在测试语料的20个词上进行，上下文窗口为

(−4，+4). 由于文献[4]只给出了 SENSEVAL-3 中 8 个词的消歧结果，为了便于比较，表 5 中只给出了 8 个词的消歧结果，而在全部 20 个词上运行本文方法的消歧结果如表 6 所示. 其中准确率定义如下：

$$\text{准确率}=\frac{\text{正确消歧的词个数}}{\text{测试文本可消歧的词个数}}\times 100\%$$

表 5　本文方法和基于词典资源的词语相似度方法[4]的消歧结果对比

Tab. 5　Comparison of the results for the method in this paper and the method based on word semantic similarity[4]

歧义词	准确率/%	
	本文方法	基于词语相似度方法[4]
材料	65.00	36.67
分子	77.78	51.92
运动	57.40	45.68
穿	64.29	41.46
日子	51.06	39.13
没有	70.00	35.56
少	61.90	41.67
地方	69.44	37.74
均值	64.61	41.22

表 6　本文方法与基于语料库的方法在 SENSEVAL-3 汉语词义消歧任务上的计算结果比较

Tab. 6　Comparison of computation results for the method in this paper and corpus-based methods on SENSEVAL-3 Chinese WSD task

方法	是否基于语料库	准确率/%
I^2R_WSD System[9]	是	60.40
HLTC HKUST Comb2[10]	是	66.20
HLTC HKUST Comb[10]	是	66.50
HLTC HKUST Me[10]	是	64.40
UMD_SST3[11]	是	51.30
本文方法	否	64.61

3.2　实验分析

3.2.1　*本文方法与基于词典资源的同类方法比较*　本文在基于词典资源方法中选择基于词语相似度方法[4]对比的原因在于两者使用了相同的测试文本，而目前检索到的其他同类方法[2,3]大部分都是使用自建的语料库进行测试，无法对比.

由表 5 结果可以看出，本文方法与基于词语相似度的方法[4]相比，准确率提高 22%. 这说明基于义原关系的消歧策略起了作用，其原因分析如下：

(1)基于部件-整体关系的消歧方法提高了对名词歧义词的消歧准确性. 例如，对“分子”一词消歧的准确率很高，主要是因为“分子”含有部件义原“physical|物质”，可进行基于部件-整体关系的词义消歧.

(2)基于属性-宿主关系和值-属性关系的消歧方法提高了对名词和形容词的消歧准确性. 比如对“少”一词消歧的准确率较高，主要是因为“少”含有表示属性值关系的义项：“aValue|属性值，age|年龄，young|幼”，借助基于属性值关系的消歧方法，此义项在消歧时的识别率很高.

3.2.2　*本文方法与基于语料库的方法比较*　这属两类方法的比较. 一般而言，基于词典资源的方法比基于语料库的方法准确率低，但无需耗费大量人力建立标注语料库，不存在数据稀疏问题. 由表 6 的结果对比可以看出，本文方法的准确率除了比表 6 的 5 种基于语料库方法[7,9~11]中的 HLTC HKUST Comb2 和 HLTC HKUST Comb 略低外，比其他三种方法都要高.

为下一步的研究工作考虑，分析本文方法准确率不能继续提高的原因如下：

(1)基于义原关系的词义消歧方法有时候起消极作用. 例如对“日子”一词消歧时，基于部件-整体关系的词义消歧方法总是在不合适的时机插入工作，导致“日子”一词的消歧准确率不高.

(2)本文的消歧方法所借助的知网中的义原关系，主要作用于名词和形容词，而对其他词性的歧义词消歧只依靠中文信息结构和相关度计算.

4　结　语

与基于语料库的词义消歧方法相比，本文方法无需人工构造规则或标注语料，也没有对语料库的依赖性. 与其他基于知网的词义消歧方法[2~4]相比，本文方法不仅是把知网当做一个知识词典来使用，而且能综合知网中所体现的自然语言中概念之间和概念的属性之间的关系，最大限度地利用知网以提高消歧的准确率.

参考文献：

[1] IDE N, VERONIS J. Introduction to the special issue on word sense disambiguation:the state of the art [J]. **Computational Linguistics**, 1998, **24**(1):1-40

[2] 杨尔弘,张国清,张永奎. 基于义原同现频率的汉语词义排歧方法[J]. 计算机研究与发展, 2001, **38**(7):833-838

[3] 闫 蓉,张 蕾. 一种新的汉语词义消歧方法[J]. 计算机技术与发展, 2006, **16**(3):22-25

[4] 余晓峰,刘鹏远,赵铁军. 一种基于《知网》的汉语词语词义消歧方法[C] // 第二届学生计算机语言学研讨会. 北京:中国中文信息学会, 2004

[5] WANG Chi-yung. Knowledge-based sense pruning using the HowNet:An alternative to word sense disambiguation [D]. Hong Kong:Hong Kong University, 2002

[6] DONG Zheng-dong, DONG Qiang. HowNet [EB/OL]. [2010-04-23] . http://www.keenage.com/

[7] 刘 群,李素建. 基于《知网》的词汇语义相似度计算[C] // 第三届汉语词汇语义学研讨会论文集. 台北:中文与东方语言信息处理学会, 2002:59-76

[8] ACL. SENSEVAL-3 [EB/OL]. [2010-04-23]. http://www.senseval.org/

[9] NIU Zheng-yu, JI Dong-hong, TAN Chew-lim. Optimizing feature set for Chinese word sense disambiguation [C] // **SENSEVAL-3:Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text**. Barcelona: ACL-SIGLEX, 2004:191-194

[10] CARPUAT M, SU Wei-feng, WU De-kai. Augmenting ensemble classification for word sense disambiguation with a kernel PCA model [C] // **SENSEVAL-3:Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text**. Barcelona:ACL-SIGLEX, 2004:88-92

[11] CABEZAS C, BHATTACHARYA I, RESNIK P. The University of Maryland SENSEVAL-3 System descriptions [C] // **SENSEVAL-3:Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text**. Barcelona: ACL-SIGLEX, 2004:83-87

Multi-strategy approach to Chinese word sense disambiguation based on sememe relations

CHE Chao[1], JIN Bo[1], TENG Hong-fei[*1,2], QU Fu-zheng[2]

(1. School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China;
2. School of Mechanical Engineering, Dalian University of Technology, Dalian 116024, China)

Abstract: Word sense disambiguation (WSD) is an important and difficult problem in natural language processing. A multi-strategy approach for WSD is proposed based on sememe relations in HowNet. The approach benefits from part-whole and attribute-host relations, which are the most basic and important sememe relations in HowNet. In addition, value-attribute relation, message structure and semantic relevancy are utilized to do disambiguation as supplement. The experimental results of SENSEVAL-3 Chinese WSD task show that the proposed method performs well in comparison with the official results.

Key words: HowNet; word sense disambiguation; sememe relation; semantic relevancy