文章编号: 1003-0077(2007)03-0099-07

中文词语语义相似度计算 ——基于《知网》2000

李 峰,李 芳

(上海交通大学 计算机科学与工程系,上海 200240)

摘 要: 词语语义相似度的计算,一种比较常用的方法是使用分类体系的语义词典(如 Wordnet)。本文首先利用 Hownet 中"义原"的树状层次结构,得到"义原"的相似度,再通过"义原"的相似度得到词语("概念")的相似度。本文通过引入事物信息量的思想,提出了自己的观点: 认为知网中的"义原"对"概念"描述的作用大小取决于其本身所含的语义信息量;"义原"对"概念"的描述划分为直接描述和间接描述两类,并据此计算中文词语语义相似度,在一定程度上得到了和人的直观更加符合的结果。

关键词: 计算机应用;中文信息处理;词语语义相似度;知网;"义原";语义信息量

中图分类号: TP391 **文献标识码**: A

An New Approach Measuring Semantic Similarity in Hownet 2000

LI Feng, LI Fang

(Department of Computer Science and Technology, Shanghai Jiao Tong university, Shanghai 200240, China)

Abstract: A basic approach for measuring semantic similarity/distance between words and concepts is to use lexical taxonomy, such as Wordnet. Hownet is a Chinese semantic dictionary, containing abundant semantic information and ontology knowledge, but has quite different construction and architecture. In this paper, we present a new approach using Hownet by drawing in the idea of information theory. We propose that the more semantic information a "sememe" take, the more powerful it in describing concepts. Then we divide "sememe" which describes a concept into two set: directly describing part and indirectly describing part. In the experiments, we demonstrate our method have improved performance in measuring semantic similarity between Chinese words.

Key words: computer application; Chinese information processing; semantic similarity; Hownet; "sememe"; semantic information

1 引言

语义相似度,在信息检索,信息抽取,词义排歧,机器翻译等都有很大的应用。词语的语义相似度的计算,主要有两类计算方法: 一类是通过树型的义类词典来获得;一类是通过词语上下文的统计背景信息获得。在一颗或几颗树上计算节点的相似度的方法研究相对比较成熟,比如Resnik's[1]、Dekang Lin[2]都给出了比较合理的计算理论和公式。

但中文词语的相似度计算并不能直接借用国外研究人员在 Wordnet 中的方法。原因在于知网并没有像 Wordnet 一样将所有的词组织在一个分类的层次体系中(树状结构中),而是精心选取了一个"语义单位"——"义原"的集合,然后用这个集合中的元素来描述中文词语/概念。"义原"被组织在几颗层次树中,可以借用在 Wordnet 的分类体系中计算词语相似度的思想。如何通过"义原"的相似度来得到词语/概念的相似度,成为利用知网计算中文词语相似度的关键所在。我们在这篇论文里提出"义原"本身所含信息量具有大小之分,而它所含有的语义信息量决定着它对概念的描述作用(区分此概念和其他概念)。另外,在"义原"对概念的描述方式上,我们也提出了自己的观点: 认为描述/定义一个概念的"义原"分为直

收稿日期: 2006-06-03 **定稿日期**: 2006-12-13

作者简介: 李峰(1983—),男,硕士,主要研究方向为自然语言处理。

接描述和间接描述两个部分。

接下来的第 2 部分,我们将首先从两个角度来简要地介绍《知网》;第 3 部分给出《知网》中词语相似度的计算归结为“概念”相似度的计算;第 4 部分讨论“概念”的相似度如何由描述它的“义原”的相似度得到;第 5 部分给出我们计算“义原”之间相似度所采用的公式。第 6 部分为我们的实验结果和分析。最后第 7 部分是我们的结论。

2 《知网》2000 介绍

《知网》[3]是我国著名机器翻译专家董振东先生逾十年功夫创建的一个知识系统。它含有丰富的词汇语义知识和世界知识,内部结构复杂。我们主要从语义词典和世界知识库两个角度对《知网》进行理解分析。

2.1 《知网》是一部语义词典

《知网》的基本形式是对中文词语的释义和描述。与一般的语义词典如 Wordnet 不同的地方有两点:

第一,词语(概念)的意义不是通过一些其他的常用词语来解释、说明,而是通过“义原”来描述、定义。比如“打”(打篮球,打太极),这个词有一项描述是:

DEF = exercise| 锻炼,sport| 体育

“锻炼”和“体育”就是两个义原。《知网》作者总共定义了 1 600 多个这样的义原 —汉语中“最基本的、不易于再分割的意义的最小单位”,然后用它们来对 3 万多个中文词语进行解释描述。义原的具体分类如下(数字标号为义原个数):

a Event| 事件 813
b entity| 实体 142
c attribute| 属性 / aValue| 属性值 433
d quantity| 数量 / qValue| 数量值 13
e SecondaryFeature| 次要特征 100
f syntax| 语法 41
g EventRole & Features| 动态角色和属性 74

《知网》作者认为义原是比词语更小一级的语义单位,但我们更倾向于这样的理解:这 1 600 多个义原是中文语言的一个核心词语集合,和词语是同一级的语义层次。《知网》用这个核心集合构成的语义内涵(语义特征)去描述所有中文词语。因此,我们认为义原分类隐含着如下的语法结构:“实体”义原,描述万物,名词的核心集合;“事件”义原,描述动作,动词的核心集合;“属性”/“属性值”义原和“数量”/“数量值”义原,描述属性(属性程度),形容词副词的核心集合;“语法”义原,对应助词、代词、介词等不含有直接语义信息或含较少语义信息的词类。“次要特征”义原,专门规定,用来描述事物类概念(名词类)的次要特征。“动态角色和属性”义原,专门规定,描述事件类概念(动词类)的内容和特征。

第二,词语不是组织在一个树状的层次体系中,而是存在一种网状关系[4]。

首先,用来描述词语的义原之间存在多种关系。我们认为在《知网》2000 中,义原之间的主要关系有:上下位关系;属性关系,指“实体”类义原(置于[ ]中,见图 1)和“事件类义原”的共性(置于{ }中);对义关系和反义关系。其中最基本的仍然是树状层次体系中的上下位关系(见图 1)

-entity| 实体
├thing| 万物 [# time| 时间, # space| 空间]
… ├physical| 物质 [!appearance| 外观]
…… ├animate| 生物 [*alive| 活着, !age| 年龄, *die| 死, *metabolize| 代谢]
…… ├AnimalHuman| 动物 [!sex| 性别, *AlterLocation| 变空间位置, *StateMental| 精神状态]
…… ├human| 人 [!name| 姓名, !wisdom| 智慧, !ability| 能力, !occupation| 职位, *act| 行动]
…… └humanized| 拟人 [fake| 伪]
……

图 1 树状的义原层次结构,Entity| 实体

其次,借助一些标识符号对概念进行描述,这些标识符体现了各种关系。(见表 1)

从表 1 的例子中可以看出,《知网》义原加标识符来定义词语的方式不但给出了词语的语义信息,比如“医院”是医疗场所,也显式地给出了概念之间的联系,比如“医治”的实施者是“医生”,受事者是“患者”,而地点是“医院”。又比如:“布”是“衣物”的原材料,而“T 恤”的定义是:DEF = clothing| 衣物, # body| 身。我们就可以推理出“T 恤”的原料是“布”。这种联系正是《知网》作者所要反映的“世

界知识”。

表 1　知网的主要标识符及其代表关系

词语	定　义	标识符	代表关系
鼾声	DEF = sound| 声, # sleep| 睡	#	相关关系
踝骨	DEF = part| 部件, % Animal Human| 动物, bone| 骨	%	部件-整体
颜色	DEF = attribute| 属性, color| 颜色, &physical| 物质	&	属性-宿主
布	DEF = material| 材料, ? clothing| 衣物	?	材料-成品
医院	DEF = InstitutePlace| 场所, @cure| 医治, # disease| 疾病, medical| 医	@	场所/时间-事件
医生	DEF = human| 人, # occupation| 职位, * cure| 医治, medical| 医	*	施事者-事件
患者	DEF = human| 人, * SufferFrom| 罹患, $cure| 医治	$	受事者-事件
得利	obtain| 得到, possession = pros| 益(注:等号左边为“动态角色和属性”类义原)	=	事件的内容和特征

2.2　《知网》是一个世界知识库 (a knowledge base system)

何谓“世界知识库”? 我们引用 Ontology 的定义来说明,“与词典和分类表类似,但包含有更详细的信息,最重要的是其组织方式能够让计算机处理和识别”。比如上文提到的“推理出‘T恤’的原料是‘布’”,《知网》借助于符号标识,让计算机具备了这个层次上的逻辑推理知识能力。《知网》的作者一再强调《知网》是“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”,但本文更多的将《知网》2000 放在语义词典的层次上使用,故不对此作深入的讨论。

3　词语语义相似度的计算

什么是词语语义相似度? Dekang Lin[2] 认为任何两个事物的相似度取决于他们的共性(Commonality)和个性(Differentces),然后从信息理论的角度给出任意两个事物相似度的通用公式:

$$sim(A,B) = \frac{\log p(common(A,B))}{\log p(description(A,B))}$$

其中分子是描述 A, B 共性所需要的信息量的大小;分母是完整的描述出 A, B 所需要的信息量大小。

刘群、李素建[4] 认为两个词语的相似度是它们在不同的上下文中可以互相替换且不改变文本的句法语义结构的可能性大小。在下文中(第 4 部分和第 5 部分)我们分别借鉴了前者的事物信息量的概念;后者的整体相似度由部分相似度合成的思想。

词语存在着一词多义的现象,知网中的一词多义表现为单个词语有多个概念,每个概念由一项定义来描述。比如:“打”在“打架”,“打太极”,“打猎”中的意义各不相同,知网中对应的概念描述分别是:

DEF = fight\| 争斗
DEF = exercise\| 锻炼, sport\| 体育
DEF = catch\| 捉住, # animal\| 兽

词语语义相似度的计算,严格来讲,应该是计算概念之间的语义相似度。本文中采用刘群[4] 的思路,认为两个孤立词语(不处在一定的上下文背景中)的语义相似度是其所有概念之间相似度的最大值。

$$Sim(W_1,W_2) = \max Sim(C_{1i},C_{2j}) \quad i = 1\cdots n, j = 1\cdots m \qquad (1)$$

其中, C_{1i} 是词 W_1 的 n 项概念(词义), C_{2j} 是 W_2 的 m 项概念。

4　概念相似度的计算

我们先假定得到了任意两个义原之间的相似度(第 5 部分介绍),现在讨论如何利用义原之间的相似度合成两个概念的相似度。假定描述两个概念的义原集合是:

$$C_1 = \{S_{11}, S_{12}, \cdots, S_{1n}\}, \quad C_2 = \{S_{21}, S_{22}, \cdots, S_{2m}\}$$

问题即为如何由集合中元素的相似度得到整体的相似度。一种比较直观的方法是先寻找最优匹配,集合中彼此最相似的元素两两组合,然后加权取均值就是整体的相似度。每组义原之间的相似度在整体相似度中的权值大小,我们遵循以下两个观点:

一、每个义原在定义概念中的作用大小不同。一个义原所携带的语义信息越丰富,对概念的描述就越具决定性作用(即越能区分此概念同其他概念),相应的其在概念相似度计算中的比重就越大。怎样判定一个义原携带的语义信息丰富与否? 我们认为一个义原所代表的语义内涵越具体,其语义信息就越丰富。反映在义原层次树上,层次越深/越靠近叶子的节点,该义原语义信息越丰富。由此得到

计算公式(假定 $m \leq n$):

$$Sim(C_1, C_2) = \sum_{i=1}^{n} w_i Sim(P_i) \quad (2)$$

其中 $P_i = \begin{cases} < A_{1i}, A_{2i} >, 1 \leq i \leq m \\ < \Phi, A_{2i} >, m < i \leq n \end{cases}$

$$w_i = \frac{d_i}{\sum_{i=1}^{m} d_i}, \quad 1 \leq i \leq n$$

$$d_i = \min(depth_{A_{1i}}, depth_{A_{2i}})$$

P_i 是一组元素配对——最大匹配取得,Φ表示和空元素对应,w_i 是这组元素的相似度在整体相似度中的权值,d_i 是两个义原在义原层次树中深度的较小值,根节点层次设为 1。若义原和空元素对应,则 $Sim(p_i)$取一较小的值 γ(参数)。

二、我们认为在概念的定义中不带符号的义原是对概念的一种直接描述,表明一种 is a 的定义关系或者是识别该概念必不可少的特征(属性);带有符号标识的义原是对概念的一种间接描述,表明概念的一些其他属性。两者对概念的描述作用大小不同,应该分成两组集合分别计算,然后再加权求均值。同时,前者的权值应该更大。公式(2)变为:

$$Sim(C_1, C_2) = \beta \times Sim(C_{11}, C_{21}) + (1 - \beta) \times Sim(C_{12}, C_{22}) \quad (3)$$

C_{i1}是 C_i 中不带符号义原的集合,C_{i2}是 C_i 中带符号义原的集合,β为调节参数。

此外,在概念的定义中,有时候会出现不用义原而直接使用其他词语来描述的情况——出现在一个括号内。比如,"佛教徒":

DEF = human|人,religion|宗教,(Buddhism|佛教)

"盟军":

DEF = arm|军队,#country|国家,*ally|结盟,military|军,desired|良,#(WWⅡ|二战)

我们统一规定此时词语和词语若相同,则相似度为 1,否则相似度为 0;词语和义原之间的相似度则统一取较小值 η。词语的"层次深度"统一设为一个较小值 h(参数)。

在实际的计算中,带符号的义原之间分组应该是带有相同符号的义原配对,如果仍旧使用最大匹配,相似度的计算成为一种相关度的计算。比如"医生":

DEF = human|人,#occupation|职位,*cure|医治,medical|医

"患者":

DEF = human|人,*SufferFrom|罹患,$cure|医治

采用最大匹配:

{〈human|人,human|人〉,〈cure|医治,cure|医治〉,〈medical|医,SufferFrom|罹患〉,〈occupation|职位,null〉}(其中的 null 表示Φ)

采用一一对应:

{〈human|人,human|人〉,〈cure|医治,SufferFrom|罹患〉,〈occupation|职位,null〉,〈medical|医,null〉,〈null,cure|医治〉}

前者的值比后者大,因为"医生"和"患者"十分相关,却不能说很相似。

5 义原相似度的计算

义原相似度的计算依据义原的层次体系(上下位关系)来计算,这种基于树状层次结构计算语义相似度的研究已经十分成熟。Resnik's[1]、Dekang Lin[2]、刘群[4]等都提出了自己的公式,BUDANITSKY[5]对基于 Wordnet 的几种计算方法进行了比较。我们认为他们的方法可以分为两大类:一种是基于两个节点之间的路径长度,一种是基于两个节点所含的共有信息大小。本文分别采用了两种公式来计算义原相似度:

a. 刘群的公式:

$$Sim(S_1, S_2) = \frac{\alpha}{\alpha + \text{distance}(S_1, S_2)} \quad (4)$$

其中,S_1,S_2 表示两个义原,distance(S_1, S_2)表示它们的路径长度,α是一个调节参数,表示相似度为 0.5 时的路径长度。

同时,我们参考吴健,吴朝晖,李莹[6]的计算词汇相似度的思路,引入节点的层次深度:

$$Sim(S_1, S_2) = \frac{\alpha \times \min(depth_{s_1}, depth_{s_2})}{\alpha \times \min(depth_{s_1}, depth_{s_2}) + \text{distance}(S_1, S_2)} \quad (5)$$

这样在路径距离相同的情况,层次越深的节点具有越高的相似度。

b. Lin 的公式:

$$Sim(S_1, S_2) = \frac{2 \times \log p(S_p)}{\log p(S_1) + \log p(S_2)} \quad (6)$$

$$p(S) = \frac{\text{the number of node attached to node S}}{\text{the total node number of the tree}}$$

其中,S_1,S_2 表示两个义原,S_p 表示离他们最近的共同祖先,$p(S)$是该节点的子节点个数(包括自己)与树中的所有节点个数的比。

由于《知网》定义的所有义原并不是在一颗树上,而是构成森林。我们统一规定,不在同一颗树上的两个义原之间的相似度取一较小值 δ(参数)。

如果两个义原之间存在对义或者反义关系(通过查表得到),我们将它们的相似度减低为原来的 n 分之一(参数),比如"大","中","小"在"属性值"这颗义原树上是兄弟节点,按上述公式计算,它们的相似度都很高,但我们认为实际的语言经验中是不会把"大"和"小"作为相似的语义概念来对待的。

6　实验结果与数据分析

6.1　实验一

为了比较,我们选取刘群、李素建[4]论文中的一组实验词语(表 3 的上半部分)并加入几组典型词语(表 3 的下半部分)来说明两种方法的区别。实验中的参数设置见表 2。

表 2　参数设置

β	0.7	不带符号义原集合在整体相似度中的权值
γ	0.0	义原(词语)和空元素之间的相似度
η	0.01	词语和义原之间的相似度
h	5	词语的"层次深度"
α	1.6	见公式(4)
δ	0.01	不在同一颗树上两个义原之间的相似度
$1/n$	5	具有对义反义关系义原相似度的减小

表 3　实验结果

词语 1	词语 2	李素建	刘群,李素建	公式(4)	公式(5)	公式(6)
男人	女人	0.668	0.833	0.815	0.849	0.940
男人	父亲	1.000	1.000	1.000	1.000	1.000
男人	和尚	0.618	0.833	0.938	0.921	0.921
男人	经理	0.351	0.657	0.545	0.530	0.530
男人	高兴	0.024	0.013	0.155	0.191	0.141
男人	收音机	0.008	0.164	0.081	0.159	0.045
男人	鲤鱼	0.009	0.208	0.217	0.357	0.374
男人	苹果	0.004	0.166	0.179	0.313	0.280
男人	工作	0.035	0.164	0.110	0.148	0.013

续表

词语 1	词语 2	李素建	刘群,李素建	公式(4)	公式(5)	公式(6)
男人	责任	0.005	0.010	0.105	0.143	0.006
跑	跳	—	0.444	0.444	0.762	0.606
发明	创造	—	0.615	0.615	0.849	0.891
珍宝	宝石	—	0.130	0.859	0.859	0.859
粉红	深红	—	0.074	0.700	0.7	0.7
出兵	出征	—	0.105	0.172	0.307	0.122
美丽	丑陋	—	0.722	0.452	0.436	0.463
中国	联合国	—	0.136	0.123	0.125	0.123
中国	美国	—	0.940	0.615	0.625	0.615
医生	患者	—	0.574	0.594	0.609	0.608
青山	苍山	—	0.600	0.467	0.467	0.467
香蕉	苹果	—	1.000	1.000	1.000	1.000

1. 比较第 4 列和第 5 列。两者上半部分基本保持一致,无大的波动,下半部分有些数据变化较大,分为以下几类:

i. "珍宝"和"宝石","粉红"和"深红"。第 4 列它们相似度很低,而第 5 列相似度比较高。其原因在于第 4 列的计算方法倚重于第一个义原,比如"珍宝"、"宝石"的定义分别为:

DEF = treasure| 珍宝,generic| 统称
DEF = stone| 土石,treasure| 珍宝

第一义原为"珍宝"和"土石",不具有很高的相似度。按刘群、李素建的方法,它们整体相似度不会很高。而我们的方法是将所有不带符号(独立义原)放在一起计算,并赋予适当的权值。这样"珍宝"和"珍宝"对应,"统称"和"土石"对应,且前一对的权重较大,因此得到整体较高的相似度。

ii. "美丽"和"丑陋"。第 4 列相似度较高,第 5 列较低。因为我们认为具有对义和反义关系的义原即便在义原树上相隔很近(这里是"美"和"丑"),也不能认为它们具有很高的相似度。类似的义原有"大"和"小"、"冷"和"热"等。

iii. "中国"和"美国"。第 4 列给出的相似度十分高,我们的相似度 0.6 多一些。《知网》中,"中国"和"美国"的定义分别为:

DEF = aValue| 属性值,attachment| 归属,# country| 国家,ProperName| 专,(US| 美国)
DEF = aValue| 属性值,attachment| 归属,# country| 国

家,ProperName| 专,(Asia| 亚洲)

据定义似乎应该具有十分高的相似度,它们的区别只有最后一项词语(具体词)描述,这部分相似度为 0。(前文中规定:出现在知网描述中的两个词语若相同,则相似度为 1;不同,则取 0)。在我们的计算中,具体词对整体的相似度影响较大,"层次深度"设为 5。因为我们认为知网描述中出现的具体词包含有较丰富、具体的语义信息,对其所描述词的性质具有直接的决定和影响。

2. 比较第 5 列和第 6 列。后者的上半部分数据略微有些整体上移,下半部分数据中有两对词语相似度明显改善,"跑"和"跳","发明"和"创造"。考察它们的主要定义,都是由单一义原组成:

DEF = run| 跑　DEF = jump| 跳
DEF = produce| 制造　DEF = create| 创造

由于只有单一的义原描述,相似度完全等同于义原之间的相似度。我们认为具有单独描述能力的义原是包含较多语义信息的义原,应该提高它们的相似度,而公式(5)恰恰提高了具有深层次的节点(包含较多语义信息的节点)之间的相似度。

3. 比较第 6 列和第 7 列,两者基本一致。但对于义原树中两个叶子节点而言,公式(6)会给出更高的相似度,比如"男人"和"女人"中的"男"和"女",公式(5)是计算得到的相似度为 0.545,公式(6)计算的相似度为 0.819。我们认为前者更合理一些。因为用 1 600 多个义原来描述所有的中文词语,从语义分布上而言,它们之间应该具有一定的间隔。公式(6)更加适合 Wordnet 这种由大量词语构成的树状体系,节点与节点之间信息相对细微紧密。

4. 虽然我们得到的结果总体来说和人的直观相似,但有些结果显然与实际经验不符。比如按照"男人"和"女人","男人"和"和尚"的相似度,我们可以认为"女人"和"和尚"也十分相似。又比如"青山"和"苍山"的相似度只有 0.467,"香蕉"和"苹果"的相似度为 1(它们的定义都是"fruit| 水果")。这种结果一方面是因为知网的描述在有些地方有待加强修改,进一步的深入细致;另一方面我们对知网的理解也有待进一步的深化。

6.2 实验二

为了更加直观的观察我们方法的效果,我们特别计算了《同义词词林》[7]中同义词语对的相似度。《同义词词林》将汉语的常用词按词义的远近和相关性分成若干词群,每个词群被编排在同一行,我们选择表示同义词群(即去掉相关词群和独立词群)的行。每行选择前两个词进行计算,得到统计结果如下(表 4):

表 4　同义词词林相似度计算

所有词对	Hownet 可计算的词对	0.1～0.2	0.2～0.3	0.3～0.4	0.4～0.5	0.5～0.6	0.6～0.7	0.7～0.8	0.8～0.9	0.9～1	1
9 959	6 759	59	117	175	327	339	374	462	453	128	4 278
	100 %	0.88 %	1.74 %	2.60 %	4.86 %	5.04 %	5.56 %	6.87 %	6.73 %	1.90 %	63.71 %

1. "Hownet 中可计算的词对"是指在《知网》中可查询的词语。有相当一部分《同义词词林》中的词语在《知网》中没有被收录,说明二者的编撰的确存在较大较别。而在可计算的词对中相似度为 1(即《知网》释义完全相同的词组)占 63.71 %,又说明两者对词语的解释基本保持符合一致。

2. 从表中我们可以看到相似度计算在 0.7～1 之间的词对占了相当一部分,说明我们的方法比较有效。但我们也看到在 0.4～0.7 之间的词对也有不少,我们认为除了《知网》本身有待进一步完善和补充外,通过义原的相似度(相对稀疏的层次结构)来反映大量词语之间的相似度(相对密集)的方法本身是否存在一定的上限是一个需要进一步深入研究的地方。

7 实验结论

《知网》含有丰富的语义信息和世界知识,理解其构建的哲学思想和义原体系,充分利用其特定的描述方式是使用《知网》的关键。

本文在参考刘群、李素建[4]的基础上,提出了自己的观点:首先认为义原携带的语义信息有大小之分,越是处于底层的节点语义信息越丰富;其次认为义原对概念定义作用的大小正是取决于其本身所携带的语义信息;最后,我们将义原对概念的描述分为直接描述和间接描述,并认为直接描述是区分概念

必不可少的语义信息,间接描述是区分概念的补充信息和世界知识。根据这三个基本观点,我们得到了自己的计算公式。最后在实验中和刘群、李素建[4]的结果作了比较,并详细分析了两者差别的地方;同时通过计算《同义词词林》中若干同词义的相似度验证了我们的方法。

在下一步的工作中,我们将改用《知网》2005 的免费版来进行研究,进一步探讨研究知网义原构建体系的特殊性以及如何利用这种特殊性得到更加合理的计算方法。

参考文献:

[1] Eneko Agirre, German Rigau. A Proposal for Word Sense Disambiguation using Conceptual Distance [A]. In: Proceedings of the First International Conference on Recent Advanced in NLP [C]. 1995.

[2] Dekang Lin. An Information-Theoretic Definition of Similarity Semantic distance in WordNet [A]. In: Proceedings of the Fifteenth International Conference on Machine Learning [C]. 1998.

[3] HowNet [R]. HowNet's Home Page. http://www.keenage.com.

[4] 刘群, 李素建. 基于《知网》的词汇语义相似度的计算[A]. 第三届汉语词汇语义学研讨会[C],台北,2002.

[5] BUDANITSKY, A. AND HIRST, G. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures [A]. In: Workshop on WordNet and Other Lexical Resources, Second meeting of the North American Chapter of the Association for Computational Linguistics[C]. 2001.

[6] 吴健, 吴朝晖, 李莹, 等. 基于本体论和词汇语义相似度的 Web 服务发现[J]. CHINESE JOURNAL OF COMPUTERS, 2005, 28 (4).

[7] 同义词词林[R]. http://www.ir-lab.org/.

首届中国语言学期刊主编论坛在绍兴举行

2007 年 4 月 21—22 日,首届中国语言学期刊主编论坛在浙江绍兴举行。会议由中国社会科学院语言所、中国人民大学文学院、商务印书馆联合主办,绍兴文理学院承办。

《中国语文》《民族语文》《当代语言学》《语言文字学》《中国术语研究》《现代外语》《语言学论丛》《汉藏语学报》等 35 家语言学期刊主编或负责人与会。中国社会科学院语言所沈家煊、北京大学陆俭明、南京大学鲁国尧、中央民族大学戴庆厦、北京语言大学赵金铭、中国人民大学殷国光等 40 位专家参加了会议。商务印书馆总经理杨德炎、绍兴文理学院院长王建华致开幕词。

根据全国 2005 年公开发行的人文、社会科学期刊目录统计,每年大约有 6 000 篇左右的语言学论文得以发表。语言学期刊如何提高质量、编出特色、相互交流、谋求合作,显得尤为重要。举办首届主编论坛旨在加强语言学期刊的交流,通过交流办刊经验,探讨期刊的定位与特色,期刊的学术规范,期刊的发展思路等,提高刊物学术质量,促进我国语言学的健康发展。

沈家煊、陆俭明、顾曰国、殷国光、苏新宁等专家分别从如何提高杂志的水平、语言研究的目的、语言学期刊规范和语言学研究学术影响力等角度做了主题发言。与会代表各抒己见,对各自期刊的特点、成功经验与发展方向做了交流。与会专家认为特色与创新是期刊发展的动力,应该加强学术规范,提高编校质量。会议就语言学期刊今后的学风、编风问题、期刊协作问题和发展方向问题等进行了深入的讨论,达成诸多共识。商务印书馆将为语言学期刊提供网络平台,交流学术规范、知识产权等方面的信息,开办商务印书馆语言学期刊网。

乔永