文章编号: 1003-0077(2008)05-0084-06

一种改进的基于《知网》的词语语义相似度计算

江 敏[1],肖诗斌[1,2],王弘蔚[1,2],施水才[1,2]

(1. 北京信息科技大学 中文信息处理研究中心,北京 100101;2. 北京拓尔思信息技术股份有限公司,北京 100101)

摘 要: 中科院刘群的基于《知网》的词语相似度计算是当前比较有代表性的计算词语相似度的方法之一。在测试中我们发现对一些存在对义或反义的词语与同义、近义词语一样具有较高的相似度,一些明显相似的词反而相似度较低,如"美丽"与"贼眉鼠眼"的相似度为 0.814 815,与"优雅"的相似度为 0.788 360,"深红"与"粉红"的相似度仅为 0.074 074,这将不利于进行词语的极性识别。基于文本情感色彩分析的需要,把词语相似度的取值范围规定为[-1,+1],在刘群论文的基础上,进一步考虑了义原的深度信息,并利用《知网》义原间的反义、对义关系和义原的定义信息来计算词语的相似度。在词语极性识别实验中,得到了较好的实验结果: P 值为 99.07%,R 值为 99.11%。

关键词: 计算机应用;中文信息处理;知网;词语相似度;义原;词语极性识别

中图分类号: TP391 **文献标识码**: A

An Improved Word Similarity Computing Method Based on HowNet

JIANG min[1], XIAO Shi-bin[1,2], WANG Hong-wei[1,2], SHI Shui-cai[1,2]

(1. Chinese Information Processing Research Center, Beijing Information Science & Technology University, Beijing 100101, China;
2. Beijing TRS Information Technology CO. LTD, Beijing 100101, China)

Abstract: Word similarity computing based on the "HowNet" of Liu-Qun is a representative method to compute the word similarity. But it is found that some words with contrastive or contradictive meanings are computed with high similarity compared those true synonymous. To resolve this defect for the word polarity analysis, we confine the value of word similarity between [-1, +1] in this paper, and enhance the word similarity computation on the basis of Liu's paper by employing sememes' depth information, the antonym and definition information of the sememe. This method produces a good performance in the word polarity recognition experiment, achieving 99.07% in accuracy and 99.11% in recall.

Key words: computer application; Chinese information processing; HowNet; word similarity; sememe; word polarity recognition

1 引言

词语语义相似度,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: 一类是通过建立各种词典来获得;一类是通过词语上下文的统计背景信息获得。在我国,以《知网》[1]为基础的词汇相似度计算是当前较好的方法之一,在业内有着一定程度的应用[2]。

《知网》是我国著名机器翻译专家董振东先生几十年工夫创建的一个知识系统。它含有丰富的词汇

收稿日期: 2008-04-15 **定稿日期**: 2008-06-02

基金项目: 国家 863 计划重点资助项目(2006AA010105);国家自然科学基金资助项目(60772081);北京市属市管高校人才强教计划项目(PXM2007_014224_044677, PXM2007_014224_044676);北京市教委科技发展计划项目(KM200710772010)

作者简介: 江敏(1983—),男,硕士生,研究方向为信息检索、中文信息处理;肖诗斌(1966—),男,高级工程师,硕导,主要研究方向为信息检索、中文信息处理;王弘蔚(1966—),男,高级工程师,主要研究方向为信息检索、中文信息处理。

语义知识和世界知识,内部结构复杂。中科院的刘群等[3]通过分析《知网》的知识描述结构,利用义原的上下位关系计算义原相似度,进而得到词语的相似度。

实际上在《知网》中,义原之间除了上下位关系外,还有其他关系,如果在计算时把它们考虑进来,可能会得到更精细的相似度度量。在义原层次结构中,义原所处的深度对相似度计算是有影响的。本文在刘群论文的基础上,进一步考虑了义原的深度信息,并利用《知网》义原间的反义、对义关系和义原的定义信息来计算词语的相似度,得到了更好的实验结果。

众所周知,《知网》的建设是一个不断完善的过程,于是文中没有考虑对未登录词的处理。本文假定读者对《知网》的知识结构有一定了解,因此未对《知网》做更多介绍。

2　什么是词语相似度

什么是词语语义相似度?

Dekang Lin[4]认为任何两个事物的相似度取决于它们的共性(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:

$$Sim(A,B) = \frac{\log p(common(A,B))}{\log p(description(A,B))} \quad (1)$$

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[3]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。

刘群的描述是 Dekang Lin 描述的具体化,是基于实例的机器翻译这一研究背景的。

下面是我们用刘群开发的软件做的一组测试。

表 1　用刘群开发的软件做的测试数据

词语一	词语二	相似度
美丽	贼眉鼠眼	0.814 815
美丽	优雅	0.788 360
出生	覆没	0.242 424
粉红	深红	0.074 074

在测试中我们发现:对于一些存在对义或反义的词语与同义、近义词语一样具有较高的相似度,一些明显相似的词反而相似度较低。按照刘群[3]对词语相似度的定义,如果把"贼眉鼠眼"一词去替换原句子中的"美丽"是不合适的,这样可能导致句子的语义就与作者原意相反了。笔者在做文本的情感分析研究的时候发现,像"贼眉鼠眼"和"优雅"与"美丽"一词的相似度都较高,但它们的倾向性却是相反的,因此这种结果不利于进行词语的极性识别分析。

基于此,本文把词语相似度的取值范围规定为[-1,+1]之间:一个词语与其本身的语义相似度为 1,如果两个词语替换后原句的语义取反,那么其相似度为 -1。

3　基于《知网》的语义相似度计算

在《知网》的结构中,词是用概念来描述的,一个词可以表达为几个概念,而概念则用义原来描述。假设 W_1 有 n 个概念 $C_{11}, C_{12}, \cdots\cdots, C_{1n}$,$W_2$ 有 m 个概念 $C_{21}, C_{22}, \cdots\cdots, C_{2m}$,本文中定义两个词语 W_1 和 W_2 的语义相似度是其所有概念之间相似度绝对值的最大值,其符号取该对概念相似度的符号:

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

这样,就把两个词语之间的相似度问题归结到了两个概念之间的相似度问题了。由于所有的概念都最终是用义原(个别地方用具体词)来表示,所以义原的相似度计算是概念相似度计算的基础。

3.1　义原相似度计算

义原相似度的计算一般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。Eneko Agirre[5]、Dekang Lin、刘群等都提出了自己的公式,BUDAN-ITSKY[6]对基于 WordNet 的几种计算方法进行了比较。李峰[7]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。

a. 基于节点间的路径长度(α 表示相似度为 0.5 时的路径长度参数):

$$Sim(S_1,S_2) = \frac{\alpha}{\alpha + dist(S_1,S_2)} \quad (3)$$

b. 基于两个节点所含的共有信息大小(S_p 表示离它们最近的共同祖先,$p(S)$ 是该节点的子节点个数与树中的所有节点个数的比):

$$Sim(S_1,S_2) = \frac{2 \times \log p(S_p)}{\log p(S_1) + \log p(S_2)} \quad (4)$$

此外,吴健[8]认为节点所处的深度对相似度计算是有影响的。同样距离的两个词语,词语相似度随着他们所处层次的总和的增加而增加,随着它们之间层次差的增加而减小。因为层次总和的增加意味着分类趋向细致,和同样词语距离的层次总和较小的词语对比,其相似程度就越高。笔者认为这种说法与实际也是比较符合的。

基于上述的考量,本文提出了如下的公式来计算义原 S_1,S_2 的相似度 $Sim(S_1,S_2)$。

$$Sim(S_1,S_2)=\alpha\times(depth(S_1)+depth(S_2))\Big/\alpha\times(depth(S_1)+depth(S_2))+dist(S_1,S_2)+|depth(S_1)-depth(S_2)| \quad (5)$$

其中,S_1,S_2 表示两个义原,$depth(S_1)$ 表示 S_1 距离根节点的层次,$dist(S_1,S_2)$ 表示它们的路径长度,α 是一个调节参数,表示深度对相似度计算影响的大小。

在《知网》中,一共描述了义原之间的 8 种关系:上下位关系、同义关系、反义关系、对义关系、属性-宿主关系、部件-整体关系、材料-成品关系、事件-角色关系。笔者认为反义关系和对义关系是除上下位关系外最重要的关系,因为其他关系能不同程度地通过上下位关系反映出来。本文在义原上下位关系的基础上,进一步利用了反义、对义关系来计算词语的相似度:

Ⅰ如果两个义原是对义或反义关系的话,它们的相似度为 - 1。

Ⅱ如果两个义原路径中存在对义或反义关系的话,它们的相似度为 $-1\times Sim(S_1,S_2)'$。

其中 $Sim(S_1,S_2)$ 是把距离这义原 S_1,S_2 最近的一对对义或反义关系的义原节点看作同一个节点后得到的路径长度根据公式(3)、(5)计算得来的。例如图 1 中假设 O_5,O_2 是一对反义词,那么 O_{10},O_{15} 的相似度 $Sim(O_{10},O_{15})$ 计算相当于把 O_{10} 连接到 O_2 作为子节点后路径长度为 3,再根据公式(3)、(4)计算得到。

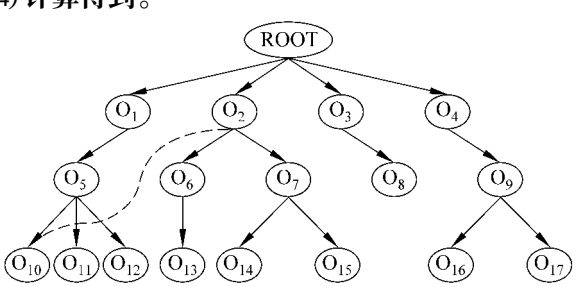

图 1 义原关系树

在《知网》的知识描述语言中,在一些义原出现的位置都可能出现一个具体词(概念),并用圆括号()括起来。所以我们在计算相似度时还要考虑到具体词和具体词、具体词和义原之间的相似度计算。为了简化起见,我们做如下规定:

Ⅰ具体词与义原的相似度一律处理为一个比较小的常数(γ);

Ⅱ具体词和具体词的相似度,如果两个词相同,则为 1,否则为 0。

3.2 概念相似度计算

我们已经得到了任意两个义原之间的相似度的计算方法,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。本文借鉴了刘群论文中概念间整体相似度由部分相似度合成的思想。

虚词与实词、虚词与虚词概念的相似度计算比较简单,本文不再叙述,可参考文献[3],本文主要介绍实词概念的相似度计算。

《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,也显式地给出了概念之间的的联系,主要有下面三种描述形式:

a) 独立义原描述式:用"基本义原",或者"(具体词)"进行描述;

b) 关系义原描述式:用"关系义原 = 基本义原"或者"关系义原 = (具体词)"或者"(关系义原 = 具体词)"来描述;

c) 符号义原描述式:用"关系符号 基本义原"或者"关系符号(具体词)"加以描述;

李峰[7]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,直接描述的作用比间接描述更大。

对于实词概念的相似度,我们将其分成如下部分计算:

Ⅰ独立义原描述部分的相似度 $Sim_1(S_1,S_2)$;

Ⅱ关系义原描述部分的相似度 $Sim_2(S_1,S_2)$;

Ⅲ符号义原描述部分的相似度 $Sim_3(S_1,S_2)$。

于是,两个概念的整体相似度记为:

$$Sim(C_1,C_2)=\sum_{i=1}^{3}\beta_i\prod_{j=1}^{i}Sim_j(S_1,S_2) \quad (6)$$

其中,$\beta_i(1\le i\le 3)$ 是可调节的参数,且有:$\beta_1+\beta_2+\beta_3=1$,$\beta_1\ge\beta_2\ge\beta_3$。

下面我们再分别讨论每一部分的相似度。

I独立义原描述部分的相似度: 独立义原可能不止一个,所以计算较为复杂。我们按照如下步骤对这些独立义原描述式分组:

a) 先把两个表达式的所有独立义原任意配对,计算出所有可能的配对的义原相似度;

b) 取相似度最大的一对,并将它们归为一组;

c) 在剩下的独立义原的配对相似度中,取最大的一对,并归为一组,如此反复,直到所有独立义原都完成分组。

II 关系义原描述部分的相似度: 我们把关系义原相同的描述式分为一组,对组内的基本义原或具体词计算其相似度。

III符号义原描述部分的相似度: 带符号的义原之间分组应该是带有相同符号的义原配对。我们把关系符号相同的描述式分为一组,并计算其相似度。

IV在以上 I、II、III的计算中,最后求平均值。

在实际计算中有一个问题: 如果某一部分的对应物为空,如何计算其相似度? 我们规定任何义原(或具体词)与空值的相似度定义为一个比较小的常数(δ)。

4　实验

4.1　实验一

为了比较,实验一中选了两组数据: 表 2 是实验参数,表 3 中的选词都是一些中性词,表 4 中的词是一些具有褒贬性的倾向性词汇。

表 2　实验参数列表

	α	β_i				γ	δ
刘群[3]	1.6	0.5	0.2	0.17	0.13	0.01	0.01
公式(3)	1.6	0.7	0.17	0.13		0.01	0.01
公式(4) (5)	0.5	0.7	0.17	0.13		0.01	0.01

表 3　实验一中性词结果

词语一	词语二	刘群[3]	公式(3)	公式(4)	公式(5)
男人	父亲	1.000 000	1.000 000	1.000 000	1.000 000
男人	和尚	0.861 111	0.814 815	0.722 222	0.833 333
男人	工作	0.113 060	0.105 263	0.057 017	0.057 017
男人	鲤鱼	0.175 652	0.178 913	0.127 642	0.127 642
男人	收音机	0.093 953	0.085 408	0.190 761	0.065 253
发明	创造	0.615 385	0.615 385	0.615 385	0.850 000
香蕉	苹果	1.000 000	1.000 000	1.000 000	1.000 000
奥运会	词典	0.093 953	0.059 843	0.048 284	0.034 590
医生	医治	0.037 407	0.004 426	0.003 571	0.004 422
珍宝	宝石	0.130 191	0.500 000	0.500 000	0.500 000
中国	联合国	0.124 651	0.257 500	0.257 500	0.257 500
粉红	深红	0.074 074	0.580 867	0.468 667	0.580 867
跑	跳	0.444 444	0.444 444	0.444 444	0.818 182
跳槽	拔脚	0.184 242	0.131 955	0.049 212	0.158 175
预料	检验	0.285 714	0.285 714	0.285 714	0.705 882
风度	面积	0.611 522	0.611 522	0.409 339	0.427 833
名声	硬度	0.617 683	0.617 683	0.396 239	0.565 357

在表 3 中,刘群[3]和公式(3)的实验结果总体上相差不大,但有一些词汇,公式(5)的结果比公式(3)好得多,比如"珍宝"与"宝石"、"粉红"与"深红"。它们在《知网》中的描述为:

珍宝 treasure| 珍宝,generic| 统称

宝石 stone| 土石,treasure| 珍宝

粉红 aValue| 属性值,color| 颜色,red| 红

深红 attribute| 属性,color| 颜色,red| 红,&physical| 物质

刘群[3]认为第一独立义原反映了概念的主要特征,因此赋予较大权重,这样当第一独立义原不同时它们之间的相似度将降低很多。本文中将独立义原统一考虑,避免了含有相同义原时词语相似度反而降低的现象。

再来看公式(3)与公式(5)的结果。两者基本是一致的,但"预料"与"检验"的相似度比"跳槽"与"拔脚"的相似度结果要大得多,分析其原因如下:

预料 predict| 预料

检验 exam| 考试

跳槽 alter| 改变,patient = affairs| 事务

拔脚 start| 开始,content = run| 跑

义原"预料"≠"考试"与"改变"≠"开始"在义原树中的路径长度都为 6,但它们在树中的深度是不一样的:depth(改变) = 5, depth(开始) = 4, depth(预料) = 14, depth(考试) = 13,这也是影响它们相似度的主要因素。

至于"医生"与"医治"的相似度较低也是符合事实的,它们只能说是相关度较高而相似度较低。

表 4 中,一些看似存在一定程度反义的词语,其实际结果也确实为负值,而且公式(5)和公式(3)中对所有的负值,其绝对值都比刘群[3]大一些,这是因为我们把对义、反义义原的路径长度调为 0(如图 1 中所示),义原间路径长度缩短了,根据公式(3)、(5),其相似度的值必然或多或少要大一些。但对"父亲"≠"母亲"、"三伏"≠"冬眠"这样的词语结果仍是正值,这主要是因为在《知网》2000 免费版中,义原"男"≠"女"、"夏"≠"冬"未列入反义、对义词表中,随着《知网》的不断完善,这种现象应该会有所好转。

表 4 实验一褒贬词结果

词语一	词语二	刘群[3]	公式(3)	公式(4)	公式(5)
美丽	贼眉鼠眼	0.814 815	- 1.000 000	- 0.500 000	- 1.000 000
美丽	优雅	0.788 360	0.6111 11	0.617 062	0.750 000
高尚	卑鄙	0.788 360	- 0.750 000	- 0.500 000	- 0.912 500
医生	患者	0.503 701	- 0.447 806	- 0.504 000	- 0.547 805
拜寿	濒临灭绝	0.072 103	- 0.052 959	0.000 000	- 0.083 581
出生	覆没	0.242 424	- 1.000 000	- 0.400 346	- 0.928 571
安康	抱恙	0.210 526	- 1.000 000	- 0.492 804	- 0.875 000
红光满面	受伤	0.186 047	- 0.615 385	- 0.400 345	- 0.749 999
舒服	残废	0.042 771	- 0.179 265	0.003 515	- 0.222 203
健康	耳聋	0.186 047	- 0.219 823	- 0.083 300	- 0.277 230
父亲	母亲	0.861 111	0.814 815	0.950 439	0.888 888
三伏	冬眠	0.044 444	0.005 855	0.004 687	0.005 838

综合两组选词实验结果,公式(4)的效果与公式(3)差不多,但对于本身就是义原的词,其相似度结果都会偏小,这是因为该方法的提出是针对 WorldNet 的,《知网》的知识结构与 WorldNet 是不同的,它只有 1 600 多个义原,树的层次关系比较稀疏简单,不像 WorldNet 是一个由大量词汇构成的树状体系。

4.2 词语极性识别实验

上面词语相似度计算的结果评价是一个非常主观的过程,最好是放到实际的应用中,观察不同的相似度计算方法对实际结果的性能的影响。

词语极性识别就是判断词语的褒贬倾向,一般做法是选定一组极性比较强烈的基准词,通过词语

相似度计算该词的极性,计算公式为:

$$Orientation(W) = \frac{1}{m}\sum_{i}^{m} Sim(base - P_i, W) - \frac{1}{n}\sum_{j}^{n} Sim(base - N_j, W)$$

其中: $base - P_i$ 表示第 i 个褒义基准词, $base - N_j$ 表示第 j 个基准词, m、n 为褒义、贬义基准词的个数。

如果 $Orientation(W) > 0$,则 W 为褒义;如果 $Orientation(W) < 0$,则 W 为贬义。

本文使用朱嫣岚论文[9]中的 40 对褒贬基准词,测试数据也是使用了 HowNet 免费版中文词表中标注"良"(褒义),"莠"(贬义)属性的词汇。总共选用 6 622 词。其中褒义词 3 264 个,贬义词 3 358 个。实验结果如表 5 所示。

表 5　词语极性识别实验结果

	朱嫣岚[9]	公式(3)	公式(4)	公式(5)
准确率	68.04 %	98.92 %	97.74 %	99.07 %
召回率	- -	99.11 %	99.11 %	99.11 %

可见,本文中将反义的相似度取为负值的计算方法可以大大提高词语褒贬性识别的准确率,而且公式(5)的识别效果最好,基本上可以满足实际需求。

5　结论

《知网》含有丰富的词汇语义知识和世界知识,内部结构复杂,是一部比较详尽的语义知识词典。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。词语语义相似度计算是文本情感分析的基础,本文在参考刘群[3]的基础上,对词语相似度的计算方法进行了如下改进:

1) 义原间的相似度大小不仅跟它们之间的路径长度有关,还与它们所处的层次有关: 词语相似度随着它们所处层次的总和的增加而增加,随着它们之间层次差的增加而减小。

2) 在《知网》中义原间除上下位关系外,还有反义、对义等其他关系。本文把词语相似度的取值范围规定为[- 1, + 1]之间: 一个词语与其本身的语义相似度为 1,如果两个词语替换后原句的语义取反,那么其相似度为 - 1。

3) 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,在计算相似度时他们的权重是不同的。

在试验中利用改进的相似度计算方法,与刘群[3]、Dekang Lin[4]的结果进行了比较,实验证明该方法能有效影响词语的相似度计算。

在利用该算法进行词语极性识别的实验中,准确率有了大幅度提高,基本上可以满足实际需求。

参考文献:

[1] 董振东, 董强. 知网[DB/OL], http://www.keenage.com

[2] 夏天,樊孝忠,刘林. 基于 ALICE 的汉语自然语言接口[J]. 北京理工大学学报,2004,24(10): 885-889.

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

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

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

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

[7] 李峰,李芳. 中文词语语义相似度计算 ——基于《知网》2000[J]. 中文信息学报,2007,21(3): 99-105.

[8] 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).

[9] 朱嫣岚,闵锦,周雅倩,黄萱菁,等. 基于 Hownet 的词汇语义倾向计算[J]. 中文信息学报, 2006, 20(1): 14-20.