文章编号: 1003-0077(2008)04-0039-04

语义对立度及其计算模型的研究

麦范金[1] ,王 挺[2]

(1. 桂林工学院 现代教育技术中心 ,广西 桂林 541004; 2. 桂林工学院 电子与计算机系 ,广西 桂林 541004)

摘 要: 人类的思维离不开语言 ,联想思维主要通过相关、相似和对立三种方式。现阶段有关语义的相关和相似的研究已比较多 ,而有关对立的研究却比较少。文章把负值引入到相似度计算中 ,提出对立度等概念和相关的计算模型 ,将它们运用到语义对立程度的计算中 ,并通过仿真试验论证了这些概念模型和计算方法的可行性和有效性。

关键词: 计算机应用 ;中文信息处理 ;语义 ;相似度 ;对立度 ;反义词

中图分类号: TP391 **文献标识码**: A

A Computational Model of The Contrary Degree Between Two Words

MAI Fan-jin[1], WANG Ting[2]

(1. Modern Education Technology Center, Guilin University of Technology, Guilin, Guangxi 541004, China;
2. Department of Electronic and Computer Science, Guilin University of Technology, Guilin, Guangxi 541004, China)

Abstract: Languages play an important part in the thinking of human beings. Relativity, similarity, and contrariety are three ways of the thinking association. At present, the researches of relativity and similarity are well touched while the contrary degree is less studied. This paper introduces the negative value into the similarity calculation and puts forward the conception of the contrary degree as well as a computational model. The feasibility and validity of the proposed conceptual model and computational model are demonstrated with a simulated test.

Key words: computer application; Chinese information processing; semantic meaning; similarity; contrary degree; antonym

1 引言

联想是人类思维的一种方式 ,主要通过相关、相似和对立(相反)三种方式进行[1] 。人类的思维离不开语言 ,因此 ,词语意义的相似度、相关度、对立度的计算对于研究思维 ,特别是联想思维具有重要的价值和意义。

现在 ,在自然语言处理领域中 ,语义相似度和相关度的计算已经成为一个热点。但是 ,关于语义对立度的研究却相当不足。比如 ,“父亲”和“爸爸”是两个意义十分接近的词。“父亲”和“父母”也是意义相近的词。因此 ,我们非常容易从“爸爸”联想到“父母”。同样 ,由于对立意义的存在 ,我们也能很容易地从“父母”联想到“孩子”。如果在未来的某一天 ,机器人不能做到这一点 ,那么这将是令人遗憾的。因此 ,对立作为一种与相关、相似同样重要的联想方式 ,应该受到研究的重视 ,才有利于自然语言处理、人工智能的应用和发展。

本文将首先提出阳系语义与阴系语义的概念 ,据此提出类石墨语义模型来区分两种不同性质的语义。然后在此模型上 ,基于语义距离的概念提出语义位移 ,进而将负值引入到相似度计算中 ,产生对立度的概念 ,并计算对立度 ,最后通过仿真实验验证该模型及其计算方法的正确性 ,为进一步完善基于词义的语义关系计算理论奠定基础。

收稿日期: 2007-06-06 **定稿日期**: 2007-11-19

基金项目: 广西教育厅资助项目(桂教科研[2006]26 号)

作者简介: 麦范金(1963 —) ,男 ,副教授 ,硕导 ,主要研究方向为人工智能 ,自然语言处理;王挺(1981 —) ,男 ,硕士研究生 ,主要研究方向为人工智能 ,自然语言处理。

2 义系

德语、法语、俄语等欧洲语言的词汇都有阴性和阳性之分。汉语的形成与中国古代的易学也有着重要的联系[2]。易学对汉语的影响使其产生了类似德语、法语的词汇阴阳性之分,但又不完全相同。因此,为了避免误解,我们把词汇按其意义分为属阳义系和属阴义系,简称阳系、阴系。属于阳系的词语称为阳系词语,属于阴系的词语称为阴系词语。阳系词语与阴系词语之间有些有对应词语,有些没有对应词语。通常,不同义系之间的对应,表示一种反义或对立的关系。

一般来说,客观的、褒义的、实的、确定的、雄性的、白天的、过去的、物质的、暖色系的,方位上在正的、上面的、凸出的词义都属于阳系;相反,主观的、贬义的、虚的、未定的、雌性的、黑夜的、未来的、精神的、冷色系的,方位上在反面、下面、凹下的词义都属于阴系。

3 语义相似度和对立度的计算模型

据前所述,语义按其所属义系的不同,分成了阳系和阴系两个层面。在这两个层面上的词语,两两互相联系,组成了阳系和阴系两个语义网络。每层语义网络上又有部分词语与另一层语义网络上的词语相对应,它们之间具有对立的关系。如果把词语看成原子,把词语间的关系看成原子之间的相互作用,那么这两层语义网络及词语间的相互关系所组成的结构,可以看作类似两层相临的石墨结构。

石墨由碳元素构成,其结构是由许多平行于基面的层面连续叠合而成的。每一层内,碳原子排列呈正六边形,三个相邻的碳原子以共价键连接,成为一个二度空间无限伸展的网状平面,称为基面。叠在一起的相邻基面依次错开,相邻基面之间的碳原子依靠较弱的范德华力相连,如图 1 所示[3]。从图中我们可以看到,由于相邻基面依次错开,部分原子在其临层中有其对应原子。

抽出相邻的两层,按前述将词语看成原子,将词语间的关系看成原子间的相互作用,称此模型为类石墨语义模型。同层词语属于同一义系,意义相似的词语距离较近,反之较远。层之间的对应词语互为反义,与反义词距离较近的词语是与其意义不完全对立的词语。有些原子在另一层上并没有对应原子,这说明这些词语本身没有反义词。由于词汇较碳原子拥有更大的自由度,所以,同一层内,与某一词语相邻的词语个数没有限制,可以是一个或多个。图 2 为类石墨语义模型的示意图。从图中可以看到,阳系词语 a、b、c、d 分别对应阴系词语 a'、b'、c'、d',它们互为反义词。而阳系词语 y 没有对应的阴系词语,阴系词语 x 没有对应的阳系词语,但它们却与 a、b、c、d、a'、b'、c'、d' 有着一定的相似或对立关系。而且,在同一层中,与词语 b、c、b'、c' 直接相临的词语个数多于 3。

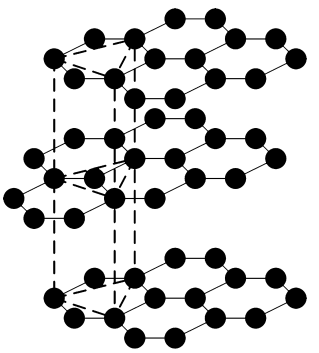

图 1 石墨结构示意图

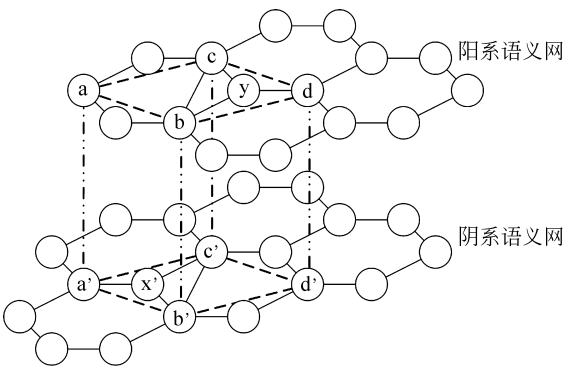

图 2 类石墨语义模型示意图

4 语义相似关系的度量与计算

下面给出几个概念,用于衡量语义之间的相似性关系。

4.1 语义距离

语义距离可以用来反映两个词语语义之间的相似程度。语义距离的计算可以通过根据某种世界知识(Ontology)或大规模的语料库的统计来进行。一般认为,基于世界知识的方法简单有效,也比较直观、易于理解[4]。而且上文所提出的类石墨模型本身就是一种网状的世界知识语义词典,因此,本文采

用基于类石墨模型来计算语义距离。

定义 1: 词语全集为 $I, a \in I, b \in I$,记 $Dist(a, b)$ 为 a 与 b 的语义距离, $Dist(a, b) \in [0, +\infty)$。若 a 与 b 为相邻节点,则 $Dist(a, b) = 1$;若 a 与 b 为完全对立节点,则 $Dist(a, b)$ 为一个无穷小量。

4.2　语义位移

语义距离虽然可以用于描述两个词语之间的意义差距,但无法描述两个语义之间的对立。语义位移正是为了能够在有效地描述两个词语之间意义差距的同时,说明语义之间的对立状况而提出的。

定义 2: 词语全集为 I, A 为阳系词语集合, B 为阴系词语集合, $A \subset I, B \subset I$,若 $a \in A, b \in A$ (或 $a \in B, b \in B$),则称词语 a 与词语 b 同层;若 $a \in A, b \in B$ (或 $a \in B, b \in A$),则称词语 a 与词语 b 不同层。

定义 3: 记 $Disp(a, b)$ 为词语 a 与词语 b 之间的语义位移,

$$Disp(a,b) = \begin{cases} Dist(a,b) & \text{当 a 与 b 同层时} \\ -Dist(a,b) & \text{当 a 与 b 不同层时} \end{cases} \quad (1)$$

显然, $Disp(a, b) \in (-\infty, +\infty)$。

通过语义位移的正负值,可以准确地标示两个语义分属的义系,以及它们之间是否存在对立关系。值得注意的是,语义位移不能直接进行加减计算。要计算距离较远的语义位移,必须先将其转化为语义距离,然后通过加减语义距离来得到语义位移的值。

4.3　语义相似度与对立度

语义位移能够描述两个语义之间是否存在对立关系。但这种描述往往是粗糙的,更细致的描述可以通过语义相似度(或语义对立度)来描述。

定义 4: 词语全集为 $I, a \in I, b \in I$,记 $Sim(a, b)$ 为 a 与 b 的相似度, $Sim(a, b) \in [-1, 1]$。当 $Sim(a, b) = 1$ 时, a, b 为同义词语,即 $a = b$;当 $Sim(a, b) \in (0, 1)$ 时, a, b 为近义词语;当 $Sim(a, b) = 0$ 时, a, b 无相似性;当 $Sim(a, b) \in (-1, 0)$ 时, a, b 为意义不完全相对的反义词语;当 $Sim(a, b) = -1$ 时, a, b 为意义完全相对的反义词语。特别地,当 $Sim(a, b) \in [-1, 0)$ 时, $Sim(a, b)$ 可以称为词语 a 和 b 的对立度,记为 $Con(a, b)$,此时 $Con(a, b) = Sim(a, b)$。

关于对立度的理解,通俗地说,当 $Sim(a, b) \in (-1, 0)$ 时,说明 a 与 b 的反义词相似;当 $Sim(a, b) = -1$ 时,说明 a 等价于 b 的反义词,也就是 a, b 互为反义词。

4.4　语义位移与语义相似度

语义位移与语义相似度之间有着密切的关系: 两个语义位移的绝对值越大,其相似度的绝对值越小;反之,两个语义位移的绝对值越小,其相似度的绝对值越大。

定义 5: 词语全集为 $I, a \in I, b \in I$,语义相似度与语义位移之间的关系如下:

$$Sim(a,b) = \begin{cases} \lambda^{Disp(a,b)}, & Disp(a,b) \in [0, +\infty) \text{ 时} \\ Con(a,b) = -\lambda^{-Disp(a,b)}, & \\ & Disp(a,b) \in (-\infty, 0) \text{ 时} \end{cases} \quad (2)$$

其中, λ 为一个常量参数, $\lambda \in (0, 1)$,表示当语义距离为 1 时的语义相似度。特别地,当 $Con(a, b) = -1$ 时, $Disp(a, b) = -\varepsilon, Dist(a, b) = \varepsilon$,其中 ε 为一个无穷小量。

其实,早就有研究发现,在词语相似关系中不仅包含了词语的同义或近义关系,还包含有反义关系和一些其他相关关系[5]。在式(2)中,当 $Disp(a, b) \in (-\infty, 0)$ 时,语义相似度的计算实际就是语义对立度的计算。因此,语义对立度的计算与语义相似度的计算应该是一个统一的概念。

5　仿真实验与结果

根据上述计算方法和模型,将文献[6]中所列的词语及一些常用专业术语共计约 4 100 条写入数据库中,人工标注这些词语的阴阳。其中,一些中性词标注为阳系词语。然后根据文献[6]中对词语的分类和文献[7],人工标注词语间的相似或对立关系,确定了一部分同义词和反义词。然后以这部分词为关键节点,构建类石墨语义模型。按上一节最后所述的方法进行语义相似度和对立度计算。在实验中,设定(2)式中的参数 $\lambda = 0.950$,采用广度优先算法测量两个词语之间的语义距离。为了比较,我们将部分结果与文献[4]中的实验结果一起列入表 1 中。

从表 1 中可以发现,类石墨语义模型及其计算方法不仅可以有效地计算两个语义之间的相似度,而且可以通过计算结果的正负得知两个词语之间是否存在对立。这是使用其他方法所不能实现的。同时,实验结果与人类平时的思想也比较贴近。

表 1 语义相似度和对立度计算结果

词语 1	词语 2	李素建的方法	刘群、李素建的方法	本文的方法
男人	女人	0.668	0.833	- 0.902 5
男人	父亲	1.000	1.000	0.814 5
男人	母亲	0.668	0.833	- 0.814 5
男人	和尚	0.668	0.833	0.814 5
男人	经理	0.351	0.657	0.773 8

6 结论

人类的思维离不开语言。联想思维主要通过语义的相关、相似和对立等关系进行。现有的自然语言处理技术在语义关系方面的研究偏重于相似和相关计算,有关语义对立关系的研究较少。基于此,本文首先提出了语义的阳系与阴系的概念,同时提出了类石墨语义模型,利用此模型将负值引入到语义相似度计算中,产生语义位移、对立度等概念,不仅可以计算词语的正相似度,也可以计算词语的负相似度,即对立度。最后,通过仿真实验验证了该模型及其计算方法的可行性和有效性。

参考文献:

[1] 商艳芝. 提高联想能力,加强听力记忆[J]. 外语电化教学, 2002,86:12-15.

[2] 孔刃非. 汉字全息学[M]. 北京:华艺出版社,2005:223-237.

[3] 宋正芳. 碳石墨制品的性能及其应用[M]. 北京:机械工业出版社,1987:1-3.

[4] 刘群,李素建. 基于《知网》的词汇语义相似度计算[J]. Computational Linguistics and Chinese Language Processing, 2002,7(2):59-76.

[5] 胡俊峰,俞士汶. 唐宋诗中词汇语义相似度的统计分析及应用[J]. 中文信息学报,2002, 16(4):39-44.

[6] 郑林曦. 普通话三千常用词表[M]. 北京:文字改革出版社,1987. 12.

[7] 李志江,宋惠德,曹兰萍. 汉语同义词反义词词汇[M]. 社会科学文献出版社,1991.

(上接第 23 页)

[3] Lynn Carlson, Daniel Marcu, and Mary E. Okurowski. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory [C]//Jan van Kuppevelt and Ronnie Smith, editors, Current Directions in Discourse and Dialogue. Kluwer Academic Publishers. 2003.

[4] Manfred Stede. The Potsdam Commentary Corpus. [C]//Proceedings of the ACL 2004 Workshop Discourse Annotation', Barcelona. 2004.

[5] R. Soricut and Daniel Marcu. Sentence level discourse parsing using syntactic and lexical information [C]//Proceedings of Human Language Technology and North American Association for Computational Linguistics Conference (HLTNAACL'2003). Edmonton, Canada.

[6] J. Burstein and Daniel Marcu. A machine learning approach for identification of thesis and conclusion statements in student essays [J]. Computers and the Humanities. 2003,37(4), 455-467.

[7] Benjamin K. T'sou, Lin H. L., Ho H. C., Lai T. and Chan T. Automated Chinese Full-text Abstraction Based on Rhetorical Structure Analysis [J]. Computer Processing of Oriental Languages. 1996,10(2): 225-238.

[8] 张益民,陆汝占,沈李斌. 一种混合型的汉语篇章结构自动分析方法 [J]. 软件学报,2000,(11).

[9] YUE Ming. Discursive Usage of Six Chinese Punctuation Marks [C]//Proceedings of COLING/ACL-2006 Student Research Workshop. Sydney. July 2006. 43-48.

[10] 王维贤,张学成,卢曼云,等. 现代汉语复句新解[M]. 华东师范大学出版社. 1994.

[11] 吴为章,田小琳. 汉语句群[M]. 北京: 商务印书馆. 2000.

[12] 吴启主. 汉语构件语篇学[M]. 岳麓书社. 2001.

[13] 邢福义. 汉语复句研究[M]. 北京: 商务印书馆. 2002.

[14] Daniel Marcu. The Theory and Practice of Discourse Parsing and Summarization [M]. The MIT Press. 2000.