中文语义相关度计算模型研究

王红玲,吕 强,徐 瑞
WANG Hong-ling,LV Qiang,XU Rui

苏州大学 计算机科学与技术学院 江苏省计算机信息处理技术重点实验室,江苏 苏州 215006
Jiangsu Provincial Key Lab of Computer Information Processing Technology,School of Computer Science and Technology,Soochow University,Suzhou,Jiangsu 215006,China
E-mail:redleaf@suda.edu.cn

WANG Hong-ling,LV Qiang,XU Rui.Computation model of Chinese semantic relevancy based on howNet.Computer Engineering and Applications,2009,45(7):167-170.

Abstract: The current computation models of Chinese semantic relevancy variate due to various definitions of relevancy degree. These models are almost based on a similarity computation model,which much limits its application.This paper presents a new definition of semantic relevancy,and defines two words semantically related if the two concepts associated with the two words respectively have the relationship in a semantic knowledge base like HowNet.By mining all the direct and indirect relationships,a new computation model of semantic relevancy is proposed.This model can be generalized to any semantic knowledge bases constructed by the knowledge dictionary mark-up language.Finally,it applies the computation model to the word sense disambiguation problem,and therefore validates that this computation model can be efficiently employed to many applications.
Key words: semantic relevancy;mining semantic relationships;HowNet

摘 要:现有的中文语义相关度计算模型对相关度的定义并不明确和统一,且计算方法多以相似度计算为基础,导致应用语义相关度存在局限。提出了一个新的语义相关的定义,认为两个词所表达的概念之间,如果存在用类似"知网"的知识描述体系所描述的语义关系,那么这两个概念之间就是语义相关的。通过挖掘这些直接或间接的关系,提出了一种新的语义相关度的计算模型,适用于所有类似知网的知识体系中语义相关度的计算。最后将该计算模型应用于词义排歧,验证了该计算模型的有效性。
关键词:语义相关度;语义关系挖掘;知网
DOI:10.3778/j.issn.1002-8331.2009.07.050 **文章编号**:1002-8331(2009)07-0167-04 **文献标识码**:A **中图分类号**:TP391

1 引言

相关度计算是各种信息系统中极为重要的基本技术。语义相关度计算可以为许多领域,如文本聚类、词义消歧、语句对齐等研究工作,提供具体而有效的支持。目前中文语义相关度的计算研究不多,在现存的计算模型中,对相关度计算所度量的关系不明确。同时,相关度的计算方法多以相似度计算为基础。这导致了相关度计算的应用方面的局限性。提出了一种新的中文语义相关度计算模型,明确定义语义相关的概念,并基于知网提出通过关系挖掘来计算语义相关度的计算模型,从而可以真正地在语义层次上度量了两个语言现象(本文中只是词)之间的相关程度。

2 相关研究

语义相关度有两类常见的计算模型,第一类是统计方法,采用概率统计,参数估计和特征获取等学习模型获得词语共现规律,计算语义的相关程度;第二类是根据语言知识和分类体系计算,通常的方法是利用知网或《同义词词林》等知识源提供的分类体系,去计算词语间的语义相关程度。

在统计方法中,通过对词语共现频率的计算,确定词语间的相关程度,认为共现频率越高的词语,他们之间的相关程度也越高。虽然共现频率在一定程度上反映了词语之间的关联程度。但统计方法无法对词语做进一步的语义解析。对于词语间的具体关系如何,关联的语义依据是什么等问题,基于统计的词语共现频率计算模型是无法解决的。而相关是从语义的角度提出的指标,如果单纯用频率统计作为语义的衡量标准,这是不完全合理的。在分类体系方法中,文献[1]给出了领域模型内基于实体非相关度的本体语义相关度的计算模型,事实上,这种具体方法的本质是解析用领域本体描述语言 DODL 描述的领域模型。文献[2]给出了基于概念层次网络 HNC 的两个词语语义相关度的计算模型,实际上,语义知识库是一个树结构。文献[3]根据《同义词词林》中对每个词的归类关系,给出了词语间相关度的计算模型。这种归类关系本质上是搭配关系的一种模

基金项目:2006 年高等学校博士学科点专项科研基金(No.20060285008)。
作者简介:王红玲(1975-),女,博士研究生,主要研究方向:自然语言处理;吕强(1965-),男,教授,主要研究方向:智能计算、中文信息处理;徐瑞(1980-),女,硕士研究生,主要研究方向:中文信息处理。
收稿日期:2008-01-21 **修回日期**:2008-03-31

糊体现。文献[4]提出了基于语义相似度及知网语义信息,计算语义相关度的方法。该方法以义原的相似度计算为基础,结合知网中义原间的上下位关系,作为词语间纵向联系的依据,辅以实例信息,综合这三个方面的影响因素进行相关度的计算。这种计算方法充分利用了相似度高,相关度也高的原则,但无法覆盖相关度高,可能相似度低的情况。对于词语之间相互关联的关系,该计算方法只考虑了可能存在的上下位关系,但对于其它关系并没有充分考虑,这将影响计算结果的准确性。文献[5]在语句的相关度计算中提到利用《同义词词林》、知网两种语言资源计算词与词之间的相关度,继而得到语句间的相关度。其中所使用的相关度计算方法虽与上述方法不同,但仍然将相关度计算构建在相似度计算的基础之上,也没有充分考虑词语间的关系,因此同样存在相关计算不完整的缺点。

3 语义相关度计算模型

3.1 语义相关概念

定义 1 两个语言单位之间存在概念相关,就认为它们语义相关。所谓概念相关是指概念之间存在知网定义关系中的一种或多种关系。概念之间存在的知网关系越多,认为概念相关度就越高。所谓相关度,是对概念相关程度的数量刻画。

本文定义的相关是基于知网的,更准确地说是基于类似于知网的知识表达体系的,因而,使用知网知识描述语言 KDML 所表达的知识均可以用以下公式计算概念间的相关度,所以又可以认为本文的相关定义具有一定的通用性。关于两个语言单位 w_1 和 w_2 的语义相关度 SR 计算概念模式(computational conceptual model)如下:

$$SR(w_1,w_2)=\sum_{r_i\in R} p_i \cdot r_i(w_1,w_2) \quad (1)$$

其中,$r_i(w_1,w_2)$计算了 w_1 和 w_2 是否存在概念相关(直接关系或间接关系)关系:是,返回 1,否则返回 0。p_i 既可以理解为关系 r_i 的权值,又可以理解为关系 r_i 对相关度贡献程度,所以有 $\sum P_i=1$。

事实上,上面概念模式只是示意了相关度计算应该考虑的属性,具体地在应用中,可以有许多特定的实现。见下面的计算实例。

3.2 语义关系挖掘依据

定义 2 语义关系挖掘是指,借助类似知网提供的语义知识库,找出概念的所有可能发生的语义关系。

将计算模型涉及的知网关系分为两类:直接关系与间接关系。所谓直接关系是知网明确定义的 16 种关系;而间接关系是指根据语义关系挖掘规则获得的关系。在知网定义的关系中,义原关系是其描述关系的重要体现,通过对义原关系的计算,获得概念间的关系。知网义原关系是通过特征文件建立起来的。整体为树状结构,每一个节点代表了一个主义原,后面所跟方括号里的内容是该义原的一些解释义原,每个解释义原通过关系符号表示与主义原的关系。解释义原和主义原都具有上下位节点,因此不同特征集中的义原产生关联关系[6]。一棵树中上下位义原之间的关系,称为纵向关系。而在义原构成的体系中,每个义原和不在同一个树中的义原也可能有关联,这样就为义原体系的树状层次结构增加了横向联系,从而使整个义原体系呈现网状结构[7]。在知网定义的关系结构中,为这些关系分别定制了一系列语义关系挖掘的规则。

3.3 挖掘语义关系规则

3.3.1 部分-整体关系挖掘规则

在上下位结构中,根据义原的继承特性,下位义原继承上位义原的所有解释义原所描述的特性。因此以下位义原为主要特征的概念与原概念也构成部分与整体的关系。例如有义原“头”的描述:

头:part|部件,%AnimalHuman|动物,head|头

而在上下位关系中,“动物”的下位义原包括:“人”、“兽”,而“兽”的下位义原又包括“走兽”、“牲畜”、“鱼”、“禽”,那么认为所有以“动物”的下位义原为主要特征的概念与“头”构成部分-整体关系。如概念:“牛”、“食草动物”、“鱼”、“鸡”、“多足动物”等都与“头”构成部分-整体关系。

3.3.2 相关关系挖掘规则

定义 1 中的相关关系可能是与概念相关,也可能是与义原相关。如果相关元素是概念,显然,原概念与对应概念之间存在相关关系;如果相关元素是义原,则以相关义原为主要特征或第二特征的概念与原概念相关。

知网的相关关系:这里指的是具体的知网描述体系中的(小)相关关系,不是语义相关的(大)相关关系。具有传递性,定义如下:

定义 3 相关关系的传递性,若概念 A 与义原 B 存在相关关系,且概念 C 与义原 B 也存在相关关系,认为概念 A 与概念 C 存在相关关系。

根据相关关系的传递性,认为具有相同相关元素的概念是相关的。例如:

暑假:time|时间,@rest|休息,@WhileAway|消闲,#education|教育

作业:fact|事情,#education|教育

学院:aValue|属性值,attachment|归属,InstitutePlace|场所,#education|教育

教师:humal|人,#occupatio|职位,*teach|教,education|教育

学生:humal|人,*study|学,education|教育

教授:humal|人,#occupatio|职位,*teach|教,education|教育

“暑假”的相关元素是义原“教育”,根据相关的传递性,“作业”、“学院”与“暑假”相关,而“教师”、“学生”、“教授”均以“教育”为第二特征,因此也与“暑假”构成相关关系。

3.3.3 材料-成品关系挖掘规则

成品义原的解释义原体现了材料概念所具有的属性。因此认为成品的解释义原与材料概念相关。例如:

丝绸:material|材料,?clothing|衣物,?tool|用具

材料“衣物”的解释义原包括“样式”、“颜色”、“穿戴”、“性别”。而“用具”的解释义原为“利用”。因此认为以这些解释义原为主要特征的概念与“丝绸”相关,如概念“衣物”,“样式”,“颜色”等。

3.3.4 施事/经验者/关系主体-事件或工具-时间关系挖掘规则

(1)概念中的事件义原表示伴随施事/经验者概念可能发生的事件。我们认为以相同事件义原为施事/经验者的概念相互关联。例如:

讲师:human|人,#occupation|职位,*teach|教,education|教育

教科书:readings|读物,*teach|教,$study|学,education|教育(受事者)

老师:human|人,*teach|教,education|教育(施事者)

概念“讲师”与义原“教育”构成实施事件关系,而“教科书”是“教育”的受事者,因此概念“讲师”与“教科书”相关。“老师”同样是“教育”的实施者,“老师”与“教师”也相关。

(2)知网为每一个事件都标识一个角色框架。在框架中所列出的角色是该事件的必要绝对角色,换言之,角色与事件紧密相关。因此以事件为角色的概念与施事/经验者概念相关。例如:

“教”事件的角色框架:实事者、受事者、内容、结果、事件发生时间、事件地点

而概念“开讲”:start|开始,content=teach|教(内容)

在“开讲”事件的角色框架中,其内容角色为“教”,而“教”的实施者是“老师”,则认为“开讲”与“老师”相关。

(3)事件发生的时间、地点与事件相关。因此与事件构成时间/地点-事件关系的概念与施事/经验者概念相关。例如:

讲台:facilities|设施,space|空间,@teach|教

“讲台”是“教”事件发生的地点,两者为事件-地点关系,认为“讲台”与“讲师”概念相关。

(4)概念的主要特征是对概念最为深刻的描述,以事件义原为主要特征的概念与施事/经验者概念相关。例如:

教育:teach|教(“教育”的主要特征是“教”)

“教师”概念与“教育”概念之间存在施事者-事件关系。

3.3.5　时间/地点-事件关系挖掘规则

(1)时间/地点-事件关系中的事件发生需要施事者、受事者等角色,因此与事件存在施事、受事关系的概念与时间/地点概念相关。例如:

国庆:time|时间,day|日,@congratulate|祝贺,#country|国家(“祝贺”事件发生的时间是“国庆”)

烟花炮竹:tool|用具,*WhileAway|消闲,*congratulate|祝贺(“祝贺”的施事者是“烟花炮竹”)

“国庆”-“祝贺”-“烟花炮竹”,通过事件“祝贺”,概念“国庆”与概念“烟花炮竹”相互关联。

(2)事件角色框架中的各角色与概念紧密相关,因此以事件为角色的概念与时间/地点概念相关。例如:

观礼:include|纳入,ResultWhole=congratulate|祝贺(“观礼”事件的结果角色为“祝贺”)

“观礼”-“祝贺”-“国庆”,通过事件“祝贺”,概念“观礼”与概念“烟花炮竹”相互关联。

(3)以事件义原为主要特征的概念与时间/地点概念相关。例如:

喜庆:congratulate|祝贺(“喜庆”的主要特征为“祝贺”)

概念“喜庆”与“国庆”相关。

3.3.6　受事/内容/领属物等-事件关系挖掘规则

由于受事/内容/领属物等-事件关系中描述的事件发生的客体,而实施/经验者/关系主体-事件关系描述了事件发生的主体,两种关系均围绕事件关系展开,因此对两种关系使用了类似的语义关系挖掘原则。

(1)与事件存在施事、受事关系的概念与原受事概念相关;

(2)以事件为角色的概念与受事概念相关;

(3)与事件构成时间/地点-事件关系的概念与受事概念相关;

(4)以事件义原为主要特征的概念与受事概念相关。

3.3.7　宿主-属性关系挖掘规则

(1)由于上下位关系中的下位义原继承上位义原的解释义原所描述的各种属性,因此以宿主为主要特征的概念与原概念之间同样存在宿主-属性关系。例如:

理解力:attribute|属性,ability|能力,#understand|领会,&AnimalHuman|动物(属性“理解力”的“动物”)

动物的下位义原有“人”、“兽”、“兽”的下位义原包括“走兽”、“牲畜”、“鱼”、“禽”。因此“人”、“牛”、“食草动物”、“鱼”、“鸡”、“多足动物”等与“理解力”构成属性-宿主关系。

(2)知网为每个属性都定义了可能的取值,构成属性-值关系,因此以该属性所对应的属性值概念与原概念具有属性-值关系。例如,属性“能力”的取值:“能”、“庸”,他们与概念“理解力”之间存在属性-值关系。

3.3.8　语义关系挖掘规则总结

至于知网中同义、对义、反义关系,是通过知网提供程序包实时计算得到的;而上下位关系、事件-角色关系在上述的关系挖掘规则中已经充分考虑到了;实体-值、值-属性关系是在知网中直接标注的,使得关系的识别变得将较为困难,因此在本文中就不再为其制定挖掘规则了。

综上所述,将本文涉及所有间接关系挖掘规则总结如表1所示。

表1　知网间接关系挖掘规则

直接关系	关系符	间接关系挖掘规则
部分-整体	%	以下位义原为主要特征的概念与原概念也构成部分与整体的关系
相关关系	#	(1)以相关义原为主要特征或第二特征的概念与原概念相关 (2)具有相同相关元素的概念是相关的
材料-成品关系	?	成品的解释义原与材料概念相关
施事/经验者/关系主体-事件或工具-时间关系	*	(1)以相同事件义原为施事,经验者的概念相互关联 (2)以事件为角色的概念与原概念相关 (3)与事件构成时间/地点-事件关系的概念与原概念相关 (4)以事件义原为主要特征的概念与原概念相关
时间/地点-事件关系	@	(1)与事件存在施事、受事关系的概念与原概念相关 (2)以事件为角色的概念与原概念相关 (3)以事件义原为主要特征的概念与原概念相关
受事/内容/领属物-事件	$	(1)与事件存在施事、受事关系的概念与原概念相关 (2)以事件为角色的概念与原概念相关 (3)与事件构成时间/地点-事件关系的概念与原概念相关 (4)以事件义原为主要特征的概念与原概念相关
宿主-属性	&	(1)以宿主为主要特征的概念与原概念之间同样存在宿主-属性关系 (2)以该属性所对应的属性值概念与原概念具有属性-值关系

表 2 一个词义消歧的计算实例

概念	字典序号	定义	相关度 1	相关度 2
迈向/v	31 240	walk|走	0	0
充满/v	7 257	exist|存在,quantity=many|多	0	0
希望/n	48 957	expect|期望	0.436 0	0.119 0
的/u	10 948	{DeChinese|构助}	0	0
新/a	50 861	aValue|属性值,newness|新旧,new|新,desired|良	1.045 0	
	50 862	aValue|属性值,time|时间,now|今	1.947 6	
世纪/n	41 806	time|时间	0.236 9	0.854 6
--/w			0	0
一九九八年/t			0	0
新年/t	50 896	time|时间,festival|节,@congratulate|祝贺	0.372 0	0.974 0
讲话/n	23 456	speak|说	0	0

3.4 相关度计算模型

下面给出一种具体的语义相关度的计算实例。设 w_1 和 w_2 为任意的两个词,在知网中,w_1 有 p 个义项:s_{1_1};:::;s_{1_p};w_2 有 q 个义项:s_{2_1};:::;s_{2_q}。那么 w_1 与 w_2 的相关度是两词中相关度最高的两个义项的相关度,即:

$$SR(w_1,w_2)=\max(CR[s_{1i},s_{2j}]),i=1,\cdots,p,j=1,\cdots,q \quad (1)$$

其中,$SRel[s_{1_i};s_{2_j}]$为义项 s_{1_i} 与 s_{2_j} 的相关度。下面来计算两个义项 s_1 和 s_2 之间的相关度。

s_1 与 s_2 的知网形式化表示为:

$s_1:=r_{1_1}ss_{1_1},\cdots<r_{1_m}ss_{1_m},r_{1_i}\in R,ss_{1_i}\in SS,i=1,\cdots,m$

$s_2:=r_{2_1}ss_{2_1},\cdots<r_{2_n}ss_{2_n},r_{2_i}\in R,ss_{2_i}\in SS,i=1,\cdots,n$

可以定义相对相关度 RR 如下:$RR_i(s_1,s_2)$代表概念 s_1 相对于概念 s_2 在 r_{1_i} 关系上的相对相关度,即如果能将 r_{1i} 遵循关系挖掘规则转换为 r_{2_j},那么定义:

$$RR_i(s_1,s_2)=\begin{cases}1,\text{if}(ss_{1_i}=ss_{2_j})\\0,\text{otherwise}\end{cases} \quad (3)$$

注意,由于 s_1 和 s_2 所对应的语义关系各自独立,所以,显然除了 $RR_{r_i}(s_1,s_2)=RR_{r_i}(s_2,s_1)$,对于其它的 $r_i\in R$,未必有 $RR_\#(s_1,s_2)=RR_\#(s_2,s_1)$,这完全取决于由表 1 所定义的语义关系挖掘规则,以及 s_1 和 s_2 的定义。这也就是相对相关度中"相对"的含义。

至此,可以给出一种具体的语义相关度的计算方法:在概念 s_1 与概念 s_2 相对相关度基础上加权求和可得概念的相关度,即:

$$CR(S_1,S_2)=[\sum_{i=1}^{m}w_iRR_i(S_1,S_2)+\sum_{j=1}^{n}w_jRR_j(S_2,S_1)]/2 \quad (3)$$

其中,w_i 为关系 r_{1_i} 对应的权值,w_j 为关系 r_{2_j} 对应的权值,当然有 $\sum_{i\in R}w_i=1$。

4 应用实例

由于相关度评价工作主观性非常强,独立地评价其计算结果显得较为困难。因此这里将提出的相关度计算方法应用于词义排歧,通过对排歧结果的考察,验证相关度计算方法的有效性。

词义排歧[8]是指根据一个多义词在文本中出现的上下文环境来定其词义代码。这个代码既可以是该词义在一部普通词典中的义项号,也可以是它在一部义类词典中的义类代码。选择知网作为知识词典,也就是要确定一个多义词,在一个上下文语境中表现出的唯一的一个词义。从 1998 年人民日报半年中取 1 000 句共 58 022 个词作为实验数据,分为两组每组 500 句:第一组句子包含的词数小于 15,第二组句子包含的词数大于等于 15。每个句子中均含有一个多义词,即该词在知网中存在多个义项,以一句作为上下文语境,计算该多义词与同在本句中的其他词语的相关度,排除不合适的义项,达到消除歧义的目的。

表 2 中是一个实验例句,"迈向/v 充满/v 希望/n 的/u 新/a 世纪/n-/w 一九九八年/t 新年/t 讲话/n",其中"新"字是多义词,对应两个义项编号 50 861、50 862。"相关度 1"是与"新"的 50 861 义项相关度计算结果,而"相关度 2"是与"新"的 50 862 义项相关度计算结果。"新"的 50 861、50 862 义项分别对应的相关度,是所有其他词语与该义项的相关度计算结果之和。通过比较,可以排除相关度较小的 50 861 义项,唯一确定"新"在本句中的正确语义标号应该是 50 862。

表 3 词义消歧实验结果

句子长度	句子总数	正确消歧数	正确消歧比例/(%)
<15 个词	500	384	76.8
≥15 个词	500	406	81.2

使用该方法,为两组共 1 000 个句子作词义排歧,实验结果如表 3,从实验结果可以发现排歧的正确率还是比较理想的,在同等条件下,长句的排歧效果要高于短句。这是因为在长句中上下文语境的内容更为丰富,为正确义项的选择提供了更多的语义知识,因而优化了相关度的计算结果。

5 结束语

针对以往相关度计算中相关关系不明确的问题,重新定义相关性概念,明确相关关系为知网定义直接关系与通过关系挖掘规则获得的间接关系。同时,摒弃了以相似度为基础计算方法,为相关关系的度量提出了通用的基于关系挖掘计算模型,该模型适用于所有使用知网描述语言表达的知识体系,也就是说即使更换了知识库,提及的相关度计算模型仍然是适用的。在将该计算模型应用到词义排歧中,也取得了比较好的结果。

当然本模型也存在一些有待改进的问题,在为概念挖掘关系时,该方法分别为各种知网关系制定了挖掘规则,由于这些挖掘规则的形成多为经验性的,因此有进一步扩充和优化的可能。但本文的模型为将来对相关度进一步研究提供了一个很好的可计算化的框架。

参考文献:

[1] 陈刚,陆汝钤,金芝.基于领域知识重用的虚拟领域本体构造[J].软件学报,2003,14(3):350-355.

(下转 180 页)