基于《知网》的词语相似度计算方法①

孙润志[1,2], 于 放[2]

[1](中国科学院大学, 北京 100049)
[2](中国科学院 沈阳计算技术研究所, 沈阳 110168)

摘 要: 词语相似度计算中常用的一种方法是基于某种语义词典的计算. 首先介绍《知网》中的基本概念和层次体系结构, 借鉴刘群、李素建在词语相似度方面的基础理论, 利用《知网》的义原层次体系结构计算出其中的义原相似度, 再计算出概念的相似度, 最后得到词语的相似度. 还对其中的计算方法做出适当的改进调整, 使其计算出的结果更加符合实际情况.

关键词: 词语相似度; 知网; 义原; 义原相似度

Word Similarity Computing Method Based on HowNet

SUN Run-Zhi[1,2] , YU Fang[2]

[1](University of Chinese Academy of Sciences, Beijing 100049, China)
[2](Shenyang Institute of Computing Technology, Chinese Academy of Sciences, Shenyang 110168, China)

Abstract: A common method of word similarity computing is based on a semantic dictionary. This paper introduces the "HowNet" in the basic concepts and hierarchical architecture, drawing lessons from Liu Qun, Li Sujian basic theory in word similarity. It makes use of "HowNet" sememe hierarchical architecture to calculate sememe similarity, then calculates the similarity of the concept, and finally gets the similarity of the words. This paper also makes appropriate adjustment to the computing method, so that the calculated results can be more in line with the actual situation.

Key words: word similarity; HowNet; sememe; sememe similarity

词语相似度在信息的检索处理、语义分析、机器翻译等各个方面都有重要的作用. 词语相似度的计算主要有两类计算方法: 一类是基于语义词典的计算方法; 一类是基于语料库统计的计算方法. 其中基于语义词典的计算方法根据词典对词语概念的分析描述, 根据语义关联的方法来计算词语相似度, 计算结果通常符合人工评测. 在我国汉语方面经常使用的语义词典有《同义词词林》和《知网》等. 其中《知网》凭借独特的知识描述形式和丰富的词汇语义知识成为研究汉语词语相似度计算的典型平台.

本文通过充分利用《知网》的知识描述结构和义原的层次体系关系. 在计算义原相似度方面添加考虑了义原节点的深度和密度信息, 使得计算结果更加合理, 实验数据的验证了这一方面.

1 《知网》

1.1 《知网》概述

《知网》(HowNet)知识库是由我国著名机器翻译专家董振东先生设计和开发的. 他提出应首先建立一种可以被称为知识系统的常识性知识库. 它以通用的概念为描述对象, 建立并描述这些概念之间的关系. 《知网》是一个各类概念为描述对象的知识系统, 它是把概念与概念之间的关系以及概念的属性与属性之间的关系形成一个网状的知识系统. 《知网》中的词汇语义知识和世界知识非常丰富, 在自然语言处理和机器翻译等方面发挥了巨大的作用.

1.2 《知网》组织结构

《知网》的结构中"概念"和"义原"是两个主要概念, 其中"概念"也叫做"义项", 是对词汇语义的一种

① 收稿时间:2014-11-13;收到修改稿时间:2014-12-29

描述, 并且每一个词语同时可以表达为几个不同的概念; “概念”是用“义原”来描述的, “义原”是描述一个“概念”的最小意义的单位, 也是《知网》的词汇结构中用来描述其他词汇不可再分的基本元素. 《知网》作为一个知识系统, 同时也是一部语义词典, 是以网状结构存在, 它与一般的词汇数据库有本质不同, 重点反映的是概念的共性和个性的关系, 以及概念之间和概念属性之间的复杂关系. 在这里文章主要是根据义原层次树结构来计算其相似度, 接下来将主要介绍《知网》的知识描述语言与义原层次体系结构.

(1)知识描述语言

在《知网》中, 对具体概念的描述是比较复杂的, 对每一个词语表达的具体概念以及对概念的描述会形成一个记录, 其记录由 4 项内容组成, 其中每一项由两部分组成, 并由“=”来连接两部分, 其中数据的域名在“=”的左侧, 数据的值在其右侧, 它们排列如下:

W_X= 词语

G_X=词语词性

E_X=词语例子

DEF=概念定义

例如: 在知网中对于词汇“打球”具体表述如下:

NO.=024038

W_C=打球 G_C=V [da3 qiu2] E_C=

W_E=play a ball G_E=V E_E=

DEF={exercise|锻炼:instrument={tool|用具}}. 其中: NO 为概念编号, W_C 表示汉语中的词语, G_C 表示汉语中的词性, E_C 表示汉语中的例子, W_E 表示英语中的词语, G_E 表示英语中的词性, E_F 表示英语中的例子, DEF 表示义项, 是该概念在《知网》中的定义.

表 1 《知网》知识描述语言实例

词语	概念编号	义项
除夕	020254	{time|时间:TimeSect={night|夜}}
处决	020464	{punish|处罚:means={kill|杀害}}
成家立业	017605	{GetMarried|结婚};{start|开始:content={affairs|事务}}
报纸	005769	{material|材料:{wrap|包扎: instrument={~}}, {write|写:LocationFin={~}}}
报纸	005770	{publications|书刊:{publish|出版: ContentProduct={news|新闻}, LocationFin={~}}}

(2)义原层次体系结构

《知网》结构中“义原”是描述一个“概念”的最小意义的单位. 同样根据描述作用的不同, “义原”可分为以下三类: 描述单个概念的语义特征的称为基本义原; 描述词语的语法特征(主要指词性)的称为语法义原; 描述概念间关系的称为关系义原. “义原”的具体分类如图 1 所示.

1) Event|事件
2) entity|实体
3) attribute|属性值
4) aValue|属性值
5) quantity|数量
6) qValue|数量值
7) SecondaryFeature|次要特征
8) syntax|语法
9) EventRole|动态角色
10) EventFeatures|动态属性

图 1 《知网》义原类别

《知网》中还描述了义原之间的多种关系, 如上下位关系、部件-整体关系、同义关系、反义关系、对义关系、属性-宿主关系与相关关系等. 各类义原之间关系种类多样, 其中最为常见的是义原的上下位层次关系. 根据这种义原的上下层次关系关系, 可以构建一棵包含了所有基本义原的义原组织结构树. 我们可以通过组织结构树作为工具计算出义原的相似度, 进而计算出概念、词语的相似度.

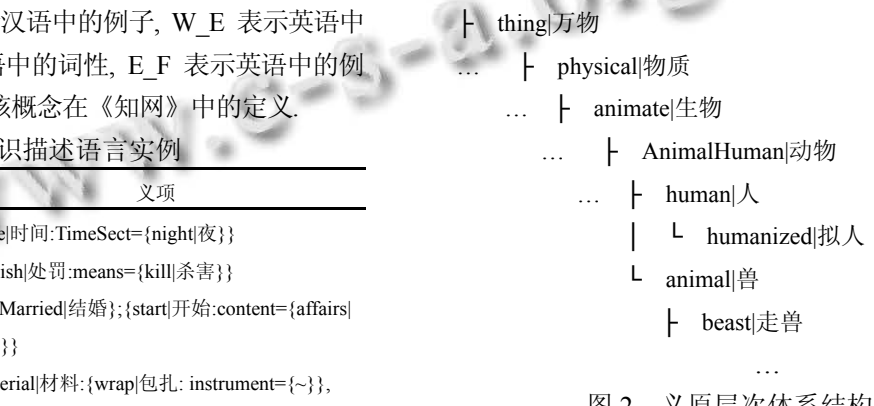

图 2 义原层次体系结构

2 词语相似度及计算方法

2.1 词语语义相似度

Dekang Lin[3]认为任何两个事物的相似度取决于它们的共性(Commonality)和个性(Differences), 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 个性所需要的信息量大小.

刘群[2]等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.

2.2 词语相似度计算

本文的词语相似度计算是借鉴刘群、李素健[2]在词语相似度的计算基础上进行的, 即根据《知网》中词语、概念与义原的关系, 将词语相似度的计算转化为概念相似度计算, 再有概念相似度计算转化为义原相似度来计算.

(1) 词语相似度计算

在《知网》的结构中词语是用概念来描述的, 一个词可以表达为几个概念, 而概念则用义原来描述.

假设词语 W_1 有 n 个概念 C_{11}、C_{12}、…、C_{1n}, 词语 W_2 有 m 个概念 C_{21}、C_{22}、…、C_{2n}, 本文中两个词语 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:

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

通过此公式可将两个词语之间的相似度问题转化到两个概念之间的相似度问题.

(2) 概念相似度计算

概念是通过义原描述的, 然而不同类型的义原对概念描述作用的大小不同. 据《知网》中对概念的具体描述, 概念的相似度计算可以由以下三个部分的相似度计算来得到.

独立义原描述式: 独立义原作为对概念的直接描述, 对概念的相似度有主要影响, 将两个概念的这一部分的相似度记为 $Sim_1(p_1,p_2)$;

关系义原描述式: 关系义原是表明是一种 is a 的定义关系或者是识别概念必不可少的特征属性, 对概念的相似度有一定影响, 将两个概念的这一部分的相似度记 $Sim_2(p_1,p_2)$;

符号义原描述式: 符号义原是对概念的一种间接描述, 表明概念的一些其它属性, 对概念的描述作用小于前面两种, 将两个概念的这一部分的相似度记 $Sim_3(p_1,p_2)$;

于是两个概念的整体相似度记为:

$$Sim(C_1,C_2)=\sum_{i=1}^{3}\beta_i\prod_{j=1}^{i}Sim_j(p_1,p_2)$$

其中, β_i $(1\le i\le 3)$ 是可调节的参数, 且有: $\beta_1+\beta_2+\beta_3=1$, $\beta_1\ge\beta_2\ge\beta_3$, 同时反映了不同义原描述式对概念描述作用的不同.

(3) 义原相似度计算

我们将概念相似度计算转化为义原相似度计算, 而义原的相似度计算需要利用《知网》的义原层次体系结构. 通过《知网》中义原的上下位关系确定义原在层次结构中的语义距离, 借此通过语义距离的方法来计算义原相似度. 刘群与李素健提出了通过语义距离来计算义原相似度的公式, 即

$$Sim(p_1,p_2)=\frac{\alpha}{d+\alpha}$$

其中 p_1、p_2 分别表示两个不同的义原, d 表示其在义原层次树中的路径长度, 是一个正数, α 是调节因子, 表示相似度为 0.5 时的路径长度.

两个义原相似程度很大程度取决于他们之间的语义距离, 也可通过数据验证证明上述公式是较为不错的义原相似度计算方法. 但在进一步的研究中我们会发现, 义原相似度计算不仅与语义距离有关系, 也跟其他的因素有一定关联. 两组义原在义原层次树中所处的深度不一样, 节点所在的密度不一样, 即使有相同的语义距离也会导致两组义原相似度的差距. 因此在计算义原相似度考虑加入了义原节点深度和节点密度, 对他人提出的公式进行了改进[4,5]:

$$Sim(p_1,p_2)=\frac{a\alpha(dp(p_1)+dp(p_2))+b\alpha(ds(p_1)+ds(p_2))}{d+\alpha}$$

其中, p_1、p_2 分别表示两个不同的义原, $dp(p_1)$、$dp(p_2)$ 表示两个义原在义原树中的深度, $ds(p_1)$、$ds(p_2)$ 表示两个义原节点在义原树中的密度, a、b、α 为可调节参数且 $a+b=1$, α 表示相似度为 0.5 时的路径长度, d 表示其在义原层次树中的路径长度.

3 实验结果

根据以上词语相似度计算理论的研究, 我们使用以下三种方法来计算词语相似度, 并对计算结构比较分析.

方法 1. 使用李素建的基于语义计算的词语相似

度计算方法[6];

方法 2. 使用刘群与李素建的基于《知网》的词语相似度计算方法;

方法 3. 本文中的词语相似度计算方法;

在本次实验中, 其实验结果如表 2 所示:

表 2 词语相似度计算结果

词语 1	词语 2	方法 1	方法 2	方法 3
男人	父亲	1.000	1.000	1.000
男人	和尚	0.861	0.861	0.765
男人	鲤鱼	0.007	0.176	0.127
男人	收音机	0.006	0.094	0.063
医生	医治	0.014	0.037	0.034
跳槽	拔脚	0.006	0.184	0.259
风度	面积	0.315	0.612	0.365
深红	粉红	0.013	0.074	0.518

将方法 1、方法 2 和方法 3 得到结果相比较, 总体来看方法 1 的结果比较粗糙, 方法 2 的结果比方法 1 更细腻一些, 但有些词语相似度的结果也不太合理, 比如“深红”和“粉红”的相似度明显太低, “风度”和“面积”的相似度也显得过高. 根据人工主观评判标准, 可以看出方法 3 的结果更为合理.

方法 3 的计算方法在一定程度上提高了词语相似度的准确度, 但仍有两方面的问题需要考虑: 第一个问题是对于有些词语优化效果不明显, 需要进一步提高优化; 第二个问题是对于节点密度和节点深度的加入, 造成存储节点信息的增加, 提高节点存储效率也是我们要考虑的问题.

4 结语

本文通过研究《知网》知识库的知识描述语言和义原体系结构, 更加深入了解基于《知网》的词语相似度计算方法的相关原理. 在基于刘群等人的研究基础上, 对词语相似度的计算方法进行了如下部分改进, 使计算方法更加合理. 在下一步的工作中,我们将进一步研究《知网》的体系结构, 并将词语相似度的计算应用到文本相似度[7]的计算中.

参考文献

1 董振东,董强.知网,http: //www.keenage.com

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

3 Lin DK. An Information-Theoretic Definition of Similarity Semantic distance in WordNet. Proc. of the Fifteenth International Conference on Machine Learning, 1998.

4 夏天.汉语词语语义相似度计算研究.计算机工程,2007, 33(6):191–194.

5 吴思颖,吴扬扬.基于中文 WordNet 的中英文词语相似度计算.郑州大学学报(理学版),2010,42(2):66–69.

6 李素建.基于语义计算的语句相关度研究.计算机工程与应用,2002,38(7):75–83.

7 肖志军,冯广丽.基于《知网》义原空间的文本相似度计算. 科学技术与工程,2013,13(29):8651–8656.