基于知识图的汉语词语间语义相似度计算

张晓孪[1,2],张 蕾[1],王西锋[1,2]
ZHANG Xiao-luan[1,2],ZHANG Lei[1],WANG Xi-feng[1,2]

1.西北大学 信息科学与技术学院,西安 710127
2.宝鸡文理学院 计算机科学系,陕西 宝鸡 721007
1.College of Information Science & Technology,Northwest University,Xi'an 710127,China
2.Department of Computer Science,Baoji University of Arts & Sciences,Baoji,Shaanxi 721007,China
E-mail:bjwlxyzxl@126.com

ZHANG Xiao-luan,ZHANG Lei,WANG Xi-feng.Semantic similarity computation based on knowledge graph between chinese words.Computer Engineering and Applications,2007,43(8):160-163.

Abstract: Semantic similarity is one of the basic relations between words,the research of semantic similarity between Chinese words takes an important role in many Chinese language processing applications.In this paper,the semantic similarity based on knowledge graph between words is presented,therefore,knowledge graph,which belongs to the semantic network,is used in the Chinese information processing.The experiment result given in the end proves this method is effective for word semantic similarity computation.

Key words: semantic computation;Chinese word;similarity;knowledge graph

摘 要:语义相似是词语间的基本关系之一,汉语词语间语义相似的研究对于许多自然语言处理的应用具有重要的指导意义。提出了一种基于知识图的词语间语义相似度计算的方法,把知识图这种属于语义网络范畴的知识表示方法应用于汉语信息处理中。实验结果表明该方法对词语间语义相似度计算是有效的。

关键词:语义计算;汉语词语;相似度;知识图
文章编号:1002-8331(2007)08-0160-04 **文献标识码**:A **中图分类号**:TP391

1 引言

词语是汉语最基本的语法和语义的单位,词语的语义相似度计算是计算句子相似度的基础。词语相似度是一个主观性较强的概念,没有非常明确的客观标准可以衡量。因为词语之间的关系非常复杂,其相似或差异之处很难用一个简单的数值来进行度量。脱离具体的应用背景谈论词语相似度,很难得到一个统一的定义。相似度这个概念,涉及到词语的词法、句法、语义甚至语用等方方面面的特点。其中,对词语相似度影响最大的应该是词的语义。语义相似度是将词语间的种种不同的直接或间接语义关系影射为一个表示词语间语义相关的紧密程度的数值。词语相似度对信息检索领域中的查询扩展,词语歧义的消除,以及提高信息检索结果的精度和召回率等方面都有着重要的指导意义。

2 词语间语义相似度计算方法

从国内外研究情况看词语间语义相似度的计算方法大体上可以分成两类:一类是依赖于概念间结构层次关系组织的语义词典的方法, 主要是根据概念之间的上下位关系和同义关系,利用概念距离方法(Conceptual Distance)来计算;另一类是基于大规模语料库统计的方法,将词语的上下文信息的概率分布作为词语间语义相似度计算的参照。

基于语义词典的方法,主要是根据这类语言学资源中概念之间的上下位关系和同义关系来计算。这种方法建立在两个词语具有一定的语义相似性当且仅当它们在概念间的结构层次网络中存在一条通路(上下位关系)这一假设基础之上。基于语义词典的方法通常依赖于较完备的概念间结构层次关系组织的大型语义词典,如 Wordnet[1]等等,在汉语处理领域中可以使用知网[2]、同义词词林[3]等语义词典。

国外有人[4,5]通过计算在 Wordnet 中词节点之间上下位关系构成的最短路径来计算词语之间的语义相似度,还有人[6]提出根据两个词的公共祖先节点的最大信息量来衡量两个词的语义相似度等等。国内王斌[7]利用《同义词词林》来计算汉语词语之间的相似度。刘群、李素建[8]研究了知网中知识描述语言的语法,分析描述一个词义所用的多个义原之间的关系,区分这些义原在词语相似度计算中所起的不同作用,进而提出了利用知网进行词语相似度计算的算法。这种基于语义词典方法一般都要利用概念距离。概念距离是语义相似度一种递降函数,也就是说两个概念越相似,这两个概念间的距离就越短。概念层次树中的最短路径长度只存在上下位关系时才可以被用来作

基金项目:陕西省教育厅专项科研基金(No.HD01302)。
作者简介:张晓孪,女,硕士研究生,研究方向:人工智能及自然语言理解;张蕾,女,硕士生导师,博士,教授,研究方向:人工智能及自然语言理解。

为概念距离的度量。

基于语料库统计的方法用大规模的语料来统计,利用词语的上下文信息的概率分布来计算词语间语义相似度。例如可以利用词语的相关性来计算词语的相似度。事先选择一组特征词,然后计算这一组特征词与每一个词的相关性,一般用这组词在实际的大规模语料中在该词的上下文中出现的频率来度量。于是对于每一个词都可以得到一个相关性的特征词向量,然后利用这些向量之间的相似度(一般用向量的夹角余弦)来计算并作为这两个词的相似度。

国外如 Brown[9]的基于平均互信息的方法,Lillian Lee[10]的基于相关熵的方法等等,这类方法建立在两个词语具有某种程度的语义相似当且仅当它们出现在相同的上下文中这一假设的基础上。国内关毅[11]等提出基于统计的汉语词语间语义相似度计算,主要研究了汉语中的实词(名词、动词、形容词),他首先采用模糊集合中的隶属函数定义语义相似度的数学模型,然后以同义词词林的词语分类体系为基础,提出了基于相关熵的汉语词语间语义相似度的计算方法。

基于语义词典方法比较直观而且简单有效,但是构造汉语语义词典却是一件规模浩大的系统工程。基于统计的定量分析方法能够对词语间的语义相似性进行较精确和有效的度量。但是这种方法比较依赖于训练所用的语料库,不仅计算量大而且计算方法复杂,受数据稀疏和数据噪声的干扰较大,有时会出现明显的错误。

本文提出的词语间语义相似度计算方法是基于知识图和知网进行语义计算,改进了传统的知识图表示方式,根据知网中概念项的抽取结果对词语的义项进行表示,用词图的相似度来表示相应词语的语义相似度。实验结果表明该方法对词语间语义相似度计算是有效的。

3　背景知识

3.1　知识图(Knowledge Graph)

知识图是一种属于语义网络范畴的知识表示方法。它使用节点表示概念,使用有向弧表示概念之间的关系。知识图的基本思想[12]是:每个词的词义可以由称作词图的知识图来表示,进而通过合并词图而组成短语图,再通过合并短语图而组成句子图,最后通过合并句子图而组成文本图。因而构造词图成为应用知识图的基础和核心工作。本文以《知网》为语义知识资源提出了一种构造词图的方法。其中知识图的定义如下:

定义 设 C 为概念的集合,T 为关系类型的集合,$G=<N,A,ln,la>$是知识图。其中:

N 表示节点的集合。

A 表示弧的集合。

ln 表示节点集合到概念集合的映射,即 $ln:N\rightarrow C$。

la 表示弧的集合到关系类型集合的映射,即 $la:A\rightarrow T$。

关系类型 T 有 12 种,分别为:EQU、SUB、ALI、DIS、CAU、ORD、PAR、SKO、FPAR、NEGPAR、POSPAR 和 NECPAR,详细概念请参考文献[13]。

3.2　知网(HowNet)

知网[2]是 1999 年 3 月发布于网上的一个知识资源,是一个以汉语和英语的词语所代表的概念为描述对象,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。它是一个网状的有机的知识系统。这些关系都隐含在知网的知识词典和义原的特征文件中。义原在知网中是个重要的概念,它是从所有汉词中提炼出的可以用来描述其它词语的不可再分的基本元素。知网使用了一种知识词典的描述语言(KDML)对所有概念进行定义,从而使概念定义形式化,有效地保证了描述的复杂度和描述的一致性。知网中的义原分类树把各个义原及它们之间的联系以树的形式组织在一起,为进行语义计算提供了方便。

4　基于知识图的词语间语义相似度计算

4.1　基本思想

本文提出了一种基于知识图的词语间语义相似度计算的方法,主要思想是以知识图为知识表示方法,利用《知网》为语义知识资源,先为词语构造词图,然后根据词图相似度的计算方法计算出词图的相似度,词语间语义相似度就是通过词语对应的词图的相似度得到的。

4.2　对知识图表示方法的改进

在用知网作为语义知识资源时,对于每一个词语的语义描述由多个义原组成,其中各个义原之间并不是平等的,它们之间有着复杂的关系。因此根据语义计算的需要把知识图中的节点分为义原节点和非义原节点。义原节点表示词语义项中义原的节点,非义原节点是除了义原节点之外的其它节点。义原节点根据在相似度计算过程中所扮演的角色不同分为中心义原节点和非中心义原节点。中心义原节点在语义计算过程中起主要作用,代表整个知识图所要表达的主要含义。非中心义原节点在语义计算过程中起着辅助作用,是用来辅助说明中心义原节点的。

4.3　词图的构造

在用知识图作为知识表示方法时,词图的构造[14]是关键之一,本文提出的词图的构造方法是根据知网中词典的概念项(DEF)构造的,首先根据概念项的描述方式,总结出了概念项的形式化描述:

```
<动态角色>::=<知网中的动态角色>
<次要特征角色>::=<知网中的次要特征角色>
<角色>::=<动态角色>|<次要特征角色>
<主要义原>::=<实体义原>|<事件义原>|<属性义原>|<属性值义原>
<次要义原>::=<次要特征义原>
<关系表达式>::=<角色>={<主要义原>}
                |<角色>={<次要义原>}
                |<角色>=<DEF>
                |<角色>={<义原标志>}
                |"EventRole"={<动态角色>}
<义原标志>::="~"|"$"|"*"|"#"|"%"
<DEF>::={<主要义原>}
        |{<主要义原>:<DEF>[,<关系表达式>][,<DEF>]
        [,<关系表达式>}
        |{<主要义原>:<关系表达式>[,<DEF>]
        [,<关系表达式>][,<DEF>]}
```

|{<次要义原>:<关系表达式>}

从 DEF 的形式化描述看概念的描述分为两大类：一类称为一般概念，它所描述的词语称为一般词语；另一类称为功能概念，它所描述的词语称为功能词语。一般概念满足以下三个特征：

(1)DEF 中至少有一个主要义原；

(2)DEF 是“{”和“}”两个符号之间的描述字符串；

(3)DEF 是层次的、嵌套的。

功能概念的描述中没有主要义原，它所对应的词语相当于汉语语法中的虚词。这些特点为自动构造词图提供了依据。

4.3.1 一般词语词图的构造

显然知网中给出的概念是一个网状形式，本文在构造词图时，对于一般词语选取了其概念项中的主要部分，即对其概念项进行自动抽取，使其成为如下形式：

<DEF>={<主要义原>

|{<主要义原>:<关系表达式>[,<关系表达式>]}

这里的关系表达式为上面给出的概念项形式化描述中关系表达式的前两种形式，在构造词图的过程中，为概念项中“{”和“:”之间的“主要义原”建立中心义原节点；为“关系表达式”中的“主要义原”建立非中心义原节点；为“关系表达式”中的“次要义原”建立非中心义原节点；在非中心义原节点和中心义原节点之间建立从非中心义原节点到中心义原节点的弧，弧的关系类型为相应关系表达式中的“角色”。

例如对于下列词语：

计算机：DEF={电脑}，其词图如图 1 所示。

男人：DEF={人：Belong={家}，Modifier={男}}，其词图如图 2 所示。

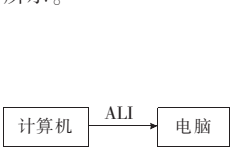

图 1 “计算机”的词图

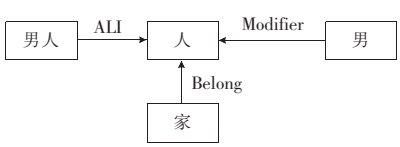

图 2 “男人”的词图

4.3.2 功能词语词图的构造

功能词语是一种特殊的词语，一般都不能充当句法成分，并且大都以表示语法意义为主，所以其词图构造也很特殊。功能概念有两种描述形式，一种描述形式为：{功能词：adjunct={<次要义原>}}；另一种描述形式为：{功能词：EventRole=<动态角色>}。

对于第一种描述形式，构造词图时，仅仅构造一个中心义原节点。如：

也：DEF={功能词：adjunct={也}}，其词图为图 3 所示。

对于第二种描述形式，构造词图时，构造两个子图节点且这两个子图节点均指向 NULL，两个节点之间的弧的关系类型为概念项中相应的动态角色。如：

被：DEF={功能词：EventRole={agent}}，其词图如图 4 所示。

图 3 “也”的词图

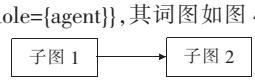

图 4 “被”的词图

通过构造词图，就可以以图的形式来存储词义信息，从而也使词义信息层次化、明了化，这为构造短语图、构造句子图打下了良好的基础。

4.4 构造词图的算法

```
CreatWordGraph(struct Word *word)
{ // 初始化图的顶点信息和弧信息
  char *semantic;  // 建立指向义原开始处的义原指针
  IntiQueue(Q);  // Q 为包含顶点 id 的栈
  建立词语顶点;
  建立中心义原的节点;
  EnQueue(nodeId);
  建立词语节点与中心义原节点之间的弧;
  While(*semantic! =NULL)
  { // 开始建立词图
    if(*semantic=='{')  // 在建立词图的过程中义原的信息从知网的概念词典获得
    { semantic++;
      char *fuhao_mao=strchar(semantic,':');  // 考虑到知网词典 DEF 项的表示方法
      if(fuhao_mao==NULL)
        fuhao_mao=strchar(semantic,')');
      strncpy(data,semantic,fuhao_mao);  // data 为义原节点的内容信息
      GetTop(Q,nodeId);
      nodeId++;
      建立义原节点 nodeId;
      EnQueue(nodeId);
      semantic=fuhao_mao+1;
    }
    if(*semantic==':'|| *semantic==',')
    { semantic++;
      判断是否建立弧，如需建立，则建立相应的弧;
      char *fuhao_deng=strchar(semantic,'=');
      strncpy(role-relation,semantic,fuhao_deng);
      GetTop(Q,nodeId);
      建立弧;
     semantic=fuhao_deng+1;
    }
    if(*semantic==')')
    {
      栈顶元素取出，准备下一个节点;
    }
  }
}
```

4.5 词图相似度的计算

4.5.1 义原间的语义相似度

(1)义原 a 与义原 b 的语义距离 $Distance(a,b)$：

$Distance(a,b)$=a 与 b 在义原分类树上的最短距离 (1)

(2)义原 a 与义原 b 的语义相似度 $Sim(a,b)$：

$$Sim(a,b)=1-\frac{Distance(a,b)}{\text{义原分类树树高}\times 2} \quad (2)$$

4.5.2 词图相似度的计算

若 G_1 与 G_2 均为词图，则计算它们的相似度分为以下三步：

(1)按公式(2)计算 G_1 与 G_2 的中心义原节点对应的义原相

似度,记为 $Sim_1(a,b)$。

(2)按如下步骤计算 G_1 与 G_2 的非中心义原节点对应的义原相似度,记为 $Sim_2(a,b)$。

由于非中心义原节点不止一个,所以计算较为复杂。作者按照把整体相似度还原为部分相似度的加权平均的思想,按照如下步骤对这些非中心义原节点分组:

①检查 G_1 与 G_2 的非中心义原节点与中心义原节点所连接的弧的关系类型是否有相同的,如果有相同的关系类型,则先计算具有相同的关系类型节点的义原相似度;

②对 G_1 与 G_2 中的弧的关系类型不同的非中心义原节点任意配对,计算出所有可能的配对的节点义原相似度;

③取相似度最大的一对,并将它们归为一组;

④在剩下的非中心义原节点对应的义原的配对相似度中,取最大的一对,并归为一组,如此反复,直到所有非中心义原节点都完成分组。如果某一部分的对应物为空,如何计算其相似度,处理方法是将任何义原与空值的相似度定义为一个比较小的常数 δ;

⑤对所有的非中心义原节点分组,最后求加权平均,其中各分组取相等的权值。

(3)于是,两个词图的相似度记为

$$Sim(G_1,G_2)=Sim_1(a,b)\times\beta_1+Sim_1(a,b)\times Sim_2(a,b)\times\beta_2 \quad (3)$$

其中 β_1、β_2 为两个参数,$\beta_1+\beta_2=1$,$\beta_1>0.5$。

5　实验结果与讨论

本文设计了两个对比实验。第一个实验,采用本文中提出的词语相似度计算方法,计算一组词和另外任意选取的一组词的相似度,由人来判断这组词和另外一组词的相似度大小是否与人的直觉相符合;第二个实验,使用了两种方法来计算词语相似度,并把它们的计算结果进行比较:

方法 1　仅使用《知网》语义表达式中第一独立义原来计算词语相似度。

方法 2　使用本文中介绍的语义相似度计算方法。

$\beta_1=0.6$　　$\beta_2=0.4$　　$\delta=0.1$

两个实验结果如表 1 所示:

表 1　实验结果对比

词语 1	词语 2	词语 2 的语义	方法 1	方法 2
工人	学生	人,*学,教育	1.000	0.960
工人	男人	人,家,男	1.000	0.671
工人	保姆	人,#职位,照料,家庭	1.000	0.563
工人	教师	人,#职位,*教,教育	1.000	0.782
工人	经理	人,#职位,官,商	1.000	0.681
工人	学校	场所,教,学,教育	0.211	0.150
工人	机器	机器	0.186	0.132
工人	计算机	电脑	0.186	0.102
工人	草莓	水果	0.286	0.151
工人	车间	部件,%场所,厂,工	0.186	0.147
工人	推卸	拒做	0.074	0.054

考察实验 1 的结果,也就是上面方法 2 的结果,可以看到,“工人”和其他各个词的相似度与人的直觉是比较符合的。

考察实验 2 的结果,也就是将方法 2 和方法 1 的结果相比较,可以看到:方法 1 的结果比较粗糙,只要是人,相似度都为 1,显然不够合理;方法 2 较细腻一些,能够区分不同人之间的相似度。

6　结束语

本文提出了基于知识图的词语间语义相似度计算,使这种属于语义网络范畴的知识表示方法在汉语信息处理中得以应用;建立了知识图知识表示方法和《知网》知识描述语言的对应关系,为语义计算提供了形式化工具;利用了知识图这种知识表示方法和《知网》这种语义资源对汉语的词语间语义相似度进行计算。实验结果表明该方法对词语间语义相似度计算是有效的。(收稿日期:2006 年 7 月)

参考文献:

[1] Miller G,Beckwith R,Fellbaum C,et al.Introduction to WordNet:an online lexical database[J].International Journal of Lexicography,1990,3(4):235-244.

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

[3] 梅家驹,竺一鸣,高蕴琦,等.同义词词林[M].上海:上海辞书出版社,1983.

[4] Rada R,Hafedh M,Bicknell E,et al.Development and application of a metric on semantic nets[J].IEEE Transactions on System,Man,and Cybernetics,1989,19(1):17-30.

[5] Lee J H,Kim M H,Lee Y J.Information retrieval based on conceptual distance in IS-A hierarchies[J].Journal of Documentation,1993,49(2):188-207.

[6] Resnik P.Using information content to evaluate semantic similarity in a taxonomy[C]//Proceedings of IJCAI,1995.

[7] 王斌.汉英双语语料库自动对齐研究[D].中国科学院计算技术研究所,1999.

[8] 刘群,李素建.基于《知网》的词语语义相似度计算[C]//第三届汉语词语语义研讨会,台北,2002.

[9] Brown P,Pietra S D,Pietra V D,et al.Word sense disambiguation using statistical methods[C]//Proceedings of the 29th Meeting of the Association for Computational Linguistics(ACL-91),Berkley,C A,1991:264-270.

[10] Lee L.Similarity-based approaches to natural language processing[D].Cambridge,M A:Harvard University,1997:11-97.

[11] 关毅,王晓龙.基于统计的汉语词语间语义相似度计算[C]//全国第七届计算语言学联合学术会议论文集,2003:221-227.

[12] Hoede C,Willems M.Knowledge graphs and natural language[M].The Netherlands:University of Twente,1989.

[13] Hoede C,Li X.Word graphs:the first set[C]//Eklund P W,Elliis G,Mann G.Conceptual Structures:Knowledge Representation as Interlingua:Auxiliary Proceedings of the 4th International Conference on Conceptual Structures,Sydney,Australia,1996:81-93.

[14] 张瑞霞.基于语义的汉语句法分析系统的研究与实现[D].西北大学,2005.