词汇语义相似度计算中相关技术的分析

余超[1] 蔡东风[1] 张桂平[1]

(1.沈阳航空工业学院 自然语言处理研究室 沈阳 110034)

摘 要： 本文扼要介绍了目前国内外关于词汇语义相似度计算的研究现状，阐述并分析了几种具有代表性的计算方法，并对目前较为流行的基于 SUMO、HowNet 的词汇语义相似度计算进行了较详细的介绍，对计算方法的特点进行分析，并在对知网概念的层次关系进行解析的基础上设计了一种形成概念-义原树结构的方法，该概念-义原树以树结构形式形象的表明了概念义原间的关系，方便计算机进行计算处理。

关键词： WordNet；HowNet；本体；概念-义原树

The Analysis to the related technology of Word Meaning Similarity Computation

Yu Chao[1], Cai DongFeng[1], Zhang GuiPing

(1.Natural Language Processing laboratory, Shenyang Institute of Aeronautical Engineering, Shengyang 110034)

Abstract: The paper introduces in brief the word meaning similarity computation situation in the world and analyzes several representative computation methods. It also expatiates the word meaning similarity computation based on SUMO and HowNet which is popular recently. We analyze the feature of these methods and parse the hierarchy relation of concepts in HowNet, on base of it we design a concept-sememe tree structure which can make it clearer to realize the relations between the sememes in concepts. The tree structure describes the relation between sememes in concept visually in form of tree and conveniences the computation through computer.

Keywords: WordNet; HowNet; ontology; concept-sememe tree

1 引言

词汇语义相似度(后面简称为词汇相似度)计算在机器翻译、信息检索、信息抽取、词义排歧等领域都有着广泛的应用，针对词汇级的相似度计算已经有不少学者做了大量的工作。目前词汇相似度的计算方法大体上可以分成两类，一类是基于规则的方法，另一类是基于统计的方法。基于规则的方法往往需要借助与某种世界知识来计算，主要是基于概念间结构层次组织的语义词典的方法，根据资源中概念间的语义关系来计算词汇间的相似度，这里所指的概念是词汇的语义描述，我们假定在词汇语义已经确定的情况下计算其相似度。基于统计的方法常借助与大规模语料的训练来判断两个词汇所出现在的上下文是否具有相似的相关词集合。其理论依据在于两个相似的词它们的相关词汇集合相近，将相关词汇集合向量化并计算向量夹角的余弦值来计算词汇间的相似度，本文重点介绍基于世界知识的词汇相似度计算方法。

基金资助：国防基础科研项目（K0504020515）资助

作者简介：余超(1977-) 男，武汉，硕士在读,E-mail:yc089067@sina.com.

尽管词汇相似度计算经过了多年的研究，但是它是很多研究领域的基础，由于词汇数目的庞大并且相似度是主观性很强的概念，目前尚未有令人满意的计算结果，回顾词汇相似度计算的研究历史，总结其研究现状，将有助于这方面工作的向前发展。

2 研究现状

由于词汇相似性由人为判断而具有较强的主观性，因此通常先计算词汇间的语义距离然后再用公式转化为相似度值。一般而言词汇间的语义距离是一个大于等于 0 的实数，数值越大相似度值越小，如果词汇间距离为 0，则相似度值为 1。基于语义词典的方法通常依赖于比较完备的大型语义词典。一般语义词典都是将所有的词组织在一棵或几棵树状的层次结构中，比如 WordNet 和同义词词林。

WordNet[1]是普林斯顿大学的心理学家，语言学家和计算机工程师联合设计的一种基于认知语言学的英语词典，以 synsets(在特定的上下文环境中可互换的同义词的集合)为单位组织信息。Miller 在 1985 年解释这样的思想：WordNet 使用同义词集合（synset）来代表词汇概念，并描述词汇矩阵，即在词的形式和意义之间建立起映射关系。

2.1 基于语义词典计算词汇相似度

Rada et al(1989)认为语义图中两节点间的路径越短，两节点语义越接近，并将此思想应用于 MEDLINE (联机医学文献分析和检索系统)，在这个系统中建立了 15000 个语义节点、9 个层次用来检索专业论文，其理论基础在于该系统层次结构中两节点间的边数就是两节点的语义距离。尽管其语义距离计算的思想简单，但是这个检索系统却取得了令人惊奇的好结果。

随后 Resnik, P. 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。Agirre & Rigau (1995) 在利用 WordNet 计算词语的语义相似度时，除了考虑结点间的路径长度外，还考虑到概念层次树的深度和宽度的因素。

王斌(1999)采用节点间路径长度来衡量其语义距离的方法，利用《同义词词林》来计算汉语词语之间的相似度。其中，同义词词林按照树状的层次结构把所有收录的词条组织到一起，把词汇分成大、中、小三类，其中大类 12 个，中类 97 个，小类有 1400 个，小类中随着级别的递增，语义刻画越来越细，形成了一个庞大的树状体系结构。

颜伟、荀恩东(2004)从 WordNet 中提取同义词并采取向量空间方法[2]计算英语词语的相似度，其向量包括三个方面：WordNet 的同义词词集、类属信息、意义解释。

徐德智、郑春卉，K. Passi(2006)基于 SUMO(建议上层共享知识本体)计算概念的语义相似度，其计算结果同人类的主观判断较为吻合。

2.2 基于本体的概念相似度计算

朱礼军、陶兰、刘慧借鉴计算语言学中的语义距离思想，提出了针对领域本体的概念相似度计算方法。该方法可以定量地分析资源描述框架[3]RDF (Resource Description Framework)构词所描述的概念、特征之间的相似度。设 C_1 和 C_2 是领域本体中的两个概念，$Sim(C_1, C_2)$ 表示这两个概念之间的相似度，见公式 1：

$$Sim(C_1,C_2)=\sum_{i=1}^{n}\delta_i(C_1,C_2)\theta_i \qquad (1)$$

其中，n 是概念 C_1、C_2 在领域本体中所具体有的最大深度：θ_i 是权重(可简单地取 $\theta_i=\frac{1}{n}$)；

$\delta_i(C_1,C_2)=\begin{cases}1.....当C_1,C_2前I个父类相同时\\0.....当C_1,C_2前I个父类代码不同时\end{cases}$，上述公式的权值可以根据实际需要进行调整。其语义计算的基本流程如图 1 所示：

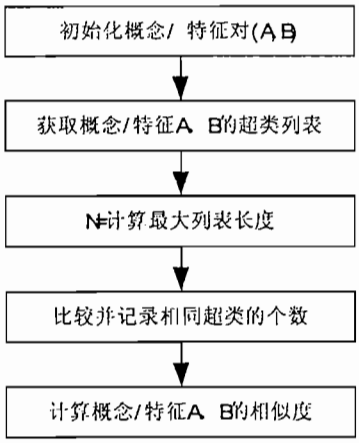

图 1 概念/特性相似度计算过程

2.3 基于统计方法的概念相似度计算

利用统计方法计算概念相似度通常是利用词语的相关性来计算词语的相似度。相似和相关是既互有联系又有区别的关系，两个较相似的词很可能有近似的相关词，比如“医生”和“护士”，和它们的较相关的词经常是“医院、病人、打针、吃药” 等；另一方面两个相关的词往往不一定具有相似性，如“吃”和“面包”、“医生”和“手术”等。基于统计的方法计算概念相似度的理论假设是：凡是语义相近的词，他们的上下文也应该相似。因此统计的方法对于两个词的相似度计算常建立在计算它们的相关词向量的相似度基础上。首先要选择一组特征词，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，一般用向量夹角余弦的计算结果作为这两个词的相似度。李涓子(1999)利用这种思想来实现语义的自动排歧；鲁松(2001)研究了如何利用词语的相关性来计算词语的相似度。Dagan(1999)使用了更为复杂的概率模型来计算词语的距离，P. Brown et al 采用平均互信息来计算词语之间的相似度。基于统计的定量分析方法能够对词汇间的语义相似性进行比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，计算量大，计算方法复杂，另外，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。

3 关键技术

3.1 语义层次树深度和宽度对语义计算的影响

尽管基于语义词典的相似度计算方法较多，但是概述起来常用的方法是形成概念层次树，那么概念层次树的深度和宽度对语义距离有什么影响呢？怎样综合考虑概念节点所在的深度和宽度对语义距离计算的影响呢？下面我们借鉴徐德智所写的文献[4]一文进行介绍。

SUMO 是由 IEEE 标准上层知识本体工作小组所建置的，其目的在于发展标准的上层知识本体，力求促进数据互通性、信息搜索和检索、自动推理和自然语言处理。目前 SUMO 已经和英语词汇网络 WordNet1.6 版本作初步的连结[5]，一个 SUMO 概念对应相关的 Wordnet 同义词集。

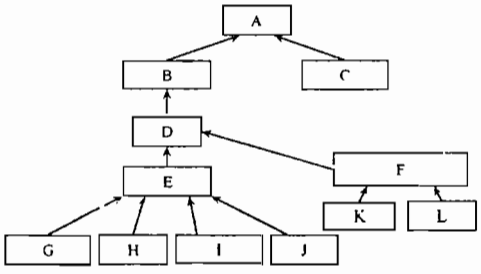

图 2 SUMO(片断)树状层次结构图

在未考虑深度和宽度影响之前，以两个概念在树中的最短路径距离来表示它们的语义距离，即两概念 C_1，C_2 的语义距离 Dist(C_1，C_2)为连接它们的最短路径上 n 条边的权值的总和，即：

$$Dist(C_1,C_2)=\sum_{i=1}^{n} weight_i \tag{2}$$

其中 $weight_i$ 是连接 C_1,C_2 的最短路径上第 i 条边的权值，起初树中所有边的权值都赋值为 1，即 $weight_i$=1。可以计算出 Dist(B,C)=2, Dist(G,H)=2。

1) 考虑概念层次树深度的影响

人们观察发现，处于层次树中离根较远的概念间的相似度(Dist(G,H))应该比离根近的概念间相似度(Dist(B,C))大些。这是因为在概念层次树中，自顶向下，上层的节点相对下层的节点稀疏，概念的描述也较粗糙。于是这里定义一个概念 C 在树中的深度 Depth(C)等于该概念与树根 A 的最短路径中所包含的边数，见公式 3、公式 4：

$$Depth(C)=\sum_{i=1}^{n} 1 \tag{3}$$

$$weight(C)=\frac{1}{2^{Dep(C)}} \tag{4}$$

其中，weight(C)表示从概念 C 引出的边的权值，按照这个公式，随着概念节点在树中深度的增加，以它引出的所有边的权值将减小。这就使得具有较大深度的概念间的语义距离相对较小，从而相似度相对较大。

2) 考虑概念层次树宽度的影响

另外，根据人们的主观判断，同一层的概念节点，兄弟节点越多，概念的描述越详尽，它们间的语义距离将越接近。比如世界上有大约 380 种鲨鱼，那么同层概念“大白鲨”、“鲸鲨”、“虎鲨”的语义距离就应该很小，而某些区域的概念描述又比较粗糙，相对而言语义距离应该较大。图 2 中，概念节点 K 和 L 的语义距离应该大于 G 和 H 的语义距离。所以概念的分类细致程度也应该是计算语义距离时应考虑的因素。下面考虑深度因素进行修正，如果用 Wid(C)表示概念 C 的宽度，即其孩子节点的数目，则概念 C 的权值为：

$$weight(C)=\frac{1}{Wid(C)}\times\frac{1}{2^{Dep(C)}} \tag{5}$$

这样，处于相同深度的概念宽度越大，其权值就越低，反之越高。

最后综合考虑深度、宽度对概念节点间语义距离的影响，修改公式为：

$$weight(C)=\begin{cases}\dfrac{1}{Wid(C)} & \text{.....C为根}\\ \dfrac{1}{Wid(C)}\times\dfrac{1}{2}\times weight(parent(C)) & \text{.....C为其他节点}\end{cases} \tag{6}$$

其中 parent(C)表示概念 C 的父节点，这样就既保证了概念在树中所处的深度由浅入深，概念的权值由大变小，又保证了概念的分类从粗糙到细致，概念的权值也由大变小，概念间的相对语义距离也随之减小，具体计算过程可参考文献[4]。在语义距离计算得出后再将语义距离转化为相似度值，该方法得到的结果和人们的判断比较吻合。

3.2 基于知网的概念相似度计算

基于知网的概念相似度计算是近年来研究的热点，并应用于语义排歧、问答系统等领域，其概念相似度计算的关键在于如何计算表示概念语义的义原(sememe)间的相似度，表示描述概念语义的义原间有着较为复杂的关系。

刘群教授在文献[6]一文中所提出的实词概念相似度计算的基本方法是将实词概念分成四个部分：首义原描述式、其他独立义原描述式、关系义原描述式与符号义原描述式，这四个部分的相似度分别计为：$Sim_1(S_1, S_2)$、

$Sim_2(S_1, S_2)$ 、$Sim_3(S_1, S_2)$ 、$Sim_4(S_1, S_2)$ ，S_1，S_2，表示两个实词概念，β_i是可以调整的权值系数，其和为 1，然后用公式 7 计算：

$$Similarity(S_1, S_2) = \sum_{i=1}^{4} Sim_i \times \beta_i \qquad (7)$$

具体的算法描述可参阅文献[6]，这里我们介绍其他独立义原描述式的计算方法，其他独立义原即语义概念中除第一义原外的所有义原。在其他独立义原较多的情况下，刘群教授将这些独立义原描述式分组计算其相似度值，步骤如下：

1) 将两个表达式的所有独立义原两两计算其义原相似度，义原相似度可以用知网提供的树状义原层次体系采用计算路径距离的方法求得，这里不在赘述；
2) 将相似度值最大的一组进行组合；
3) 在剩余的独立义原中再进行最大相似度值计算并配对组合直到其中一个表达式的独立义原全部完成配对，剩余未配对的义原与空值计算得到一个比较小的相似度值；

在部分相似度值已知的情况下采用整体加权平均的方法就能得到两个概念的相似度值。上述的步骤能轻松的计算两独立义原描述式间的整体相似度，由于知网中概念义原的数目有限，该方法通常能很快的计算出较大的相似度值。

知网的创造者董振东教授将词汇概念相似度公式表示为如公式 8 所示：

$$S(c1,c2)=p1*\beta 1+p2*\beta 2+p3*\beta 3+p4*\beta 4 \qquad (8)$$

其中 c1，c2 表示两概念，p1、 p2、 p3、 p4 分别表示两个概念中的首义原的相似度、两个概念中首义原框架的相似度、两个概念义原之间的相似度、完全包含的概念之间的相似度，详细计算步骤可以参考董振东教授所出版的书[7]。这里我们介绍两个概念义原之间的相似度(p3)计算方法，它的计算对象是概念的整体描述式，即刘群教授的首义原描述式和其他独立义原描述式的结合。其计算公式 9 如下所示：

$$p3 = Ns*2/(Nc1+Nc2) \qquad (9)$$

Ns 表示：两个概念中相同的概念节点的总数；

Nc1 表示：概念 1 中概念节点的总数；

Nc2 表示：概念 2 中概念节点的总数；

所谓概念节点是指：“动态角色={值}”这样的一个对。另外，这样的概念节点的结构必须是相同的才能说找到一对相同的概念节点。尽管书[7]中已对是否相同的概念节点进行了举例说明，在实际计算中要很好的确定概念节点是否具有相同的结构却并不容易，因此在本文中我们将知网的概念描述式转化为概念-义原树的形式，根据概念-义原树的层次结构我们就很直观的认识概念描述的层次结构，同时方便计算机计算处理。例如，知网中“医生”和“护士”的概念描述分别为：

“医生”：DEF={human|人:HostOf={Occupation|职位},domain={medical|医},{doctor|医治:agent={~}}}

“护士”：DEF={human|人:HostOf={Occupation|职位},domain={medical|医},{TakeCare|照料:agent={~}}}

我们根据分析“医生”和“护士”的概念，通过对知网知识系统描述语言 KDML 的分析，编程解析概念形成如下图 3、图 4 所示的概念-义原树：

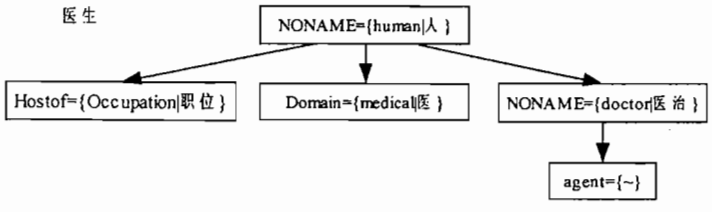

图 3 “医生”概念-义原树

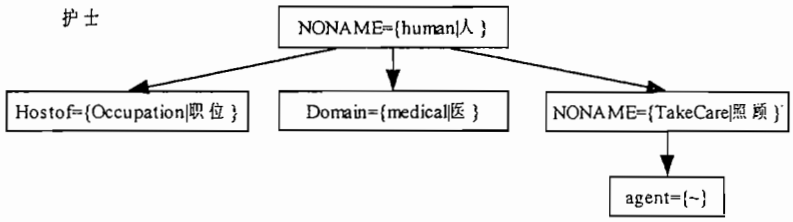

图 4 “护士”概念-义原树

概念-义原树的构建

1) 首先将词汇概念节点 Node(i)从词汇概念中抽取出来并记录下每个概念节点在概念中所处的位置 $Pos(N_i)$，其中 i 是概念节点的序号，$Pos(N_i)$表示序号为 i 的节点 Node 在概念描述式中的位置；
2) 将概念描述式中冒号 j 的位置记录在数组 PosFuhao[j]中，其中 j 表示冒号的序号，PosFuhao[j]表示序号为 j 的冒号在概念描述式中的位置；
3) 概念描述式的首节点为根节点，是所有概念节点的祖宗节点；
4) 将概念节点分为两类，将“动态角色={值}”描述完整的归为第一类，将其描述不完整的归为第二类，如“医生”的概念描述式中“HostOf={Occupation|职位}”属于第一类，而“human|人”和“doctor|医治”属于第二类；
5) 第一类概念节点的父节点寻找方法：从 Node(i)所在位置 Pos(Ni)向前查找，即 $Pos(N_i)$--，找到与其最近的冒号 j 所在位置 PosFuhao[j]，若在区间[PosFuhao[j], $Pos(N_i)$]中，符号“{”和“}”的个数相等，则冒号 j 所在位置 PosFuhao[j]前面最近的概念节点为 Node(i)的父节点，否则 $Pos(N_i)$--，找下一个冒号所在位置，继续相同的判断；
6) 第二类概念节点的父节点寻找方法：从 Node(i)所在位置 $Pos(N_i)$向前查找，即 $Pos(N_i)$--，找到与其最近的冒号所在位置 PosFuhao[j]，若在区间[PosFuhao[j], $Pos(N_i)$]中，符号“{”的个数比“}”的个数多 1，则冒号 j 所在位置 PosFuhao[j]前面最近的概念节点为 Node(i)的父节点，否则 $Pos(N_i)$--，找下一个冒号所在位置，继续相同的判断；
7) 对概念描述式中的个别情况如引号，左括号和右括号相连的情况“}{”等需要另外制定规则处理，这里不再详述；
8) 在所有概念节点的父节点找到的情况下，我们就能很容易的用概念-义原树形象的描述概念间义原的关系，方便计算机的计算处理；

可以看出“医生”和“护士”的概念树各有 5 个概念节点，从字符形式上看有 4 对概念节点一样，它们分别是：NONAME={human|人}、Hostof={Occupation|职位}、Domain={medical|医}和 agent={~}，但是我们可以发现“医生”概念-义原树中概念节点 agent={~}的父节点是 NONAME={doctor|医治}，而“护士”概念-义原树中概念节点 agent={~}的父节点是 NONAME={doctor|照顾}，即它们所对应的父节点不相同，因此概念节点 agent={~}不能认为是两概念“医生”和“护士”所具有的相同结构的概念节点，由此我们很快可以计算出 Ns=3, Nc1=5, Nc2=5,根据公式 p3 = Ns*2/(Nc1+Nc2)=3*2/(5+5)=0.6，即概念“医生”和“护士”的概念义原相似度为 0.6，而刘群教授的方法计算出概念“医生”和“护士”的概念义原相似度为 0.948。

从计算公式上可以看出董振东教授计算知网的方法体现出很强的结构性，在计算 p3 相似度时对只有完全相同的概念节点做加 1 处理，否则视为 0；而不采用将其义原抽出来计算其在树状义原层次体系中路径距离的方法。知网还提供了反义和对义词表，当两个概念属于反义或对义时，它们相似度为 0。例如“男人”和“女人”的概念相似度值为 0，而使用刘群教授的方法就能得到一个较大的相似度值 0.8611。

我们注意到，知网是一个很大的常识知识宝库，但在具体使用时还需要根据不同的使用领域进行加工才能取得最好的效果，在这方面已有学者做了不同的尝试，例如石晶、戴国忠在基于知网的文本推理一文中对知网概念进行改造和扩展以便于知识推理，余正涛在文献[8]中基于领域知网分析用户输入的问题进行受限领域的自动问答，均取得较好的实验结果。

4 结束语

词汇相似度计算是很多重点研究领域的基础，尽管已经有很多学者进行了大量的工作，但是由于汉语词汇语义表达的复杂性，词汇语义概念较强的主观性、具体应用领域的专业性等因素，目前仍将是计算机语言学深入研究的内容，随着人们研究的逐步深入、世界知识的不断完善及广阔的应用前景，相信在不久的将来会取得更好的计算结果。

参考文献：

[1] Miller G. Wordnet: An On-line Lexical Database[J]. International Journal of Lexicography,1990,3(4)

[2] 颜伟，荀恩东 基于 WordNet 的英语词语相似度计算[A] 2004 年全国计算语言学学生会议论文集[C] p282-288

[3] Dan Brickley, R V Guha. Resource Description Framework (RDF) model and syntax specification, W3C Working Draft , 2002. http://www.w3.org/TR/rdf-schema/

[4] 徐德智，郑春卉 K.Passi 基于 SUMO 的概念语义相似度研究 计算机应用[J] Vol.26 No.1 page180-184

[5] IEEE Working Group. Suggested upper Merged Ontology [DB/OL]. http://www.ontologyportal.org

[6] 刘群，李素建，基于知网的词汇相似度计算. http:/ /www.keenage.com, 2002.

[7] Dong ZhengDong, HowNet and Computation of Meaning[M] Singapore：World Scientific press, 2006. p197-206

[8] 余正涛，樊孝忠；宋丽哲 基于问句语料库的受限领域自动问答系统 计算机工程与应用[J] 2003.36 page28-30