

基于 HowNet 句子相似度的计算

闫红¹, 李付学¹, 周云²

(1. 营口理工学院 机电工程系, 辽宁 营口 115014;
2. 辽宁科技大学 软件学院, 辽宁 鞍山 114051)

摘要:汉语句子的相似度计算在自然语言处理领域中是一项基础而又重要的工作,它直接决定着相关领域的研究发展状况。在词语相似度计算的基础上,针对目前句子相似度计算方法的不足,文中提出一种基于 HowNet 的计算句子相似度的方法。在《知网》的词汇语义相似度计算基础上,加入了词语定义义原间的反义、对义关系、单义原的否定和符号义原、定义信息来计算词语的相似度。计算句子相似度前加入词语的消歧,在计算句子相似度时考虑了词语定义的关系义原与待比较的词定义的某个义原相等的情况,并加大了关系义原的权重。实验结果表明,在同等测试条件下,所提出的句子相似度计算方法可以提高句子相似度的计算精度,更符合人的直观感觉。

关键词:知网;词语相似度;义原;句子相似度

中图分类号:TP391.1

文献标识码:A

文章编号:1673-629X(2015)11-0053-05

doi:10.3969/j.issn.1673-629X.2015.11.011

Calculation of Sentence Similarity Based on HowNet

YAN Hong¹, LI Fu-xue¹, ZHOU Yun²

(1. Department of Electrical and Mechanical Engineering, Yingkou Institute of Technology, Yingkou 115014, China;
2. College of Software, Liaoning University of Science and Technology, Anshan 114051, China)

Abstract: Chinese sentence similarity computation is a fundamental and important work in the natural language processing. It directly determines the status of research and development for certain related fields. Based on the word similarity computing, for the shortcoming of current sentence similarity computing methods, present a method to compute sentence similarity Based on HowNet. Based on the lexical semantic similarity calculation of HowNet, antonyms, negative single sememe, symbol sememe, and definition information are demonstrated to calculate word similarity. In this method, word disambiguation is completed before the calculation of sentence similarity. The situation of the similarity between the relation sememe of a word definition and a certain sememe of the given word are considered, and the relation sememe weight is added. Under the same test conditions, the experimental results show that the proposed method can improve the computational accuracy of sentence similarity and it is much closer to the people's comprehension to the meanings of the sentences.

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

0 引言

计算句子相似度是信息处理中非常重要的工作,而词语之间的相似度计算对于句子间相似度计算的处理起着至关重要的作用。计算词语的相似度目前主要有两种方法:一种是基于语义字典的相似度计算方法(如同义词词林、知网、WordNet等);另一种是基于统计的相似度计算方法(如TF-IDF等)。目前国内,以《知网》^[1]为基础的词语相似度计算是当前较好的方法之一。

中科院刘群的基于《知网》的词语相似度计算^[2]

是利用义原的上下位关系来计算义原相似度,进而得到词语的相似度。朱征宇考虑概念描述式中各义原之间的线性关系,提出一种位置相关的权重分配策略^[3]。王石等提出一种基于搭配的中文词汇语义相似度计算方法^[4]。

文中在刘群等研究的基础上,计算词语间相似度时加入了词语定义的关系义原间的反义关系、对义关系、单个义原的否定和符号义原(如“^”和“~”)的定义信息。在计算句子相似度时,考虑了词语定义的关系义原与具体义原相关的情况。若词语定义的关系义原与

收稿日期:2014-12-11

修回日期:2015-03-04

网络出版时间:2015-11-04

基金项目:辽宁省教育科学研究一般项目(L2013539)

作者简介:闫红(1984-),女,硕士,讲师,研究方向为自然语言处理、物联网。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20151104.0949.028.html>

待比较词定义的某个义原相等,说明两个词语之间存在一定程度的同义,则词语的相似度也就高。

1 相似度简介

Lin Dekang^[5]从信息论的角度出发,定义了影响两个事物相似度的因素:即事物间的共性及其个性,事物A和B相似度定义公式如下:

$$\text{Sim}(A, B) = \frac{\log p(\text{common}(A, B))}{\log p(\text{description}(A, B))} \quad (1)$$

其中, $\log p(\text{common}(A, B))$ 是事物A,B的共性的信息量的大小; $\log p(\text{description}(A, B))$ 是事物A,B所含信息量的大小。

1.1 词语语义相似度

文献[2]的研究主要以基于实例的机器翻译为背景,认为词语之间相似度越高,它们在不同的上下文中就可以互相替换,而且不会改变语义。文献[6]在此基础上,考虑义原间的深度信息,并利用《知网》义原间的反义关系、对义关系和义原定义信息^[6]。文中在此基础上,在计算词语的相似度时考虑了单义原的否定(义原相似度取反)、加大了符号义原“~”和“~”的权值,并对第一义原有符号“~”的词语相似度的值取反。把词语相似度的取值范围设为 $[-1, +1]$,若词语的定义相同,则定义其相似度为+1;反之,若两个词语的定义相反,那么定义其相似度为-1。

1.2 句子语义相似度

目前基于HowNet的句子相似度的计算很大程度上依赖于词语相似度的计算结果,并且HowNet词语定义比较全面,而且考虑到语义层面的句子相似度,是目前比较好的方法之一。程传鹏等构造出义原的语义层次树,由各个义原在树中的相对位置,计算出义原之间的相似度^[7]。李佳媛提出了一种融合多种句子特征的汉语句子相似度计算方法^[8]。文中基于HowNet计算句子相似度,计算前加入了词语的消歧,计算时考虑了词语定义的关系义原与具体义原相关的情况。如果词语定义的关系义原与待比较词定义的某个义原相等时,说明两个词语之间存在一定程度上的同义,相互替换的机率比较大,必然词语的相似度也较高。

总结起来,用于确定句子相似度的方法主要有4类^[9-12],分别是基于字重叠(Word Overlap)的方法、基于语料库统计(TF-IDF)的方法、基于语言学(Linguistic)的方法和混合方法。

2 基于HowNet的词语相似度的计算

在《知网》的知识库,词是用义项来描述的(又称概念),一个词可以具有多个义项,而一个义项又可能由多个义原来描述。所以将两个词语间相似度的问题

转换为两个义项的相似度计算。当然,文中这里考虑的是两个孤立词语的相似度计算。因为在《知网》中所有的义项都是用义原来描述的,所以义项的相似度计算可以转化为义原间的相似度计算。

2.1 义原相似度计算

文献[2]中义原间计算相似度的公式如下:

$$\text{Sim}(p_1, p_2) = \frac{\alpha}{d + \alpha} \quad (2)$$

式中, p_1, p_2 分别代表两个义原; d 的取值为一个正整数,其描述的是义原 p_1 和 p_2 在义原结构层次体系中的路径长度; α 为一个动态的可调节参数。

文献[6]中义原间计算相似度的公式如下:

$$\text{Sim}(p_1, p_2) = \frac{\alpha \times w}{\alpha \times w + \text{dist}(p_1, p_2) + |v|} \quad (3)$$

其中, $w = \text{depth}(p_1) + \text{depth}(p_2)$; $v = \text{depth}(p_1) - \text{depth}(p_2)$ 。同式(2)中的 p_1, p_2 所描述的含义一样,表示两个义原; $\text{depth}(p_1)$ 描述的是义原 p_1 到根节点的深度; $\text{dist}(p_1, p_2)$ 描述的是两个义原间的路径长度; α 为一个动态的可调节参数,描述了深度信息对相似度计算结果的影响程度。

文中在此基础上,加入单义原的否定时两个义原相似度为相反数。提出了计算义原相似度的公式。

$$\text{Sim}(p_1, p_2) = \pm \frac{\alpha}{(d + d') + \alpha} \quad (4)$$

$$d = d(p_1, r) + d(p_2, q) \quad (5)$$

其中, d 是义原层次体系中的路径长度; r 和 q 分别是 p_1 和 p_2 路径上存在的反义义原。

如果 p_1, p_2 是反义义原,则 $d' = 0, d = 0, \text{Sim}(p_1, p_2) = -1$; 如果 p_1, p_2 其中一个存在否定时,则 $d' = 0, \text{Sim}(p_1, p_2) = -\frac{\alpha}{d + \alpha}$ 。

《知网》的知识库,有些专有名词(也可称为具体词)可能会在本来是义原的位置上(这些词会用圆括号括起来),因此在计算相似度时,需要特别处理一下专有名词与专有名词之间、专有名词和义原之间的相似度计算。因此,文中做如下规定:

(1) 专有名词 U 和义原 P 之间的相似度固定设置为一个较小的常数 γ ;

(2) 专有名词 u_1 和 u_2 之间的相似度计算规则为: u_1 和 u_2 为同一个词,则 $\text{Sim}(u_1, u_2) = 1$; 若 u_1 和 u_2 为不同词,则 $\text{Sim}(u_1, u_2) = 0$ 。

2.2 虚词概念的相似度计算

在实际的文本中,虚词一般不表达具体的语义信息,因此,认为实词概念和虚词概念之间的相似度为零。在《知网》的知识库中,虚词概念采用“{句法义原}”或“{关系义原}”两种方式进行描述。因此,虚

词概念之间的相似度计算就可以转化为计算其对应的句法义原或关系义原之间的相似度。

2.3 实词概念的相似度计算

对于实词概念间的相似度计算,可以分为四种情况:

(1)第一独立义原:定义这两个实词概念之间的相似度 $\text{Sim}_1(S_1, S_2)$;

(2)其他独立义原:表达式中除了第一独立义原以外的其他独立义原(或具体词),即非第一独立义原,定义这两个实词概念之间的相似度为 $\text{Sim}_2(S_1, S_2)$;

(3)关系义原:表达式中所有采用关系义原来描述的,定义这两个实词概念之间的相似度为 $\text{Sim}_3(S_1, S_2)$;

(4)符号义原:表达式中所有采用符号义原描述的,定义这两个实词概念之间的相似度为 $\text{Sim}_4(S_1, S_2)$ 。

综上所述,定义两个实词概念语义表达式的相似度为:

$$\text{Sim}(S_1, S_2) = \sum_{i=1}^4 \beta_i \prod_{j=1}^i \text{Sim}_j(S_1, S_2) \quad (6)$$

其中, $\beta_i (1 \leq i \leq 4)$ 是动态可调节的参数,且满足条件:

$$\beta_1 + \beta_2 + \beta_3 + \beta_4 = 1, \beta_1 \geq \beta_2 \geq \beta_3 \geq \beta_4$$

从第二个条件可以看出,从 Sim_1 到 Sim_4 对总体相似度结果的贡献程度逐个减少,这是因为一个概念最重要的属性和特征主要是由第一独立义原反映的,因此将其权值设定为 0.5。

下面对四种情况的相似度计算进行阐述。

1)第一独立义原相似度:概念开始描述的义原间的相似度,只要代入式(3)计算;与此同时,文中加入了符号义原“^”,当两个第一独立义原中只要有一个存在“^”时,则将计算的相似度结果取反。

2)其他独立义原相似度:在表达式中除了第一独立义原外,其他独立义原描述式可能多个,所以这种情况下的计算比较复杂。文中将这种情况下整体相似度的计算分解为部分相似度,然后再加权平均。按照如下步骤对这些独立义原描述式进行计算:

(1)将两个表达式中的其他独立义原任意组合,并计算出所有可能组合的义原相似度;

(2)按照计算相似度的结果降序排列,取相似度最大的一对,将它们划分一对,记录其相似度;

(3)在剩余的已经计算的独立义原相似度的组合中,再次取出相似度最大的一对,记录其相似度,重复步骤(3),直到所有独立义原都完成分组,剩余的未配对的独立义原文中规定其相似度为 0.2;

(4)最后求平均值作为其他独立义原的相似度。

3)关系义原相似度:把关系义原相同的分为一组,并计算其相似度;未配对的关系义原文中规定其相似度为 0.2(在计算句子相似度时,考虑了关系义原与待比较词定义的某个义原相等的情况,最后加权平均)。

4)符号义原相似度:符号义原的分组采用和关系义原分组相同的策略,将关系符号相同的划分在一起。文中对“^”或者“~”,符号相似度的总值做减 1 或者 0.5 处理,最后加权平均各个相似度结果。

3 消歧与句子相似度的计算

3.1 词语的消歧

根据 HowNet 的多义词消歧^[13-14],假设句子 s 中的词 t 是多义词, s 的其他词表示为 $s = (s_0, s_1, \dots, s_m)$, 多义词 t 有 n 个概念 k_1, k_2, \dots, k_n , 其他 m 个词也分别有 r_1, r_2, \dots, r_m 个概念。其中,第 s_i 的 r_i 个概念分别为 $k_1^i, k_2^i, \dots, k_{r_i}^i$; T 与 s_i 的概念相似度的最大值为:

$$\max(t, s_i) = \max(\lim_{1 \leq a \leq n, 1 \leq b \leq r_i} (k_a, k_b^i))$$

则 t 的 n 个概念与 m 个词的 $r_1 + r_2 + \dots + r_m$ 个概念的相似度最大值为:

$$\max(t) = \max(t, s_i) = \max(\lim_{1 \leq i \leq m} \lim_{1 \leq a \leq n, 1 \leq b \leq r_i} (k_a, k_b^i)) = \max(\lim_{1 \leq a \leq n, 1 \leq b \leq r_i} (k_a, k_b^i))$$

取 $\max(t)$ 所对应的某个概念 k_i 作为 t 的概念输出, $k_i = \arg\max(t)$, 即多义词 t 的概念是 k_i 。

3.2 基于知网的句子相似度计算

以词语相似度计算和消歧方法为基础,句子相似度的计算方式如下:

设句子 A 和 B , A 分词和预处理后的词序列为 $A(A_1, A_2, \dots, A_m)$, B 分词和预处理后的词序列为 $B(B_1, B_2, \dots, B_n)$, 定义句子中任意两个词 $A_i (1 \leq i \leq m)$ 和 $B_j (1 \leq j \leq n)$ 的相似度为 $\text{Sim}(A_i, B_j)$ 。句子 A 和 B 之间的语义相似度 $\text{Sim}(A, B)$ 为:

$$\text{Sim}(A, B) = (\frac{\sum_{i=1}^m a_i}{m} + \frac{\sum_{j=1}^n b_j}{n}) / 2 \quad (7)$$

式中, $a_i = \max(\text{Sim}(A_i, B_1), \text{Sim}(A_i, B_2), \dots, \text{Sim}(A_i, B_n))$; $b_j = \max(\text{Sim}(B_j, A_1), \text{Sim}(B_j, A_2), \dots, \text{Sim}(B_j, A_m))$ 。

文中在计算句子相似度时加入了词语定义的关系义原(与具体义原相关)与比较词定义的某个义原相等时的情况,认为两个词语之间存在一定程度上的同义,相互替换的机率比较大,则词语的相似度也较高,并在关系义原相似度总值加 0.5。通过实验对比发现,关系义原权重应该和其他义原的权重一样设为

0.2 时,能得到较好的句子相似度计算结果。

4 实验分析

在词语相似度计算时,加入了词语定义的义原间的反义关系、对义关系、单义原的否定和符号义原“^”和“~”的定义信息,通过实验证明该方法能有效提高词语相似度的精度。在句子相似度计算前,进行了词语消歧,并加入关系义原与待比较的词的定义的某个义原相同的情况,进而提高句子相似度计算的准确率。

表 1 词语相似度实验结果

词语 1	词语 2	文献[2]方法	文献[6]方法	文中方法
成功	失败	0.242 424	-1	-1
没有	有	0.444 444	-1	-1
丑	美丽	0.814 815	-1	-1
三伏	冬眠	0.044 444 4	0.058 38	-0.032 637
父亲	母亲	0.861 111	0.088 888	0.861 111
粉红	深红	0.074 074	0.580 367	0.074 074
跳槽	拔脚	0.184 242	0.158 175	0.184 242
出生	覆没	0.242 424	0.092 857 1	-1
健康	耳聋	0.186 047	-0.277 23	-0.253 26
舒服	残废	0.042 771	-0.222 203	-0.210 32
不得	不了		1	1
不得	得到		0.144 781	-1
待	不等		0.231 564	-1

从表 1 可以看出,文献[2]对一些有区别的词如出生和覆没、成功与失败等的相似度仍然很高,文献[6]对三伏(N time|时间,summer|夏,hot|热)和冬眠(V sleep|睡,#winter|冬)的区分度不是很大,原因在于计算相似度时,没有对单面义原里否定义原相似度取反。文中在计算义原时考虑单面义原否定相似度,得到的结果更为合理。例如,待(wait|等待)和不等(^wait|等待)对于义原中有符号“^”,文献[2]方法出现异常。由于没有把第一义原的“^”加入计算值,文献[6]得到的结果为正数。文中结合了知网符号“^”对义原的意思取反的特点,对第一义原相似度取反得到的结果更符合语义。

词语消歧实验如表 2 所示。

表 2 词义消歧

句子	分词去停用词	消歧的结果
我给他去了封信	去 信	去 V send 发送 信 N letter 信件 喜欢 V FondOf 喜欢 打 V associate 交往 老婆 N human 人, family 家, female 女
他喜欢打老婆	被当选 市长 被选拔 市长	

从表 2 分析,消歧时第一个结果比较合理准确,第

进行词语相似度计算时发现:每个概念中的义原个数不相等,有时候出现单个义原没有配对的情形。文中为了实验的简便,将义原(或具体词)与空值的相似度结果设置为一个较小的常数 δ 。在计算词语间相似度的实验中,参数分别设为: $\alpha = 1.6;\beta_1 = 0.5,\beta_2 = 0.2,\beta_3 = 0.17,\beta_4 = 0.13;\gamma = 0.2;\delta = 0.2$ 。文中所采用的词语相似度计算方法与文献[2,6]采用方法的实验对比结果见表 1。

二个就不是很合理,这只是根据词语间的相似度的最大值选取词语的定义,没有考虑词语间的相关性。

在词语相似度计算和词语消歧的基础上,文中对句子相似度计算方法进行了实验。经过对关系义原相似度权重的选取值的对比,最后确定为 0.2。几个参数的取值如下: $\alpha = 1.6;\beta_1 = 0.5,\beta_2 = 0.2,\beta_3 = 0.2,\beta_4 = 0.1;\gamma = 0.2;\delta = 0.2$ 。

句子相似度实验如表 3 所示。

表 3 句子相似度

句子	预处理结果	文献[2]方法	文中方法
我夸奖他 我谴责他	夸奖 谴责	0.242 424	-1
他被当选为市长 他被选拔为市长	被当选 市长 被选拔 市长	0.073 142 8	0.839 048

从表 3 分析,文献[2]方法计算夸奖和谴责的相似度是 0.242 424。那么“我夸奖他”和“我谴责他”的相似度是 0.242 424。而文中方法计算结果为-1。实验表明,在计算义原相似度时加入了反义义原,句子相似度和语义相似度更加相符。

当选和选拔的相似度是 0.194 286,文献[2]方法

的计算结果为 0.073 142 8。其实这两句话语义是相同的。

:当选 V win|获胜,scope=select|选拔

:选 V choose|选择/N publications|书刊/V select|选拔

文中加入了关系义原,由于当选的 scope 与 select|选拔相关,而选的第三个概念的定义有 select|选拔,在加大了关系义原的权值后句子相似度变为 0.839 048,结果比较合理准确。

5 结束语

《知网》是一个结构复杂,含有丰富的词汇语义知识和世界知识的语义知识词典。文中借鉴相关文献知识,提出了基于 HowNet 句子相似度的计算方法。

(1)对义原相似度的公式进行改进,并给含有单反义义原的相似度取反。

(2)在计算词语的相似度时加大符号义原“^”和“~”的权值,并对第一义原有符号“^”和“~”的词语相似度的值取反。实验结果表明,对一些用正面词加“^”来表示其正面意思的词效果很好。

(3)在计算句子相似度时,考虑了词语定义的关系义原(与具体义原相关)与比较的词的定义的某个义原是相等的情况。

但是,文中仍存在不足之处,词语消歧准确率不高,会影响句子相似度的计算结果,这是以后研究需要解决的问题。

参考文献:

- [1] 董振东,董 强,郝长伶. 知网的理论发现[J]. 中文信息学报,2007,21(4):3-9.
- [2] 刘 群,李素建. 基于《知网》的词汇语义相似度的计算[C]//第三届汉语词汇语义学研讨会. 台北:出版地不详,2002.

- [3] 朱征宇,孙俊华. 改进的基于《知网》的词汇语义相似度计算[J]. 计算机应用,2013,33(8):2276-2279.
- [4] 王 石,曹存根,裴亚军,等. 一种基于搭配的中文词汇语义相似度计算方法[J]. 中文信息学报,2013,27(1):7-14.
- [5] Lin Dekang. An information-theoretic definition of similarity semantic distance in WordNet[C]//Proceedings of the fifteenth international conference on machine learning. [s. l.]:[s. n.],1998.
- [6] 江 敏,肖诗斌,王弘蔚,等. 一种改进的基于《知网》的词语语义相似度计算[J]. 中文信息学报,2008,22(5):84-89.
- [7] 程传鹏,吴志刚. 一种基于知网的句子相似度计算方法[J]. 计算机工程与科学,2012,34(2):172-175.
- [8] 李佳媛. 汉语句子相似度计算技术及其应用[D]. 北京:北京信息科技大学,2013.
- [9] Jacob B,Benjamin C. Calculating the Jaccard similarity coefficient with map reduce for entity pairs in Wikipedia[EB/OL]. 2008. <http://www.infosci.cornell.edu/weblab/papers/Bank2008.pdf>.
- [10] Allan J,Bolivar A,Wade C. Retrieval and novelty detection at the sentence level[C]//Proceedings of SIGIR. [s. l.]:[s. n.],2003:314-321.
- [11] Li Y,McLean D,Bandar Z A,et al. Sentence similarity based on semantic nets and corpus statistics[J]. IEEE Transactions on Knowledge and Data Engineering,2006,18(8):1138-1150.
- [12] Chukfong H,Masrah A A M,Rabiah A K,et al. Word sense disambiguation based sentence similarity[C]//Proceedings of the 23rd international conference on computational linguistics. [s. l.]:[s. n.],2010:418-426.
- [13] 杨思春. 一种改进的句子相似度计算模型[J]. 电子科技大学学报,2006,35(6):956-959.
- [14] 刘小宇. 基于语义理解的中文常问问答系统的研究[D]. 大连:大连理工大学,2006.

基于HowNet句子相似度的计算

作者：[闫红](#)，[李付学](#)，[周云](#)，[YAN Hong](#)，[LI Fu-xue](#)，[ZHOU Yun](#)

作者单位：[闫红, 李付学, YAN Hong, LI Fu-xue\(营口理工学院 机电工程系, 辽宁 营口, 115014\)](#)，[周云, ZHOU Yun\(辽宁科技大学 软件学院, 辽宁 鞍山, 114051\)](#)

刊名：[计算机技术与发展](#)

英文刊名：[Computer Technology and Development](#)

年，卷(期)：2015, 25(11)

引用本文格式：[闫红](#). [李付学](#). [周云](#). [YAN Hong](#). [LI Fu-xue](#). [ZHOU Yun](#) [基于HowNet句子相似度的计算](#)[期刊论文]-[计算机技术与发展](#) 2015(11)