

基于《知网》的词汇语义相似度计算¹

刘群^{†‡} 李素建[†]

{liuqun,lisujian}@ict.ac.cn

[†] 中国科学院计算技术研究所

[‡] 北京大学计算语言学研究所

摘要：

《知网》是一部比较详尽的语义知识词典。在基于实例的机器翻译中，词语相似度计算是一个重要的环节。不过，由于《知网》中对于一个词的语义采用的是—种多维的知识表示形式，这给词语相似度的计算带来了麻烦。这一点与 WordNet 和《同义词词林》不同。在 WordNet 和《同义词词林》中，所有同类的语义项（WordNet 的 synset 或《同义词词林》的词群）构成一个树状结构，要计算语义项之间的距离，只要计算树状结构中相应结点的距离即可。而在《知网》中词语相似度的计算存在以下问题：

1. 每一个词的语义描述由多个义原组成，例如“暗箱”—词的语义描述为：
part|部件,%tool|用具,body|身，“写信”—词的语义描述为：
#TakePicture|拍摄 write|写,ContentProduct=letter|信件；
2. 词语的语义描述中各个义原并不是平等的，它们之间有着复杂的关系，通过—种专门的知识描述语言来表示。

我们的工作主要包括：

1. 研究《知网》中知识描述语言的语法，了解其描述—词义所用的多个义原之间的关系，区分其在词语相似度计算中所起的作用；
2. 提出利用《知网》进行词语相似度计算的算法；
3. 通过实验验证该算法的有效性，并与其他算法进行比较。

关键词：《知网》 词汇语义相似度计算 自然语言处理

1 引言

在基于实例的机器翻译中，词语相似度的计算有着重要的作用。例如要翻译“张三写的小说”这个短语，通过语料库检索得到译例：

- 1) 李四写的小说 / the novel written by Li Si
- 2) 去年写的小说 / the novel written last year

通过相似度计算我们发现，“张三”和“李四”都是具体的人，语义上非常相似，而“去年”的语义是时间，和“张三”相似度较低，因此我们选用“李四写的小说”这个实例进行类比翻译，就可以得到正确的译文：

the novel written by Zhang San

¹ 本项研究受国家重点基础研究计划（973）支持，项目编号是G1998030507-4 和G1998030510。

如果选用后者作为实例，那么得到的错误译文将是：

* the novel written Zhang San

通过这个例子可以看出相似度计算在基于实例的机器翻译中所起的作用。

在基于实例的翻译中另一个重要的工作是双语对齐。在双语对齐过程中要用到两种语言词语的相似度计算，这不在本文所考虑的范围之内。

除了基于实例的机器翻译之外，词语相似度计算在信息检索、信息抽取、词义排歧等领域都有着广泛的应用。

2 词语相似度及其计算的方法

2.1 什么是词语相似度

什么是词语相似度？

我们认为，词语相似度是一个主观性相当强的概念。脱离具体的应用去谈论词语相似度，很难得到一个统一的定义。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。从某一角度看非常相似的词语，从另一个角度看，很可能差异非常大。

不过，在具体的应用中，词语相似度的含义可能就比较明确了。例如，在基于实例的机器翻译中，词语相似度主要用于衡量文本中词语的可替换程度；而在信息检索中，相似度更多的要反映文本或者用户查询在意义上的符合程度。

本文的研究主要以基于实例的机器翻译为背景，因此在本文中我们所理解的词语相似度就是两个词语在不同的上下文中可以互相替换使用而不改变文本的句法语义结构的程度。两个词语，如果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越大，二者的相似度就越高，否则相似度就越低。

相似度是一个数值，一般取值范围在 $[0,1]$ 之间。一个词语与其本身的语义相似度为 1。如果两个词语在任何上下文中都不可替换，那么其相似度为 0。

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。

2.2 词语相似度与词语距离

度量两个词语关系的另一个重要指标是词语的距离。

一般而言，词语距离是一个 $[0, \infty)$ 之间的实数。

一个词语与其本身的距离为 0。

词语距离与词语相似度之间有着密切的关系。

两个词语的距离越大，其相似度越低；反之，两个词语的距离越小，其相似度越大。二者之间可以建立一种简单的对应关系。这种对应关系需要满足以下几个条件：

- 1) 两个词语距离为 0 时，其相似度为 1；
- 2) 两个词语距离为无穷大时，其相似度为 0；
- 3) 两个词语的距离越大，其相似度越小（单调下降）。

对于两个词语 W_1 和 W_2 ，我们记其相似度为 $Sim(W_1, W_2)$ ，其词语距离为

$Dis(W_1, W_2)$ ，那么我们可以定义一个满足以上条件的简单的转换关系：

$$Sim(W_1, W_2) = \frac{\alpha}{Dis(W_1, W) + \alpha} \quad \dots\dots (1)$$

其中 α 是一个可调节的参数。 α 的含义是：当相似度为 0.5 时的词语距离值。这种转换关系并不是唯一的，我们这里只是给出了其中的一种可能。

在很多情况下，直接计算词语的相似度比较困难，通常可以先计算词语的距离，然后再转换成词语的相似度。所以在本文后面的有些章节，我们只谈论词语的距离，而没有提及词语的相似度，读者应该知道这二者是可以互相转换的。

2.3 词语相似度与词语相关性

度量两个词语关系的另一个重要指标是词语的相关性。

词语相关性反映的是两个词语互相关联的程度。可以用这两个词语在同一个语境中共现的可能性来衡量。

词语相关性也是一个[0,1]之间的实数。

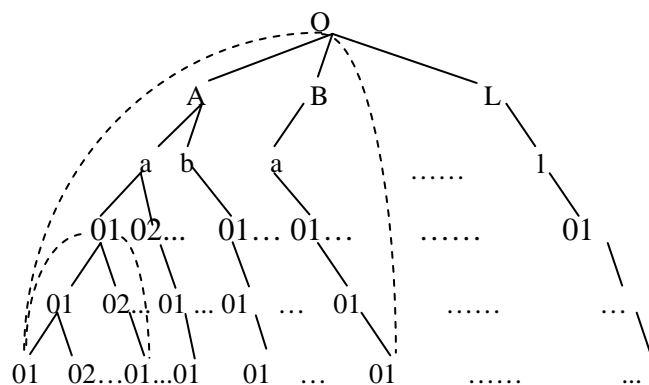
词语相关性和词语相似性是两个不同的概念。例如“医生”和“疾病”两个词语，其相似性非常低，而相关性却很高。可以这么认为，词语相似性反映的是词语之间的聚合特点，而词语相关性反映的是词语之间的组合特点。

同时，词语相关性和词语相似性又有着密切的联系。如果两个词语非常相似，那么这两个词语与其他词语的相关性也会非常接近。反之，如果两个词语与其他词语的相关性特点很接近，那么这两个词一般相似程度也很高。

2.4 词语相似度的计算方法

词语距离有两类常见的计算方法，一种是根据某种世界知识（Ontology）来计算，一种利用大规模的语料库进行统计。

根据世界知识（Ontology）计算词语语义距离的方法，一般是利用一部同义词词典（Thesaurus）。一般同义词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。我们知道，在一棵树形图中，任何两个结点之间有且只有一条路径。于是，这条路径的长度就可以作为这两个概念的语义距离的一种度量。



虚线用于标识某上层节点到下层节点的路径

图 1 《同义词词林》语义分类树形图

王斌(1999)采用这种方法利用《同义词词林》来计算汉语词语之间的相似度(如图1所示)。有些研究者考虑的情况更复杂。Agirre & Rigau (1995)在利用Wordnet计算词语的语义相似度时,除了结点间的路径长度外,还考虑到了其他一些因素。例如:

- 1) 概念层次树的深度:路径长度相同的两个结点,如果位于概念层次的越底层,其语义距离较大;比如说:“动物”和“植物”、“哺乳动物”和“爬行动物”,这两对概念间的路径长度都是2,但前一对词处于语义树的较高层,因此认为其语义距离较大,后一对词处于语义树的较低层,其语义距离更小;
- 2) 概念层次树的区域密度:路径长度相同的两个结点,如果位于概念层次树中高密度区域,其语义距离应大于位于低密度区域。由于Wordnet中概念描述的粗细程度不均,例如动植物分类的描述及其详尽,而有些区域的概念描述又比较粗疏,所以加入了概念层次树区域密度对语义距离的影响。

另一种词语相似度的计算方法是大规模的语料来统计。例如,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)作为这两个词的相似度。这种做法的假设是,凡是语义相近的词,他们的上下文也应该相似。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。

这两种方法各有特点。基于世界知识的方法简单有效,也比较直观、易于理解,但这种方法得到的结果受人的主观意识影响较大,有时并不能准确反映客观事实。另外,这种方法比较准确地反映了词语之间语义方面的相似性和差异,而对于词语之间的句法和语用特点考虑得比较少。基于语料库的方法比较客观,综合反映了词语在句法、语义、语用等方面的相似性和差异。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。

本文主要研究基于《知网(Hownet)》的词语相似度计算方法,这是一种基于世界知识的方法。

3 《知网(Hownet)》简介

按照《知网》的创造者董振东先生自己的说法(杜飞龙,1999):

《知网》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。

《知网》中含有丰富的词汇语义知识和世界知识,为自然语言处理和机器翻译等方面的研究提供了宝贵的资源。不过,在我们真正试图利用《知网》来进行计算机处理时,发现还是会遇到不少困难。我们的感觉是,《知网》确实是一座宝库,但另一方面,《知网》的内容又非常庞杂。尽管《知网》的提供了详细的文档,但由于这些文档不是以一种形式化的方式说明的,很多地方多少显得有些混乱。当我们阅读这些文档时,很容易一下子陷入大量的细节之中,而很难对《知

网》有一个总体的把握。这使得我们在进行计算的时候觉得很不方便。因此，我们在试图利用《知网》进行计算的过程中，也在逐渐加深我们对于《知网》的认识，并试图整理出一个关于《知网》的比较清晰的图象。

本节中，我们对于《知网》的描述是按照我们自己的语言来组织的，很多地方加入了我们的理解，并不一定都是《知网》文档中描述。我们希望通过这种方法，使读者更快地了解《知网》，对《知网》有一个比较清晰而全面的印象。当然，我们的理解也难免有错误和遗漏之处，欢迎《知网》的作者和其他读者批评指正。

3.1 《知网》的结构

董振东先生反复强调，《知网》并不是一个在线的词汇数据库，《知网》不是一部语义词典。

在介绍《知网》的结构之前，我们首先要理解《知网》中两个主要的概念：“概念”与“义原”。

“概念”是对词汇语义的一种描述。每一个词可以表达为几个概念。

“概念”是用一种“知识表示语言”来描述的，这种“知识表示语言”所用的“词汇”叫做“义原”。

“义原”是用于描述一个“概念”的最小意义单位。

与一般的语义词典（如《同义词词林》，或 Wordnet）不同，《知网》并不是简单的将所有的“概念”归结到一个树状的概念层次体系中，而是试图用一系列的“义原”来对每一个“概念”进行描述。

《知网》一共采用了 1500 义原，这些义原分为以下几个大类：

- 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

对于这些义原，我们把它们归为三组：第一组，包括第 1 到 7 类的义原，我们称之为“基本义原”，用来描述单个概念的语义特征；第二组，只包括第 8 类义原，我们称之为“语法义原”，用于描述词语的语法特征，主要是词性（Part of Speech）；第三组，包括第 9 和第 10 类的义原，我们称之为“关系义原”，用于描述概念和概念之间的关系（类似于格语法中的格关系）。

除了义原以外，《知网》中还用了一些符号来对概念的语义进行描述，如下表所示：

,	多个属性之间，表示“和”的关系
#	表示“与其相关”

%	表示“是其部分”
\$	表示“可以被该‘V’处置，或是该“V”的受事，对象，领有物，或者内容
*	表示“会‘V’或主要用于‘V’，即施事或工具
+	对V类，它表示它所标记的角色是一种隐性的，几乎在实际语言中不会出现
&	表示指向
~	表示多半是，多半有，很可能的
@	表示可以做“V”的空间或时间
?	表示可以是“N”的材料，如对于布匹，我们标以“?衣服”表示布匹可以是“衣服”的材料
{}	(1) 对于V类，置于[]中的是该类V所有的“必备角色”。如对于“购买”类，一旦它发生了，必然会在实际上有如下角色参与：施事，占有物，来源，工具。尽管在多数情况下，一个句子并不把所有的角色都交代出来 (2) 表示动态角色，如介词的定义
()	置于其中的应该是一个词标记，例如，(China 中国)
^	表示不存在，或没有，或不能
!	表示某一属性为一种敏感的属性，例如：“味道”对于“食物”，“高度”对于“山脉”，“温度”对于“天象”等
[]	标识概念的共性属性

表 1: 《知网》知识描述语言中的符号及其含义

我们把这些符号又分为几类，一类是用来表示语义描述式之间的逻辑关系，包括以下几个符号：~ ^，另一类用来表示概念之间的关系，包括以下几个符号：# % \$ * + & @ ? !，第三类包括几个无法归入以上两类的特殊符号：{ } () []。

我们看到，概念之间的关系有两种表示方式：一种是用“关系义原”来表示，一种是用表示概念关系的符号来表示。按照我们的理解，前者类似于一种格关系，后者大部分是一种格关系的“反关系”，例如“\$”我们就可以理解为“施事、对象、领有、内容”的反关系，也就是说，该词可以充当另一个词的“施事、对象、领有、内容”。

义原一方面作为描述概念的最基本单位，另一方面，义原之间又存在复杂的关系。在《知网》中，一共描述了义原之间的8种关系：上下位关系、同义关系、反义关系、对义关系、属性-宿主关系、部件-整体关系、材料-成品关系、事件-角色关系。可以看出，义原之间组成的是一个复杂的网状结构，而不是一个单纯的树状结构。不过，义原关系中最重要的是的上下位关系。根据义原的上下位关系，所有的“基本义原”组成了一个义原层次体系（如图2）。这个义原层次体系是一个树状结构，这也是我们进行语义相似度计算的基础。

```

- entity|实体
  thing|万物
  ...   physical|物质
        ...   animate|生物
              ...   AnimalHuman|动物
                    ...   human|人
                          humanized|拟人
                                animal|兽
                                      beast|走兽
                                          ...

```

图 2 树状的义原层次结构

从表面上看，其他的语义词典，如《同义词词林》和 Wordnet，也有一个树状的概念层次体系，好像《知网》和它们很相似，但实际上有着本质的不同。在《同义词词林》和 Wordnet 种，概念就是描写词义的最小单位，所以，每一个概念都是这个概念层次体系中的一个结点。而在《知网》中，每一个概念是通过一组义原来表示的，概念本身并不是义原层次体系中的一个结点，义原才是这个层次体系中的一个结点。而且，一个概念并不是简单的描述为一个义原的集合，而是要描述为使用某种专门的“知识描述语言”来表达的一个语义表达式。也就是说，在描述一个概念的多个义原中，每个义原所起到的作用是不同的，这就给我们的相似度计算带来了很大的困难。下面我们就对这个描述概念的知识描述语言进行一些考察。

3.2 《知网》的知识描述语言

《知网》对概念的描述是比较复杂的。在《知网》中，每一个概念用一个记录来表示，如下所示：

```

NO.=017144
W_C=打
G_C=V
E_C=~网球，~牌，~秋千，~太极，球~得很棒
W_E=play
G_E=V
E_E=
DEF=exercise|锻炼,sport|体育

```

其中 NO.为概念编号，W_C，G_C，E_C 分别是汉语的词语、词性和例子，W_E、G_E、E_E 分别是英语的词语、词性和例子，DEF 是知网对于该概念的定义，我们称之为一个语义表达式。其中 DEF 是知网的核心。我们这里所说的知识描述语言也就是 DEF 的描述语言。

在《知网》的文档中，对知识描述语言做了详尽的介绍。不过，由于该文档过于偏重细节，不易从总体上把握。本节中我们试图对于这种知识描述语言给出一个简单的概括。

我们看几个例子：

打	017144	exercise 锻练,sport 体育
男人	059349	human 人,family 家,male 男
高兴	029542	aValue 属性值,circumstances 境况,happy 福,desired 良
生日	072280	time 时间,day 日,@ComeToWorld 问世,\$congratulate 祝贺
写信	089834	write 写,ContentProduct=letter 信件
北京	003815	place 地方,capital 国都,ProperName 专,(China 中国)
爱好者	000363	human 人,*FondOf 喜欢,#WhileAway 消闲
必须	004932	{modality 语气}
串	015204	NounUnit 名量,&(grape 葡萄),&(key 钥匙)
从良	016251	cease 停做,content=(prostitution 卖淫)
打对折	017317	subtract 削减,patient=price 价格,commercial 商,(range 幅度=50%)
儿童基金会	024083	part 部件,%institution 机构,politics 政,#young 幼,#fund 资金,(institution 机构=UN 联合国)

表 2 : 《知网》知识描述语言实例

从这些例子我们可以看到,《知网》的知识描述语言是比较复杂的。我们将这种知识描述语言归纳为以下几条:

- 1) 《知网》收入的词语主要归为两类,一类是实词,一类是虚词;
- 2) 虚词的描述比较简单,用“{句法义原}”或“{关系义原}”进行描述;
- 3) 实词的描述比较复杂,由一系列用逗号隔开的“语义描述式”组成,这些“语义描述式”又有以下三种形式:
 - a) 独立义原描述式:用“基本义原”,或者“(具体词)”进行描述;
 - b) 关系义原描述式:用“关系义原=基本义原”或者“关系义原=(具体词)”或者“(关系义原=具体词)”来描述;
 - c) 符号义原描述式:用“关系符号 基本义原”或者“关系符号(具体词)”加以描述;
- 4) 在实词的描述中,第一个描述式总是一个基本义原,这也是对该实词最重要的一个描述式,这个基本义原描述了该实词的最基本的语义特征。

4 基于《知网》的语义相似度计算方法

从上面的介绍我们看到,与传统的语义词典不同,在《知网》中,并不是将每一个概念对应于一个树状概念层次体系中的一个结点,而是通过用一系列的义原,利用某种知识描述语言来描述一个概念。而这些义原通过上下位关系组织成一个树状义原层次体系。我们的目标是要找到一种方法,对用这种知识描述语言表示的两个语义表达式进行相似度计算。

利用《知网》计算语义相似度一个最简单的方法就是直接使用词语语义表达式中的第一独立义原,把词语相似度等价于第一独立义原的相似度。这种方法好处是计算简单,但没有利用知网语义表达式中其他部分丰富的语义信息。

Li Sujian, et al. (2002)中提出了一种词语语义相似度的计算方法,计算过程综

合利用了《知网》和《同义词词林》。在义原相似度的计算过程中，不仅考虑了义原之间的上下文关系，还考虑了义原之间的其他关系。在计算词语相似度时，加权合并了《同义词词林》的词义相似度、《知网》语义表达式的义原相似度和义原关联度。这种算法中，《同义词词林》和《知网》采用完全不同的语义体系和表达方式，词表也相差较大，把它们合并计算的合理性值得怀疑。另外，把语义关联度加权合并计入义原相似度中，也未必合理。

4.1 词语相似度计算

对于两个汉语词语 W_1 和 W_2 ，如果 W_1 有 n 个义项（概念）： $S_{11}, S_{12}, \dots, S_{1n}$ ， W_2 有 m 个义项（概念）： $S_{21}, S_{22}, \dots, S_{2m}$ ，我们规定， W_1 和 W_2 的相似度各个概念的相似度之最大值，也就是说：

$$Sim(W_1, W_2) = \max_{i=1..n, j=1..m} Sim(S_{1i}, S_{2j}) \quad \dots\dots (2)$$

这样，我们就把两个词语之间的相似度问题归结到了两个概念之间的相似度问题。当然，我们这里考虑的是孤立的两个词语的相似度。如果是在一定上下文之中的两个词语，最好是先进行词义排歧，将词语标注为概念，然后再对概念计算相似度。

4.2 义原相似度计算

由于所有的概念都最终归结于用义原（个别地方用具体词）来表示，所以义原的相似度计算是概念相似度计算的基础。

由于所有的义原根据上下位关系构成了一个树状的义原层次体系，我们这里采用简单的通过语义距离计算相似度的办法。假设两个义原在这个层次体系中的路径距离为 d ，根据公式(1)，我们可以得到这两个义原之间的语义距离：

$$Sim(p_1, p_2) = \frac{\alpha}{d + \alpha} \quad \dots\dots (3)$$

其中 p_1 和 p_2 表示两个义原（primitive）， d 是 p_1 和 p_2 在义原层次体系中的路径长度，是一个正整数。 α 是一个可调节的参数。

用这种方法计算义原相似度的时候，我们只利用了义原的上下位关系。实际上，在《知网》中，义原之间除了上下位关系外，还有很多种其他的关系，如果在计算时考虑进来，可能会得到更精细的义原相似度度量，例如，我们可以认为，具有反义或者对义关系的两个义原比较相似，因为它们在实际的语料中互相可以互相替换的可能性很大。对于这个问题这里我们不展开讨论，留给以后的研究工作来处理。

另外，在知网的知识描述语言中，在一些义原出现的位置都可能出现一个具体词（概念），并用圆括号（）括起来。所以我们在计算相似度时还要考虑到具体词和具体词、具体词和义原之间的相似度计算。理想的做法应该是先把具体词还原成《知网》的语义表达式，然后再计算相似度。这样做将导入函数的递归调用，甚至可能导致死循环，这会使算法会变得很复杂。由于具体词在《知网》的语义表达式中只占很小的比例，因此，在我们的实验中，为了简化起见，我们做如下规定：

- 具体词与义原的相似度一律处理为一个比较小的常数 ()；
- 具体词和具体词的相似度，如果两个词相同，则为 1，否则为 0。

4.3 虚词概念的相似度的计算

我们认为，在实际的文本中，虚词和实词总是不能互相替换的，因此，虚词概念和实词概念的相似度总是为零。

由于虚词概念总是用“{句法义原}”或“{关系义原}”这两种方式进行描述，所以，虚词概念的相似度计算非常简单，只需要计算其对应的句法义原或关系义原之间的相似度即可。

4.4 实词概念的相似度的计算

由于实词概念是用一个语义表达式来描述的，因此其相似度计算变得非常复杂。

如何计算两个语义表达式的相似度呢？

我们的基本设想是：整体相似要建立在部分相似的基础上。把一个复杂的整体分解成部分，通过计算部分之间的相似度得到整体的相似度。

假设两个整体A和B都可以分解成以下部分：A分解成 A_1, A_2, \dots, A_n ，B分解成 B_1, B_2, \dots, B_m ，那么这些部分之间的对应关系就有 $m \times n$ 种。问题是：这些部分之间的相似度是否都对整体的相似度发生影响？如果不是全部都发生影响，那么我们应该如何选择那些发生影响的那些部分之间的相似度？选择出来以后，我们又如何得到整体的相似度？

我们认为：一个整体的各个不同部分在整体中的作用是不同的，只有在整体中起相同作用的部分互相比拟才有效。例如比较两个人长相是否相似，我们总是比较它们的脸型、轮廓、眼睛、鼻子等相同部分是否相似，而不会拿眼睛去和鼻子做比较。

因此，在比较两个整体的相似性时，我们首先要做的工作是对这两个整体的各个部分之间建立起一一对应的关系，然后在这些对应的部分之间进行比较。我们把这种做法比喻成古代的战场的两军对垒：兵对兵、将对将，捉对厮杀。

还有一个问题：如果某一部分的对应物为空，如何计算其相似度？我们的处理方法是：

- 将任何义原(或具体词)与空值的相似度定义为一个比较小的常数()；整体的相似度通过部分的相似度加权平均得到。

对于实词概念的语义表达式，我们将其分成四个部分：

- 1) 第一独立义原描述式：我们将两个概念的这一部分的相似度记为 $Sim_1(S_1, S_2)$ ；
- 2) 其他独立义原描述式：语义表达式中除第一独立义原以外的所有其他独立义原(或具体词)，我们将两个概念的这一部分的相似度记为 $Sim_2(S_1, S_2)$ ；
- 3) 关系义原描述式：语义表达式中所有的用关系义原描述式，我们将两个概念的这一部分的相似度记为 $Sim_3(S_1, S_2)$ ；
- 4) 符号义原描述式：语义表达式中所有的用符号义原描述式，我们将两个

概念的这一部分的相似度记为 $Sim_4(S_1, S_2)$ 。
于是，两个概念语义表达式的整体相似度记为：

$$Sim(S_1, S_2) = \sum_{i=1}^4 \beta_i Sim_i(S_1, S_2) \quad \dots\dots (4)$$

其中， $\beta_i (1 \leq i \leq 4)$ 是可调节的参数，且有： $\beta_1 + \beta_2 + \beta_3 + \beta_4 = 1$ ， $\beta_1 > \beta_2 > \beta_3 > \beta_4$ 。后者反映了 Sim_1 到 Sim_4 对于总体相似度所起到的作用依次递减。由于第一独立义原描述式反映了一个概念最主要的特征，所以我们应该将其权值定义得比较大，一般应在 0.5 以上。

在实验中我们发现，如果 Sim_1 非常小，但 Sim_3 或者 Sim_4 比较大，将导致整体的相似度仍然比较大的不合理现象。因此我们对公式(4)进行了修改，得到公式如下：

$$Sim(S_1, S_2) = \sum_{i=1}^4 \beta_i \prod_{j=1}^i Sim_j(S_1, S_2) \quad \dots\dots (5)$$

其意义在于，主要部分的相似度值对于次要部分的相似度值起到制约作用，也就是说，如果主要部分相似度比较低，那么次要部分的相似度对于整体相似度所起到的作用也要降低。

下面我们再分别讨论每一部分的相似度。

- 1) 第一独立义原描述式：就是两个义原的相似度，按照公式(3)计算即可；
- 2) 其他独立义原描述式：由于其他独立义原描述式不止一个，所以计算较为复杂。我们还是按照上面的思想，把整体相似度还原为部分相似度的加权平均。困难在于，各个独立义原描述式之间没有分工，所以很难找到对应关系。我们按照如下步骤对这些独立义原描述式分组：
 - a) 先把两个表达式的所有独立义原（第一个除外）任意配对，计算出所有可能的配对的义原相似度；
 - b) 取相似度最大的一对，并将它们归为一组；
 - c) 在剩下的独立义原的配对相似度中，取最大的一对，并归为一组，如此反复，直到所有独立义原都完成分组。
- 3) 关系义原描述式：关系义原描述式的配对分组较为简单，我们把关系义原相同的描述式分为一组，并计算其相似度；
- 4) 符号义原描述式：符号义原描述式的配对分组与关系义原描述式类似，我们把关系符号相同的描述式分为一组，并计算其相似度。
- 5) 在以上 2)、3)、4) 的计算中，最后求加权平均时，各部分取相等的权值。

到此为止，我们已经讨论了基于《知网》的词语相似度计算的所有细节，具体的算法我们不再详细说明。

5 实验及结果

根据以上方法，我们实现了一个基于《知网》的语义相似度计算程序模块。

词语相似度计算的结果评价，最好是放到实际的系统中（如基于实例的机器翻译系统），观察不同的相似度计算方法对实际系统的性能的影响。这需要一个完整的应用系统。在条件不具备的情况下，我们采用了人工判别的方法。

我们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,我们计算一个词和另外任意选取的一组词的相似度,由人来判断这个词和这一组词的相似度大小是否与人的直觉相符合;

第二个实验,我们使用了三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:Li Sujian et al. (2002) 中使用的词语语义相似度计算方法;

方法 3:本文中介绍的语义相似度计算方法;

在实验中,几个参数的取值如下:

$\alpha = 1.6;$

$\beta_1 = 0.5, \beta_2 = 0.2, \beta_3 = 0.17, \beta_4 = 0.13$

$\gamma = 0.2$

$\delta = 0.2$

两个实验结果如下表所示:

词语 1	词语 2	词语 2 的语义	方法 1	方法 2	方法 3
男人	女人	人,家,女	1.000	0.668	0.833
男人	父亲	人,家,男	1.000	1.000	1.000
男人	母亲	人,家,女	1.000	0.668	0.833
男人	和尚	人,宗教,男	1.000	0.668	0.833
男人	经理	人,#职位,官,商	1.000	0.351	0.657
男人	高兴	属性值,境况,福,良	0.016	0.024	0.013
男人	收音机	机器,*传播	0.186	0.008	0.164
男人	鲤鱼	鱼	0.347	0.009	0.208
男人	苹果	水果	0.285	0.004	0.166
男人	工作	事务,\$担任	0.186	0.035	0.164
男人	责任	责任	0.016	0.005	0.010

表 3: 实验结果

考察实验 1 的结果,也就是上面方法 3 的结果,我们可以看到,“男人”和其他各个词的相似度与人的直觉是比较相符合的。

考察实验 2 的结果,也就是将方法 3 和方法 1、方法 2 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 的结果比方法 1 更细腻一些,能够区分不同人之间的相似度,但有些相似度的结果也不太合理,比如“男人”和“工作”的相似度比“男人”和“鲤鱼”的相似度更高。从可替换性来说,这显然不合理,至少“男人”和“鲤鱼”都是有生命物体,而“工作”只可能是一个行为或者一个抽象事物。方法 2 出现这种不合理现象的原因在于其计算方法把部分相关度数值加权计入了相似度中。另外,方法 2 的结果中,“男人”和“和尚”的相似度比“男人”和“经理”的相似度高出近一倍,而方法 3 的结果中,这两个相似度的差距更合理一些。

6 结论

与传统的语义词典不同,《知网》采用了 1500 多个义原,通过一种知识描述语言来对每个概念进行描述。

为了计算用知识描述语言表达的两个概念的语义表达式之间的相似度,我们采用了“整体的相似度等于部分相似度加权平均”的做法。首先将一个整体分解成部分,再将两个整体的各个部分进行组合配对,通过计算每个组合对的相似度的加权平均得到整体的相似度。通过对概念的语义表达式反复使用这一方法,可以将两个语义表达式的整体相似度分解成一些义原对的相似度的组合。对于两个义原的相似度,我们采用根据上下位关系得到语义距离并进行转换的方法。

实验证明,我们的做法充分利用了《知网》中对每个概念进行描述时的丰富的语义信息,得到的结果与人的直觉比较符合,词语相似度值刻划也比较细致。

参考文献：

- Agirre E. and Rigau G. (1995), *A proposal for word sense disambiguation using conceptual distance*, in International Conference "Recent Advances in Natural Language Processing" RANLP'95, Tzigras Chark, Bulgaria,.
- Dagan I., Lee L. and Pereira F. (1999), Similarity-based models of word cooccurrence probabilities, Machine Learning, Special issue on Machine Learning and Natural Language, 1999
- Li Sujian, Zhang Jian, Huang Xiong and Bai Shuo (2002), Semantic Computation in Chinese Question-Answering System, Journal of Computer Science and Technology (Accepted)
- 李涓子(1999), 汉语词义排歧方法研究, 清华大学博士论文
- 王斌(1999), 汉英双语语料库自动对齐研究, 中国科学院计算技术研究所博士学位论文
- 鲁松(2001), 自然语言中词相关性知识无导获取和均衡分类器的构建, 中国科学院计算技术研究所博士论文
- 董振东, 董强 (1999), “知网”, <http://www.keenage.com>
- 杜飞龙 (1999), 《知网》辟蹊径, 共享新天地——董振东先生谈知网与知识共享, 《微电脑世界》杂志, 1999 年第 29 期