文章编号：1003-0077（2011）00-0000-00

基于语义构词的汉语词语语义相似度计算*

康司辰[1]，刘扬[2]

（1. 北京大学中国语言文学系，北京 100871；2. 北京大学计算语言学研究所，北京 100871）

摘要：汉语词语语义相似度计算，在中文信息处理的多种应用中扮演至关重要的角色。基于汉语字本位的思想，我们采用词类、构词结构、语素义等汉语语义构词知识，以“语素概念”为基础，计算汉语词语语义相似度。这种词义知识表示简单、直观、易于拓展，计算模型简洁、易懂，采用了尽可能少的特征和参数。实验表明，本文方法在典型“取样词对”上的表现突出，其数值更符合人类的感性认知，且在全局数据上也表现出了合理的分布规律。

关键词：词语语义相似度计算 语义构词 词义知识表示 语素概念

中图分类号：TP391　　**文献标识码：**A

Semantic Word-formation Based Chinese Word Similarity Computing

Kang Sichen[1], Liu Yang[2]

(1. Department of Chinese Language and Literature, Peking University, Beijing 100871;
2. Institute of Computational Linguistics, Peking University, Beijing 100871)

Abstract: Chinese word similarity computing plays an important role in the application of Chinese information processing. Based on the notion of character-orientation, Chinese semantic word-formation knowledge, including word POS, word-formation pattern and morphemic concepts, is employed to compute Chinese word similarity. This lexical knowledge representation is simple, intuitive and easy to expand and the model is straight-forward, with characteristics and parameters adopted as less as possible. Experimental results show that the approach is promising for the typical sampling word pair. Also, the numerical values of similarity are more in line with human cognition and present a reasonable distribution of the global data.

Key words: Chinese word similarity computing; Chinese semantic word-formation; lexical knowledge representation; morphemic concepts

1 引言

在自然语言处理领域中，词语语义相似度计算长久以来都具有很高的理论和应用价值，对词义消歧、查询识别、机器翻译等应用起着尤为重要的作用。

在此前研究中，汉语词语语义相似度的计算方法可归为两类，一类利用语言知识库中的知识，另一类利用语料中的上下文特征，并依据不同的算法计算相似度。第一类方法[1,2,3,4,5,6,7,8,9]采用的知识包括《同义词词林》、《知网》、知识图、概念图和百度百科等，其方法往往依赖于特定的词义知识表示，可称为基于知识的方法；第二类方法[10,11,12,13,14]对语料进行上下文分析，提取词向量做相似度计算，可称为基于语料的方法。

目前，这两类方法都存在问题：基于知识的方法，大体以理性方法为主，偏重考察“取样词对”语义相似度的合理性，主要通过增加参数、调节公式中的系数等手段，力图提升限定取样数据的计算结果，这导致相似度计算的方法越来越趋于繁琐；基于语料的方法，大体以经验方法为主，主要通过模型选取、特征优化、降噪处理等手段，以获得更理想的全局数据计算结果，其优点是词语的覆盖面广，但在“取样词对”上的表现往往不佳。

基于以上分析，我们希望建立一套新的汉语词义知识表示及词语语义相似度计算方法，并满足如下特征：在词义知识表示方面，符合人类对汉语语言的一般认知，其表达形式也更加直观、有效；此外，建立在该知识表示上的语义相似度计算方法简洁、易懂，能够在“取样词对”上表现优异，同时，在全局数据上也表现出合理的分布规律。

众所周知，汉语语言以字为自然单位，苏宝荣[15]等多位语言学家阐述了汉语的构词结构对词义理解至关重要的观点，这表明从构词结构出发，进而表达词义的手段是可行的；此外，苑春法、黄昌宁[16]的研究也证实“只有很少一部分的语素在构词时意义发生了变化”，而绝大多数词义可由语素义直接导出。结合以上观点，我们认为，以汉语的语义构词（包含构词结构、语素义等知识）作词义知识表示是有可靠的语言学依据的，对词义研究和相关计算有

* **收稿日期：**　　**定稿日期：**

基金项目：国家社科基金一般项目（16BYY137）、国家重点基础研究发展计划资助项目（2014CB340504）、国家社科基金重大项目（12&ZD119）

可能产生重要价值和积极意义。这样一来，语义相似度计算也有了更为直观的知识表示，而其算法有可能趋于简化并表现出好的特性。

2 汉语的语义构词知识表示与获取

凡是对词的理解有意义的构词知识，在中文信息处理应用中都是有用的。因此，本文所讲的构词知识，涵盖词性、构词结构、语素义等，是广义的语义构词知识。我们以这些知识为基础，进行汉语词语语义相似度计算并做评估。

课题组研发多年并计划推出的北京大学《汉语概念词典》（以下简称《概念词典》，英文名称 the Chinese Object-Oriented Lexicon, 简称 COOL）在生成词库理论（GLT 理论）[17]、面向对象思想（OO 思想）[18]、WordNet 理论[19]等观点指导下，以《现代汉语词典（第 5 版）》（以下简称《现汉》）刻画的汉语的语素及语素义为依据，采用“同义语素集”来表征“语素概念”并建立“语素概念体系”；在此基础上，详尽描述汉语词的构词结构，并实现构词结构下的构词成分（即语素）对“语素概念体系”中的“语素概念”的严格绑定，以此来诱导和表达汉语词义，并提供多种应用程序接口。

《概念词典》中包含的这些语义构词知识，构成本文工作的一个数据基础。

2.1 词类知识

《概念词典》为收录的词都标注了词性，其中，51454 个二字词的情况如表 1 所示。

表 1 《概念词典》中二字词词性统计表

词性	数量	比率	例词
名词	25720	49.99%	丈夫
动词	18679	36.30%	上升
形容词	5543	10.77%	严峻
副词	905	1.76%	临时
数词	57	0.11%	好多
量词	90	0.17%	公尺
介词	36	0.07%	为了
代词	114	0.22%	咱们
助词	23	0.04%	不得
叹词	10	0.02%	呜呼
拟声词	115	0.22%	乒乓
连词	162	0.31%	不但
合计	51454	100.00%	

2.2 构词结构知识

在语言学界有两种主流的构词结构体系，一种注重表达构词语素间的语义关系（如主体、客体等），而另一种体系注重表达构词语素间的语法关系（如主谓、述宾等）。相对而言，后一种构词体系更为精简，与句法结构有天然的相似性，相关研究更为成熟，有利于词语相似度计算，本研究采用这种构词体系。实际上，由于在后续，还要求构词成分对“语素概念体系”中的“语素概念”严格绑定，我们获得的依然是广义的语义构词知识。

我们参考杨梅[0]和北京大学中文系郭锐教授对构词结构的研究成果，构建了基于语法的构词体系，并为《概念词典》中所有二字词按义项区分标注了构词结构，共计 52108 个。为保证构词结构知识的可靠性，请三位专家对同一词项进行标注，两人以上标注结果相同的一致率为 93.46%。标注结果的具体情况见表 2。

表 2 《概念词典》二字词构词结构统计表

构词结构	数量	比率	例词
主谓	524	1.01%	年轻
连谓	1709	3.28%	进攻
联合	11414	21.90%	丰满
述宾	8141	15.62%	选材
述补	630	1.21%	提高

定中	19581	37.58%	红旗
状中	4215	8.09%	热爱
介宾	157	0.30%	从小
重叠	310	0.59%	哥哥
名量	78	0.15%	纸张
数量	56	0.11%	一些
方位	189	0.36%	野外
复量	20	0.04%	场次
前附加	698	1.34%	老虎
后附加	2308	4.43%	忘却
单纯词	2078	3.99%	克隆
合计	52108	100.00%	

需要说明的是，该构词体系可以方便地拓展到多字词的情形。以“化学反应”为例，“化学反应”为定中结构，构词成分分别为“化学”、“反应”；“化学”为定中结构，构词成分分别为“化”、“学”；“反应”为后附加结构，构词成分分别为“反”、“应”。

2.3 语素义知识

语言学上的“语素”指的是“最小的音义结合体”，在本文中，为方便起见，汉语语素暂且限定为一个汉字。借鉴 WordNet 理论，课题组成员陆顾婧[21]在其硕士论文中用“语素特征”（现在称其为“语素概念”）来称谓汉语中可计算的最小意义单元，并采用“同义语素集”的形式来加以表示，该集合中的元素为具有相同或基本相同意义（即语素义）的那些语素，其中的每个语素都携有独特的“语素义编码”。例如，语素“选”有多个语素义，其中的一个语素义的“语素义编码”为“选 1_04_01”，这表明：它是该单字在词典中的第 1 次条目出现（即“选 1”），该条目共有 4 个义项（即“选 1_04”），当前为第 1 个义项（即“选 1_04_01”）。

目前，对《现汉》中全部语素所表达的 20175 个语素义，我们按释义计算相似度，形成初步的“同义语素集”，并经反复的人工校对、核对，获得了 5113 个“语素概念”。在这些“语素概念”之间，我们进一步构建了初步的上、下位语义关系，形成了一个树状结构的“语素概念体系”。在后续的知识表示中，如果确定了特定语素的语素义，携有了“语素义编码”，就意味着特定语素在该体系中绑定了一个“语素概念”，并接受该体系的意义表达和约束。

以表达“选择、挑选”意义的动语素“语素概念”X 为例，X={刷 3_01_01，抡 1_01_01，拔 1_08_03，拣 1_01_01，择 1_02_01，择 2_02_01，挑 1_02_01，擢 1_02_02，调 4_02_02，选 1_04_01，遴 1_01_01，铨 1_02_01}，在“语素概念体系”中，其所处的“语素概念”位置如图 1 所示。

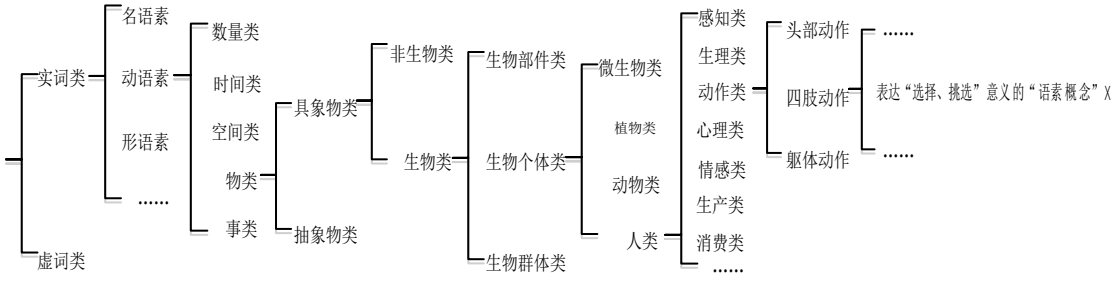

图 1 树状结构的“语素概念体系”示例

在标注《概念词典》中二字词的构词结构后，我们继续为所有二字词的前、后语素标注其在《现汉》中的语素义，并按其语素义与对应的“语素义编码”挂钩。于是，二字词的前、后语素与它们在“语素概念体系”中的“语素概念”就建立了严格的绑定关系。

这样一来，在构词结构因素之下，进一步地，每个语素义拥有更丰富的、便于交流和计算的意义形式。每个语素义携有唯一的“语素义编码”，每个“语素义编码”对应唯一的“语素概念”，每个“语素概念”在“语素概念体系”中有唯一确定的位置。这些位置表达了“语

素概念”间的距离，而词语概念（即词义）之间距离与此相关，这为词语语义相似度的计算带来极大方便。

3 汉语词语语义相似度计算模型

3.1 基本思路

从本质上讲，词语语义相似度是词语概念（即词义）间的距离的描述。在本研究中，汉语词表达的词语概念由语义构词知识表示，包括词类知识、构词结构、语素义知识等三项内容。其中，词类知识(part of speech，简称 POS)代表了词语概念跨词类的惩罚代价，构词结构知识(word-formation pattern，简称 WFP)、语素义知识(morpheme knowledge，简称 MK)则表达了语素义对词语概念（即词义）的贡献情况。

因此，对于词对 A、B，有如下定义：

定义 1：词语概念距离 D(A,B)定义为词对 A、B 的三元关系：D(A,B)=R<POS, WFP ,MK>。

定义 2：词语语义相似度 sim(A，B)定义为词语概念距离 D(A,B)的函数：sim(A，B)= f(D(A,B))。

3.1.1 词性知识的利用

此前，对于词语语义相似度，基于方便考虑，多数研究只考察相同词类的情形，我们将其推广到不同词类上去。我们认为，在词语概念的内涵保持基本不变的情况下，不同词类的词语概念距离应该大于相同词类的词语概念距离。

我们采用距离惩罚方式调整词语概念距离，进而调整词语相似度计算结果。在本研究中，系数约定遵循的一般思路为：实词与虚词之间的惩罚系数相对较高，虚词之间惩罚系数大致相同；实词之中，体词和谓词之间惩罚系数相对较高，而体词与体词之间、谓词与谓词之间的惩罚系数相对较低。

在本文中，动词、名词、形容词等开放词类之间的惩罚系数见表 3。该取值采用经验值，可依应用需求灵活调整。其它词类的情况不再赘述。

表 3 开放词类之间的惩罚系数

词类 1	词类 2	惩罚系数
名词	动词	1.2
动词	形容词	1.1
名词	形容词	1.2

3.1.2 构词结构知识的利用

汉语词的构词结构反映了在不同构词结构下，各语素对于整体词义的不同贡献。例如，在联合结构中，两成分对整体词义的贡献基本相同，而在定中结构中，中心语成分对整体词义的贡献更大一些。我们用贡献系数衡量不同结构下的各语素对于整体词义的不同贡献，在本文中，其取值情况见表 4。该取值采用经验值，可依据应用需求灵活调整。

表 4 贡献系数取值情况

构词结构	前语素贡献系数	后语素贡献系数
联合	0.5	0.5
定中	0.3	0.7
名量	0.4	0.6
数量	0.4	0.6
方位	0.4	0.6
状中	0.3	0.7
连谓	0.5	0.5
复量	0.5	0.5
介宾	0.4	0.6
述宾	0.6	0.4
述补	0.7	0.3
主谓	0.4	0.6
前附加	0.9	0.1
后附加	0.1	0.9
重叠	0.5	0.5

该系数取值同样可拓展至多字词。例如，对于前面分析的多字词“化学反应”，通过迭代过程，即可求得该词中的不同语素的贡献系数，分别为：“化”0.09（即 0.3*0.3）、“学”0.21（即 0.3*0.7）、“反”0.07（即 0.7*0.1）、“应”0.63（即 0.7*0.9）。

3.1.3 语素义知识的利用

在树状结构的“语素概念体系”中，考虑上位概念表达的语义颗粒度大于下位概念的因素，在计算时，本文采取边加权的方式计算路径长度。目前，“语素概念体系”的最大深度为 10 层，约定根节点下的边为第 1 层，我们对层数为 c 的边的权值 w 设定如下：w=1.0+(10-c)*0.1，各层的边的权值如表 5 所示。该取值采用经验值，可依应用需求灵活调整。

表 5 各层的边的权值设定

层数	权值	层数	权值
1	1.9	6	1.4
2	1.8	7	1.3
3	1.7	8	1.2
4	1.6	9	1.1
5	1.5	10	1.0

此外，在语素义未明确指定的情况下，语素存在多义性，在《概念词典》中有多个“语素概念”与之绑定，相应的，有多个位置与之对应。在计算语义相似度时，按照惯例原则，我们取能够促成“语素概念”a、b 之间保持最短距离的位置 P_a、P_b。

3.2 语义相似度算法描述

形式上，设二字词 $A=a_1a_2$，二字词 $B=b_1b_2$，则词对 A、B 的词语语义相似度计算方法如下所述：

1、计算语素对 a_i、b_j 之间的语素概念距离 $d(a_i,b_j)$

对于语素集合 $E=\{a_1,a_2,b_1,b_2\}$ 中的语素 e，在《概念词典》中取该语素的不同语素义对应的所有“语素概念”，这些“语素概念”在“语素概念体系”中的全部位置构成位置集合 $P_e=\{p_{e1},p_{e2},\cdots\cdots,p_{em}\mid e\in E\}$，其中，m 是语素 e 在“语素概念体系”中对应的“语素概念”个数。在该表示下，语素对 a_i、b_j 之间的语素概念距离 $d(a_i,b_j)$，即为语素 a_i 的位置集合 P_{ai} 和语素 b_j 的位置集合 P_{bj} 之间构成的多条路径中的最短路径的路径长度 $|V<P_{ai},\cdots\cdots,P_{bj}>|_{min}$。

简而言之，语素概念距离 $d(a_i,b_j)=|V<P_{ai},\cdots\cdots,P_{bj}>|_{min}=|V<P_{ais},\cdots\cdots,P_{bjt}>|$，其中 P_{ais}、P_{bjt} 为能够促成“语素概念”a_i、b_j 之间保持最短距离的位置。

2、构造词对 A、B 之间的贡献系数集合 $C=\{c_{11},c_{12},c_{21},c_{22}\}$

记词 A 的构词结构类型为 $S_A=<m_{a1},m_{a2}>$、词 B 的构词结构类型为 $S_b=<m_{b1},m_{b2}>$，m_{a1}、m_{a2} 为 S_A 结构下的前、后语素贡献系数，m_{b1}、m_{b2} 为 S_B 结构下的前、后语素贡献系数，它们的取值见表 4 中的约定。在该表示下，$c_{11}=m_{a1}*m_{b1}$，$c_{12}=m_{a1}*m_{b2}$，$c_{21}=m_{a2}*m_{b1}$，$c_{22}=m_{a2}*m_{b2}$，这些取值下的 c_{11}、c_{12}、c_{21}、c_{22} 构成集合 $C=\{c_{11},c_{12},c_{21},c_{22}\}$。

3、计算词对 A、B 之间的词语概念距离 D(A,B)

原则上，D(A,B)是 $d(a_i,b_j)$、C、α 等参数的函数，即 $D(A,B)=f(d(a_i,b_j),C,\alpha)$，其中，$d(a_i,b_j)$由步骤 1 得到，C 由步骤 2 得到，α 为词类惩罚系数，见表 3 中的约定。

在本文中，$f(d(a_i,b_j),C,\alpha)$采用如下函数计算：

$$D(A,B)=f(d(a_i,b_j),C,\alpha)=\alpha*\sum_{i=1,\ 2}d(a_i,b_j)*C_{ij}$$

4、计算词对 A、B 的语义相似度 Sim(A,B)

考虑 D(A,B)的分布特性，约定 $Sim(A,B)=f(d(A,B))=\begin{cases}1, & d(A,B)=0\\ \frac{1}{1+e^{(a*d(A,B)+c)}}, & d(A,B)>0\end{cases}$

其中，a用于调整函数的整体趋势，c用于调整函数的对称中心，本文取a=0.5，c=-15。

需要说明的是：在本计算模型中，词语概念距离转化为语义相似度的公式采用 logistic 曲线。其原因在于，词语概念距离在整体上基本满足正态分布，考虑数据稠密程度，logistic 曲线能使距离分布密集区的函数取值得到平滑。

此外，本计算模型具有一般性，可以方便地拓展至汉语 n 字词的计算。对于多字语 A、B，记语素集合 E={a_1,……,a_n,b_1,……,b_n}，而贡献系数集合 C={c_{11},c_{12},……,c_{1n},……,c_{ij},……,c_{nn}}可由二字词贡献系数迭代得到，再依照 D(A,B)=f(d(a_i,b_j),C, α) 计算 Sim(A,B)=f($D(A,B)$)。

4 实验结果与数据分析

4.1 关于评价方法的讨论

之前的研究与评价标准，往往倾向于挑选一些同类词的“取样词对”，我们认为这不具有随机性，也缺乏客观性的，相似度计算的需求可以存在于任意词对之间，与是否属于同类词无关。

此外，对于汉语词语语义相似度计算方法的评价，实际上应包含两个部分，即语义相似度取值在局部数据上的表现优劣,以及，语义相似度在全局数据上的分布规律是否合理。只有这两部分均表现优异的方法，才能在实际应用中获得有效采用。对于特定方法，如果只满足于“取样词对”上的相似度结果优良，而不满足全局数据上的分布合理，可以认为该方法存在对“取样词对”过拟合的倾向；反之，如果保持了全局数据上的分布规律，而在“取样词对”上的计算结果欠佳，可以认为该方法不具有典型性和精确性，同样不足为取。

4.2 与基于知识的其它方法比较

基于《知网》计算汉语词语语义相似度的研究很多，往往能达到局部最优，正如刘杰[7]所言，这类方法使得部分词汇的相似度更为合适，从而符合人们的主观判断。

对于该类方法，我们选取刘群、刘素建（2002）[10]最早基于《知网 2000》的计算结果（方法 1、方法 2），以及最近刘杰[7]分别基于刘群、李素建方法的计算结果（方法 3、方法 4），刘杰另外给出了基于《知网 2008》的李素建、刘群计算结果（方法 5、方法 6），方法 7 为本文所用方法的计算结果。这些计算结果的比较见表 5，其中，表中 Null 代表未能获得相关数据。

表 5“取样词对”相似度比较

词语 1	词语 2	方法 1	方法 2	方法 3	方法 4	方法 5	方法 6	方法 7
男人	女人	0.668	0.833	0.861	0.910	0.684	0.692	0.716
男人	父亲	1.000	1.000	1.000	0.899	0.646	0.654	0.701
男人	母亲	0.668	0.833	0.861	0.890	0.569	0.576	0.701
男人	苹果	0.004	0.166	0.171	0.470	0.029	0.067	0.442
男人	责任	0.005	0.001	0.126	0.283	0.021	0.040	0.281
男人	高兴	0.024	0.013	Null	Null	Null	Null	0.096
旅程	旅行	Null	0.074	0.000	0.090	0.585	0.737	0.073
战争	打仗	Null	0.040	0.000	0.225	0.552	0.732	0.746
爱情	恋爱	Null	0.044	0.000	0.700	0.450	0.737	0.494
十分	特别	Null	0.624	0.624	0.750	0.044	0.750	0.807
灵敏	敏捷	Null	0.881	0.881	0.400	0.021	0.400	0.782
美丽	动人	Null	1.000	1.000	0.500	0.029	0.500	0.045

不难发现，在“男人”与其它词的相似度计算中，无论哪种方法，都在“人类”和“非人类”之间的关系上表现良好，“男人”和“女人、父亲、母亲”的相似度高，而和“苹果、责任、高兴”的相似度低。但是，对于“非人类”的“苹果、责任、高兴”，由“生物”和“非生物特征”来看，“男人”和“苹果”应该更近一些，在本文方法中，该特征得以体现。对于“旅行”和“旅程”，我们认为体现得更多的是相关性，而不是相似性，所以在跨词类的惩罚系数下，该相似度得以降低。本文方法对“美丽”和“动人”处理不好，这是因为“动人”在语义构词中发生了意义转变，对于这种情况，本文方法目前不做进一步的处理。但如苑春法、黄昌宁[16]所言，汉语中的这种情况极少，所占比例少于 2%，我们在构词结构标注中采取了较严格的方案，发现这类词占比为 4%。此外，本文方法在“战争、打仗”、“十分、特别”、“灵敏、敏捷”等词对上的表现突出，计算结果优于其余方法。

此外，我们注意到，受《知网》数据限制，一些词语的相似度无论如何调整算法，都是无法优化的。比如，对于具有相同概念定义的词语，如“成败、成效、得失、功利、功效、胜败、胜负、输赢、损益、盈亏”等词语具有相同定义:“attribute|属性,effect|效用,&event|

事件”，则它们之间的相似度只能为 1，但是其中“功效、胜负、盈亏”等词语在感觉上是不应该相似的。这是用《知网》进行相似度计算需要解决的一个问题，其它不再赘述。

4.3 与基于语料的方法比较

在基于语料的方法中，我们采用时间最近、效果较好的王石方法[11]进行对比，该方法覆盖所有词语，并且对较大的词表进行了评估。王石方法的相似度取值范围是{-1}∪[0,1]，对于“-1”取值情形，文献未给解释。王石对词语相似度做了 4 次迭代计算，我们取效果最好的第 2 次迭代结果。由于这类方法相似度取值普遍偏低，我们只能从相似度相对序的角度来进行分析。名词词对相似度比较的情况如下：

表 6 名词词对相似度比较

词语 1	词语 2	本文方法	王石方法
宝石	珠宝	0.815	0.367
珠宝	正午	0.353	0.007
正午	中午	0.936	0.383
男人	母亲	0.701	0.165
男人	工作	0.380	0.031
森林	林地	0.501	0.177
苹果	香蕉	0.788	0.255
森林	手机	0.688	0.105
医院	诊所	0.624	0.191
汽车	轿车	0.977	0.383
汽车	飞机	0.814	0.205
汽车	医院	0.459	0.195
手机	电话	0.830	0.370
电话	电视	0.410	0.167
椅子	凳子	0.972	0.453
房子	桌子	0.880	0.130
电影	邮票	0.844	0.072

在名词词对相似度相对序上，王石方法相似度高的词对从高到低为：“椅子、凳子”、“汽车、轿车”、“正午、中午”，本文方法相似度高的词对从高到低为：“汽车、轿车”、“椅子、凳子”、“正午、中午”，结果基本一致，对于词对“椅子、凳子”和“汽车、轿车”相似度高低的判断，不同人有不同理解。王石方法相似度较低的词对从低到高为：“珠宝、正午”、“男人、工作”、“电影、邮票”，本文方法相似度低的词对从低到高为：“珠宝、正午”、“男人、工作”、“电话、电视”，结果基本一致。

表 7 动词词对相似度比较

词语 1	词语 2	本文方法	王石方法
抚摸	触摸	0.968	-1.000
鞠躬	微笑	0.883	-1.000
抚摸	担心	0.586	-1.000
忧虑	担心	0.112	0.233
担心	放心	0.592	0.179
鞠躬	听见	0.962	-1.000
体会	感觉	0.848	0.215
购买	销售	0.987	0.162
考虑	思考	0.980	0.285
思考	问候	0.963	-1.000
发明	创造	0.905	0.017
停留	运动	0.955	0.000
衰老	告诉	0.006	0.000

在动词词对上，本文方法优于王石方法。王石方法中的很多动词词对缺乏有效的取值，本文方法不存在这类问题。

表 8 形容词词对相似度比较

词语 1	词语 2	本文方法	王石方法
聪明	寒冷	0.715	0.012
聪明	机智	0.968	0.146
高兴	粉红	0.064	0.000
高尚	陡峭	0.472	0.009
高兴	开心	0.167	0.038
炎热	干燥	0.867	0.114
初级	基础	0.240	0.110
初级	高级	0.742	0.135
陡峭	崎岖	0.942	-1.000
崎岖	平坦	0.901	-1.000

在形容词词对上，对于“聪明、机智”、“炎热、干燥”、“初级、高级”、“陡峭、崎岖”、“崎岖、平坦”等词对，本文方法占优，其余词对大致持平。

此外，百度 CW 算法[14]和王石算法有类似问题，这里不再赘述。

实际上，基于语料的相似度计算方法，其相似度取值普遍偏低，在相似度数值的合理性方面，本文方法更优。此外，基于语料方法的特征提取依赖上下文环境，而在上下文中出现的词语体现的不一定是相似性，有可能是相关性，这会造成较大的干扰。

4.4 关于语义相似度分布的讨论

语义相似度分布体现特定模型在全局数据上的分布合理性。目前，《概念词典》中有 52108 个二字词，它们之间词对组合的数量达到了 $2.72*10^9$。考虑计算代价问题，我们对二字词采取十分之一随机抽样，该取样并不影响整体分布。本文方法对 5211*5211 个词对的相似度计算结果满足正态分布，即对于整体的汉语词语，可以表述为“特别相似”或“特别不相似”的情形相对较少。

百度 CW 算法基于词向量计算语义相似度，利用百度公司 NLPC 小组提供的计算工具，我们也得到了该 5211*5211 个词对的相似度计算结果，同样满足正态分布。

实验表明，上述两组数据在置信度 95%区间上进行正态分布拟合，R 方值达到 0.9 以上，具有很强的说服力。这种情况也符合人类对于语言的一般认知。

5 结语

在汉语词语语义相似度计算领域，因为知识表示欠缺、数据匮乏等原因，完全采用语义构词知识的方法前人还未曾实践过。

基于汉语字本位的思想，我们尝试采用词类、构词结构、语素义等汉语语义构词知识，以“语素概念”为基础，并结合其在“语素概念体系”上的意义表达和约束，借助这些密集的构词知识来计算语义相似度，该词义知识表示具有简单、直观、易于拓展等优良特性。

建立在这种词义知识表示上的相似度计算模型简洁、易懂，在算法中采用了尽可能少的特征和参数，实验表明，其在典型“取样词对”上的表现突出，相似度数值更符合人类的直观感觉，且在全局数据上也表现出合理的分布规律。

当然，本文方法还存在一些不尽人意的地方。比如，汉语单纯词的语义与语素义无关，部分合成词存在转义、隐喻等现象，这些问题目前尚没有加以考虑和处理，虽然它们在所有词中占比不高；此外，词语概念距离如何转化为语义相似度，如何选取更合适的函数模型，技术细节也还有待探索和深入。

后续要开展的工作包括“语素概念体系”的修订完善、多字词构词结构和意义标注、以及语义相似度算法的优化等，以进一步提高相似度计算的准确率和覆盖面，并将其应用于实际的应用系统。

最后，感谢北京大学中文系郭锐教授对汉语构词结构工作的指导，感谢百度公司 NLPC 团队对相似度计算研究的大力支持和 KRR 小组关于相似度应用实用性问题的启发。

参考文献

[1]张亮,尹存燕,陈家骏.基于语义树的中文词语相似度计算与分析[J].中文信息学报,2010,06:23-30.

[2]李峰,李芳.中文词语语义相似度计算——基于《知网》2000[J].中文信息学报,2007,03:99-105.

[3]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,2008,05:84-89.
[4]张瑞霞,朱贵良,杨国增.基于知识图的汉语词汇语义相似度计算[J].中文信息学报,2009,03:116-120.
[5]王小林,王东,杨思春,等.基于《知网》的词语语义相似度算法[J].计算机工程,2014,12:177-181.
[6]张沪寅,刘道波,温春艳.基于《知网》的词语语义相似度改进算法研究[J].计算机工程,2015,02:151-156.
[7]刘杰,郭宇,汤世平,等.基于《知网》2008 的词语相似度计算[J].小型微型计算机系统,2015,08:1728-1733.
[8]何夏燕.基于汉语概念图的词汇语义相似度计算[D].上海交通大学,2010.
[9]詹志建,梁丽娜,杨小平.基于百度百科的词语相似度计算[J].计算机科学,2013,06:199-202.
[10]刘群,李素建.基于《知网》的词汇语义相似度计算[C].第三届汉语词汇语义研讨会,台北,2002.
[11]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,2013,01:7-14.
[12]蔡东风,白宇,于水,等.一种基于语境的词语相似度计算方法[J].中文信息学报,2010,03:24-28.
[13]关毅,王晓龙.基于语料的汉语词汇间语义相似度计算[A].语言计算与基于内容的文本处理——全国第七届计算语言学联合学术会议论文集[C],2003:7.
[14]Ronan Collobert, Jason Weston, L´eon Bottou, Michael Karlen, etc. Natural Language Processing (Almost) from Scratch. Journal of Machine Learning Research, 12 (2011) 2493-2537.
[15]苏宝荣.汉语复合词结构义对构词语素意义的影响[J].语文研究,2013,01:1-4.
[16]苑春法,黄昌宁.基于语素数据库的汉语语素及构词研究[J].语言文字应用,1998,03:86-91.
[17]Pustejovsky, J. The Generative Lexicon[M]. Mass: MIT Press,1994.
[18]Grady Booch, Robert A. Maksimchuk, Michael W. Engle, etc. Object-Oriented Analysis and Design with Applications, 3rd Edition[M]. Addison-Wesley Professional,2007.
[19]Fellbaum C. WordNet: An Electronic Lexical Database [M].Mass: MIT Press,1998.
[20]杨梅.现代汉语合成词构词研究[D].南京师范大学,2006.
[21]陆顾婧.汉语构词分析与词义知识表示研究[D].北京大学,2013.

作者简介：

康司辰（1993-），男，本科生，主要研究领域为应用语言学、语言知识工程、中文信息处理。
Email: 1008_frank@sina.com

刘扬（1971-），男，博士，副教授，主要研究领域为语言知识工程、中文信息处理。
Email: liuyang@pku.edu.cn

联系方式：

康司辰 北京大学中国语言文学系 北京 100871 18601110655 1008_frank@sina.com
刘扬 北京大学计算语言学研究所 北京 100871 13021117630 liuyang@pku.edu.cn