# 词距离的计算方法

鲁松　白硕
（ 中国科学院计算技术研究所软件研究室　100080 ）
Email: lusong@ict.ac.cn

**摘要：** 无监督地构建以词距离知识形式表示的自然语言词语相关性知识库是本文的研究目标。作为一种定量化的知识表示方式，词间距离可以为统计方法数据稀疏的平滑和基于相似性计算自然语言处理和信息检索等定量方法提供一个基础性的支持。基于向量空间模型，本文将词语依据词语上下文映射到向量空间中，经过“上下文词语相关性分析权重”方法和主成分分析的降维和消除噪音后，在保持词语相对距离关系不变的情况下，进行 2 维直观显示验证；在计算词间欧氏距离构建词间相关性知识库后，将其引入 Memory-Based Learning 的属性值距离计算，针对汉语词义消歧进行实验，其效果得到初步验证。

**关键词：** 词相关性知识，词距离计算，噪音，噪音消除

# To Calculate the distance between words

LU Song　BAI Shuo
Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080
Email: lusong@ict.ac.cn

**Abstract**: The goal in this paper is to construct Knowledge Base for Relevance between words, depending on their distance, without human labor. As one kind of word knowledge, the distance between words can greatly support lots of application, such as smoothing in statistic method, problems based on word similarity measure in Natural Language Processing, Information Retrieval. This paper formalizes the words into high-dimensions vector space as vectors, clear up the noise with calculating the weight for every feature, Principle Component Analysis and word sense disambiguation, and at last calculates the Euclidean distance between them. The effectiveness of them is tested in 2-dimension visualization and Memory-Based Learning for Chinese Word Sense Disambiguation.

**Keywords**: knowledge of relevance between words, calculating the distance between words , noise, clearing up the noise

## 1. 背景介绍.

自然语言处理中，词语通常作为最基本的语言处理单位。因此围绕着词语知识(word knowledge)的表示、获取和应用，进行了大量极有意义的工作。依据知识获取方式，分为[1] 手工获取，如：中文的《同义词词林》(梅家驹,1983)、英文的 WordNet (Fellbaum, 1998)和中文的 HowNet (董,1999)；以及[2] 机器获取，主要有基于上下文语境获取动词类内划分的逻辑聚类方法(白硕,1995)和凭借机器可读词典(Machine Readable Dictionary)自动获取词语知识的 MindNet (Richardson, 1998)，根据适用范围的不同，可分为面向特定问题的统计方法(如：n-gram)和面向通用领域、基于词距离的词语知识表示方法(Zavrel,1996; Dagan,1999)。

尽管定量表示只能反映词语间简单的相关性关系，无法涉及定性方法表示出来的深层信息和语义关系，但定量方法易于计算和操作的优点是现有定性方法无法比拟的，而且统计方法在语音识别(Lawrence,1989)和自然语言文本处理中(Doug,1992)的成功在一定程度上反映了自然语言个性知识定量方法的优势。

本文的研究对象是词距离的计算方法，即词距离作为词语知识的机器获取方法，目标是采用定量化方法构建一个以词间距离为知识表示形式的通用型词语相关性知识库。

问题是复杂的，但思想是简单的。不仅要形式化词语之间相关与否，而且还要用距离在 0 和 1 之间连续区域取值对词语间相关程度进行形式化。如此的距离表示词语相关性知识正是本文的目标。

词汇空间(lexical space)(Zavrel,1996)借鉴词空间(word space)(schutze,1993)的向量空间模型(Vector Space Model)提出了词间 cosine 距离表示词相似度的方法，但仅用上下文词频来向量化词语是极为粗糙的；另一类方法是以信息论为基础的统计方法获取词距离，并将其用于平滑数据稀疏问题 (Dagan,1999)和进行汉语未登陆词词类的自动判定(孙茂松,2000)，并在各自应用问题中都取得较好的实验结果，但都没有提升到构建词语通用知识库的层次上来讨论问题。此外，自然语言本身固有的灵活性，在形式化过程中可以被认定为是难以处置的“噪音”，对此，上述两种方法均未涉及。

本文同样借鉴了源于信息检索领域(Information Retrieval)的向量空间模型，依据上下文语境将词语映射到高维空间中，经过“上下文词语相关性分析权重”方法和主成分分析(PCA)方法极为有效地去除噪音和降维之后，求解词语间欧氏距离，最后在保持词语间相对距离关系不变的情况下，计算显示 95 个汉语词的 2 维显示图，并基于最近词距离对(k-NN word pair)给出在汉语多义词消歧中的验证实验结果。

本文以下章节安排如下：在 2.1.节给出词语的向量化方法，2.2 节定义了词距离计算过程中的噪音类别，有针对性的给出了去噪和降维的详细描述；2.3 节给出了选择距离计算的方法；在 2.4 节给出了词距离计算的整个框架和计算步骤；3.1.节的 2 维显示验证；3.2 节的 Memory-Based Learning 属性值距离改造；4 节给出结论和讨论。

## 2. 词距离计算方法.

词的相似性计算必须要解决的问题依次为词的形式化表示、噪音消除和距离的计算。本小节将顺沿这样的路线进行介绍。

### 2.1. 词语的形式化定义

数据和知识的表示问题始终是人工智能中的关键问题之一。词语如何形式化？如何实现可计算性是本节讨论的重点。

沿用信息检索领域中经典的向量空间模型(Vector Space Model)(Salton, 1968)，我们在认同词语的上下文(context)可以为词语定义提供足够信息这一假设的情况(朱德熙,1985;Miller and Charles,1991;白硕,1995)下，定义了核心词词矩阵(Word Matrix)概念，并在核心词词矩阵和信息检索中文档做了个等价的类比，即：核心词词矩阵集合倒排表(Inverted Index)中的一个上下文词语作为高维向量空间中的一维，形式如下：

$$FW_i = < w_{term(1)}, w_{term(2)}, \cdots, w_{term(k)}, \cdots, w_{term(n)} > \quad (1)$$

向量中每一分量为上下文词语 $term_{(k)}$ 在核心词词矩阵 $FW_i$ 中的权重值计算方法 tf.idf(term frequency, inverse document frequency)，计算公式如下：

$$w_{ik} = \frac{tf_{ik} \log(N/n_k + 0.01)}{\sqrt{\sum_{k=1}^{t} (tf_{ik})^2 [\log(N/n_k + 0.01)]^2}} \quad (2)$$

公式(2) 中，$w_{ik}$ ：<上下文词语 $term_{(k)}$> 在<$FW_i$ 词矩阵>中的权重；$tf_{ik}$ ：<上下文词语 $term_{(k)}$>在< $FW_i$ 词矩阵>中出现的频率；$log(N/n_k+0.01)$ ：<上下文词语 $term_{(k)}$>在所有

核心词矩阵中分布情况的量化；N ：核心词数目；$n_k$ ：上下文中出现过<上下文词语 $term_{(k)}$>的核心词数目。

## 2.2. 降维与噪音

### 2.2.1. 噪音的问题

自然语言的灵活性，诸如同义词、多义词或小概率词语等问题，在非此即彼的离散符号空间中，必然给上下文统计性分布差异的准确计算带来无法忽视的扭曲。这就是噪音。

应更明确指出的是，在词距离计算中，噪音出现的唯一原因就是自然语言的灵活性，更确切的讲，就是词语上下文环境的灵活性。这是词语知识无监督学习，其中包括：词类划分、词语聚类难以回避的问题。在本文中，噪音被定义为以下 3 点：

**[噪音 1]** 相关性分析无关词噪音：针对词语的相关性分析，对信息获取和词语甄别无效的词语。可以人工建立停用词表或采用模式识别中经典的特征选择方法确定对甄别核心词最为有效的上下文词语。但手工建立停用词耗时耗力，且标准模糊。

**[噪音 2]** 同义/近义词噪音：上下文中的同义和近义词现象，会使相关核心词语倾向于无关。因此，本文也将同义或近义词噪音称为相关性噪音。

**[噪音 3]** 多义词噪音：由于上下文中的多义词现象，使不同义项刻画的核心词语在这一特征上倾向于相关。作为自然语言处理中的老牌儿问题，多义词消歧是解决多义词问题最为直接和有效的方法。

鉴于多义词消歧已成为一个专门的研究课题，本文仅涉及相关性分析无关词噪音和同义/近义词噪音的消除问题。

### 2.2.2. 降维与噪音的消除.

**[1]** 降维与相关性分析无关词噪音的消除

统计数据揭示了这样一个现象：汉语中的绝大部分内容可以被词集中为数不多的高频词所覆盖。在依靠上下文语境定义的核心词词矩阵时，由于这一现象，给词距离计算带来了 2 个严重问题：[1] 数据稀疏问题，即如果不考虑手工建立停用词表，[公式 1]中的绝大多数分量均会为 0；[2]计算复杂度问题，庞大的上下文词语带来数量巨大的维数，给计算的时间和空间复杂度带来了巨大的负担。

在信息检索领域向量空间模型中基于 Zipf's Law 的词语权重计算假设为：最高频和最低频的词语对甄别文档无用。这个假设同样适用于通过词距离计算获取词的相关关系，即在核心词语上下文矩阵中，分布频率最高和最低的上下文词语对考察词间相关关系价值不大。由此，为满足核心词语间相关性甄别要求，本文给出确定“上下文词语相关性分析权重”的统一标准，公式如下：

$$W(context\_word) = \log_{10}(n) \times IG(context\_word) \quad (3)$$

[公式 3]中：

W(*context_word*) ：上下文词语 *context_word* 的“上下文词语相关性分析权重”，即 context _word 对考察词间相关关系的价值大小；

$\log_{10}(n)$ ：为摒弃分布频率最低的上下文词语对 W(*context_word*)影响过大而添加的调整因子，其取值范围为 $[0, \log_{10}(m)]$，$m$ 为核心词语数量；其中 n 为核心词词矩阵

中出现过该 *context_word* 的核心词数量，取以 10 为底对数的目的是降低该调整因子的增长速度。

IG(*context_word*)：用信息增益(Information Gain = IG)概念指示该 *context_word* 在甄别核心词语中的分类能力，其中数学解释如下：

$$IG(context\_word) = H(FW) - H(FW|context\_word) \quad (4)$$

$$H(FW) = -\sum_{fw \in FW} P(fw) \times \log_2 P(fw) \quad (5)$$

$$P(fw) = \frac{|wordset(fw)|}{\sum_i |wordset(fw_i)|} \quad (6)$$

$$H(FW|context\_word) = -\sum_{fw \in FW} P(fw|context\_word) \times \log_2 P(fw|context\_word) \quad (7)$$

所有上下文词语按照“上下文词语相关性分析权重”从大到小进行排序，选择 W(*context_word*)大于阈值的上下文词语 *context_word* 作为特征，完成特征选择。在降低特征维数的同时，实现相关性分析无关词噪音的消除。

**[2]** 同义词、近义词噪音的消除及相关性分析无关词噪音的进一步消除

特征选择得到的特征组合不一定是最佳的组合(王碧泉,1989)。特征之间相关性的存在是这一问题出现的最主要原因。在本文词距离计算中，由于自然语言中存在大量应用的同义词和近义词，在离散值判定仅有相同和不同的二值判定条件下，使这个问题已经成为词相关性获取中最为严重的噪音之一。

本文应用线性映射方法中的主成分分析(Principle Component Analysis = PCA)来完成相关性噪音的消除。

由[公式 1]可以定义一个 $m$ 维核心词，维数为经过特征选择上下文词语个数为 $n$ 的数值矩阵 $X_{n\times m}$，其中矩阵 $X_{n\times m}$ 中每一列向量为一个核心词向量。PCA 的目标是找到一个矩阵 $U_{n\times k}$，使矩阵 $X_{n\times m}$ 经过矩阵 $U_{n\times k}$ 线性变换后得到新的矩阵 $Y_{k\times m}$：

$$Y_{k\times m} = U_{n\times k}' \times X_{n\times m} \quad (8)$$

通过[公式 8]的线性变换，矩阵 $X_{n\times m}$ 中的核心词语从原来的 $n$ 维空间被映射到新的 $k$ 维空间中，并且在新的特征空间中保证 $Y_{k\times m}$ 的特征协方差矩阵 $S_Y$ 为对角阵，即在新空间中，各特征之间的协方差均为 0。协方差为 0 意味着新空间中特征之间已无相关性。可见，PCA 可以达到消除相关性的目的。

另一方面，利用 PCA 中“累计方差贡献率”的概念可以保证得到 k 个方差最大，即信息量最大的新特征集合，在降维的同时，可以极为有效的消除相关性分析无关词带来的噪音。在本文词距离计算中，“累计方差贡献率”被经验地定义为 45%，详细解释见本文 3.1 节。

又是空间复杂度问题！PCA 计算的第一步是求解矩阵 $X_{n\times m}$ 中特征的协方差矩阵，但由于特征数量巨大，即使经过特征选择，仍有 $n \geq 10^4$，正如上文所述，通常情况下，硬件仍难以承受。由于 $n >> m$，必须使用对应分析，通过矩阵 $X_{n\times m}$ 中核心词语间的协方差矩阵来求矩阵 $U_{n\times k}$，限于篇幅，详细描述参见文献(王碧泉，1989)。

**[3]** 小结

“上下文词语相关性分析权重”和 PCA 两种方法都同时承担了 2 个任务：去噪和降维，只不过二者的侧重点不同：“上下文词语相关性分析权重”方法是降维为主，去噪为辅；PCA 方法是去噪为主，降维为辅。在其中，PCA 去噪的作用最大，效果也最明显，这在 3.1 节的词语 2 维平面显示实验中可以得到有力的验证。

## 2.3. 距离的选择.

在词的向量化过程中，不存在量纲问题；经过噪音的有效消除，特征间不存在相关性引起的信息重叠问题；而且由于 euclidean 距离计算方法简便，所以本文选用了 euclidean 距离方法，向量 $X=<x_1,x_2,\ldots,x_n>$和向量 $Y=<y_1,y_2,\ldots,y_n>$的 euclidean 距离公式形式如下：

$$\Delta(X,Y)=\sqrt{\sum_{k=1}^{n}(x_k-y_k)^2} \tag{9}$$

基于向量空间模型的 euclidean 测度满足距离在数学中的经典性质：正定，对称和三角不等式。

## 2.4. 词距离计算步骤.

至此，可以给出词距离计算整体框架：

[1] 确定核心词词集；

[2] 从 97 年一年《人民日报》分词语料中提取核心词的词矩阵；

[3] tf.idf 方法(2.2.1 节)计算每个核心词词矩阵中每个上下文词语的权重，将核心词映射到向量空间中，完成核心词的形式化工作；

[4] 应用“上下文词语相关性分析权重”方法降维和消除相关性分析无关词语噪音；

[5] 应用 PCA 方法消除词相关性和相关性分析无关词语 2 种噪音；

[6] 计算核心词集中两两核心词间，经过归一化的 euclidean 距离，构成两两词间距离的对称矩阵；

[7] 在两两词间距离的对称矩阵中，为每个核心词提取距离最小的 10 个词对(10-NN word paris)。

核心词集中并不是所有两两核心词都是语义详尽或相关的，此时距离经归一化后应为 1，但正是因为归一化和微小共性的存在，必然使大量无关核心词之间距离小于 1。在核心词词集为汉语词语全集的前提下，不会出现问题，但由于计算复杂度的原因，这个前提无法满足。所以，必须要采取步骤[7]，限定保留距离最小的 10 个词对。举例说明：在核心词词集中存在“烟”和“精神”两个词，词集外有“香烟”，若无步骤[7]，势必造成：

$$\delta(\text{“烟”},\text{“精神”})<\delta(\text{“烟”},\text{“香烟”})=1$$

显然，上式是不合理的。这一问题的严重性在实验中得到了验证。

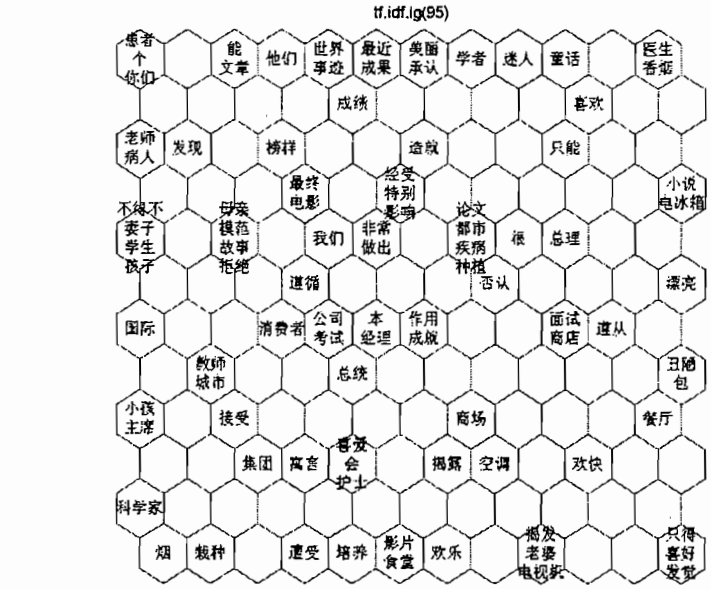
## 3. 验证实验.

为考察词距离计算，即词语相关性知识获取的效果，本文给出了 2 个验证实验：[1] 对高维空间的词向量进行 2 维显示，对比噪音消除前后直觉上的效果；[2] 针对汉语多义词消歧问题，在 Memory-Based Learning 的词语属性值距离计算中引入词语相关性知识库中的词距离，考察词距离的引入是否有价值，以及价值的大小。

### 3.1. 2 维显示验证.

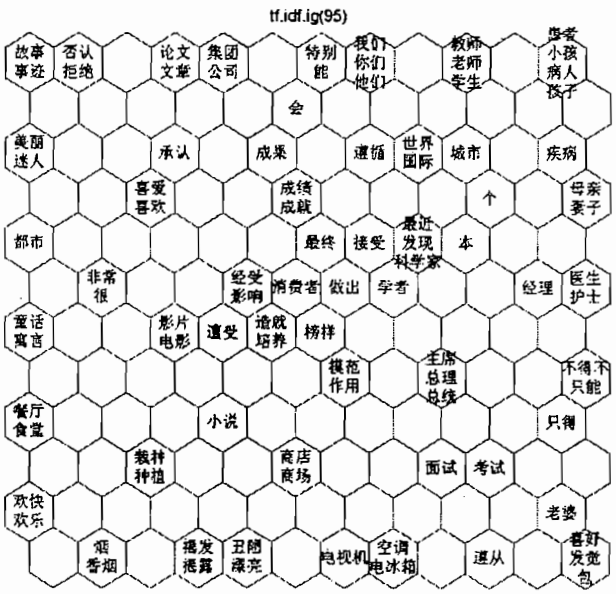
为了形象说明噪音的消除效果，需要保持词语在高维空间中相互距离关系不变的情况下，给出高维空间中词向量的一个直观反映。

本文应用了可进行高维数据 2 维显示的一种神经网络方法、自组织映射(Self-Organizing Maps = SOM)( Kohonen,1995)。下面给出噪音消除前后的 2 维 SOM 显示图，为了突出说明相似性效果，95 个核心词语是经过挑选的。挑选原则是 95 个核心词语集合中存在语义相同或相近的词语，用以反映词语的相对距离关系。




word visualization(PCA 100%) by Lu,Song

[图 1] 噪音消除前的 2 维 SOM 显示图




word visualization(PCA 45%) by Lu,Song

[图 2] 噪音消除后的 2 维 SOM 显示图

[图 2]的 2 维显示图是在 PCA 方法中特征协方差矩阵特征值由大到小排序后，“累计方差贡献率”达到 45%的词语显示结果。

为了考察“累计方法贡献率”的设定对噪音消除效果的影响，我们将“累计方差贡献率”从 95%开始进行 9 次实验，每次将其值递减 10%。“累计方法贡献率”设定为 45%，直观效果是令人振奋的！

无疑，45%是一个经验值。依据 2 维 SOM 显示图，我们主观经验地假设“累计方差贡献率”达到 45%时，在保留足够词语信息的情况下，可以极为有效地消除相关性噪音和相关性分析无关词噪音。这个假设，可以在[图 2]中得到直观验证。

通常，PCA 在模式识别中“累计方差贡献率”定为 85%，即丢失 15%信息量情况下去除噪音就可以得到一组满意特征集(方开泰,1989)，但本文中“累计方差贡献率”仅为 45%，可以从另一个角度反映和说明自然语言的灵活性和难以形式化的特点。

## 3.2. 自然语言处理问题中的验证.

2 维显示可以直观反映噪音消除的效果，但我们更重要的目标是，希望通过词距离计算得到的词语相关性知识可以对自然语言处理问题提供支持。所以，还需在自然语言处理问题中进行验证。本文选取的问题是汉语多义词的消歧问题。

必须指出的是，词相关性知识的应用基础是词的相似性计算。这样的运算或推理机制是这种知识应用的必要条件。分类机器学习方法 Memory-Based Learning(MBL) (Daelemans,1998)最为基本的操作是属性值间的距离计算。

本文给出基于词距离的词语属性值计算方法，公式如下：

$$\delta(w_1,w_2)=\begin{cases}0 & w_1=w_2\\ dis(w_1,w_2) & \text{若 } w_1\neq w_2,\text{且}dis(w_1,w_2)\in\text{知识库};\\ 1 & w_1\neq w_2,\text{且}dis(w_1,w_2)\notin\text{知识库};\end{cases} \qquad (10)$$

针对 6 个汉语多义词，可以得到如下比较性实验结果：

| | 训练样本数 | 测试样本数 | MBL | [公式 10] | |
|---|---|---|---|---|---|
| | | | | 2164 个核心词; PCA(45%) | 7149 个核心词; PCA(45%) |
| 容易 | 1446 | 392 | 81.68% | 81.68% | 82.20% |
| 可靠 | 633 | 365 | 92.55% | 90.43% | 91.49% |
| 方向 | 3234 | 288 | 95.14% | 95.14% | 95.49% |
| 健康 | 2591 | 293 | 90.10% | 90.10% | 90.10% |
| 发表 | 2174 | 296 | 90.54% | 90.54% | 90.87% |
| 造就 | 398 | 338 | 99.11% | 99.11% | 99.11% |
| AVE | - | - | 91.52% | 91.16% | 91.54% |

[表 1] 验证实验中多义词消歧开放实验结果数据表

针对[表 1]的实验结果分析：

从总的平均正确率看，使用词语相关性知识库的平均正确率(91.54%)高于不使用词语相关性知识库的(91.52%)。尽管仅提高了 0.02%，但应该注意到，仅就其中单个多义词消歧正确率来讲，使用词语知识库与不使用正确率比较结果是：3 个高(容易/方向/发表)，2

个平(健康/造就)和 1 个低(可靠)。一定程度上，这说明了词语相关性知识库的价值。而且，核心词数量由 2164 增加到 7149 后，正确率上有明显改进！

需要指出的是，核心词数量由 2164 增加到 7149 后，单个正确率和平均正确率均有所改进，但测试时属性值距离计算覆盖率无论是 2164 时的 17.0750%还是 7149 时的 22.6107%，都是很低的。所以尽管是高频词语集，但小规模知识库低覆盖率造成了其作用范围不大的负面结果。即使在这样的情况下，IB-RK 能够在平均正确率和单个正确率个数上优于 IB1，一定程度上体现了词相关性知识的价值。

## 4. 结论和讨论

针对通过词距离计算获取词语相关性知识的问题，本文借鉴信息检索中的向量空间模型的词语权重计算方法，形式化定义了核心词语(2.1 节)。针对自然语言灵活性给相关性分析带来的巨大障碍，定义了 3 类噪音(2.2.1 节)，并就其中的“相关性分析无关词噪音”、“同义/近义词噪音”以及空间维数过高的问题，给出了“上下文词语相关性分析权重”方法和主成分分析方法(2.2.3 节)予以了有效地解决。经过 2 维显示和 Memory-Based Learning 解决汉语多义词的 2 个验证实验，效果得到了初步的证明。

通过词距离计算获取词语相关性知识过程的特点可以体现在如下 3 点：

[1] 整个计算过程是在无监督的条件下完成的，所依靠的资源仅是分词处理后(无词性标注)的上下文语境；

[2] 得到的词语相关性知识库是通用的、基础性的自然语言词语间知识，所以可以为诸如信息检索中的搜索引擎的查询扩充、问答系统中的问答词语匹配、基于实例机器翻译和词语聚类研究等诸多自然语言处理问题提供支持；

[3] 整个计算和噪音消除方法是在汉语进行实验的，但方法是与语言无关的通用方法。

与此同时，仍有以下问题需要进一步完善和解决：

[1] 通过本文的实验，有理由相信，随着核心词集的进一步扩大，实验效果会随之得到提高。但本文词距离计算方法所需高昂的硬件开销使这点难以实现。所以，若要实用，必须对其进一步优化。

**[2]** 验证实验的问题。计算得到的汉语词语知识库，其价值和适用范围，仍需进一步在其他自然语言处理应用问题中加以验证。

## 参考文献


[1] 白硕.1995.语言学知识的计算机辅助发现.科学出版社.

[2] 孙茂松,左正平,邹嘉彦.2000.基于 k-近似的汉语词类自动判定.计算机学报.Vol.23.No.2.p.166-170.

[3] 任若恩,王惠文.1997.多元统计数据分析——理论、方法、实例.国防工业出版社.

[4] 方开泰.1989.实用多元统计分析.华东师范大学出版社.上海.

[5] Christiane Fellbaum (ed.).1998.WordNet: An Electronic Lexical Database. Cambridge, Massachusetts and London, England: The MIT Press, Massachusetts Institute of Technology.

[6] Richardson, S.D.; Dolan, W.B. and L. Vanderwende, "MindNet: Acquiring and Structuring Semantic Information from Text," ACL'98: 36th Annual meeting of the Association for Computational Linguistics and 17th International conference on computational linguistics, ACL, 1998, CONF 17, Vol. 2, pp. 1098-1102.

[7] Ido Dagan, Lillian Lee, Fernando C. N. Pereira.1999.Similarity-Based Models of Word Cooccurrence Probabilities. 43-69.Volume 34, Numbers 1-3.Kluwer Academic Publishers.Boston/U.S.A; Dordrecht/Holland; London/U.K.

[8] Jakub Zavrel.1996.Lexical Space: learning and using continuous linguistic representations . Ph.D. Thesis. Utrecht Univesity.
[9]Schutze,Hinrich.1993.Word Space.in S.J.Hanson, J.D.Cowan & C.L.Giles(eds.). Advances in Neural Inoformation Processing Systems.Vol.5.San Mateo.CA.Morgan Kaufmann.pp.895-902.
[10] Lawrence R.Rabiner.1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE,77(2):257-285.
[11] Doug Cutting, Julian Kupiec, Jan Pedersen, and Penelope Sibun.1992.A practical part-of-speech tagger. In 3rd conference on applied natural linguistic processing(ANLP-92),pages 133-140.
[12]Salton,G..1968.Automatic Information Organization and Retrieval.McGraw-Hill,New York.
[13]朱德熙.1985.语法问答.北京:商务印书馆.
[14]Miller,George A. and Walter G. Charles.1991.Contextual correlates of semantic similarity. Language and Cognitive Processes,6(1):1-28.
[15]Daelemans, W., Van den Bosch, A., Zavrel, J. 1998. `Forgetting Exceptions is Harmful in Language Learning.` In Cardie,Claire.(ed.).Machine Learning 11:1-3, Special Issue on Natural Language Learning .
[16] 王碧泉,陈祖荫.模式识别——理论、方法和应用.1989.北京:地震出版社.
[17]Kohonen,T.1995.Self-Organizing Maps.Springer.Berlin.1995.
[18] Vesanto, J., Himberg, J., Alhoniemi, E. & Parhankangas, J. Self-Organizing Map in Matlab: the SOM Toolbox.1999.Matlab DSP Conference 1999, Espoo, Finland, November 16-17, 1999. pp. 35-40
[19] 梅家驹,竺一鸣,高蕴琦,殷鸿翔.1983.同义词词林.上海:上海辞书出版社.