# 基于区分词的汉语隐喻短语识别

符建辉[1,2] 曹存根[2] 王 石[3]
(首都师范大学计算机科学联合研究院 北京 100037)[1]
(中国科学院计算技术研究所智能信息处理重点实验室 北京 100080)[2]
(中国科学院研究生院 北京 100049)[3]

**摘 要** 隐喻识别是自然语言处理的一个重要研究分支。目前人们越来越清楚地认识到隐喻在思维及语言中所处的中心地位。从计算语言学和自然语言处理的角度来考虑,隐喻问题若不能得到很好的处理,语言理解和机器翻译的效果都会受到影响。通过观察隐喻短语和非隐喻短语在汉语中的上下文发现,有一批词可用于有效地识别隐喻短语,称之为区分词。首先从Web中自动抽取了一部分区分词,进而提出了一种基于区分词的隐喻短语识别方法。实验表明基于区分词的识别方法是有效的。

**关键词** 隐喻识别,汉语隐喻短语,自然语言理解

## Approach to Recognizing Chinese Metaphorical Phrases Based on Distinction Words

FU Jian-hui[1,2] CAO Cun-gen[2] WANG Shi[3]
(Joint Institute of Computer Science, Capital Normal University, Beijing 100037, China)[1]
(Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China)[2]
(Graduate University of Chinese Academy of Sciences, Beijing 100049, China)[3]

**Abstract** The research of metaphors is an important branch of natural language processing. People have realized that metaphors have taken a central position in thoughts and language. From the view of computational linguistics and natural language processing, language understanding and machine translation would be affected badly if the problem of metaphor was not well dealt with. Through observing the context of metaphorial and non-metaphorial phrases respectively, we found that some words can be used to identify metaphorial phrases effectively from Chiese phrases and we call these words distinction words. In this paper, we extracted a number of distinction word form the Web, and then proposed a distiction word based method for recognizing metaphorial phrases. Experimental results show the effectivity of our method.

**Keywords** Metaphor recognition, Chinese metaphorical phrase, Natural language understanding

## 1 引言

作为自然语言处理一个重要分支的隐喻处理研究,最近几年逐渐被引起重视,人们越来越清楚地认识到隐喻在思维及语言中所处的中心地位。隐喻是一个复现的语言现象,其定义多种多样。有人认为隐喻是一种修辞手段,是一个能取得特殊效果的单词或者词组,使得它不再具有其通常的含义或者字面意义;也有人认为隐喻是比喻的一种,但不含有"如"、"似"、"像"等比喻词,而用"是"、"就是"、"成为"等词,把某事物比拟成和它有相似关系的另一种事物。戴维森在《隐喻意味着什么》一文中提出"隐喻的含义无非就是其所涉及的那些词语的最严格的字面上的解释"[1]。莱柯夫和约翰森两人合著的《Metaphors we live by》中这样界定隐喻的本质:通过另一件事情来理解、经验某事,例如通过战争、战斗来理解、经验辩论。谈到一场辩论,我们会有如下说法:攻击某个薄弱环节,击中要害,摧毁了他的论点;采用某种战略,赢得或输掉了一场辩论,等等[2]。

总结对隐喻界定的观点,主要有两种认识思路:一种是将隐喻作为修辞现象看待,另一种是将隐喻作为一种认知现象看待。分别介绍如下:

隐喻作为一种修辞现象。

亚里士多德认为隐喻是通过把一个事物的词语给予另一个事物而构成的表达方式,提出了著名的"对比论"。不过他的"对比论"和后来的"替换论"都认为,就基结构和形式来看隐喻是正常语言的一种偏离,隐喻的功能只是一种修辞作用。20世纪30年代,Richard的"互动论"指出隐喻是一种新意义的创新过程,是两个主词的词义相互作用的结果。突破了把隐喻仅仅看作一种词汇层次的修辞现象的局限,后来Black发展和完善了"互动论"。他们把隐喻作为一种语义现象,放到句子层面考察的方法为认知语言学研究提供了启示。

到稿日期:2009-11-09 返修日期:2010-01-25

符建辉(1985—),男,主要研究方向国文本挖掘,E-mail:fjh5228203@126.com;曹存根(1964—),男,研究员,博士生导师,主要研究方向为知识获取与共享、文本挖掘;王 石(1981—),男,助理研究员,主要研究方向为知识获取、文本挖掘。

隐喻作为一种认知现象。

20 世纪 80 年代，G. Lakeoff 和 M. Johnson 从认知角度提出了全新的隐喻概念：隐喻不仅仅是一种语言修辞手段，而且是一种思维方式，是人们对客观世界的一种认知形式，是文化的反映，隐喻在思维及语言中的中心地位逐步被确立了。其本质是“用一种事情或经验理解和经历另一种事情或经验”[3]。之后出现了隐喻理解的多种认知模型，如结构映射匹配理论、现代隐喻理论、概念映射模型等。此外，隐喻的研究还受到了语用学家的重视，Searle 从言语行为理论角度，提出了隐喻的识别和理解的 8 项原则，对隐喻理解有着重要意义。

## 2 相关工作

自 20 世纪 70 年代以来，出现了一些隐喻计算模型。Fass[3] 提出了可以处理隐喻、转喻、字面义，反常表达的隐喻理解模型 MET5 系统。Martin[4] 提出了识别和解释常规隐喻的 MIDAS 系统。Kintsch[5,6] 给出了一个基于向量空间的隐喻解释计算系统，即 CI-LSA 框架，能处理形如“X si(a) Y”的隐喻。Mason[7] 利用大规模语料动态提取优先参数来识别特定领域的隐喻表达。

在中国，王治敏[8] 等人通过最大熵方法加辅助规则结合上下文对词汇层隐喻进行抽取，取得了一定的效果。张威[9] 从解决逻辑全知问题和隐喻的语义真值角度，提出了一种汉语隐喻逻辑系统。其主要思想是参考局部框架理论，采用池空间概念来替代可能世界，引入理解算子 Up、关系符＜以及格式塔规则。

在汉语中，许多隐喻更倾向于出现在词汇中。比如“知识海洋”，《现汉》中“海洋”指“地球表面连成一体的海和洋的统称”，没有隐喻含义，但是在此处当“海洋”和“知识”组合在一起成为复合词时，整个词就含有了隐喻含义，表示知识像海洋一样的宽广。类似的短语还有理想翅膀、生活风帆、时代航船、希望灯塔、改革春风等。复合词是汉语词汇的重要组成部分，其中蕴含着大量的含隐喻的词汇，要解决汉语隐喻问题，这部分是不可忽略的。王治敏重点提出了这个问题，并根据汉语隐喻短语本身以及上下文语境的特征，利用最大熵模型进行学习和测试，取得了比较好的结果。

通过以前的工作，我们获取了大规模的汉语短语库，其目的是要将词汇库中的隐喻词语抽取出来。因为汉语短语库中并不包含每个短评的上下文信息，针对每一个汉语短语，都去寻找它的上下文也是非常耗时的，所以我们很难利用王治敏提出的方法。针对这一情况，我们设计了一种新颖的方法，本方法利用区分词来识别隐喻短语和非隐喻短语。我们利用少量的句模从 Web 中获取一定的候选区分词，然后利用一批训练集从候选区分词中抽取出区分词，最后通过利用这些区分词集来识别隐喻和非隐喻短语。最后结果显示了本方法的有效性。

本文第 3 节介绍对汉语隐喻短语的分析；第 4 节介绍汉语隐喻短语算法；第 5 节介绍试验结果与分析；最后是总结。

## 3 对汉语隐喻短语的分析

认知语言学认为一个概念隐喻包含两个部分：一个“始源域(source domain)”，一个“目标域(target domain)”。“源域”通常是熟知的比较具体直观、容易理解的一些概念范畴，而“目标域”通常是后来才认识的、抽象的、不太容易理解的概念范畴[9]。这里沿用“源域”和“目标域”的说法，将能够在短语中作为“源域”出现的词称为源域词，比如杀手、大军、海洋等都可以称为源域词。

通过对含隐喻意义的汉语隐喻短语的分析，发现一个隐喻短语的最后一个词(简称尾词)常为该短语的源喻。比如祖国〈花朵〉、电脑〈病毒〉、爱情〈杀手〉。当然也有出现在首部的情况，比如〈花样〉年华。我们从语料中抽出了 300 个隐喻短语，其中只有 10 个隐喻短语其源喻是出现在前面。本文重点考查源喻词出现在隐喻短语的末尾的情况。

下面，用源域词“杀手”为例来说明本方法的思想。我们说“政府杀手”不是隐喻，因为“政府杀手”本身就是一个“杀手”，或者说“政府杀手”的上位是“杀手”。“政府杀手”是一个人，拿枪或刀进行杀人活动。而“电脑杀手”并不是一个“杀手”，它实际指的是一种对电脑有破坏作用的病毒或其它事物。在语料中考查“政府杀手”和“电脑杀手”发现，“政府杀手”作为一个“杀手”经常和“谋杀”、“枪杀”、“犯人”等词同时出现，这类词在一定程度上反映出“杀手”的本义，我们把这类词称为“杀手”的本义相关词集。在这些本义相关词集中，有许多词并不和“电脑杀手”一起出现，或者出现得很少。也就是说，一个源喻词，当它在汉语短语中不表现出隐喻意义时，该汉语短语常与源喻词的相关词集高频共现。相反，一个源喻词，当它在汉语短语中表现出隐喻意义时，该汉语短语与源喻词的相关词集一般共现次数很少。我们试图利用这种本义相关词集来对隐喻短语和非隐喻短语进行识别。

通过考察发现，有的本义相关词集和隐喻短语的共现也非常高。比如上例中的本义相关词“杀死”，虽然“杀死”这个词一般是指将一个人打杀致死，但该词却经常能够和“电脑杀手”一起出现，比如：

“这种‘电脑杀手’能够在 10s 内‘杀死’你的台式机，并……”。

在这个句子中，“杀手”这个词实际上也作为一种隐喻的形式出现。对于源喻词 $S$，我们将能够区分以 $S$ 为结尾的隐喻短语和非隐喻短语的相关词称为区分词。我们的方法是，针对某一源喻词，首先找该源喻词的区分词集，然后利用区分词集来对短语进行判断。

## 4 汉语隐喻短语识别算法

### 4.1 候选区分词的抽取

针对某源喻词 $S$，我们利用搜索引擎来寻找 $S$ 的候选区分词。为了保证用引擎搜索的网页中尽可能多地出现 $S$ 的本义相关词，首先构建下面几种形式的句模：

“$S$ 是一个|种|些|类|场|股”

“是一个|种|些|类|场|股 * 的 $S$”

其中，* 为通配符，表示间隔一到两个词。

然后利用引擎对上面的几种句模进行检索，取出每次检索出来的前 50 个网页。首先将这些文本进行分词，利用一个停用词表滤掉像“他”、“的”、“一个”之类的词。我们把这些剩下的词作为候选区分词集 $CW=\{w_1, w_2, \cdots, w_n\}$，其中 $n$ 为候选词集的大小。

比如，利用“是一个 * 的杀手”进行检索，得到下面的一条检索结果：

"2007年8月2日…'黑狐'的公开身份是一个山货站的掌柜,实则是一个凶残的杀手。人民解放军进入东北后,我公安机关启动了一个反特的'霹雳计划',以郑昊天为代表的我公安…"

对于上面这条查询结果,首先进行分词,并过滤掉像数字、单字、"的"、"一个"这类词后,还剩下{公开,身份,山货站,掌柜,凶残,人民解放军,进入,东北,公安,机关,启动,霹雳,计划,代表}。将这些词作为候选区分词集。

### 4.2 区分词的抽取

利用一批训练集从候选区分词集中抽出区分词。首先从语料库中抽出以源喻词 $S$ 结尾的短语,构成训练集 $TS=\{t_1,t_2,\cdots,t_k,t_{k+1},\cdots,t_m\}$,其中 $m$ 为训练集的大小,$t_1-t_k$ 为非隐喻短语,$t_{k+1}-t_m$ 为隐喻短语。对于候选区分词集 $CW$ 中的 $w_i$,计算出一组数值 $P=\{p_{i1},p_{i2},\cdots,p_{ij},\cdots,p_{im}\}$,其中 $p_{ij}$ 由下式计算得出。

$$P_{ij}=\frac{search\_freq(t_j+w_i)}{search\_freq(t_j)} \tag{1}$$

式中,$search_freq(X)$表示用搜索引擎检索 $X$ 的条数。式(1)的直观理解是:短语 $t_j$ 和词 $w_i$ 一起出现的次数除以短语 $t_j$ 出现的次数,计算出在 $t_j$ 出现的情况下 $w_i$ 出现的概率。短语 $t_j$ 和词 $w_i$ 共现次数以及短语 $t_j$ 出现的次数都用搜索引擎检索出的记录条数代替。本文从语料中抽取以"杀手"结尾的短语共54个,如下:

$t_1-t_5$:连环杀手|冒牌杀手|政府杀手|全职杀手|职业杀手

$t_6-t_{10}$:黑道杀手|变态杀手|美国杀手|冷血杀手|无情杀手

$t_{11}-t_{15}$:蒙面杀手|残忍杀手|乱世杀手|专业杀手|中国杀手

$t_{16}-t_{20}$:少女杀手|软件杀手|木马杀手|煤气杀手|玫瑰杀手

$t_{21}-t_{25}$:马路杀手|马铃薯杀手|龙虾杀手|经济杀手|进程杀手

$t_{26}-t_{30}$:金融杀手|健康杀手|公路杀手|电子杀手|电视杀手

$t_{31}-t_{35}$:电脑杀手|办公室杀手|劲舞杀手|品类杀手|灵魂杀手

$t_{36}-t_{40}$:手机杀手|密码杀手|键盘杀手|广告杀手|情场杀手

$t_{41}-t_{45}$:股市杀手|婚姻杀手|封面杀手|航母杀手|富豪杀手

$t_{46}-t_{50}$:动物杀手|价格杀手|局域网杀手|脂肪杀手|少男杀手

$t_{51}-t_{54}$:庄家杀手|硬件杀手|鲜花杀手|细胞杀手

其中,$t_1-t_{15}$为非隐喻短语;$t_{15}-t_{54}$为隐喻短语。以上词构成了"杀手"的训练集。当 $w_i$="凶残"时,$search\_freq(t_1+w_i)=search\_freq$("连环杀手"+"凶残"),利用搜索引擎检索得"连环杀手"+"凶残"的查询结果有96,100条,所以 $search\_freq$("连环杀手"+"凶残")=96100。$search\_freq$("连环杀手")=229,000。则 $P_{1j}=96100/229,000=0.4197$。将 $p_{1j}-p_{54j}$ 按照上面的方法计算,并图形化,如图1所示。

一般,如果 $w_i$ 是区分词,若短语 $t_j$ 是非隐喻短语,源喻词 $S$ 在 $t_j$ 中表现出自己的本义,此时 $p_{ij}$ 的值较大。相反,若短语 $t_j$ 是隐喻短语,$p_{ij}$ 的值较小。或者说,如果 $w_i$ 是区分词,我们能够设定一个阈值 $T$ 将以 $S$ 结尾的短语分成两份:当 $p_{ij}>T$ 时,就认为 $t_j$ 是隐喻词,否则认为 $t_j$ 是非隐喻词。利用这种原理,首先假设 $w_i$ 是区分词,看是否能够寻找出阈值 $T$ 将训练集 $TS$ 的隐喻短语和非隐喻短语分开。用下式来计算划分后的正确率:

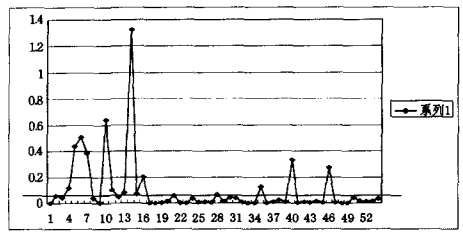


图1 "凶残"的曲线图

$$\text{正确率 } PP=\frac{\text{判断正确的短语个数}}{m} \tag{2}$$

我们通过寻找不同的阈值 $PT$,穷举所有对 $P$ 不同的划分,计算出其中最大的正确率 $PP_{max}$。然后用式(3)判断 $w_i$ 是否是一个区分词。

$$f(w_i)=\begin{cases}0, & PP_{max}>PT\\ 1, & \text{else}\end{cases} \tag{3}$$

式中,$f(w_i)=0$ 表示 $w_i$ 是一个区分词,否则,认为 $w_i$ 不是区分词。$PT$ 为阈值。

例如,上面关于"杀手"的数据中,从图1可以看出,选择 $T=0.04$ 时,将 $TS$ 分成两份,图中用浅色直线标出。此时,正确率 $PP=0.816$。在本实验中,设定的 $PT=0.89$,因为 $0.816<0.89$,认为"凶残"不是"杀手"的区分词。

针对候选区分词集 $CW=\{w_1,w_2,\cdots,w_n\}$ 中的每个词,都用上面的方法进行判断。如果是区分词,就保留,否则就删去。最后得到区分词集 $DW=\{dw_1,dw_2,\cdots,dw_h\}$,$h$ 是区分词集的大小。

对于 $DW$ 中的每一个词,都已经利用式(1)算得一组值。现在用矩阵表示如下:

| | $dw_1$ | $dw_2$ | … | $dw_h$ |
|---|---|---|---|---|
| $t_1$ | $p'_{11}$ | $p'_{21}$ | … | $p'_{h1}$ |
| $t_2$ | $p'_{22}$ | $p'_{23}$ | … | $p'_{h2}$ |
| … | | | | |
| $t_m$ | $p'_{1m}$ | $p'_{2m}$ | … | $p'_{hm}$ |

再计算一组值:$SP=\{SP_1,SP_2,\cdots,SP_i,\cdots,SP_m\}$,其中 $SP_i=p'_{1i}+p'_{2i}+\cdots+p'_{hi}$。直观的理解是利用累加来综合所有区分词的贡献。同样,我们寻找一个阈值 $ST$,将 $SP$ 分成两份。当 $SP_i>ST$ 时,认为 $t_i$ 是非隐喻短语,否则认为 $t_i$ 为隐喻短语。同理,我们用式(2)判断划分的正确率。通过选择不同的阈值 $ST$,求出其中最大的正确率 $PP'_{max}$。在 $PP'_{max}$ 处的阈值用 $ST_{max}$ 表示。

$ST_{max}$ 和 $S$ 的区分词集构成了用来判断 $S$ 结尾短语的模型。下面用这个模型来判断新的短语。

对于没有在训练集中出现过的且以源喻词 $S$ 结尾的汉语短语 $t_x$,利用 $S$ 的区分词集和式(1)算得关于 $t_x$ 的一组值 $P_x$,然后将 $P_x$ 中的值相加求得 $SP_x$,再用下式来判断 $t_x$。

$$F(t_x)=\begin{cases}0, & SP_x>ST_{max}\\ 1, & \text{else}\end{cases} \tag{4}$$

式中，$F(t_x)=0$ 表示 $t_x$ 被判断为非隐喻短语，否则，认为是隐喻短语。

## 5 实验与分析

### 5.1 实验

选取“暴雨”、“潮水”、“大军”、“杀手”、“沙漠”、“天空”这6个词源喻词进行测试。从语料库中抽出以这些词结尾的短语，结成训练语料，共有389条短语，隐喻短语210条，非隐喻短语179条，其中具体的结构如表1所列。

表1 训练集的结构

| 源喻词 | 隐喻短语个数 | 非隐喻短语个数 |
|---|---|---|
| 暴雨 | 21 | 20 |
| 海洋 | 20 | 25 |
| 大军 | 45 | 40 |
| 杀手 | 39 | 15 |
| 沙漠 | 31 | 32 |
| 天空 | 51 | 50 |

在寻找区分词时，根据经验将阈值 $PT$ 定为0.7。针对上面6个源语词，通过训练寻找出的区分词以及区分阈值如表2所列。

表2 训练出来的模型

| 源喻词 | 区分词 | $PP'_{max}$ | 区分阈值 $ST_{max}$ |
|---|---|---|---|
| 暴雨 | 洪灾，降水量，降水，雷电，雷雨，大风，降雨，降雨量，积水，雷击，强风 | 0.924 | 0.295 |
| 海洋 | 渔业，生物，能源，海岸，石油，海狮，岛屿，沉积物 | 0.931 | 0.555 |
| 大军 | 战役，战例，战火，战法，战场，前线，率领，军队，军人，军方，攻击，敌人，部队 | 0.968 | 2.250 |
| 杀手 | 追捕，鲜血，死亡，杀人，杀戮，谋杀，残忍 | 0.875 | 0.658 |
| 沙漠 | 沙，绿洲，荒漠，沙滩，风沙，戈壁，大漠，骆驼，气候，沙地，沙丘，沙子，热风，荒滩，植被，黄沙，尘土，沙尘暴 | 0.865 | 0.835 |
| 天空 | 湛蓝，云层，云彩，月亮，雨雪，雨天，雨滴，雨，阴雨天，雪天，乌云，晚霞，天边，晴空，蓝天，积雪，彩虹，白云，暴风雨，碧空 | 0.820 | 0.985 |

从表2可以看出，所找出的区分词都是与源喻词本身意义关联很紧的词，通过这些词，可将源喻词本身的意义很好地表示出来。另外，$PP'_{max}$ 的值是非常高的，这说明在训练集上这些词已能非常好地辨别出隐喻和非隐喻短语。

本实验采用下面的评价标准：

$$\text{隐喻判断正确率 } P=\frac{\text{正确判断隐喻短语的个数}}{\text{所有判断为隐喻短语的个数}}\times 100\% \quad (5)$$

$$\text{隐喻判断召回率 } R=\frac{\text{正确判断隐喻短语的个数}}{\text{总共存在的隐喻短语的个数}}\times 100\% \quad (6)$$

$$F\text{值}=\frac{2\times P\times R}{P+R}\times 100\% \quad (7)$$

从语料库中随机抽取100个短语作为测试集。这些短语都是以上面给出的源喻词结尾，并且要求它们没有在训练集中出现。我们随机抽取10次，取平均值。测出的正确率 $P=86.8\%$，召回率 $R=78.2\%$，$F$ 值 $=77.5\%$。

### 5.2 分析

从上面的结果来看，本方法在一定程度上能够体现出隐喻和非隐喻短语的差别。针对测出的错误短语进行分析，总结原因如下：

1)本方法中采用的训练集规模并不大。随着训练集的增大，找到的区分词会更加准确，相信结果会更好。

2)在测试集中存在着一些姓名和商标名，比如李海洋、赵大军等这种词。因为这些词本身不是隐喻短语，但源喻词在其中又不作为本义出现，所以用本方法对它们进行识别常得出错误的结果。

3)有些词本身就有二意性。比如“少女杀手”，该词既可以表示专杀少女的杀手，也可以表示获得少女芳心的情场高手。这种词的存在造成区分界线不明显，给结果带来一定的影响。

4)搜索引擎自身的问题。以“低沉天空”为例进行说明，本文使用的是Google搜索引擎，在Google中搜索〈“低沉天空”〉出现的记录条数是167条。而搜索〈“低沉天空”+“月亮”〉出现的条数是597条。包含“低沉天空”的网页数少于包含“低沉天空”和“月亮”一起出现的网页数，这显然是不合理的。

如果能够很好地解决上面的问题，相信能够进一步提高测试的精度。另外，从上面的步骤可以看出，本实验要求有一批训练集，而隐喻的训练集本身是很难获取的。有些尾词的隐喻短语数量也是非常有限的。比如源喻词“班长”，在语料中只出现“心灵班长”。这显然是不能用上面方法进行识别的。另一方面，在寻找区分词上，本实验的耗时也非常大，有待进一步改进。

通过本实验能得到一些判断隐喻的启示：不同短语因其意义不同所处上下文也差别很大。但是同处于一个上位下的短语，它们的上下位本身却存在着一定的相似性。比如“公路”和“马路”，它们的上位都是“路”，在它们的某些上下文中，也必然会出现与“路”关联很紧的词，像“行驶”、“车子”等。另外，在汉语中，一个短语往往是其自身尾词的下位，则具有相同尾词的短语因为上位相同，它们的上下文也存在着一定的相似性。如果不同，说明这条短语不是其后缀的下位。这时很可能是隐喻短语，当然也可能是姓名、团体机构名等。比如“人民大学”和“北京大学”，这两个词的上位都是“大学”，它们在语料中的上下文也非常相似。从本实验的结果来看，也很好地反映了这一点。

本文下一步工作将是针对这些存在的问题寻求解决的办法。

**结束语** 汉语隐喻处理在中文信息处理领域是一个新的研究方向。本文从计算机处理的角度，提出了一种新颖的方法，以发现汉语短语中的隐喻短语。本方法首先通过对汉语短语的分析，利用少量的句模从Web中获取一定的候选区分词，然后利用一批训练集从候选区分词中抽取出区分词，通过利用这些区分词集来识别隐喻和非隐喻短语。最终的实验结果表明本方法是有效的。最后分析了目前此方法中仍存在的问题以及通过本次实验得到的一些启示，认为本方法还有很大的提升空间。

## 参考文献


[1] A. P. 马蒂尼奇. 语言哲学[M]. 北京：商务印书馆，2004

[2] Lakoff G, Johnson M. Metaphors We Live By[M]. Chicago: The University of Chicago Press, 1980


(下转第232页)

图像的 FWSNR，实验流程见图 12，实验数据见表 1。我们发现该指标表明加网的频率越高，其生成的半色调化图像的质量越好。这与胶印生产的实际情况是不相符的。按现有的工艺条件，一般以 175LPI 的频率加网质量较好，超过这个频率，生产工艺的缺陷会使产品的质量下降。因为在这个质量评价实验中没有考虑胶印工艺的特点和缺陷，所以表 1 给出的数据得出这样的结论就可以理解了。

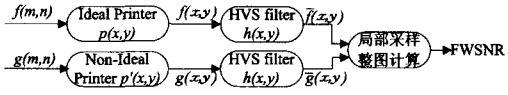


图 11　基于理想的打印机模型质量评价流程

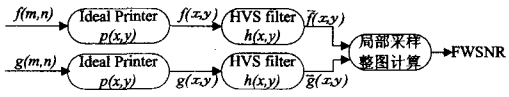


图 12　基于理想的打印机模型质量评价流程

表 1　理想打印机模型

| FWSNR | 125LPI | 150LPI | 175LPI | 200LPI |
|---|---|---|---|---|
| 局部计算 | 4.146 | 5.557 | 6.084 | 7.806 |
| 整图计算 | 4.157 | 5.544 | 6.067 | 7.794 |
| FWSNR | 225LPI | 250LPI | 275LPI | 300LPI |
| 局部计算 | 10.287 | 13.005 | 16.680 | 16.680 |
| 整图计算 | 10.320 | 13.016 | 16.707 | 16.707 |

为了更真实地反映胶印的生产工艺，我们在第二阶段的质量评价实验中引入非理想打印机模型。按图 11 流程计算得出的数据见表 2。

表 2　非理想打印机模型

| FWSNR | 125LPI | 150LPI | 175LPI | 200LPI |
|---|---|---|---|---|
| 局部计算 | 4.360 | 5.030 | 5.120 | 5.048 |
| 整图计算 | 3.445 | 3.666 | 3.671 | 3.537 |
| FWSNR | 225LPI | 250LPI | 275LPI | 300LPI |
| 局部计算 | 4.584 | 4.047 | 3.635 | 3.635 |
| 整图计算 | 3.098 | 2.691 | 2.385 | 2.385 |

因为实验中在计算 SNR 前做了 HVS 滤波处理，相当于对人眼的频率特性做加权处理，这样的质量指标我们也称为频率加权的信噪比（FWSNR，Frequency Weighted Signal Noise Ratio）。为了验证 FWSNR 对于评价半色调化图像质量的有效性，对一个渐变连续调图像按胶印要求加网做半色调化，假定输出设备的分辨率为 2400DPI，在 PhotoShop 中用 Euclidean 网点做从 125LPI 到 300LPI 共 8 个不同频率的加网。

非理想打印机模型参数的选择。现有的 175LPI 加网的印刷工艺，一般能保证 2%的网点不丢失，98%的网点不实地化。2400DPI 的输出分辨率的设备以 175LPI 加网，其网点单元为 14×14 的像素点阵。2%的网点为 3.92 个像素点，因此过滤小网点的门槛值 Max_number 设定为 4。

人眼对黑白图像最小分辨率空间角为 1 角分；转换成弧度值为：3.14 * 1/(60 * 180)，如果观察距离为 12 英寸(30cm)，则对应 2400DPI 的输出设备的 8.837 个像素点的长度。对彩色图像的分辨率约为黑白图像的 1/5。因此选择 9×9到 45×45 的像素点阵采样计算信噪比都是合理的。表 1 和表 2 都是采用的 20×20 像素点阵来计算 SNR。

**结束语**　从实验结果可以看出，用理想打印机模型计算 FWSNR 评价半色调化图像的质量，加网频率越高，FWSNR 值越大，也就是质量越好。这与胶印工艺实际情况不一致。而用非理想打印机模型计算 FWSNR 则在 200LPI 内，加网频率越高质量越好，超出这个范围，随着频率升高，质量反而下降，这与胶印工艺实际情况一致。当然质量变化的频率拐点与打印机模型选择的参数有关。实验中我们选择了与胶印工艺匹配的参数，并得出了与胶印工艺一致的相关结论。如果要用在凹印或柔印工艺中，模型的参数要做相应的调整。

进一步的研究工作可以尝试在非理想的打印机模型加入油墨颜色预测方程，将本研究拓展到彩色印刷中。在半色调化图像中用集束网点的叠加方式预测 Moirê 的产生和用颜色方程[7]预测印刷品的颜色也是对生产实践非常有价值的工作。本项目前期所做的均衡集束网点密度的研究[8]也可以用上一节的质量评价方法为工具去探索性能更好的优化算法。

## 参 考 文 献


[1] Sheikh H R, Sabir M F, Bovik A C. A statistical evaluation of recent full reference image quality assessment[J]. IEEE Transaction on Image Processing, 2006, 15(1): 3440-51

[2] Ghanem B, Resendiz E, Ahuja N. Segmentation-based perceptual image quality assessment (SPIQA) [C]// 15th IEEE International Conference on Image Processing. San Diego, CA, USA, 2008: 393-396

[3] Axelson P E. Quality measures of halftoned images (a review) [D]. Norrköping Sweden, Institute of Technology, Linköping University, 2003

[4] Kim S H, Allebach J P. Impact of HVS models on model-based halftoning[J]. IEEE Transactions on Image Processing, 2002, 11(3): 258-269

[5] Damera-Venkata N, Kite T D, Geisler W S, et al. Image quality assessment based on a degradation model[J]. IEEE Transactions on Image Processing, 2000, 9(4): 636-649

[6] Gooran S, Li Yang. Frequency Modulated Halftoning and Dot Gain[R]. Department of Science and Technology, Linköping University, Sweden

[7] Amidror I, Hersch R D. Neugebauer and Demichel: Dependence and Independence in n-Screen Superpositions for Colour Printing [J]. Color research and application, 2000, 25(4): 267-277

[8] 徐国梁，谭庆平. 均衡集束网点密度的伪随机混合抖动半色调化算法的研究[J]. 计算机学报，2009，32(8)：1550-1559


(上接第 196 页)


[3] Fass D. Met *: A Method for Discriminating Metonymy and Metaphor by Computer[J]. Association for Computational Linguistics, 1991, 17(1): 46-49

[4] Martin J H. A Computational Model of Metaphor Interpretation [M]. NY: Academic Press, 1990

[5] Kintsch W. Metaphor comprehension: A computational theory [J]. Psychonomic Bulltin and Review, 2000, 7(2): 257-266

[6] Kintsch W, Bowles A R. Metaphor comprehension: What makes a metaphor difficult to understand? [J]. Metaphor and Symbol, 2002, 17(4): 249-262

[7] Mason Z J. CorMet: A Computational, Corpus-based Conventional Metaphor Extraction System [J]. Computational Linguistics, 2004, 30(1): 23-44

[8] 王治敏. 汉语名词短语隐喻识别研究[D]. 北京：北京大学，2006

[9] 张威，周昌乐. 汉语隐喻理解的逻辑描述初探[J]. 中文信息学报，2004，18(5)：23-28