文章编号：1003-0077(2008)06-0027-06

“像”的明喻计算

李 斌[1]，于丽丽[1]，石 民[1]，曲维光[2]

(1. 南京师范大学 文学院，江苏 南京 210097；2. 南京师范大学 计算机系，江苏 南京 210097)

摘 要：汉语隐喻计算是一项难度很大的工作，明喻由于带有明显的标志(比喻词)成为计算机自动识别的基础类型。该文着力于典型的比喻词“像”的比喻义及相关比喻成分的自动识别。首先，人工标注了 1 586 句语料，分析了明喻句的基本特点。然后，使用最大熵模型对“像”的比喻义和非比喻义进行分类，开放测试 F 值达到了 89%。最后，用条件随机场模型识别出比喻的本体、喻体和相似点，F 值分别达到了 73%、86%和 83%。

关键词：计算机应用；中文信息处理；隐喻计算；明喻；明喻识别

中图分类号：TP391　　**文献标识码**：A

Computation of Chinese Simile with “Xiang”

LI Bin[1], YU Li-li[1], SHI Min[1], QU Wei-guang[2]

(1. School of Chinese Language and Literature, Nanjing Normal University, Nanjing, Jiangsu 210097, China;
2. Department of Computer Science, Nanjing Normal University, Nanjing, Jiangsu 210097, China)

Abstract: The computation of metaphors in Chinese is certainly a challenging issue. Simile, with an obvious mark word, is a good start for automatic processing of metaphor. This paper is focused on automatic identification of Chinese simile phrases with the word “xiang”. Altogether 1 586 sentences containing “xiang” are first retrieved from the corpus and manually tagged and analyzed. Then the Maximum Entropy Model is applied to detect the simile meaning of “xiang”resulting in a F-score of 89% in open test. Finally, Conditional Random Fields (CRFs) Model is used to identify tenor, vehicle and similarity in the simile, achieving an acceptable F-score of 73%, 86% and 83%.

Key words: computer application; Chinese information processing; metaphor computation; simile; simile identification

1 引言

隐喻是一切语言中都普遍存在的现象[1]，也是文本中常见的表达方式。近年来，随着认知语言学背景下的隐喻理论逐步发展，隐喻计算成为计算语言学的新兴课题。国内外相关研究主要集中于以下三个方面。一是探讨隐喻理解的规则建立模型，如 Fass 提出的可以处理隐喻、转喻、字面义反常表达的隐喻理解模型 Mat5 系统[2]，Martin 利用概念之间的关系映射提出的识别和解释常规隐喻的 MIDAS 系统[3]，杨芸、周昌乐的汉语隐喻分类和逻辑推理模型[4]。二是建立隐喻知识库，如 Martin 提出的面向自然语言应用的经验诱导和理论驱动相结合的隐喻知识库 MetaBank[5]。三是对隐喻表达进行抽取和识别，如 Mason 利用大规模语料库进行特定模式的隐喻抽取[6]，王治敏对汉语“N＋N”短语的隐喻识别[7]，W. G. Qu 对“黄金”等词语比喻义的识别[8]。总的说来，隐喻计算特别是汉语的隐喻计算还处于起步阶段，隐喻的复杂性和多样性使得对真实文本中的句子进行隐喻识别的工作还比较薄弱。因此，本文尝试对隐喻中带有明显标记的“明喻”用

收稿日期：2008-06-11　**定稿日期**：2008-09-10

基金项目：国家自然科学基金资助项目(60773173)；国家 973 资助项目(2004CB318102)；国家社科基金资助项目(07BYY050)；江苏省社科基金资助项目(06JSBYY001)

作者简介：李斌(1981—)，男，博士生，研究方向为计算语言学；于丽丽(1983—)，女，硕士生，研究方向为计算语言学；石民(1984—)，男，硕士生，研究方向为计算语言学。

法进行识别。

“明喻”,无论在传统修辞学和认知隐喻学中都是一个重要的研究内容。区别于其他的隐喻模式,明喻带有比喻词,基本模式为“X像Y(一样Z)”,X,为本体;像,为比喻词;Y为喻体;Z为相似点。明喻由于带有明显的比喻标志而成为研究隐喻计算的一个较理想的突破口,如果可以识别出带有明喻用法的句子,特别是识别出X和Y,就可以建立起X和Y之间的映射关系,从而理解和分析句子的深层意义,也可以为其他类型的隐喻识别积累资源。Veale利用搜索引擎搜集了“as ADJ as NOUN”的短语实例,人工逐一校对后,根据相似点ADJ区分出明喻和反讽手法,并建立了ADJ和NOUN的关系数据库[9]。明喻的自动识别将有助于这类工作的展开。

汉语明喻的比喻词是较为丰富的,有“像”、“好像”、“仿佛”、“宛如”、“犹如”等等。我们在《人民日报》1998年上半年语料(由北京大学计算语言学研究所人工标校了分词和词性标记信息,下文简称“R98”)上进行了统计,选取了“像”这个非常典型且频率较高的比喻词作为研究对象,收集了1 000多句含有词语“像”的句子,区分了比喻义和非比喻义,并人工标注了本体、喻体和相似点,进行了初步统计分析。然后,尝试使用最大熵模型区分“像”的比喻义和非比喻义,用条件随机场模型识别本体、喻体和相似点。

2 比喻义的界定和语料考察

2.1 “像”的比喻义的界定

“像”的比喻义和非比喻义,在一般情况下是比较容易辨认的,即本体和喻体的语义类别不同,如“姑娘像朵花儿”。但真实文本中存在着一些难以判定的用例,特别是有不少本体和喻体语义类别相差不大甚至相同的情况。语言学界对此争论较多,说法各异。为了界定比喻义并尽可能地减少主观性误判,我们参考并扩展了盛若菁[10]和崔应贤[11]的观点,结合R98中的实例,对难于区分的情况制定了如下标准:

(1) 用于比喻的甲乙两事物所属的语义类别差别小或同类,两者具有区别性语义特征,且在区别性语义特征上作比,定为比喻句。相反,没有区别性语义特征,且乙事物是确指的,则是一种比较,定为非比喻句。

① 他高兴得像个孩子似的。 比喻

② 老人……越看越觉得像他分别40多年的朋友石田东四郎。 非比喻

例①中,“他”和“孩子”,同属人的语义类,但句子凸显的是孩子的天真活泼的区别性语义特征,是比喻句。例②中,对象是确指的,专指老人的朋友石田东四郎,二者不具有区别性语义特征,只是长相相似,为非比喻句。

(2) 甲乙两事物属同类,但乙事物带有明显的典型性,且句子突出的是这种典型性的语义特征,将其视为比喻句。

③ 周围的群众连连称赞:一一〇的民警像雷锋。 比喻

④ 她……不用像记者那样整天在外面跑新闻。 非比喻

例③的两事物同属人的语义类,用“雷锋”主要是借其“做好人好事”方面的典型性,故构成比喻。而例④中的“记者”虽然不是具体的人,却是具体活动的主体,和例②一样是一种比较而非比喻。

(3) “像”在句子结构中处于谓词性短语(VP)之前,用VP来表述某种特殊的状态或方式,一般定为比喻句。如果仅表示人对事物不太有把握的一种推测,相当于“似乎”的意思,则定为非比喻。

⑤ 活动像滚雪球一样产生了巨大的反响。 比喻

⑥ 双手抚着跪地战友的左肩,像是劝慰他不要太悲伤。 非比喻

2.2 语料标注和统计

使用R98语料,我们对常用的20个比喻词根据与比喻相关的词性做了词频统计,发现“像”的频次是比较高的(见图1)。“像”在R98语料中有名词、动词、副词、介词四种词性,有比喻义的只有后面三种词性。为方便论述,下文中提到的“像”一律不包括名词词性。我们从R98中抽取出1 586个“像”

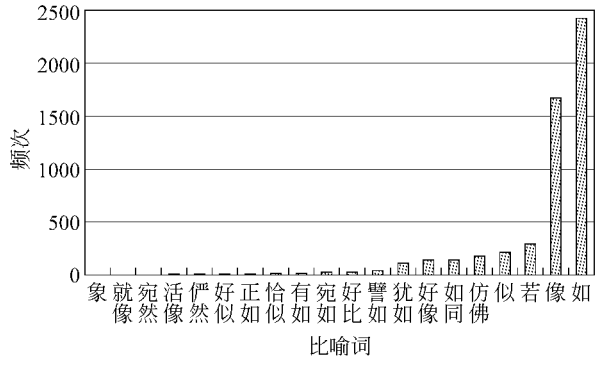

图1 常用比喻动词分布图

的句子。句子长度(包含词语的个数)范围是 4 至 108 个词,平均句长为 31 词。

按照 2.1 节的区分原则,我们采用多人互校的方式区分了是否比喻义,是比喻义的则进一步标注其比喻成分,即本体、喻体和相似点。对于比喻成分,我们采用的原则是标注包含中心词在内的整个短语。标注样例如下:

让/v [我们/r 两/m 国/n 之间/f 的/u 关系/n]b [像/p]c [黄金/n]y 一样/u [珍贵/a]s ,/w [像/p]c [钻石/n]y 一样/u [坚强/a]s 。/w

其中,b,表示本体;c,表示带有比喻义的喻词"像";y,表示喻体;s 表示相似点。

人工标注后,得到比喻句 512 个,非比喻句 1 074 个。我们统计了"像"的比喻模式、比喻成分的构成长度等信息。表 1 给出了比喻句的模式分布,典型的比喻模式"本体—喻词—喻体"占了绝大比例,而"喻词—喻体—本体"也有不少,如"像小山一样的棉花堆"。一个比喻词也可能带有不止一个本体或者喻体。这表明,虽然是明喻,其结构模式在真实语料中依然是多样的。

表 1　比喻模式在句子中的分布

比喻模式	频次	比例
本体—喻词—喻体	365	71.29%
本体—喻词—喻体—喻体	26	5.08%
喻词—喻体—本体	24	4.69%
本体—喻词—喻体—相似点	21	4.10%
喻词—喻体(句中承上省略了本体)	18	3.52%
本体—相似点—喻词—喻体	13	2.54%
其他(频次小于 10 的模式)	45	8.79%
合计	512	100.00%

表 2 给出了比喻成分的构成长度分布。构成长度即每个比喻成分所含词语的个数。本体和喻体的构成长度集中在 5 个词以内,少数比较长,最长的则达到 20 个词。本体和喻体的平均构成长度分别为 3.73 和 3.5 个词。较长的比喻成分对自动识别而言具有较大难度。

表 2　比喻成分的构成长度分布

长度	1	2	3	4	5	6	7	8	9	10	>10	合计	平均长度
本体	170	96	94	52	38	24	15	15	17	11	30	562	3.73
喻体	163	73	108	71	57	42	17	17	9	5	15	577	3.50
相似点	30	20	2	2	1	0	0	0	0	0	0	55	1.62

此外,我们注意到,本体部分和比喻词"像"有的时候距离比较大。统计显示,两者之间的距离从紧邻(相距 0 个词)到最多相隔 29 个词,平均距离为 2.73 个词。喻体和"像"之间最多只相距 1 个词。这可能给本体和喻体的自动识别带来一些影响,本体的识别效果可能会更低一些。

3　实验及分析

3.1　实验语料

我们把 2.2 节介绍的手工标注了比喻义和比喻成分的 1 586 个句子(含 1 647 个"像")作为实验语料。表 3 给出了用于实验的语料划分情况。对于比喻义和非比喻义的"像",分别随机抽取出 50 个和 100 个作为测试语料,其他的作为训练语料。

表 3　训练测试语料情况

类　别	训练语料	测试语料	合计
比喻义的"像"	493 个	50 个	543 个
比喻句	462 句	50 句	512 句
非比喻义的"像"	1 004 个	100 个	1 104 个
非比喻句	974 句	100 句	1 074 句

3.2　基于最大熵模型的"像"的比喻义识别

对于"像"的比喻义识别,我们将其转化为单点的二值分类问题,用最大熵模型进行分类。实验采用 Zhang Le 开发的最大熵工具包①。在表 3 所示的 1 436 句训练语料和 150 句测试语料上,我们选取了"像"的上下文不同窗长的分词和词性作为特征

① 下载地址:http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html

进行对比实验，结果见表4。

表4 最大熵模型识别结果(F值)

词语窗口 \ 词性窗口	0	[−1,1]	[−2,2]	[−3,3]
[−1,1]	0.786 7	**0.893 3**	0.866 7	0.813 3
[−2,2]	0.800 0	0.873 3	0.833 3	0.840 0
[−3,3]	0.766 7	0.846 7	0.846 7	0.826 7

实验结果表明，只需要较短的上下文就可以较好地区分出“像”的比喻义和非比喻义。在窗口为[−1,1]时，综合F值达到最高。此时的分项成绩也是最高的，对比喻义的“像”识别的正确率、召回率、F值分别达到0.904 8、0.76、0.826 1；对非比喻义的“像”识别的正确率、召回率、F值分别达到0.888 9、0.96、0.923 1。

3.3 基于CRFs模型的比喻成分识别

我们在512个比喻句(即确定为比喻义的“像”的句子)上进行了比喻成分的识别，将识别过程转化为一个序列标注问题。我们采用了条件随机场(Conditional Random Fields，CRFs)模型，使用了Taku Kudo编写的工具包“CRF++0.51”进行训练和测试①。实验语料为表3中的比喻句语料，462句训练，50句测试。在标记集上，我们曾尝试过用BIO的标记方式来分别标注每一类比喻成分，即给每种比喻成分都标记出开头、中间和结尾。但由于语料规模有限，数据稀疏严重，精确率较低。因此，我们不再做此区分，而把本体、喻体和相似点的每个词语都用各自的符号标注，连续的标记串则成为一个比喻成分，这样就可以确定成分的边界和类型。语料的具体格式参见表5。

表5 训练和测试语料实例

词 语	词性	是否比喻词	标准答案	自动标注结果
延安	ns	F	b	X
的	u	F	b	X
牡丹	n	F	b	b
，	w	F	X	X
就	d	F	X	X
像	v	T	c	c
延安人	n	F	y	y
的	u	F	y	y
灵魂	n	F	y	y
。	w	F	X	X

在特征选择上，我们选取了不同窗长的词语、词性标记、同现词语和同现词性。同时，增加了“是否比喻”这一列特征来区分语料中的“像”是否已判定为比喻义，而且给比喻词“像”增加一个标记“c”，以和三个比喻成分形成更好的标记序列。由于本体、喻体构成长度较大，自动识别的结果可能难以准确地锁定边界，所以我们给出了比喻成分完全匹配和部分匹配两种结果进行观察分析。实验结果见表6。

表6 CRFs识别比喻成分的结果(F值)

比喻成分	评测模式	1W+1P	2W	2W+2P	3W+2P	3W+2P+2W'	3W+2P+2W'+2P'	4W+4P+2W'+2P'	5W+5P+2W'+2P'
本体48个	完全匹配	0.318 8	0.567 6	0.571 4	0.607 6	0.623 4	0.642 9	0.712 6	**0.735 6**
	部分匹配	0.579 7	0.648 6	0.753 2	0.784 8	0.753 2	0.833 3	**0.873 6**	**0.873 6**
喻体50个	完全匹配	0.702 1	0.83	0.804 1	0.816 3	0.816 3	0.845 4	0.836 7	0.866
	部分匹配	0.893 6	0.936 2	0.927 8	**0.938 8**	**0.938 8**	0.927 8	**0.938 8**	0.927 8
相似点6个	完全匹配	0.5	0.5	0.769 2	0.769 2	0.769 2	0.769 2	0.769 2	**0.833 3**
	部分匹配	0.5	0.5	0.769 2	0.769 2	0.769 2	0.769 2	0.769 2	**0.833 3**

注：W表示词语，1W表示上下文窗口[−1,+1]的词语，nW就是窗口[−n,+n]的词语；P表示词性，1P表示[−1,+1]的词性，nP表示窗口[−n,+n]的词性；2W'表示词语的共现，即$W_{-1}W_0$，W_0W_1，$W_{-1}W_1$；2P'表示词性共现，即$P_{-1}P_0$，P_0P_1，$P_{-1}P_1$。

实验结果表明：

(1) 本体的识别效果随着观察窗口的增加而提高。在2.2节语料统计时就已经指出，有些本体距离比喻词比较远，短距离的上下文不足以较好地识别出远距离的本体。

(2) 喻体的识别效果随着观察窗口的增加并没有显著提高，在常用的左右各两个词上下文条件下已经可以达到85%的完全匹配率。2.2节的观察也

① 下载地址：http://crfpp.sourceforge.net/

已指出，喻体距离比喻词较近，因此短距离的上下文就可以取得较好成绩，其识别效果也明显好于本体。

(3) 本体和喻体在边界确定上存在一些问题，这是由于本体喻体的构成长度较大所造成的。本体的构成长度更长，所以在边界上所受的影响也比喻体大。

(4) 相似点由于经常伴随喻体或"一样、似的"等词语出现，识别效果较好。但评测数据少，还不足以说明问题。相似点内部往往只有一两个词语，所以在边界识别上问题较小。

3.4　最大熵和 CRFs 依次处理

本文的目标是对含有"像"的句子，既做比喻义的区分，又完成比喻成分的识别。为此进行了两个对比实验。

(1) 直接使用 CRFs 进行比喻义和比喻成分识别。参照 3.3 节的最佳实验方案，我们使用 CRFs 模型，使用 1 436 句训练，150 句测试。表 7 给出的实验结果表明，相比最大熵模型，在比喻义的识别上由 82.61%提高到 93.62%，但是在比喻成分的识别方面比在 3.3 节已知比喻句的条件下有所降低，特别是本体的 F 值降低了近 30%。

表 7　CRFs 模型同时识别比喻义和比喻成分的结果

比喻成分	评测模式	P	R	F
本体 48 个	完全匹配	0.866 7	0.270 8	0.412 7
	部分匹配	0.933 3	0.291 7	0.444 4
喻体 50 个	完全匹配	0.864 9	0.640 0	0.735 6
	部分匹配	0.973 0	0.720 0	0.827 6
相似点 6 个	完全匹配	1.000 0	0.166 7	0.285 7
	部分匹配	1.000 0	0.166 7	0.285 7
比喻词 50 个	完全匹配	1.000 0	0.880 0	0.936 2

(2) 用最大熵识别出"像"的比喻义，再用 CRFs 识别出比喻句中的比喻要素。依次使用前面 3.2 和 3.3 节的算法和最优参数，得到表 8 的结果。最大熵模型在 150 个句子中，判定 42 句为比喻句，其中 4 句错误。这 4 句造成了识别率的略微下降，如果以这 42 句进行评测，识别的结果还是比较理想的。如果以全部 50 个比喻句进行评测，则由于漏识的 8 句导致召回率迅速下降，甚至在喻体的 F 值上低于 CRFs 同时识别的结果。

表 8　最大熵＋CRFs 依次处理的结果(上：用 42 句评测；下：用 50 句评测)

比喻成分	评测模式	P	R	F
本体 36 个	完全匹配	0.750 0	0.583 3	0.656 3
	部分匹配	0.928 6	0.722 2	0.812 5
喻体 38 个	完全匹配	0.815 8	0.815 8	0.815 8
	部分匹配	0.944 4	0.894 7	0.918 9
相似点 6 个	完全匹配	0.833 3	0.833 3	0.833 3
	部分匹配	0.833 3	0.833 3	0.833 3
本体 48 个	完全匹配	0.718 8	0.479 2	0.575 0
	部分匹配	0.937 5	0.625 0	0.750 0
喻体 50 个	完全匹配	0.810 8	0.600 0	0.689 7
	部分匹配	0.945 9	0.700 0	0.804 6
相似点 6 个	完全匹配	0.714 3	0.833 3	0.769 2
	部分匹配	0.714 3	0.833 3	0.769 2

这两个实验表明，对比喻义和比喻成分的识别可以有两种不同的策略，一种是直接把比喻词纳入分类的范畴，一种是采用层叠的方法进行分步识别，两种策略得到的效果基本持平而各有特点。使用CRFs模型同时识别四种成分时，由于有了多类信息，对比喻义的区分得到了提高；同时由于缺少了比喻词的确定标志，使得距离喻词较远的本体的识别效果严重下降。而采用分步策略时，本体的识别效果会好一些，特别是在最大熵识别出来的那些比喻句上，各项指标都比较高。我们可以根据实际工作的需要采用不同的识别策略。

4 结论及未来工作

明喻是隐喻计算的一个基础类型，本文对常用比喻词"像"及其比喻成分的识别问题进行了探讨。通过对1 000多句语料的手工标注、分析和计算，得出如下结论：(1)"像"的比喻模式中，"本体—喻词—喻体"的典型框架占绝对优势，但模式类型在真实文本中还是多样的。(2)使用最大熵模型对"像"的比喻义识别，仅使用分词和词性标注信息，在较短的上下文窗口条件下，F值可以达到89.33%。(3)使用CRFs模型对比喻句中的本体、喻体等比喻成分进行识别，较大的上下文窗口可以提高精度效果，本体、喻体和相似点完全匹配的F值分别达到了73.56%、86.60%和83.33%。本体、喻体的长度比较大，给边界的判定带来一定影响，不完全匹配(部分匹配)的F值则可以达到87.36%、92.78%。这可以应用于比喻成分的探测上。本体与喻词之间距离较大，使得本体的识别效果最低。(4)对本体、喻词、喻体、相似点的识别可以有两种策略，一种是使用CRFs模型同时识别这四类成分，一种是采用最大熵模型和CRFs依次处理。前者在比喻义的区分和喻体识别上具有一些优势。而后者在整体上比较稳定，特别是对于最大熵模型已判定为比喻句的句子上识别效果较好。这些都可以对后续的隐喻研究工作提供参考。

我们下一步的工作主要有：(1)在充分利用分词与词性标注信息的基础上，尝试其他的句法语义信息作为特征，并采用浅层分析技术捆绑比喻成分，进一步提高识别效果。(2)识别其他常用的比喻词，从而能够从文本中正确识别出明喻句及本体、喻体和相似点，为隐喻研究建立源域向目标域的映射关系等工作奠定基础。

参考文献：

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

[2] Fass, D., met*: A Method for Discriminating Metonymy and Metaphor by Computer [J]. Computational Linguistics, 1991, 17(1): 49-90.

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

[4] 杨芸，周昌乐，李剑锋. 基于隐喻角色依存模式的汉语隐喻计算分类体系[J]. 语言文字应用，2008，(3)：125-133.

[5] Martin, J. H. MetaBank: A Knowledge-Base of Metaphoric Language Convention [J]. Computational Intelligence, 1994, 10(2): 134-149.

[6] Mason, Z. A Computational, Corpus-Based Metaphor Extraction System [D]. Brandeis University, 2002.

[7] 王治敏. 汉语名词短语隐喻识别研究[D]. 北京大学博士论文，2006.

[8] W. G. Qu, Z. Sui, G. Ji, et al, A Collocation-Based WSD Model: RFR-SUM[C]// H. G. Okuno and M. Ali (Eds.). IEA/AIE-2007, LNAI 4570, Springer-Verlag Berlin Heidelberg, 2007, 23-32.

[9] Tony Veale, Yanfen Hao. Learning to Understand Figurative Language: From Similes to Metaphors to Irony[C]// Proceedings of CogSci 2007, Nashville, USA, 2007.

[10] 盛若菁. 比喻构成中的类与语义区别[J]. 修辞学习，2002，(6)：23-24.

[11] 崔应贤. 也谈比喻和比较的区别[J]. 修辞学习，2005，(6)：56-60.