基于 CRF 的汉语动词“像”的比喻义识别*

李斌 于丽丽 石民

南京师范大学文学院，南京 210097

E-mail:gothere@126.com

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比喻标志（比喻词），成为一种较理想的用于计算机自动处理的比喻类型。本文着力于对动词“像”的比喻义自动识别，首先，利用程序提取出语料库中带有动词“像”的句子，人工判断是否为比喻句；然后用 CRF 模型进行训练和测试，开放测试 F 值达到了 83.3%，为隐喻计算的后续工作的展开奠定了的基础。

关键词：比喻义识别；隐喻计算；自然语言处理

Simile Identification of Chinese Verb “像” Based on CRF

LI Bin，YU Lili ，SHI Min

School of Chinese Language and Literature, Nanjing Normal University, Nanjing 210097

E-mail:gothere@126.com

Abstract: The computation of metaphors in Chinese is certainly an formidable task. Simile, with an obvious mark word, is an ideal type for automatic processing. This paper is focused on research on the automatic simile identification of phrases with the Chinese word “xiang”. Sentences with the verb “xiang” are firstly retrieved from Corpus and manually marked. Then the model of CRFs is applied to train and test on the data. Experiment achieved an acceptable F-score of 83.3% in open test. The research lays a sound foundation for further studies on metaphor computation.

Key Words: Simile Identification; Metaphor Computation; Natural Language Processing

一 引言

近二、三十年以来，认知语言学发展迅速，其研究方法大大拓宽了语言研究的视野，彻底改变了传统修辞学的研究思路，带来了一批重要的研究成果。“隐喻”被看做是一切语言中普遍存在的现象，也已成为当前研究的热点。它突破了传统修辞学的定义，不仅仅是一种修辞手段，更被视为是一种思维方式，其实质就是用一种事物来理解和表达另一种事物，进而把两种知识领域对应起来。

“明喻”一直是现代汉语修辞研究的一个热点，也因其带有明显的比喻标志而成为研究隐喻的一个较理想的突破口，无论是从传统修辞学角度，还是从现代认知隐喻立场，对这类比喻句的结构、特点、形成机制等各个方面的研究都有了重大突破进展。但从汉语信息处理、服务于计算机的角度对比喻句进行专门研究的文献尚不多见。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。“像”是较为典型的比喻词，出现的频率高，因此我

*本文系国家自然科学基金汉语隐喻理解关键技术研究（60773073）的研究成果之一。

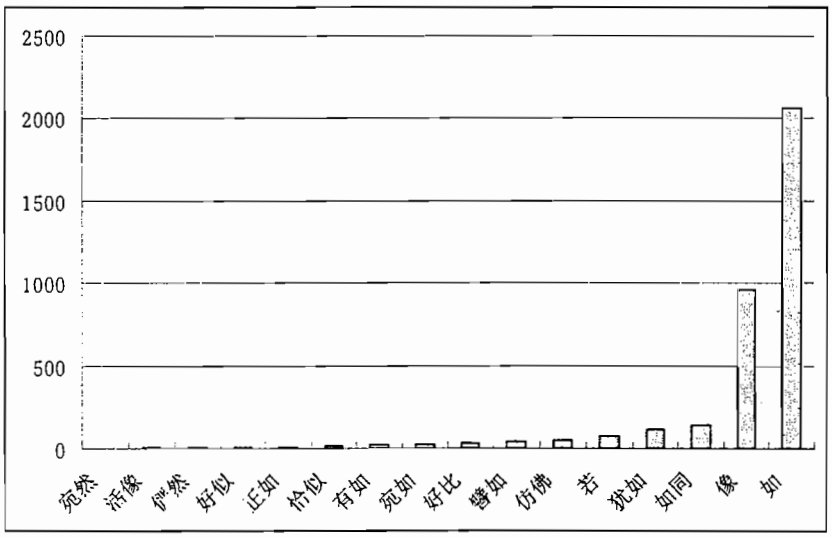
作者简介：李斌（1981—），男，汉族，博士，研究方向为计算语言学；于丽丽（1983—），女，汉族，硕士，研究方向为计算语言学；石民（1984—），男，汉族，硕士，研究方向为计算语言学。

们尝试用计算机自动识别出喻词为“像”的比喻句。该工作可以转化为动词“像”的比喻义和非比喻义的二值分类问题。这是对比喻句进行全面分析的基础，只有分析出真正的比喻句之后，才可以自动提取出句中的本体和喻体，进而有利于分析出从源域向目标域的认知转移的理据以及规律等，这对计算机理解句子，进行句法分析等具有十分重要的意义。其次，也有利于厘清汉语中概念隐喻的系统构成，为隐喻的认知机理的研究提供充分的语言学证据，进而促进计算机对自然语言的理解。

二 语料考察与比喻义的界定

我们所使用的语料是人民日报1998年上半年语料（下文简称“1998上”），该语料由北京大学计算语言学研究所人工标注了分词和词性标记信息。我们对常用的比喻动词做了词频统计，经观察发现，“像/v”的频次是比较高的。没有选择频次最高的“如/v”，只是出于“像”在人们心目中的典型性较高而已。

图1：常用比喻动词分布图



在《现代汉语词典》（第五版）中，“像”有如下词性和义项：

[1]名词，比照人物制成的形象：画像 塑像 雕像

[2]名词，从物体发出的光线，经平面镜，球面镜，透镜，棱镜等反射或折射后形成的与原物相似的图景，分为实像和虚像

[3]动词，在形象上相同或有某些共同点：他的面貌像他哥哥

[4]副词，好像，好像要下雨了

为此词性时，并不能充当喻词，在句中常常可以由”仿佛，好像，似乎”等词来替换。

[5]动词，比如、如。像大熊猫这样的动物要加以保护

作此词性解释时，由于是比照义，所以这类句子是非比喻句。

[6]名词，姓。

尽管“像”的词性和意义很复杂，但通过语料考察，我们知道只有“像”为动词时，才有可能构成比喻句，且动词“像”拥有比喻义和非比喻义两个大的义项。通过程序提取出里面所有带“像/v”的句子，然后人工标注出哪些是比喻义的，哪些是非比喻义的，作为训练和测试语料。

我们知道，现代汉语中大量的句子含有“像”字，但他们却构不成比喻。同时比喻成立与否，历来各家争论不休。综观前人对比喻句的研究大多是为人们理解比喻句子的内容、形式等服务的，讨论比较激烈，对于比喻句的成立条件及其判断标准，我们往往更多的是依靠语感来判断，因而带有极大的主观性。又因计算机缺乏人类所具有的关于语言和客观世界的知识及推理能力，困难更大。

根据盛若菁（2002）和崔应贤（2005）的观点，制定了本研究判别比喻句的标准，并给出了“1998 上”中相应句子作为例句：

1. 用于比喻的甲乙两事物所属之类差别小、距离近，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，定为非比喻句。

①贾里就像我们隔壁弄堂的初中生。

②小伙子长得白净、秀气，像个姑娘。

例①中，对象体是确指的，专指我们隔壁弄堂的那个初中生，二者不具有区别性语义特征，为非比喻句。例②中小伙子和姑娘性别不同，同属“人”的下位层次的分类，且小伙子本应健壮刚强的，与姑娘的白净秀气，具有区别性语义特征，这里正好作比形成反差，是比喻句。

2. 甲乙两事物属同类，但对象体带有典型性，具有极端性区别语义特征，且极端性受社会、历史、文化的影响，我们将其视为比喻句。

③周围的群众连连称赞:110 的民警像雷锋。

④毒物排放者，一个个像爵爷，在他们眼里，人命不值一顿筵席

⑤这姑娘长得像西施一样漂亮。

虽然这几例的两事物同属人类，但“雷锋”在“做好人好事”方面、“爵爷”在“蛮横”方面、“西施”在“漂亮”方面具有典型性，且具有极端性语义特征，两事物存在着极端义与非极端义的区别特征对立，故构成比喻。

3. “像”在句子结构中处于谓词性短语（VP）之前，只是用 VP 来表述某种形近的状态，一般定为非比喻句。

⑥雨后的大地像洗过一样。

⑦布勒伊拉造船厂鼓乐齐鸣，像过节一样。

⑧我的心里像悬着块石头，提心吊胆。

4. 两事物以现实的等同性作为认知基础，其中的“像”可以作为“像有”来理解。

⑨他像大山一样高。

⑩他像电线杆一样高。

对于这种类型，崔应贤（2005）遵从朱德熙（1981）论证“跟..... 一样”对同形词采用的同义注释转化的方法，认为“像”有“好似”义。“像”字本身带有心理预测的不确定性，以上两例可以用“有”或“像有”来理解，因此，在训练语料时我们将其定为非比喻句。

三 实验结果及分析

本文的实验使用了条件随机场（Conditional Random Fields, CRFs）模型，具体采用了 Taku Kudo 编写的工具包“CRF++0.50”进行训练和测试（下载地址：http://crfpp.sourceforge.net/）。

实验的语料是从中抽取出的 910 个含有“像/v”的例句，共包含 921 个“像/v”。其中，是比喻义的 497 个，不是比喻义的 419 个。随机抽取出 810 句做训练，100 句做测试。在测试语料中，有 50 句是比喻句，含有 52 个比喻义的“像/v”；50 句不是比喻句，含有 50 个非比喻义的“像/v”。

我们把“像/v”的比喻义和非比喻义的分类问题转化为序列标注问题。对于句中除“像”外的其他词语，标注为“X”；对于“像/v”的比喻义，标注为“Y”；对于“像/v”的非比喻义，标注为“N”。

我们分别尝试使用不同窗口长度的分词和词性标注信息进行对比试验，结果如下表：

表 1： “像/v”的比喻义和非比喻义识别结果

编号	特征	模板	比喻义（%）			非比喻义（%）			综合 F (%)
			P	R	F	P	R	F	
<1>	分词 2 元	W_i (i=-1,0,1), W_iW_{i+1}(i=-1,0), $W_{-1}W_1$	74.2	94.2	83.1	91.7	66.0	76.7	80.4
<2>	分词 3 元	W_i (i=-2,-1,0,1,2), W_iW_{i+1}(i=-1,0) $W_{-1}W_1$	74.2	94.2	83.1	94.4	68.0	79.1	81.4
<3>	分词 2 元 +词性 2 元	<1>+ T_i (i=-1,0,1), T_iT_{i+1}(i=-1,0), $T_{-1}T_1$	78.7	92.3	85.0	90.2	74.0	81.3	83.3
<4>	分词 2 元 +词性 3 元	<1>+ T_i (i=-2,-1,0,1,2), T_iT_{i+1}(i=-1,0), $T_{-1}T_1$	79.7	90.4	84.7	88.4	76.0	81.7	83.3
<5>	分词 3 元 +词性 3 元	<2>+ T_i (i=-2,-1,0,1,2), T_iT_{i+1}(i=-1,0), $T_{-1}T_1$	78.0	88.5	82.9	86.0	74.0	79.6	81.4

其中，W 表示分词后的词语，T 表示词语的词性类别。

在表 1 中，我们分别给出了“像/v”的比喻义和非比喻义的识别结果，并给出了综合 F 值。由于“像/v”只有两种标记状态“Y”和“N”，所以计算正确率 P 和召回率 R 时的分母是相同的，因此综合 P 值、R 值和 F 值也是相同的。综合 P、R、F 值=识别正确的个数/答案中正确的个数

*100%。

表 1 显示：一、比喻义的识别效果略好于非比喻义的识别效果。二、比喻义识别的召回率高、非比喻义识别的精确率高。这两点可能是由于训练语料中比喻义的数据比非比喻义的多造成的。三、综合 F 值在“分词 2 元+词性 2 元”的模板条件下最高，更复杂的模板反而使得各项指标下降，说明过多的语境信息未必好用。

需要补充的是，我们在上述实验中增加了一条后处理规则，以减少不必要的错误。仅使用分词信息标注时会发生一种特殊的情况，即 CRF 有可能标注出不是动词的“像”，这主要是由于词性标记错误造成的。如，“主格调/n 褐色/n 像/p 铺/v 满/a 了/u 鹅卵石/n 的/u 河床/n”。

词语	词性	答案标记	CRF 标注
主格调	n	X	X
褐色	n	X	X
像	p	X	Y
铺	v	X	X
满	a	X	X
了	u	X	X
鹅卵石	n	X	X
的	u	X	X
河床	n	X	X

四 结论及未来工作

本文基于 CRF 模型，对动词“像”比喻义的自动识别进行了初步的探讨和研究，通过赋予模型不同的上下文特征进行语料训练和测试。实验结果表明，基于分词二元和词性标记二元特征进行机器学习进而测试语料的方法具有较高的识别比喻正确率，综合 F 值达到 83.3%。

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，挖掘文本中其他的句法语义特征、局部统计特性等，进一步提高识别的效果。（2）在前面研究基础上，扩大比喻词的研究范围，能够从文本中正确识别出其中的比喻句，并力求进而提取出其中的本体与喻体，以为隐喻研究中的由源域向目标域的映射等结构建立起语义模式和框架体系。

参考文献

[1] 崔应贤. 也谈比喻和比较的区别[J]. 修辞学习, 2005(6).

[2] 冯广义. 汉语比喻研究史[M]. 湖北：湖北教育出版社，2002.

[3] 李济中. 比喻论析[M]. 河北:河北大学出版社, 1995.

[4] 李行健编. 现代汉语规范字典[M]. 北京：语文出版社，1998.

[5] 刘大为. 比喻、近喻与自喻辞格的认知性研究[M]. 上海：上海教育出版社，2001.

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

[7] 王治敏. 隐喻的计算研究与进展[J]. 中文信息学报. 2006(4).

[8] 袁晖. 比喻[M]. 合肥:安徽人民出版社, 1982.

[9] 张明冈. 比喻常识[M]. 北京：北京出版社，1985.
[10] 中国社会科学院语言研究所词典编辑室编. 现代汉语词典(第五版)[M]. 北京:商务印书馆, 2005.
[11] 朱德熙. 说“跟……一样”[J]. 汉语学习. 1981(1).