中 文 信 息 学 报
第 18 卷 第 4 期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol. 18 No. 4

文章编号:1003 - 0077(2004)04 - 0031 - 06

基于机器理解的汉语隐喻分类研究初步*

杨 芸,周昌乐,王雪梅,戴帅湘

(厦门大学 计算机科学系人工智能研究所,厦门 361005)

摘要:本文将汉语隐喻分类计算模型的研究引入汉语的机器理解当中,通过对大规模汉语隐喻语料的研究分析,结合汉语隐喻的认知特征,笔者构建了一套基于理解的汉语隐喻分类体系。分类主要以汉语隐喻句中本体和喻体的内在相似性作为切入点,从隐喻理解的方式、理解的难易程度以及理解所涉及的相关知识结构等方面进行综合分析,同时,辅以真实语料的统计分析,对分类的合理性作出了验证和修订,最终给出了基于理解的汉语隐喻分类体系,并对该体系作出了语言学上的比较和解释。

关键词:计算机应用;中文信息处理;隐喻;分类;计算模型;相似性

中图分类号:TP391.2 **文献标识码**:A

Research into Machine Understanding-based Classification of the Metaphor of Chinese

YANG Yun, ZHOU Chang-le, WANG Xue-mei, DAI Shuai-xiang

(Computer Science Department Xiamen University, Xiamen 361005, China)

Abstract: This paper introduces computational model of Chinese metaphor in machine- understanding of Chinese. By analyzing large-scale Chinese metaphor samples, we have classified Chinese metaphor based on understanding. The cognitive features of Chinese metaphor are also considered to improve our classification. The classification focuses on the similarity of the tenor and vehicle in a metaphor, showing the mode and difficulty of metaphor understanding. The relevant knowledge in metaphor understanding is also discussed. The classification is statistically verified. What's more, a program is developed to validate the rationality of our classification. Finally, a system of understanding-based classification of Chinese metaphor has been put forward. A comparison with other classification is listed. And a linguistic explaining of the system is given at the end of the paper.

Key words: computer application; Chinese information processing; metaphor; classification; computational model; similarity

1 引言

自然语言的隐喻性特征愈来愈受到广大学者的认同和重视。汉语作为在世界上使用最为广泛的自然语言之一,其隐喻意义的机器理解日显其极大的研究价值和意义。然而目前,我国汉语隐喻计算化的研究尚属空白,汉语隐喻的机器识别及理解的研究无疑将推动汉语机器翻译进入一个新的阶段。

* **收稿日期**:2003 - 12 - 08
基金项目:福建省自然科学基金资助(A0210005)
作者简介:杨芸(1980 —),女,研究生,研究方向为计算语言学,自然语言处理,汉语隐喻计算模型.

1.1 基于理解的汉语隐喻分类的必要性

要解决机器对汉语隐喻的自动识别和翻译问题,第一步工作就是进行隐喻的分类。在海量的汉语语言文字中直接进行隐喻识别非常困难,但是在一种合理的隐喻类别体系的基础之上,根据每一类别的基本特征去实现隐喻的机器识别则相对容易;与此同时,如果这样的类别是基于理解划分的,则对于隐喻的机器理解而言,是一项很有意义的基础性工作。

2 分类标准

2.1 隐喻的理解

隐喻的一个重要语义特征是喻体与本体之间具有明显差异或者语境之间存在尖锐冲突,完全相同的事物之间显然不能构成隐喻。另外,本喻体之间必须存在相似性,所谓"同从异出",相似性是隐喻赖以成立的基本要素。隐喻意义的理解实际上就是将源领域(喻体)的经验映射到目标领域(本体),从而达到重新认识目标领域特征的目的,而映射函数就是基于本体和喻体之间的某种相似或相关性。这里的相似性主要是指本体和喻体所指称的事物或事件属性的相似。

2.2 分类标准

基于以上的分析,我们得出影响汉语隐喻理解的四大要素:本体、喻体、相异点和相似点。基于理解的隐喻分类就是结合隐喻理解的四大要素,以本体和喻体的相似性为切入点,并结合所需访问的相关知识信息进行汉语隐喻类别划分(将不再考虑隐喻词)。

3 基于理解的汉语隐喻类别体系

3.1 研究对象

对汉语隐喻分类的研究我们采取由简入繁的方法,本文中主要研究汉语语言文字中常用和常见的隐喻句,对于较为生僻的隐喻用法暂时不予考虑。隐喻语料主要来自隐喻和修辞学著作以及中文系研究生提供的一个隐喻句库(约 8 万字)。

3.2 类别体系

经过封闭语料库的测试和统计角度的验证与修改,我们最终提出了九大类隐喻,分别阐述如下:

(1) 明确隐喻:隐喻理解的四要素全部出现在句子当中。在相似性直接在隐喻句中出现的情况下,确定相似性的位置特征是重要的(相似点在句中的位置可参见图 2 - 1)。

(2) 特性隐喻:在特性隐喻中相似性要素没有出现,但是通过喻体事物的最显著的特征体现出来。这里的相似点是认知过程中人们对喻体事物最具代表性特征的固定认知。机器在理解过程当中只要访问喻体事物的一维特征向量,进行对句子隐喻意义的推理即可。

该类别从逻辑上可形式化表示为:(∃F)(∃G)(SIM(F(T),G(V)))

其中,F 表示本体的属性,G 表示喻体的属性,T 表示本体,V 表示喻体。SIM() 为相似性计算函数。

例如:小明(T)是只猴子(V)。

由人对"猴"的固定认知可以推断出该句子的隐喻意义"小明非常活泼、聪明、灵活、好动还有调皮等等。"而不会出现"小明是灵长目动物"这样的解释。

(3) 相关隐喻:抽象和具体的事物之间的隐喻。这里相似性也是隐含的,但不是喻体的显著属性,而是喻体在某种情况下所起的作用,与本体在某种情况下所起的作用相似。

将该类别从逻辑上可形式化表示为:

$(\exists F)(\exists G)(\exists H)(SIM_H(F(T),G(V)))$

其中:F 表示本体的属性,G 表示喻体的属性,T 表示本体,V 表示喻体 H 表示某种行为范畴。

例如:岁月(T)如鞭(V)。

"岁月"是抽象的,没有长度、形状或者性格,用"鞭"的显著特征不能很好地解释之。但是"鞭"被使用是为了进行某种"监督"行为,将岁月比作"鞭"实际上就是用到"鞭"的监督、督促的功能。因而得到该隐喻句的理解:"岁月让人感到紧迫,催人奋进"。

理解此类隐喻句,需要访问知识库中更进一层的信息,并不限于该事物表面显著特征,而是涉及到该种事物在某种环境或作用于某种对象所起的作用。

(4)事件隐喻:本体或喻体是由完整的句子所描述的事件构成。在机器已经对句子的字面意思理解的基础上,分别建立两个事件相关的 n 维属性空间,从属性空间中找到本体喻体相互映射的相似点,从而进行理解。

例如:步入教堂大厅,仿佛一步踏入了幽深浩渺的苍穹。

(5)关系隐喻:本体和喻体在句中是隐含的或者不明显的,隐喻意义通过某种动作行为或表语所指示的事物状态之间的相似体现出来。实际上是一种动作或状态关系上的隐喻,是一个关系到另一个关系的映射。

句型可能为:

"x 是 F"或"xFs"　　其中 x 是主语,F 为形容词或者谓语。

形式化表示为:

$(\exists G)(\exists y)(SIM(G(x),F(y)))$

其中:G 表示某种特性,y 表示 F 动作或状态的实际施动者。

例如:这个问题太痛苦了。主语"问题"具有某种属性 G,与隐含的实际施动者 y"y 是'痛苦'的"有相似之处。

又如:船犁大海。动词"犁"的施动对象原本是土壤,而船的实际行为是航行,施动对象是海面。这里隐喻意义是通过动作与对象之间的关系存在相似而产生的,见图 1。

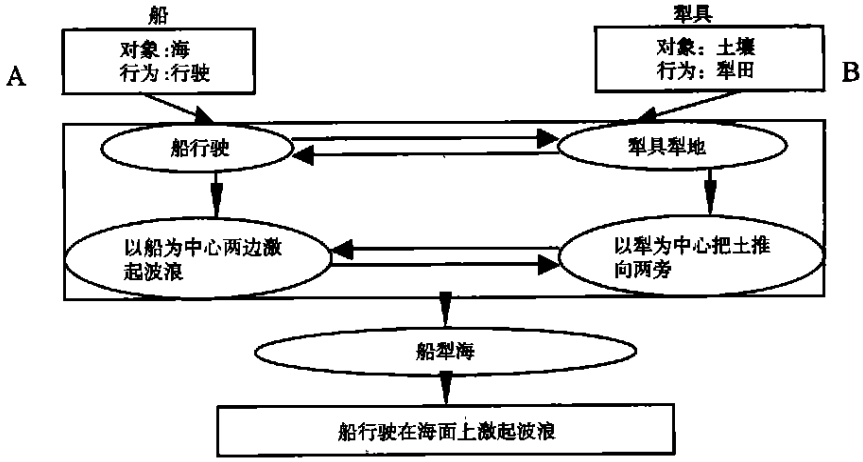

图 1　动作对象关系整合图

(6)位移式隐喻:本体不出现,隐喻关系通过喻体实际处所改变的冲突体现出来,冲突的位

置正是本体所处的位置,喻体的特征即为本喻体之间的相似性。

例如:东方的威尼斯、人间地狱等。

(7)比拟:将反映本体的语言单位在逻辑和感情意义上当作反映喻体的语言单位,按照喻体的逻辑范畴来搭配词语,产生转义。在这种情况下,本体和喻体之间是具有某种程度上的相关而形成隐喻。本体本不具备喻体的某种属性,而是将喻体中的类似属性提取出来,经过相关性的词语转换加入到本体的特征属性库中,让它具有该相关特性。

形式化表示为:

$(\exists G)(ADD(G(x),F(人)))$或者$(\exists G)(ADD(G(y),F(物)))$

其中:G表示某种特性,F表示表语形容词或谓语,x表示物,y表示人。

例如:木棉在风中起舞。他脑袋坏掉了。

(8)借喻:只有喻体出现,以显著的或有代表性的事物(喻体)直接代替不显著的或非代表性的事物(本体)的隐喻。

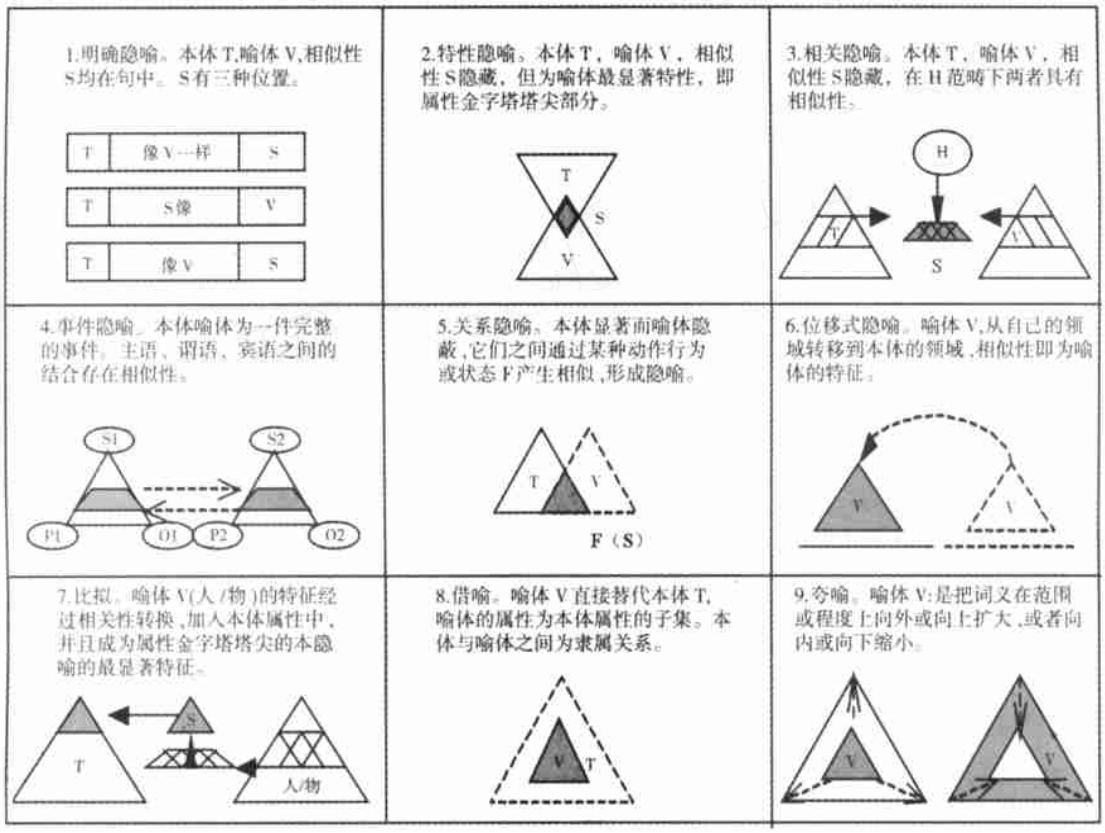

图2 图中三角形表示本体喻体事物的属性金字塔,从塔尖到塔底事物属性特征从显著到不显著变化。其中,T表示本体,V表示喻体,S和图中阴影部分表示相似性。

借喻的机器理解势必会牵涉到一个知识网,在知识网中查寻喻体特征去获取相似性。这种隐含的本体与显式喻体之间的相似性往往表现为创造性的相似,而且与句子以外的上下文有密切的关系,属于一种很特殊的隐喻。

例如:你都可以走进属于自己的土壤,或耕耘或播种。

(9)夸喻:没有本体而且喻体非常特殊,是词义跨范畴或跨认知域的隐喻。夸大是把词义

在范围或程度上向外或向上扩大 ,而缩小夸则是向内或向下缩小。夸张中本体隐去 ,所使用的夸张词语就是喻体。理解这样的隐喻句需要事先确定原词意义作用范围标准。

例如:白领丽人的泡沫星子差点从电话听筒里喷过来。

以上的类别体系是在假设机器已经能够识别句子的字面意义基础上进行的,研究的隐喻对象均较为常见。这个类别体系是一个基础的体系,更加复杂的隐喻现象可以在这个类别体系中找到原型进而进行简化和理解。图 2 为九种类别的本体、喻体和相似性关系结构图。

4 分类体系合理性验证

分类体系是否合理关系到后续对隐喻句的机器识别以及理解的顺利进行。"分类 ——统计验证 ——修改"是一个循环工作的过程。本文选择人民日报 1998 年一月标注语料库作为统计验证的样本,只保留隐喻含量丰富的文娱副刊的内容,规模约 43 万字符。

4.1 类别统计结果

我们分别对语料库中的隐喻句及其本体、喻体、喻词、类别等信息进行了标注,并发现隐喻句含量占整个文娱副刊内容的三分之一以上,一共统计了 1341 条隐喻句,它们的类别频度如表 1(限于篇幅,其他隐喻信息的统计表不在此引用):

表 1 隐喻句类型频率表

类别号	1	2	3	4	5	6	7	8	9	NUL
类别名	明确隐喻	特性隐喻	相关隐喻	事件隐喻	关系隐喻	位移式隐喻	比拟	借喻	夸喻	未知类型
使用频度	12.90%	12.61%	21.11%	12.61%	8.50%	6.45%	9.97%	9.97%	5.57%	0.29%

4.2 合理性说明

从统计结果来看,在封闭测试下,各类别隐喻句在该中等规模语料库中的分布具有一定的均衡性,即呈现一定的分布规律,以上类别基本能够穷尽,只有极少数隐喻句类别不清晰。每个类别的出现频度也较为均匀,从统计角度讲,结果具有可信度和研究价值。

4.3 存在的问题

本分类体系是对常见汉语隐喻的分类,这个划分仍然比较大,形式化存在不够充分的地方,个别类别之间存在交叉。在标注的过程中是按主要类别进行取值,忽略了这种交叉性。

整个分类在较为理想的情况下进行,对简单隐喻计算模型的设计是适用的,但是更加复杂灵活的汉语隐喻的机器理解的支持度有待提高。

5 分类系统的语言学比较

5.1 与传统修辞学隐喻类别的比较

从分类依据和作用来看,修辞学上隐喻分类主要站在作者的角度,依据的是隐喻在文章中的表达效果,追求的目标是尽可能让文章表达方式多样,引人入胜;本文的分类主要站在自然语言机器理解的角度,依据的主要是隐喻理解过程的复杂程度,追求的目标是尽可能地让机器迅速有效地从隐喻的字面意义过渡到隐含意义,从而有的放矢地对隐喻进行"阅读"和"理解"。

此外,本文的分类中把"比拟"和"夸张"也纳入到隐喻的分类体系,因为它们也是意义隐含的表达方式,也无法从字面意义上进行汉语的机器理解,因此,研究汉语隐喻机器理解的计算模型必须要将此类情况考虑进去。

5.2 与 Searle 隐喻解释原则的比较

著名语言学家 Searle 在言语行为研究中曾总结了英语隐喻解释的八条原则,分别是:

a. **定义原则**:Sam is a giant. →Sam is big.

b. **性质规则**:Sam is a pig. →Sam is filthy ,gluttonous ,and sloppy , etc.

c. **固定认知规则**:Richard is a gorilla. →Richard is mean ,nasty ,violent.

d. **感觉规则**:Sally is a block of ice. →Sally is cold.

e. **社会地位规则**:You have become an aristocrat. →You have just been promoted.

f. **状态规则**:His brain is addled. →He seems to be silly.

g. **关系规则**:Sam devours books. →Sam reads books very carefully.

h. **术语规则**:the US government →the White house

英语和汉语的隐喻思维存在较大的差别,较之英语,汉语的语法和隐喻表现手法要复杂得多,首先,a,b,c,d,e 规则之间界限存在模糊。山姆是个巨人,并不一定仅仅说他高大,或许他并不高大,但却伟大。在汉语里,说一个人是巨人,往往更多的是说他伟大。另外,pig,gorilla , ice 作为喻体不存在本质的区别,在隐喻解释的时候都是采用人们对这些事物的显著特征的固定认知。这几个原则在本文中体现为"特性隐喻"类型。f 和 g 规则实际上也可以合并,即为本文所讨论的"关系隐喻"。Searle 的八条规则中对于具体事物和抽象事物隐喻关系没有专门的说明,而且对事件、比拟、位移式隐喻、借喻和夸张都没有提及,而这些隐喻用法恰恰是汉语中常见的用法,因此不能缺少。此外,考虑到术语可以不通过推理而直接查找术语库来进行匹配消解,因而本文没有考虑"术语"类。

本文总结的汉语隐喻分类规则,从一定程度上弥补了 Searle 分类上的不足,类别之间的界限更加清晰,更贴近于汉语隐喻机器理解的研究。

6 结论

现代汉语作为当今世界最复杂的自然语言之一,其隐喻现象繁多而灵活,复杂程度也可想而知。本文首次将汉语隐喻的计算模型引入到汉语的计算机理解当中,拓宽了汉语机器理解的研究范围。把隐喻计算模型引入到汉语的机器理解中,无疑将会给整个汉语语篇机器理解带来一个质的飞跃。我们通过对具有认知功能的人的隐喻理解工作机制的剖析,初步探索了机器理解隐喻的方法,提出了基于理解的隐喻分类标准并对类别进行了合理性验证。该研究初有成效,为汉语隐喻计算模型的研究迈出了坚实的第一步。

参 考 文 献:

[1] John R. Searle. Expression and meaning: studies in the theory of speech acts [M] Foreign Language Teaching and Research Press. Cambridge University Press. 北京:外语教学与研究出版社. 2001.

[2] James H. Martin. A Computational Model of Metaphor Inerpretation [M]. Academic Press, Inc. Harcourt Brace Jovanovich, Publishers. 1990.

[3] 束定芳. 隐喻学研究[M]. 上海:上海外语教育出版社,2000.

[4] 冯广艺. 汉语比喻研究史[M]. 武汉:湖北教育出版社,2001.

[5] 杨成虎. 隐喻研究背景下修辞格的重新归类问题[J]. 四川外语学院学报. 2002(1).

[6] 俞士汶,等. 北京大学现代汉语语料库基本加工规范[J]. 中文信息学报. 2000,16(5).