面向自动理解的汉语明喻句的可计算性考察*

宋 纯[1]，李 斌[1,2]，曲维光[1,3]，陈小荷[1]

[1]南京师范大学 文学院，南京 210097

[2]南京大学 计算机软件新技术国家重点实验室，南京 210093

[3]南京师范大学 计算机科学与技术学院，南京 210097

E-mail: songchun007@163.com

摘 要：隐喻的计算语言学研究主要存在两个问题：隐喻理论多样且差异较大；隐喻知识库和语料库的可计算性不足。为解决隐喻理论与计算的衔接，寻找面向计算的隐喻分析框架，本文提出了利用易收集、本体喻体喻底易区分的明喻句作为媒介，通过分析其概念域的整合方式为其他隐喻方式的研究提供理论和计算依据的方法。语域受限的封闭语料穷尽分析试验表明，属性明喻句可通过凸显特征来计算；动作隐喻方式复杂，其可计算性比较低，并非现有知识库所能支撑。最后探讨了明喻计算的界限问题。

关键词：隐喻计算；明喻；概念隐喻；理解；生成

Computability Investigation on the Automatic Understanding of Chinese Similes

Song Chun[1], Li Bin[1,2], Qu Weiguang[1,3], Chen Xiaohe[1]

[1] School of Chinese Language and Literature, Nanjing Normal University, Nanjing 210097

[2] State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093

[3] School of Computer Science and Technology, Nanjing Normal University, Nanjing 210097

E-mail: songchun007@163.com

Abstract: It is difficult for computational linguistics to conduct researches on metaphors, due to the various metaphor theories which are hard to choose as the foundation of computation. And the current metaphor knowledge base and corpus are not fit for metaphor understanding and generation. Thus, we analyze the Chinese simile sentences which are simple in structure and easy to collect. By the sentences extracted from the limited corpus, we find the similes of attributes can be computed by salient features, while the similes of verbs are very complex which cannot be supported by the current knowledge base. At last, we discuss the limitations of metaphor computation.

Keywords: metaphor computation; simile; conceptual metaphor; understanding; generation

1 前言

20 世纪末以来，随着认知语言学的理论发展和对文本中复杂现象的关注，隐喻研究逐步成为计算语言学的研究热点，但始终未获得实质性突破，无法用于大规模语料处理。隐喻研究的难点在于两方面：首先隐喻理论多样，理论之间差异较大，让计算研究者难以选择。其次，面向计算的隐喻知识库和语料库，在构建过程中，单纯依据某一种隐喻理论，并未充分考虑到计算的要求。因此，如何将理论和计算衔接起来，寻找面向计算的隐喻分析框架，成为隐喻计算的基础性研究。本文在封闭语料基础上细致地分析了 400 多句明喻句的隐喻方式，从可计算的角度，使用主流的隐喻理论对语料进行穷尽性分析，给出明喻句的可计算范围和方法。

* 本文承国家社科基金 10CYY021、07BYY050，南京大学计算机系重点实验室招标课题 KFKT2011B03，国家自然科学基金 60773173 的资助。

2 相关研究

国际上对隐喻句的分析主要采用 Lakoff & Johnson（1980）的概念隐喻理论和 Fauconnier（1994）的概念整合理论。前者是概念域之间的互动关系，强调源域向目标域的映射。后者则认为人们在思考或交谈时不断创建的心理空间是概念的整合。Grady（1999）对比了两种理论在复杂隐喻句“The surgeon is a butcher”上的解释能力，认为概念整合理论更清晰地揭示了句子意义的具体产生过程。不过，概念隐喻理论仍适用于大多数的隐喻用法描写，目前两种理论处于并存状态。

Lakoff（1980）针对英语明喻句的研究，建立隐喻语料库进行专门分析的是。其局限在于仅对个案进行分析。以概念隐喻理论为指导，计算语言学界也建立了一批具有代表性的英文隐喻数据库：（1）Master Metaphor List；（2）Sense-frame；（3）MetaBank；（4）Metalude；（5）Hamburg Metaphor Database；（6）ATT-Meta。在隐喻语料库基础上创建隐喻计算模型，其方法主要有五种：（1）基于优先语义的方法；（2）基于知识表示的方法；（3）基于实例的模型；（4）基于类比推理和逻辑推理的方法；（5）基于语料库统计机器学习的方法。然而这些模型对于明喻句并没有系统处理，尤其在处理谓词短语混合型的明喻结构上遇到困难。

汉语隐喻句研究方面，杨芸（2008）基于概念隐喻理论，围绕句子级别的汉语隐喻计算模型进行研究，不过这种分析在颗粒度上不够细致。贾玉祥（2009）也采用了概念隐喻理论，采集目标域的大量属性词语。国内隐喻计算研究是在依据某一隐喻理论或句法理论基础取得了一定成果，研究套路遵循标注语料训练参数和模型的机器学习，隐喻计算的基础性研究仍有待深入。从已有研究成果可以看出，概念隐喻理论和概念整合理论进行分析的根本假设是：人的隐喻模式是多样的。因而要针对不同的模式建立不同的分析和计算方法。隐喻数据库和隐喻计算所使用的语言理论主要是概念隐喻理论，概念整合理论在形式化方面存在较多困难，所以暂时不能进行细致地分析和计算。

3 明喻句的数据来源和标注原则

汉语的明喻句模式简单，表现为“本体+喻词+喻体+喻底”，喻底为喻体的凸显特征，在句中可出现可不出现。喻词较丰富，由“像”、“好像”、“仿佛”、“宛如”、“犹如”等充当。喻词为“像”的句子（以下简称“像”型明喻句）出现频率最高，但其中的绝大多数并不是明喻句，而是比较句，如“小明像他爸爸一样高。”李斌（2008）从北京大学计算语言学研究所人工标注的《人民日报》上半年语料中，抽取了 1586 个带“像”句，手工筛选出明喻句 512 个，并标注了每个句子的本体、喻体和喻底。其制定的标注原则是：本体、喻体、喻底是名词结构时，分别标为 b、y、s，如“[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”；本体、喻体是动词结构时，分别标为 bv、yv，如“使/v [读者/n 在/p 接受/v 国策/n 教育/vn 的/u 时候/n]bv 就/d [像/v]c [看/v 卡通/n 漫画/n]yv 一样/u [轻松/a]s”；当本体、喻体做句中谓词主语，且为名词性成分时，标为 bs、ys，如“[我们/r 新/a 的/u 线路/n 和/c 我们/r 的/u 虹桥/ns 旅行社/n]bs [像/v]c [朝阳/n]ys 一样/u [愈/d 升/v 愈/d 高/a]by ”。这样的标注区分了本体、喻体在句中地位和性质，有助于明喻句分类分析和计算机识别。

本文研究对象的选择在已有基础上，去除“它像鲜花一样”本体为代词且喻底隐藏的句子。这类语句喻体含义不确定。经过统计这类句子共 75 句。用于本文研究分析的明喻句共 437 句。

4 明喻句的分类统计及分析

4.1 概念隐喻理论对于明喻句的解释

Lakoff&Johnson（1980）认为隐喻的形成过程就是源域向目标域的投射过程。在简单的明喻句

中，概念隐喻理论只需解释源域向目标域的特征投射。如喻底在句中出现的句子，“一枝红红得像玛瑙”。源域“玛瑙”，目标域“一枝红”，源域向目标域投射的特征是“红”。在喻底隐藏的句子中，“教室像蒸笼一样”。源域“蒸笼”，目标域“教室”，源域向目标域投射的特征虽未在句中出现，然而，根据源域“蒸笼”的凸显特征，可以推断出此处映射的内容是“热”。喻底隐藏的情况比较复杂，也有存在喻底难以归纳的情况。根据语料库的统计，在所有喻底隐藏的句子中，83%的句子我们可以将喻底的凸显特征补充出来，对于剩下 17%的句子来说，包括了喻底难以用一个词来归纳的情况，如“特警的手像刀一样”，心理画面非常形象，但难以用词语描述；另一种情况，本体的视觉形象难以令读者感受，只有从整体隐喻中才能把握句子的意思，我们在下文中称此为“引入式”明喻句，如“雾中的江桥，像一道堤，雾都从堤上溢上天空。”

表1 “像”型明喻句源域、目标域投射类型统计表

喻底出现（简单和复杂两种）					喻底隐藏				
源域	目标域	凸显特征	比例	特点	源域	目标域	凸显特征	比例	特点
玛瑙	一枝红	红	0.351	简单	蒸笼	教室	（热）	0.32	喻体明确
火柴	生命	（为人民）燃烧	0.263	复杂	刀	手	难以归纳	0.006	喻底模糊
					江水漫堤	雾溢江桥	散在句中	0.06	喻底分散

用概念隐喻理论对谓词短语混合型的明喻结构句进行解释，以“生命像火柴，划着了就要为人民燃烧”为例，源域是“火柴”，目标域是“生命”，源域将“可燃性”、“发光发热”等特征投射给目标域“生命”，这些特征就是源域的凸显特征，目标域“生命”具有了源域的凸显特征后，在喻体的谓词性短语部分，按照自身特征安排宾语，就成为“划着了就要为人民燃烧”形成的原因。这类复杂类型的句子占总数的 26.3%。

4.2 明喻句的分类统计

对于明喻的理解和生成而言，明喻形式化的关键是对隐喻表达的具体特征，包括本体和喻体表达的结构形式化，本体和喻体之间语义关系的形式化。明喻一般被认为结构简单，易理解。然而，通过本文的观察发现，明喻格式并不简单。为便于分析，我们从句法形式入手，做了几点区分。（1）在明喻句中，根据喻底是否出现，区分出喻底出现的句子和喻底隐藏的句子；（2）根据喻底（包括隐藏喻底）是何种词性的短语，分为属性隐喻 AP（喻底为形容词）和动作隐喻 VP（喻底为动词短语，包括动作、事件）。

表2 “像”型明喻句喻底特征分类表

类型		句例	比例	
属性隐喻 AP	喻底出现	60	0.137	0.386
	喻底隐藏	109	0.249	
动作隐喻 VP	喻底出现	195	0.446	0.614
	喻底隐藏	73	0.168	

从表 2 可以看出，喻底是动词短语的动词性明喻句的比例是很高的，占 61.4%，而这种明喻格式在过去的研究中恰恰是偏弱的。

如上表所示，在 AP 和 VP 中都存在喻底出现和喻底隐藏的情况，喻底即为喻体的凸显特征。喻底出现，即喻体的凸显特征表现在句中，不论是 AP 或是 VP，我们在计算时，均可以根据句中的凸显特征，在知识库中提取相关内容进行匹配。并且 AP 和 VP 中，都存在先将本体主语和喻体主语做明喻，后接一个谓词性短语的句子，分为引入式明喻和谓词混合式明喻两种形式。引入式

明喻从形式上看，喻体和本体均没有完全出现，并且喻体和本体的形象分散在句子中。谓词混合式明喻即喻底在句中出现，且喻底为喻体的谓词性凸显特征，属于VP，谓词性短语中的宾语与本体有关，与喻体无关。

由于所有的明喻句中均涉及喻底的凸显特征，根据喻底的出现与否及出现位置将437条明喻句分为3大类7小类。

4.3 明喻句的分类分析

4.3.1 喻底出现

“[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。”

本体是“我们两国之间的关系”，喻体是“黄金”，喻底“珍贵”是喻体的凸显特征，在句中做谓语，同时修饰本体和喻体。在语料调查中，我们发现，喻底在句中的位置有多种可能，做谓语、定语、补语、状语。各类型所占比例如表3所示。

表3 喻底成分统计表

成分	谓语	定语	补语	状语
比例	0.823	0.064	0.106	0.007

这类明喻句，可以根据句中出现的喻底从形式上直接进行识别。具体做法为：从句中抽取喻底，搭配本体和喻底进行理解。

4.3.2 喻底隐藏

当喻底在句中不出现时，就是喻底隐藏。喻底隐藏的明喻句具体有下几种情况：

（1）补充喻底

“他们/r [像/v]c [盼/v 星星/n 、/w 盼/v 月亮/n]yv 似的/u [希望/v 看到/v 艺术团/n 的/u 演出/vn]bv 。”

该句本体是“希望看到艺术团的演出”，喻体是“盼星星、盼月亮”，隐藏的喻底可以补充出来，即“渴望”。喻底从知识库中补充得出，如该句的识别，就是在已有的知识库中抽取与喻体是“盼星星、盼月亮”相搭配的形容词或动词成分。计算机在知识库中识别与“盼星星、盼月亮”同现频率最高的为“渴望”，并且“渴望”亦能贴切地描述本体的特征，因而“渴望”即为本句省略的喻底。通过知识库的搜索，对此句进行理解。

（2）喻底无定

① 喻底难以用一个词归纳

特警/n 之/u “/w 特/Ag ”/w ，/w 就/d 在于/v [他们/r 的/u 手/n]b [像/v]c [刀/n]y 、/w [拳/n]b [像/v]c [锤/Ng]y 、/w [指/Ng]b [像/v]c [钩/n]y 。

该句有三个独立的明喻句，“手像刀”，“拳像锤”，“指像钩”，“手”、“拳”、“指”是本体，“刀”、“锤”、“钩”是喻体。喻体具有明显的凸显特征，所以句中可以省略喻底。“手像刀”、“指像钩”：本体、喻体间的形象性，与上类相同，喻体的特征可以被计算机抽取出来，这一类是可以进行计算的。然而在实际分析中，我们发现知识库中与喻体同现频率最高或较高的名词、动词、形容词成分，都不足以贴切地描述本体“手”、“指”的特征，因而这种句子，我们认为是难以用一个词来归纳本体特征的，喻底很难补充出来。因而可计算性不高。不过这类句子在封闭语料库中所占的比例很小，仅为0.6%。

② 引入式明喻

[雾/n 中/f 的/u 江/n 桥/n]b ，/w [像/v]c [一/m 道/q 堤/n]y ，/w 雾/n 都/d 从/p

堤/n 上/f 溢/v 上/v 天空/n 。/w

本体主语	本体谓语	本体宾语	喻体主语	喻体谓语	喻体宾语
雾	溢	江桥	（江水）	（漫）	堤

从形式上看，这类句子由两个主谓分句组成。“江桥-像-堤”是一个形式确定的明喻，“江桥”是本体，“堤”是喻体。但从整个句子来看，是将“江水漫堤”作为喻体，“雾溢江桥”作为本体。由于“雾溢江桥”的视觉形象难以令读者感受，故而使用了“江水漫堤”的喻体形象。从形式上看，喻体并没有全部出现，只用了“堤”，省略了“江水”和“漫”；本体也没有全部出现，或者说被分散到句子的不同部分中，只有从整体隐喻中才能把握“雾溢江桥”的意思。对于引入式明喻，目前还没有较好的算法可以处理，而这部分占有的比例为6%。

4.3.3 谓词混合式明喻

谓词混合式明喻句中，喻体谓词短语由“谓+宾”构成，谓词均为喻体凸显特征，宾语与本体、喻体的关系则有三种情况：

表4 谓词性短语中宾语搭配表

合成谓语	合成宾语	例句	统计	
			句例	比例
喻体	本体	[人民日报]bs[像]c[磁铁]ys [吸引着]yv [我的心]yo。	45	0.424
喻体	喻体	[杨育才、侦察员、战士]bs[像]c[钢刀]ys[直捣]yv[敌人的心脏。]yo	25	0.235
喻体	本体、喻体	[古枫]bs[像]c[母亲]ys[一样]u[给予]yv[我不少温暖时光。]yo	36	0.339

第一种情况，“人民日报像磁铁吸引着我的心”，本体是“人民日报”，喻体是“磁铁”，“吸引着我的心”是谓词性短语，而宾语“我的心”只能和本体“人民日报”组成语义搭配。识别这类句子时，抓住凸显特征，在例句中为“吸引”，通过凸显特征理解本体与宾语的关系。第二种情况，“杨育才、侦察员、战士像钢刀直捣敌人的心脏。”谓词性短语部分的宾语“心脏”只能和喻体“钢刀”构成语义搭配。“直捣心脏”已经超出了字面义，产生了隐喻含义，难以根据句中各成分的意义和关系进行计算。第三种情况，“古枫像母亲一样给予我不少温暖时光。”谓词性短语部分的宾语“温暖时光”既可以受本体支配，又可受喻体支配。与第一种情况类似，具有较高的可计算性。

通过上述分析，从根本上看，喻底是一种认知上的相似性，用喻体的凸显特征来唤起读者对本体的想象，以把握说话人对本体的主观理解。

5 明喻的生成和理解机制

我们对于明喻的生成和理解机制可以做简要的归纳：（1）对于喻底出现的句子，提取凸显特征为该喻底的词语；（2）在谓词混合式的部分明喻句中，利用凸显特征和宾语进行交集运算。计算机在生成时，对于谓词混合式明喻句，可以根据主引语部分的谓语，在知识库中提取相应的具有该种凸显特征的名词。该类隐喻，在概念整合上较为复杂，但在形式上却比较简单。如，以动词“吸引”为凸显特征的名词还有“磁石”、“地球”，这些名词也可完成明喻句的生成。合成宾语与本体和喻体的关系，可以在生成时进行辅助计算。如“[古枫]bs[像]c[母亲]ys[一样]u[给予]yv[我不少温暖时光。]yo”根据谓词“给予”，知识库中可以找出若干条具有“给予”凸显特征的名词，通过计算宾语“时光”和名词间的关系，筛选出相似度高的名词，从而完成匹配。

对于喻底隐藏的句子，若凸显特征可以明确地补充，计算时需要将本体的语义类和喻体凸显特征的宿主属性进行交集运算；而对于凸显特征难以用一个词进行归纳的句子、引入式明喻句，以

及谓词混合式明喻句中合成宾语的成分受喻体支配的句子，目前还没有较好的算法可以处理，这也是明喻计算的有界性。

6 结论及未来工作

明喻的使用是生动形象的，其理解是自然微妙的，然而，在自然语言处理过程中，因其复杂性而带来了诸多问题。本文对汉语中“像”型明喻句进行句法形式的分类，采用概念隐喻理论解释，分析出“像”型明喻句凸显喻体特征的句子可以进行计算，这类句子占到32.3%；喻底隐藏的句子，计算的前提是计算机提取喻体的凸显特征，这类句子占38.9%；关于谓词混合型句子和引入句子，由于混合情况复杂，目前仍没有合适的解决方法，需要从其他角度进行分析。

在今后的工作中，我们还需要在以下几个方面进行拓展：（1）对更多名词性明喻现象进行计算；（2）尝试自动识别出本体和喻体之间的关系。

参 考 文 献

[1] Fauconnier, G. Mental Spaces[M]. Cambridge Mass: MIT Press / New York: Cambridge University Press, 1994.

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

[3] Grady, J. A Typology of Motivation for Metaphor: Correlations vs. Resemblances. In R. Gibbs & G. Steen (Eds.), Metaphor in cognitive linguistics, Amsterdam: Benjamins, 1999: 79-100.

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

[5] 贾玉祥. 隐喻自动处理研究进展[J], 中文信息学报, 2009 年第 6 期.

[6] 贾玉祥, 俞士汶. 基于实例的隐喻理解与生成[J], 计算机科学, 2009 年第 3 期.

[7] 李斌, 于丽丽. “像”的明喻计算[J], 中文信息学报, 2008 年第 6 期.

[8] 汪少华. 合成空间理论对隐喻的阐释力[J], 外国语, 2001 年第 3 期.

[9] 杨芸. 汉语隐喻识别与解释计算模型研究[D], 厦门大学博士论文, 2008.