

# 基于词典的名词性隐喻识别<sup>1</sup>

贾玉祥 俞士汶

计算语言学教育部重点实验室, 北京大学计算语言学研究所, 北京, 100871

{yxjia, yusw}@pku.edu.cn

**摘要:** 隐喻是用一个事物来类比另外一个事物的语言表达, 在自然语言中非常普遍, 要实现自然语言理解隐喻处理不可避免。本文针对最基本的隐喻类型——名词性隐喻, 提出基于词典的识别方法。结合同义词词林的语义距离与HowNet的语义关系来识别隐喻, 考察隐喻与语义距离及语义关系之间的关联。最后把基于词典的方法用于隐喻新颖性的度量。

**关键词:** 名词性隐喻 隐喻识别 词典 语义距离 语义关系 隐喻新颖性

## Nominal Metaphor Recognition Based on Lexicons

Jia Yuxiang, Yu Shiwen

Key Laboratory of Computational Linguistics, Ministry of Education, Institute of Computational Linguistics, Peking University, Beijing, 100871

{yxjia, yusw}@pku.edu.cn

**Abstract:** Metaphor, an expression to describe one thing in terms of another, is pervasive in natural languages. So metaphor processing is inescapable for natural language understanding. This paper proposes a lexicon based method to recognize nominal metaphor, one of the basic metaphor types. Semantic distance by TongYiCiLin and semantic relations in HowNet are combined for metaphor recognition, and the association between metaphor and lexicon based semantic information is also discussed. Lastly, the lexicon based method is used to measure the novelty of metaphors.

**Keywords:** nominal metaphor, metaphor recognition, lexicon, semantic distance, semantic relation, novelty of metaphor

## 1 引言

认知语言学认为, 隐喻本质上是用一个事物来理解另外一个事物的认知现象[Lakoff and Johnson, 1980]。从这个角度上说, 隐喻贯穿语言发展的整个过程, 语言中的隐喻使用非常普遍, 除涵盖比喻、比拟(拟人、拟物)、移就、移觉(通感)等修辞表达外, 还有更广泛的内容。本文采纳认知语言学的观点, 将语言中的隐喻定义为“用一个事物(源域或称喻体)来表达另外一个事物(目标域或称本体)的语言现象”, 并且表达的基础是两事物之间的相似性(喻底)。这不同于传统修辞学中的“暗喻”。

按照隐喻词的语法属性可以将隐喻分为名词性隐喻、动词性隐喻、形容词性隐喻、介词性隐喻等。本文研究最基本的一种隐喻类型——名词性隐喻。所谓名词性隐喻, 就是由名词构成的隐喻, 可以在句中充当主语、表语、宾语、同位语等成分[束定芳, 2000]。其中“N是N”(知识是力量)是最典型的一种隐喻表达形式, 另一种典型的名词性隐喻是名词短语隐喻“N+N”(知识的海洋, 金融风暴)。本文对于“N是N”形式的名词性隐喻, 提

---

<sup>1</sup> 项目支持: 国家自然科学基金资助项目(60773173, 60970083)

出基于词典的识别方法（只考虑目标域和源域，暂不考虑句子上下文），该方法也适用于“N+N”的名词短语隐喻。为了区分目标域和源域，后面叙述中用“X是Y”替代“N是N”。

## 2 相关研究

[王治敏, 2006]提出基于机器学习方法的汉语名词短语隐喻识别。为特定名词的短语选择一定的语料，标记隐喻和非隐喻用法。用标记语料训练分类器，从新的名词短语中识别出隐喻。该方法的优点是机器学习方法在大规模数据的分类中具有优势，缺点是要为每个名词标注一定数量的语料，耗费人力。

对于名词性隐喻的识别，基于现有知识库的方法是一个很好的选择。[Krishnakumaran, 2008]利用 WordNet 中的上下位关系来判断“X is Y”是常规表达还是隐喻表达。如果 Y 是 X 的上位，则是常规表达；否则是隐喻表达。如“A lion is a wild animal”，animal 是 lion 的上位概念，是常规表达。而“Ali the world is a stage”，stage 不是 world 的上位概念，是隐喻表达。

然而，只考虑上下位关系还不够，如果 X 和 Y 不是上下位关系，而是兄弟关系（如朋友是律师，朋友和律师具有共同的上位“人”）或存在某种关联（如律师是职业，这里律师指代的是律师这类人所从事的职业），仍是常规表达。

[杨芸, 2008]利用同义词词林和 HowNet 来判断“X 是 Y”是常规表达还是隐喻表达。结合同义词词林的类别相似度和 HowNet 的上下义相关度定义了一个指称异常度指标。如果指称异常度大于给定阈值则为隐喻表达，否则为常规表达。

一个指称异常度数值可以判断“X 是 Y”是不是隐喻表达，但是却不能反映 X 与 Y 在词典中的位置及语义关系情况，不能清楚体现隐喻与语义距离及语义关系之间的关联。

## 3 识别方法

本文利用同义词词林的语义距离和 HowNet 的语义关系来判断隐喻表达和常规表达，考察隐喻与语义距离及语义关系的关联。只有当两部词典都判断为隐喻表达时，才认为是隐喻表达，否则认为是常规表达。

### 3.1 同义词词林

同义词词林（简称词林）[HIT-IR]的分类体系如图 1 所示。从上到下分为大类、中类、小类、词群、原子词群（同义词集合）五个层次。大类由 A 到 L 标记，分别为 A 人、B 物、C 时空、D 抽象物、E 特征、F 动作、G 心理活动、H 活动、I 现象与状态、J 关联、K 助语及 L 敬语。

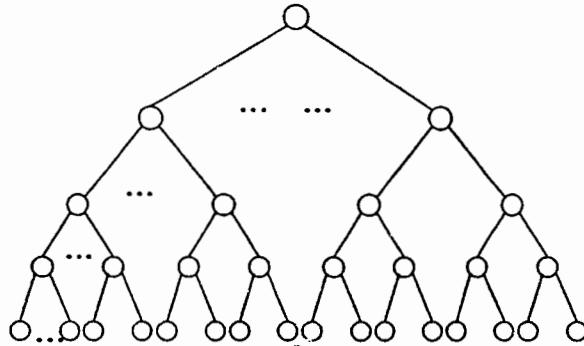


图 1 同义词词林五层结构图

由于词林是对词语的一种划分，词语均分布在树的叶子节点，中间节点没有相应的词语表示。因此，词语之间的语义距离，即路径长度，只有以下六种情况（设词语  $w_1, w_2$ ）：

$$\text{Dis}(w_1, w_2) = \begin{cases} 0 & w_1, w_2 \text{ 位于同一个原子词群} \\ 2 & w_1, w_2 \text{ 位于同一个词群} \\ 4 & w_1, w_2 \text{ 位于同一个小类} \\ 6 & w_1, w_2 \text{ 位于同一个中类} \\ 8 & w_1, w_2 \text{ 位于同一个大类} \\ 10 & w_1, w_2 \text{ 位于不同的大类} \end{cases}$$

当  $w_1$  或  $w_2$  出现在多个语义类中时，两个词之间存在多条路径，这时取最短路径作为语义距离。

一般来说，语义距离越短，是常规表达的可能性越大，如：“北京”与“首都”的语义距离为 0，词典认为二者是同义词，则“北京是首都”是常规表达；“中国”与“国家”的语义距离是 2，“中国是国家”是常规表达；“睡觉”与“休息”的语义距离是 4，“睡觉是休息”是常规表达。

反之，语义距离越长，是隐喻表达的可能性越大。如：“律师”与“狐狸”的语义距离是 10，“律师是狐狸”是隐喻表达；“知识”与“力量”的语义距离是 8，“知识是力量”是隐喻表达。

但也有例外，比如，在“A 人”这一大类里，即使两个词的距离是 8，也不是隐喻表达，如，“他”和“流氓”，“你”和“老师”的距离都是 8，但“他是流氓”，“你是老师”都是常规表达。因此，规定在“A 人”类内词之间的指称为常规表达。

对于“X 是 Y”表达，隐喻判别方法如下：

如果语义距离小于等于 6，为常规表达；

如果语义距离等于 10，为隐喻表达；

如果语义距离等于 8，

    如果 X 和 Y 都属于语义类 A，则为常规表达，

    否则，为隐喻表达。

### 3.2 HowNet

HowNet (2000 版) [董振东 董强]使用 1500 多个义原来刻画概念。义原根据上下位关系，形成树形结构，父亲节点是儿子节点的上位概念，儿子节点是父亲节点的下位概念。图 2 是 HowNet 义原树形结构的截图。

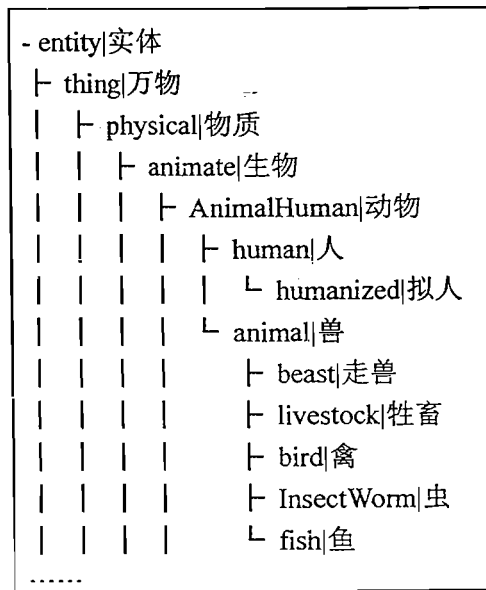


图 2 树状的义原层次结构

词的义原表示举例：

律师={human|人,#occupation|职位,police|警,#law|律法}  
 职业={affairs|事务,#earn|赚,#alive|活着,#occupation|职位}  
 鲨鱼={fish|鱼}  
 朋友={human|人,friend|友}

考虑 HowNet 中的以下语义关系：

上下位关系 (hyper)：词语定义中的第一义原是该词语的直接上位 (direct-hyper)，上位的上位称为间接上位 (indirect-hyper)，也属于上下位关系。如鱼是鲨鱼的上位，动物、生物也是鲨鱼的上位。因此，HowNet 中存在大量的上下位关系。

兄弟关系 (brother)：如果两个词语的第一义原相同，则二者称为兄弟关系，如律师、朋友，第一义原都是“human|人”。

相交关系 (intersection)：如果两个词语的义原有交集，称二者为相交关系，如律师和职业，义原定义有交集“occupation|职位”。

从语义关系的紧密性来说，上下位关系 > 兄弟关系 > 相关关系。当两个词语之间存在多种语义关系时，取紧密性最大的那个。

对于“X 是 Y”表达，隐喻判别方法如下：

如果 Y 是 X 的上位概念 (包括直接上位、间接上位), 则是常规表达。如: 律师是人。  
 如果 X 与 Y 是兄弟关系, 则是常规表达。如: 朋友是律师。  
 如果 X 与 Y 是相交关系, 则是常规表达。如: 律师是职业。  
 其他, 认为是隐喻表达。如: 律师是鲨鱼。

#### 4 实验分析

结合同义词词林和 HowNet 进行隐喻判断。词林输出 X 与 Y 的语义距离{0, 2, 4, 6, 8, 10}与 X、Y 所属大类。HowNet 输出语义关系{hyper, indirect-hyper, brother, intersection, metaphor}。只有同义词词林和 HowNet 方法同时判断为隐喻时, 才认为是隐喻表达, 否则为常规表达。[杨芸, 2008]中的 22 个例子的实验结果表 1 所示, 左边为隐喻表达用例, 右边为常规表达用例。

表 1 隐喻识别结果

隐喻用例					常规用例				
X	Y	Cilin	HowNet	隐喻	X	Y	Cilin	HowNet	隐喻
律师	狐狸	10	metaphor	y	律师	职业	10	intersection	n
手杖	伴侣	10	metaphor	y	妈妈	工人	8; A	brother	n
人生	书	8; D	metaphor	y	中国	国家	2	brother	n
眉毛	卫士	10	metaphor	y	他	流氓	8; A	metaphor	n
语言	镜子	10	metaphor	y	太阳	恒星	6	brother	n
时间	资源	10	metaphor	y	北京	首都	0	brother	n
知识	力量	8; D	metaphor	y	你	老师	8; A	metaphor	n
狮子	国王	10	metaphor	y	狮子	动物	8; B	indirect-hyper	n
打球	打仗	8; H	metaphor	y	电脑	工具	4	metaphor	n
炒股	赌博	8; H	metaphor	y	睡觉	休息	4	indirect-hyper	n
影子	画面	10	brother	n	跑步	锻炼	10	metaphor	y

从表中可以看出, 只有最后一行的两个例子做出了错误的判断, 比单独使用一个词典效果好。错误分析如下:

1、影子是画面。

词林中“影子”与“画面”的语义编码分别为:

影子 Bg03C01

画面 Dk29E01

“影子”属于自然物, 而“画面”是文教领域的抽象事物, 二者属于不同的大类, 语义距离等于 10。

HowNet 中“影子”与“画面”有以下义原定义:

影子 image|图像

画面 image|图像,shows|表演物

因此,计算出“影子”与“画面”是兄弟关系,因为二者的第一义原即直接上位相同,都是“image|图像”。其实,“影子是画面”的隐喻性并不明显,是否是隐喻还需要具体上下文来确定。

## 2、跑步是锻炼。

词林中“跑步”与“锻炼”的语义编码如下:

跑步 Fb01B01

锻炼 Hg06B01

“跑步”是下肢动作,“锻炼”属于教卫科研活动方面的术语,二者属于不同的大类,语义距离等于10。

HowNet中“跑步”与“锻炼”的义原定义如下:

跑步 run|跑

锻炼 cultivate|培养

基于HowNet现有的知识,算法无法将“锻炼”与“跑步”建立联系,因此判断为隐喻。类似的例子还有:

他 aValue|属性值,kind|类型,other|另

他 ThirdPerson|他,male|男

你 SecondPerson|你

流氓 human|人,undesired|莠,evil|恶,crime|罪

老师 human|人,teach|教,education|教育

由于人称代词与“human|人”这一语义类中的词无法从义原上建立联系,因此将“他是流氓”“你是老师”判断为隐喻。可以通过进一步丰富词典的知识描述并建立起知识之间更好的关联来解决这一问题。

## 5 隐喻新颖性考察

隐喻表达在新颖程度上有强有弱。新颖性强的隐喻表达通常是临时创造或使用时期较短,有较强的修辞效果。而新颖性弱的隐喻表达通常已经成为人们常识的一部分,不易被察觉,往往已经词汇化或进入词典。因此可以利用基于词典的方法来度量隐喻的新颖性,即目标域和源域的语义距离越大或语义关系越疏远,隐喻的新颖性越强;反之,目标域和

源域的语义距离越小或语义关系越紧密，隐喻的新颖性越弱。

对[李玉莲, 2007]中的 292 个“X 是 Y”型隐喻，分别利用同义词词林计算语义距离、HowNet 计算语义关系。表 2、表 3 分别给出语义距离为 0 和语义距离为 10 的隐喻例子。可见表 2 一般为人们熟知的隐喻表达，而表 3 的隐喻表达较新颖。

表 2 语义距离为 0 的目标域源域对

把柄	辫子	工作	饭碗	目标	靶子	人情	面子	罪名	帽子
帮手	羽翼	顾问	智囊	骗局	陷阱	人世	红尘	计划	蓝图
帮凶	爪牙	汉奸	走狗	破绽	漏洞	兄弟	手足	清官	青天

表 3 语义距离为 10 的目标域源域对

爱情	烈酒	报纸	喉舌	希望	种子	才华	光芒	老师	蜡烛
道路	血管	服务	软件	父爱	雨伞	革命	风暴	工会	桥梁
顾客	上帝	国家	机器	河流	母亲	婚姻	坟墓	理想	灯塔

## 6 总结与展望

本文对“X 是 Y”形式的名词性隐喻，提出了基于词典的识别方法。结合同义词词林的语义距离和 HowNet 的语义关系进行识别，并考察了隐喻与语义距离及语义关系之间的关联。最后将基于词典的方法用于隐喻新颖性的度量。

下一步将构建更大规模的语料，进行更充分的实验，更好地考察隐喻识别的性能，改进隐喻识别算法。完善词典知识，构建知识之间的关联也是提高识别效果的一个方向。

## 参 考 文 献

- [董振东 董强] 董振东, 董强. 知网. <http://www.keenage.com>.
- [HIT-IRLab] HIT-IRLab. 同义词词林(扩展版). <http://ir.hit.edu.cn>.
- [李玉莲, 2007] 李玉莲. 本体和喻体在中文隐喻句理解中的应用. 西南大学硕士学位论文, 2007.
- [束定芳, 2000] 束定芳. 隐喻学研究. 上海: 上海外语教育出版社, 2000.
- [王治敏, 2006] 王治敏. 汉语名词短语隐喻识别研究. 北京大学博士学位论文, 2006.
- [杨芸, 2008] 杨芸. 汉语隐喻识别与解释计算模型研究. 厦门大学博士学位论文, 2008.
- [Krishnakumaran, 2007] Krishnakumaran, S., Zhu, X.J. Hunting Elusive Metaphors Using Lexical Resources. In Proceedings of the ACL 2007 Workshop on Computational Approaches to Figurative Language, 2007, 13-20.
- [Lakoff and Johnson, 1980] Lakoff, G., Johnson, M. Metaphors We Live By. Chicago: University of Chicago Press, 1980.