基于实例的隐喻理解与生成

贾玉祥 俞士汶
(北京大学计算语言学研究所 北京 100871)

摘 要 语言中的明喻可以看作是带标记的隐喻,比较容易识别,为隐喻的理解和生成提供了很好的知识源。利用Web搜索引擎大规模获取明喻实例,自动构建明喻知识库。基于明喻知识库,考察了汉语隐喻的源域分布情况;提出了一个基于实例的隐喻自动理解和生成方法。实验结果表明,隐喻的理解和生成均取得了较高的准确率。该方法具有很好的可扩展性。明喻知识库中所表达的概念之间的组合关系也可以用于其他多种自然语言处理任务。

关键词 基于实例,隐喻理解与生成,自然语言理解,概念组合关系

Instance-based Metaphor Comprehension and Generation

JIA Yu-xiang YU Shi-wen
(Institute of Computational Linguistics, Peking University, Beijing 100871, China)

Abstract In natural language, simile, always considered as a marked metaphor, is easier to recognize and provides an ideal knowledge source for metaphor comprehension and generation. This paper proposed a strategy to acquire large-scale simile instances with Web search engine, and to construct a simile knowledge-base automatically. Based on the simile knowledge-base, the source domains of Chinese metaphors were studied and an instance-based method for metaphor comprehension and generation was put forward. Experiment results show that the method achieves high precisions for both comprehension and generation, and has a good expandability. In addition, the collocational relations in the knowledge-base are useful for many other natural language processing tasks.

Keywords Instance-based, Metaphor comprehension and generation, Natural language understanding, Conceptual collocational relations

1 引言

隐喻是一种认知现象,在我们的日常语言中普遍存在。概念隐喻[1]认为,隐喻是人的概念系统中源域(Source domain,相当于喻体)到目标域(Target domain,相当于本体)的映射,通过比较具体的源域知识来认识或表达比较抽象的目标域对象。隐喻在自然语言中的表现形式多种多样,最根本的形式是"X是Y",其中X是目标域,Y是源域。如"女人是水",可以理解为"女人像水一样温柔",用源域"水"的显著特征"温柔"来凸显"女人"的"温柔"。

隐喻是理解自然语言不可回避的问题[2],有着不同于其他语言现象的特点和困难。与自然语言理解的难点之一——歧义消解的"同中求异"不同,隐喻求解是"异中求同",找出两个不同概念之间的相似点(即喻底)[3]。隐喻处理包括隐喻识别、隐喻理解和隐喻生成三步曲。文献[4]提出的基于机器学习的汉语名词短语隐喻识别方法。本文重点关注隐喻的理解和生成。传统的隐喻理解模型是基于规则的,如Fass的Met*[5],Martin的MIDAS[6],Barnden的ATT-Meta[7]等。文献[8]提出基于隐含语义分析(Latent Semantic Analysis,LSA)的统计模型,得到广泛使用。隐喻的生成是自然语言生成的必要组成部分。具体地讲,就是给定目标域及其特征,选择恰当的源域,形成正确的隐喻表达。文献[9]将LSA模型用于隐喻的生成。汉语隐喻理解及生成的研究分别有文献[10]及文献[11],与之不同,本文在同一框架下完成隐喻理解和生成任务。

明喻和隐喻关系密切。从认知的角度说,二者同样涉及源域、目标域及喻底。明喻、新隐喻、常规隐喻是对目标域与源域之间关系的认识逐渐加深的过程。明喻和新隐喻的理解同是比较过程,而常规隐喻的理解是归类过程[12]。在语言学的层面上,明喻可以看成加上比喻词的隐喻,比隐喻较易识别,因此可以作为隐喻,尤其是新隐喻处理的理想知识源。文献[13]指出明喻源域的显著特征有助于隐喻的理解和生成。

本文使用搜索引擎从海量网页中抓取明喻实例,自动构建明喻知识库,在此基础上考察了汉语源域的分布情况,提出了基于实例的隐喻自动理解和生成方法。本文第2节详细描述了明喻知识库的构建过程;基于明喻知识库,第3,4节分别给出了汉语源域的分布情况和隐喻理解及生成方法;第5节是实验及分析,隐喻理解及生成的准确率分别达到84.73%和78.24%。最后是总结和展望。

2 明喻知识库

明喻含有比喻词,符合一定的句法模式,如"像N一样

到稿日期:2008-06-10 本文受国家973课题(文本内容理解的数据基础,编号:2004CB318102)资助。
贾玉祥 博士生,主要研究方向为计算语言学;俞士汶 教授,主要研究方向为计算语言学。

A”,“如 N 般 A”等。此处 N 指名词,即源域;A 指形容词,是 N 的显著特征,即喻底。现有语料库中明喻出现得较少,如 2000 年人民日报语料库中,符合模式“像 N 一样 A”的明喻只有 23 条。因此,使用搜索引擎从海量网页中提取明喻实例是一个很好的选择。

2.1 构建过程

本文选择模式“像 N 一样 A”,利用搜索引擎获取明喻实例,自动构建明喻知识库,整个过程如算法 1 所列。

算法 1 构建明喻知识库

1)从现代汉语语法信息词典[14](Grammatical Knowledge Base,GKB)中抽取形容词列表(3155 个形容词);对每个形容词 A,构造查询“像 * 一样 A”(* 是通配符,使用整串匹配),由搜索引擎 www.baidu.com 搜索网页;对每个查询结果,抓取前 100 个网页的片段(Snippets),一起形成原始语料库;对原始语料库进行分词、词性标注处理;抽取符合模式“像 /n 一样 /a”的串(/n、/a 分别为名词、形容词的词性标记),形成实例库 EB1。

2)从 EB1 中获取名词列表,对每个名词 N,构造查询“像 N 一样 * ”,同 1)进行网页抓取,分词、词性标注,抽取实例,形成实例库 EB2。

3)合并实例库,并用 GKB 过滤实例库中的词,消除分词错误。最后得到明喻实例(〈N,A〉对)71555 条、无重复实例 20922 条。

4)由实例库构造名词-特征(形容词)库、特征-名词库。包含名词 3666 个、形容词 1804 个。

5)使用同义词词林(哈工大信息检索研究室的同义词词林扩展版),把词映射到义项,得到义项化的、具有显著性指标的名词-特征义项库和特征-名词义项库。

搜索引擎支持通配符查询和整串匹配,通过先后构造查询“像 * 一样 A”及“像 N 一样 * ”,尽量多地抓取“像 N 一样 A”的明喻实例。实例库由模式“像 N 一样 A”中的名词、形容词对(即〈N,A〉)构成,如表 1 所列。对实例库进行整理,得到名词-特征库和特征-名词库,分别如表 2、表 3 所列示。名词-特征库列出了每一个名词所具有的显著特征,在隐喻理解时可以作为喻底的候选;特征-名词库则给出了最能凸显某一特征的一系列名词,在隐喻生成时可以作为源域的候选。

表 1 明喻实例库

N	……	水	水	水	水	水	水	水	……
A	……	安静	安宁	安稳	肮脏	博大	纯洁	纯净	……

表 2 名词-特征库

名词	特征数	特征
水	187	安静 安宁 安稳 肮脏 博大 纯洁 纯净 纯情 纯真 从容 脆弱 单纯 淡 淡漠 淡雅 ……

表 3 特征-名词库

特征	名词数	名词
温柔	107	爱心 波斯猫 晨光 春风 大地 大海 风 羔羊 鸽子 海风 海水 海豚 和风 江水 康乃馨 流水 柳树 柳絮 柳枝 鹿 猫 美人鱼 绵羊 ……

从表 2 可以看出,一个常用名词可能具有很多个显著特征(如“水”有 187 个),这是因为人们对这些名词概念特别熟悉,经常用作源域来描述其他概念,久而久之就形成了很多特征。同理,如表 3 所列,某些特征对应于很多名词(如“温柔”有 107 个),在一定程度上说明了人们对该特征比较关注,使用比较频繁。

其实,名词所对应的特征集(或特征所对应的名词集)里,有很多同义词,并且这些同义词分组与名词(或特征)之间关系的紧密程度,或称它们的显著性,是有差异的。比如,“水”的特征里,“清澈”比“甜”更显著。下面将通过义项映射,对名词对应的特征(或特征对应的名词)进行同义分组,实现同义扩展;并计算显著性,达到去除噪声的目的(如“像水一样便宜”中的“便宜”,显著性很低,可以去掉)。

2.2 义项映射

同义词词林是一部类义词典,对收录的词语按树状的层次结构进行分类,共分 5 个(义项)层次(见图 1):大类、中类、小类、词群、原子词群,逐层细分。如大类分为人、物、时间与空间、抽象事物等。物又分为动物、植物等。植物进一步分为花、蔬菜、水果等。值得一提的是,词林扩展版新增加的词条并不具有这种结构。

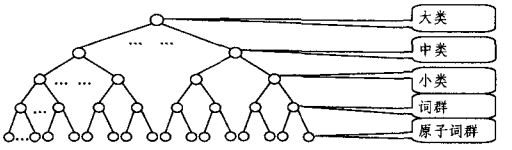

图 1 同义词词林词语分类层次结构

利用同义词词林,对名词对应的特征(或特征对应的名词)进行义项映射,映射到某一义项层;对特征(或名词)进行同义分组,计算显著性,并对分组按照显著性从高到低排序。显著性用每个同义词组中的特征(或名词)个数表示。对于一词多义的情况,映射时保留所有义项,消歧工作根据后面的显著性指标自动完成,去掉显著性较低的义项。义项映射后得到的名词-特征义项库及特征-名词义项库中的记录格式如表 4 所列,其中义项用同义词词林中的义项代码表示;根据需要,义项可以对应于任何一个义项层。

表 4 义项映射后的记录格式

名词	特征
{	{
特征义项 1:显著性	名词义项 1:显著性
特征义项 2:显著性	名词义项 2:显著性
特征义项 3:显著性	名词义项 3:显著性
……	……
}	}

例如,义项映射后,“水”(如表 2 所列)和“温柔”(如表 3 所列)的记录分别如下,其中“水”的义项层次为第 5 层(如 Ef04A01),“温柔”的义项层次为第 4 层(如 Bf02A),“//”后给出的是分组后的同义特征(或名词)。可见“安静”是“水”最显著的特征,而“温柔”的对象常被比作“春风”。

水:

{

Ef04A01 11//安静 沉静 静 静谧 冷静 宁静 清静 清净 恬静 幽静 幽深

Ed15C01 7 //平和 轻柔 柔和 温和 温柔 温软 优柔

Eb19A01 7//纯净 明澈 明净 清 清澈 清亮 清冽

Ef12A01 6//干净 洁净 净 清洁 清爽 清新

Eb10B01 5//绵软 柔 柔嫩 柔韧 柔软

Ee15A01 5//快 灵 灵动 灵活 敏感

Ee07A01 4//温存 温和 温柔 温润
Ga01A01 4// 欢快 快 快活 快乐
……
}
温柔：
{
Bf02A 7//春风 风 海风 和风 轻风 晚风 微风
Bi06D 5//羔羊 羚羊 绵羊 山羊 羊羔
Ab01B 5//妇女 娘 女儿 女人 女子
Bg01A 4//海水 江水 流水 水
Bg03B 4//晨光 阳光 月光 月色
Bh02A 4//合欢 康乃馨 山茶花 茉莉花
……
}

3 源域考察

隐喻是源域到目标域的映射。目标域是要描述的对象，可以是任何概念。而源域概念有一定的选择性，一般是人们比较熟悉、比较具体的概念，是研究隐喻的关键。基于明喻知识库，可以定量地考察汉语源域的分布情况。

源域对应的特征个数反映了其使用的频繁程度。表 5 列出了最常用的 10 个源域词及各个层次上最常用的源域义项（参见同义词词林的义项编码），可见“水（江河湖海）、家养动物（猪猫狗等）、花、阳光、天空”等常用作源域。由“人”作为源域构成的拟人也属于隐喻用法。总体上说，动物、植物、日常用品、自然现象等概念最常用作源域。

表 5 最常见的 10 个源域

词	原子词群	词群	小类	中类	大类
水	Dd15C02	Bh02A	Bh02	Bi	B
猪	Be05B01	Dd15C	Dd15	Bh	D
猫	Bg01A01	Bg01A	Dh01	Bp	A
狗	Bh02A44	Dh01A	Bg03	Bg	C
阳光	Ab04B01	Bh01A	Be05	Dk	E
大海	Ac03A01	Bg03B	Bf02	Bf	H
天空	Ah04A01	Bi11B	Bg01	Br	I
花	Bi07A01	Bf02A	Bf01	Bm	G
妈妈	Bi08A01	Bh07A	Bh01	Ah	
女人	Bf01B01	Be05B	Bi07	Be	

图 2 给出了源域与特征个数的总体分布情况。在 3666 个源域词中，特征个数≥2 的不足 75%，≥5 的约占 37%，≥10 的约占 15%，≥20 约占 4.36%，≥100 的只有 3 个。可见只有少量源域具有特别多的特征，而大部分源域的特征个数都比较少。这说明人们频繁使用的源域是很有限的，进一步印证了源域的选择性。

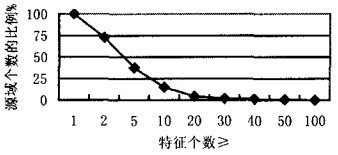

图 2 源域随特征个数的分布情况

在特征方面，人们在比喻中最常描述的 10 个特征是“大、美丽、好、可爱、漂亮、坚强、快乐、长、多、自由”，均倾向于表达正面的感情，而不是“小、丑陋、坏、脆弱、悲伤、短、少、拘束”等。从语义类上说，最常描述的特征是德才、性质、表象、境况、外形等。

4 隐喻理解与生成

隐喻往往利用源域的某一显著特征来凸显或赋予目标域这一特征，该特征就是源域和目标域之间的相似点，即喻底。隐喻理解就是寻找源域的这一显著特征。隐喻生成则是给定目标域和要描述的特征（喻底），选择一个合适的源域，以凸显这一特征。明喻知识库为隐喻的理解和生成提供了基础。

“X 是 Y”型隐喻的理解，是已知目标域 X 和源域 Y，求喻底 A，理解为“X 像 Y 一样 A”，形式化表示为 A=C(X,Y)（C 代表 Comprehension）。Y 的显著特征为 A 提供了候选，最终 A 的确定还要参照 X。坚持“X、Y 的相似点”、“Y 的显著特征”等原则，具体算法为：如果 X 和 Y 的显著特征交集不空，则 A 取特征交集；否则，看 X 和 Y 的特征义项交集，如果不空，则 A 取特征义项交集；否则，A 取 Y 的最显著特征。如果 Y 不在知识库的名词列表中，则理解失败。

下面是几个利用知识库进行隐喻理解的例子：

“女人是水。”（特征交集是“温柔”，理解为“女人像水一样温柔”。）

“人生如梦。”（特征交集是“短暂”，“梦”的最显著特征是“缥缈”，理解为“人生像梦一样短暂、缥缈”，形容世事无定，人生短促。）

“人是会思想的芦苇。”（特征交集是“脆弱”，理解为“人像芦苇一样脆弱”，强调人的脆弱性。）

隐喻生成是给定目标域 X 及要描述的特征 A，寻找能凸显该特征的源域 Y，从而得到“X 是 Y”型的隐喻，形式化为 Y=G(X,A)（G 代表 Generation）。A 对应的名词（义项）集合提供了 Y 的候选，最终 Y 的确定还需要考虑 X。原则是保证 Y 和 X 分属两个不同的概念域，否则构不成隐喻。如果 A 不在知识库的形容词列表中，则生成失败。

5 实验分析

目前的知识库中收录的形容词占形容词总数（来自 GKB）的 1804/3155≈57.18%，说明知识库能为 57.18%的形容词提供源域候选。名词比例为 3666/35162≈10.43%，说明只有约 10%的名词概念可以用作源域，体现了源域概念的选择性。

从抓取的网页中抽取符合模式“像 Y 一样 A 的 X”的串，其中 X,Y 是名词，A 是形容词，如“像大海一样广阔的胸怀”。这样选择同时具备源域（Y）、目标域（X）及喻底（A）的隐喻样本 262 个，测试隐喻自动理解和生成的效果。

隐喻理解是 A=C(X,Y)，为每一个输入的〈X,Y〉对，求出一个喻底 A。不同上下文可以对应不同的喻底，由人工判断 A 的正确性。隐喻生成是 Y=G(X,A)，为每一对〈X,A〉，选择一个源域 Y，由人工判断 Y 的正确性。这里是唯一有人工参与的地方，整个知识库构建和隐喻处理均自动实现。

隐喻理解和生成均未出现失败的情况。隐喻理解的结果如表 6 所列，总体准确率达到 84.73%，其中喻底分别来自特征交集、义项交集和最显著特征。正如所预期的，特征交集准确率最高。义项交集的准确率低，原因可能是同义词词林中的词群并不都是严格的同义词集，很多是相关词的集合。下

一步将尝试使用北京大学计算语言学研究所的中文概念词典(Chinese Concept Dictionary,CCD)进行义项影射。同 WordNet 一样,CCD 使用同义词集表示义项,不存在同义词词林中相关词集的问题;或调整理解算法,用最显著特征取代义项交集,因为前者准确率更高。

表 6 隐喻理解结果

	特征交集	义项交集	最显著特征	总体
正确样本数	106	21	95	222
样本数	116	33	113	262
正确率	91.38%	63.64%	84.07%	84.73%

隐喻生成的准确率是 205/262=78.24%。错误来源主要是返回的源域和要描述的目标域来自同一概念域,或是形成的隐喻理解起来不够直观,或语言表达上不够优美。还需要更多地考虑目标域的信息,来提高隐喻生成的准确率。当然,知识库的规模和质量也是影响隐喻处理结果的因素。因此,知识库扩展和改进将是下一步主要工作。

目前汉语隐喻处理的研究还处于起步阶段,尚无公共评测语料,因此还无法与其他研究进行横向比较。但是本文实验结果表明,文中提出的基于实例的方法是有效的。构建公共评测语料,以评测推动方法的改进,是隐喻研究的课题之一。

结束语 本文使用搜索引擎,大规模获取明喻实例,自动构建明喻知识库。在知识库的基础上,考察了汉语源喻的使用情况,提出了隐喻的自动理解和生成方法,取得了较好的实验结果。

知识库还可以用于隐喻的识别。例如,"A+N"型隐喻("红色的海洋")的识别。从知识库中可以看到,"海洋"的显著特征之一是"蓝",而"红色"违反了组合限制,因此是隐喻用法(可能是"红色的花的海洋")。知识库中的明喻实例,对于汉语学习、比喻学习等都是有帮助的。知识库中概念之间的组合关系,还可以用于词义消歧、未知词语意思推断等自然语言处理任务。

下一步将进行知识库扩展。一是模式扩展,采用模式"如 N 般 A"、"A 如 N"、"比 N 还 A"等,获取更多实例。二是词类扩展,通过模式"像 N 一样 B"、"像 N 一样 Z"、"像 N 一样 V"(像水一样流淌、像水一样蒸发)等,引入区别词(B)、状态词(Z)及动词(V)等。隐喻的表现形式多种多样,最根本的形式是"X 是 Y",其他形式都由这一根本形式衍生而成。随着知识库的扩展,将考虑处理其他形式的隐喻(如动词隐喻等)。

参 考 文 献

[1] Lakoff G. The contemporary theory of metaphor // Ortony A. Metaphor and thought. 2nd ed. Cambridge: Cambridge University Press, 1993: 202-251

[2] Carbonell J G. Metaphor: an inescapable phenomenon in natural-language comprehension// Lehnert W, Ringle M. Strategies for natural language processing. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1982: 415-434

[3] 俞士汶. 自然语言理解研究与文学表现手法//第二届文学与信息技术国际研讨会. 2005: 2-13

[4] 王治敏. 汉语名词短语隐喻识别研究. 博士学位论文. 北京大学, 2006

[5] Fass D. Met*: A method for discriminating metonymy and metaphor by computer. Computational Linguistics, 1991, 17(1), 49-90

[6] Martin J H. A computational model of metaphor interpretation. Boston: Academic Press, 1990

[7] Barnden J, et al. Reasoning in metaphor understanding: the ATT-Meta approach and system // Proceedings of COLING' 2002. 2002: 1-5

[8] Kintsch W. Metaphor comprehension: A computational theory. Psychonomic Bulletin and Review, 2000, 7(2): 257-266

[9] Abe K. A computational model for metaphor generation process //Proceedings of CogSci'2006. 2006: 937-942

[10] 苏畅,周昌乐. 基于合作机制的汉语名词性隐喻理解方法. 计算机应用研究, 2007, 24(9): 67-69

[11] 游维,周昌乐. 基于统计的汉语隐喻生成模型及其系统实现. 心智与计算, 2007, 1: 133-141

[12] Bowdle B, Gentner D. The career of metaphor. Psychological Review, 2005, 112(1): 193-216

[13] Veale T, Hao Y F. Comprehending and generating apt metaphors: A web-driven, case-based approach to figurative language //Proceedings of AAAI' 2007. 2007: 1471-1476

[14] 俞士汶,朱学锋,等. 现代汉语语法信息词典详解. 第二版. 北京:清华大学出版社, 2003

(上接第 137 页)

[2] Choi N, Song I-Y, et al. A Survey on Ontology Mapping[J]. SIGMOD Record, 2006, 35(3): 34-41

[3] 唐杰,梁邦勇,李涓子. 语义 WEB 中的本体自动映射[J]. 计算机学报, 2006, 29(11): 1956-1976

[4] Chen Huajun, Wu Zhaohui, Wang Heng, et al. RDF / RDFS-based Relational Database Integration[C] // Proc. of the 22nd Conf. on Data Enginnering. Los Alamitos, CA: IEEE Computer Society Press, 2006

[5] 郑东栋,胡伟,瞿裕忠. 一种关系数据库模式和本体间的匹配方法[C]//第二届江苏计算机大会论文集. 南京:东南大学出版社, 2006: 209-213

[6] Doan A H, Domingos P, Halevy A. Learning to Match the Schemas of Data Sources: A Multistrategy Approach[J]. Machine Learning, 2003, 50 (3): 279-301

[7] Doan A H, Madhavan J. Learning to map between ontologies on the semantic web [A] // Proceedings of the 11th International Conference on World Wide Web [C]. Hawaii, USA, 2002: 662-673

[8] Maedche A, Motik B, Silva N, et al. Mafra - A Mapping framework for distributed ontologies[A]//13th European Conference on Knowledge Engineering and Knowledge Management[C]. Lyon, France, 2002: 235-250

[9] Ehrig M, Staab S. QOM - Quick Ontology Mapping[A]// International Semantic Web Conference 2004[C]. 2004: 683-697

[10] Noy N F, Musen M A. PROMPT: Algorithm and tool for automated ontology merging and alignment[A]//Proceedings of the 2000 National Conference on Artificial Intelligence[C]. Austin, Texas, 2000: 450-455

[11] Stumme G, Madche A. FCA-Merge: bottom-up merging of ontologies[A]// Proc. of the Seventeenth Intl. Conf. on Artificial Intelligence (IJCAI '01)(C). Seattle, USA, 2001: 225-230

[12] Madhavan J, Philip A, Halevy A. Corpus-based schema matching [A]// Proceedings of the International Conference on Data Engineering (ICDE)(C). 2005: 57-68