面向自然语言处理的韩国语隐喻知识库构建研究*

徐 超，毕玉德

解放军外国语学院，洛阳 471003

E-mail: xcsuper_1986@sina.com; biyude@gmail.com

摘 要：隐喻处理是自然语言处理的一个难点问题，隐喻处理必须有隐喻知识库的支撑。本文从语言学的角度分析韩国语词汇级隐喻的分布情况，利用 WordNet 的语义网所提供的名词分类及其上下位的信息判别隐喻表达，并以此为基础，提出了一种面向自然语言处理的韩国语隐喻知识库的构建方法。

关键词：隐喻；隐喻知识库；韩国语；WordNet

A Research on the Corpus Building of Korean Metaphor Knowledge for Natural Language Processing

Xu Chao, Bi Yu-de

PLA University of Foreign Languages, Luoyang 471003

E-mail: xcsuper_1986@sina.com; biyude@gmail.com

Abstract: Metaphor processing, which requires a metaphor knowledge corpus as the starting point, is hard to deal with in the field of natural language processing. This paper makes a linguistic analysis of the distribution of Korean vocabulary metaphors, and uses the classification of nouns and their hyponymy information provided by the Semantic Web of WordNet to identify metaphors. On this basis, a method of building Korean metaphor knowledge corpus in view of natural language processing is put forward.

Keywords: metaphor; metaphor knowledge corpus; Korean; WordNet

1 引言

自然语言处理的关键就是识别与消解自然语言的歧义。人与人的交流由于有共同的知识背景，并且能领会交流的环境和过程，通常不会产生误解。但是，作为语言学研究对象的任何一个语言单位，如词、短语和句子等，如果脱离语境而孤立存在，通常都是有歧义的。当交流在人和机器之间进行时，由于机器尚不具备“背景知识”和“世界知识”，歧义现象就表现得尤为突出。

随着各种语言知识库的不断丰富和发展，自然语言处理的技术不断发展，机器翻译的质量不断提高。在各种语言知识库和大规模语料库的支持下，计算机现在可以消解绝大多数语言中的歧义现象。然而，消解了歧义是否实现了理解呢？隐喻、影射、双关、夸张、幽默、拟人以及遣词造句的技巧对自然语言处理提出了新的挑战。

无论哪一种自然语言，都普遍存在着隐喻用法，隐喻是语言运用中十分普遍的现象，也是必不可少的修辞手法，不但文学语言中是如此（Nogales，1999），日常语言、科技语言也不例外（Hoffman，1980；Hallyn，2000；Maasen，2000；胡壮麟，1996，1997；束定芳，2000）。有学者甚至认为隐喻是语言的中心问题（Lakoff，1980；Goatly，1997）。因此，不解决隐喻理解问题而仅仅局限于字面意义的获取上，要很好地解决语言理解是远远不够的（周昌乐，2000）。

目前我们正处于自然语言处理的初级阶段，但我们要记着向自然语言理解的最高境界登攀。本文针对目前自然语言理解最棘手的问题之一的隐喻问题，以韩国语为例，依托 WordNet 词典，探索隐喻判断方法，并以此为基础进行韩国语隐喻知识库的构建研究。

* 基金项目：国家自然科学基金项目（60673036）。

2 韩国语中的词汇隐喻现象

韩国语的词汇级隐喻十分发达，跨领域十分广泛，是本研究的重点，因此这里做重点阐述。

为了理解韩国语中的词汇级隐喻，把握韩国语词汇级隐喻的规律，本研究首先以东亚日报作为语料来源，从中抽取出一定数量的韩国语新闻标题进行隐喻表达的分析。分析过程中采用《동아새국어대사전》（2000，斗山东亚出版社）的词汇分类标准（即经济，建筑•土木工事，工业•工学，宇宙科学，矿物学•矿物，交通，军事，基督教•天主教，气象，农业， 动物学•动物，文学，物理，民俗•巫俗，服饰，法律•法学，不动产，佛教，生物学•生理学，数学，植物学•植物，药学，水利，言语，历史，体育，音乐，医学，政治，地理•地质，航空，海洋•船舶，别名，地名等类别），分析韩国语词汇级隐喻的分布情况。简要分析结果如下：

类别	例句	构成方式
经济	펜싱기대주 서미정 플뢰레 金 찔렀다	펜싱+기대주
建筑•土木工事	“징검다리 취업” 는다	징검다리+취업
工业•工学	“한국서 온 득점기계” 박지성 응원가 히트	득점기계
宇宙科学	세계적 성장궤도에 왜 못 오르나	성장궤도
矿物学•矿物	‘흑진주자매’ 세계랭킹 나란히 1-2위	흑진주자매
交通	내수 침체…수출 부진…경제지표 일제히 빨간불	빨간불
军事	은행 수익성 악화 정부 vs 은행 “네탓” 공방	공방
基督教•天主教	“비상이다…침착해!” 불 켜려다 총알세례	총알세례
气象	박주영 해트트릭 ‘골폭풍’	골+폭풍
农业	골가뭄 풀어야 팬 돌아온다	골+가뭄
动物学•动物	“개혁철새 안하겠다” 이성헌의원 신당行 가능성 부인	개혁철새
文学	11일 포스코 시작으로 주요기업 3분기 성적표 발표	성적표
物理	인터넷전화 서비스경쟁 점화	점화
民俗•巫俗	美 유행점치기 ‘퓨처리스트’ 뜬다…트렌드 예측해 상품화	유행점치기
服饰	“허리띠 조이자” 증권사, 예탁금 이용료을 낮춰	허리띠 조이자
法律•法学	서울대 연구센터, 남해안 적조주범 새 생물 찾았다	적조주범
不动产	‘인터넷 복덕방’ 활성화…온라인 표준계약서등 도입	인터넷 복덕방
佛教	2003년 5.18 화두는 ‘평화’	화두
生物学•生理学	맥도날드 “매장 속살까지 다 보여드립니다” 주방공개 행사	속살
数学	美“北，리비아에 核팔고 돈 받았다”/北 ‘核장사’ 드러나… ‘核 방정식’	방정식
植物学•植物	건강한 인터넷/ ‘사이버 독버섯 제거’ 제4의 혁명을	사이버+독버섯
药学	유가 연일 고공행진…정부 대책 약효 있다	약효
水利	위험수위 넘은 육군-军 검찰 갈등	위험수위
言语	내경제 10년불황 마침표 찍을까…수출-생산 호조 나타내	마침표

续表

类别	例句	构成方式
历史	자산 10조원 화교 큰손 '인천 상륙' 준비	인천 상륙
体育	韩日 동해상 대치 타결/팽팽했던 '줄다리기' 39시간	줄다리기
音乐	"NEIS 개인정보 수록 합헌" …헌재, 지루한 논란 종지부	종지부
医学	위기의 국정원, 민간인 대거 수혈	수혈
政治	해커에도 햇볕정책?	햇볕정책
地理•地质	증권업계 지각변동 시작됐다…동원-한투 6월1일 합병	지각변동
航空	青 "참여정부에 낙하산인사는 없다"	낙하산
海洋•船舶	박주영 본프레레호 합류	본프레레호
别名	아시아청소년축구 득점왕-MVP 박주영, 마라도나 빰친 '수줍은 킬러'	킬러
地名	"중국엔 만리장성…한국엔 녹색장성"	녹색장성

从以上的举例可以看出，韩国语词汇隐喻的构成主要有三种方式，即前喻式、后喻式和整体隐喻式。前喻式隐喻词是指词的前一个成分喻指后一个成分，喻体在前，本体在后，形成偏正关系的词语。后喻式隐喻词是指词的后一个成分为喻指，前一个成分为直指，即用比喻的方法来反映对象，本体在前，喻体在后，形成偏正关系的词语。整体隐喻式以整体形式构成喻体，词语真正的含义是本体，本体有时与喻体一同出现在句子中，有时省略。抽取三个例子进行说明：

(1) 펜싱기대주 서미정 플뢰레 金 찔렀다
后喻式构成：펜싱+기대주

(2) "징검다리 취업" 는다
前喻式构成：징검다리+취업

(3) "한국서 온 득점기계" 박지성 응원가 히트
整体隐喻式构成：득점기계

3 基于 WordNet 的韩国语隐喻判断方法

3.1 WordNet 简介

WordNet 是一个根据语义学理论建立起来的词义网，是以同义词集合作为基本建构单位进行组织。这些同义词集合之间又是以一定数量的关系类型进行关联的，这些关系包括上下位关系、整体部分关系、继承关系等。WordNet 中的所有词语都处于一张相互交织、相互联系的网络中，是这张网上的一个节点，与前后左右的词语都发生关系。这种数据存储结构非常有利于计算结点间的最短路径，因此在判断词语关系上十分有效。

WordNet 最早是以英语为对象的一个词汇语义网络，后来逐渐发展为一个多语言词汇语义网络。北京大学率先在 WordNet 中加入汉语词汇语义知识，构建中文概念词典（CCD: Chinese Concept Dictionary）。CCD 是一个双语 WordNet，是全球 WordNet 资源建设的组成部分，构建过程中直接复用 WordNet 的理论、方法、技术，通过一定的匹配技术，实现汉英双语概念的对应。构建完成的 CCD 知识库，反映了汉语的词汇语义特点，可以在汉语词义消岐、信息检索、文本处理等方面发挥巨大的作用。

WordNet 是一个多语言词汇语义网络，其中也包括韩国语词汇语义网络。韩国语词汇语义网络

的构建借鉴CCD的构建过程，由于CCD中具有英文和中文概念，可以利用英-韩双语词典、中-韩双语词典增加韩文概念，通过一定的匹配技术，实现英、中、韩三种语言概念的对应。我们在导师的指导下参与了韩国语词汇语义网络的构建，过程中收集到了大量的资料，本研究也将以此为基础展开。

3.2 基于WordNet的韩国语隐喻判断方法

基于构建好的同义词词典WordNet(韩国语词汇语义网络，以下类同)，我们可以实现对部分隐喻表达的判断。在WordNet中，所有的词都被组织在一棵或几棵树状的层次结构中。我们通过计算两个节点间路径的长度就可以衡量这两个词所代表的概念的语义距离。以名词性隐喻为例，利用WordNet的语义网所提供的名词分类及其上下位的信息，可以较为迅速准确地判断出语句中是否发生了隐喻现象。抽取几个事例进行说明：

(1) 그 축구 선수 "득점기계" 입니다. (那个足球选手是一个得分机器。)

通过查询WordNet，我们发现这样的上下位关系传递链：

축구 선수 (足球选手)
=>운동 선수, 스포츠맨…(athlete, jock)
=>경쟁 참대, 경연자… (contestant)
=>사람, 개인, 누군가…(person, individual, someone…)
=>생물, 유기체, 생명…(life form, organism, being…)
=>실체(entity)

득점기계 (得分机器)
=>고안물, 장치, 설비…(device)
=>기구, 공구, 도구…(instrumentality, instrumentation)
=>인공물, 가공품, 가공 유물…(artifact, artifact)
=>물건, 물체…(object, physical object)
=>실체(entity)

从查询结果我们可以看到축구선수和득점기계分处于실체根节点下的两个不同节点之下，两个词之间的距离相对较远，因此可能属于一种隐喻的用法。为了解决对于这个隐喻的理解的问题，我们再次利用WordNet对득점기계这个单词进行检索。在二次检索中发现득점기계一词拥有不止一种上下义关系，而在下列的关系中，득점기계和축구선수在"사람, 개인, 누군가 …"这一层次上得到了交叉。

득점기계 (得分机器)
=>사람, 개인, 누군가…(person, individual, someone…)
=>생물, 유기체, 생명…(life form, organism, being…)
=>실체(entity)

其中所涉及的两个对象虽然分属于不同的类别，但是在语义网中能够找到交叉点，由此我们可以推知这句话确实发生了隐喻现象。

由此可见，利用WordNet提供的分类和词汇之间的上下义关系来判断语句之中是否发生了隐喻现象是一种较为有效的手段。我们可以通过首先筛选出待判断句子中有可能涉及到隐喻使用的词语，利用WordNet查询这些词语最常见释义的上下义关系(包括层级关系，相互距离等)，根据判断结果确定该语句是否发生了隐喻现象。如果答案是肯定的，可以继续利用WordNet对相关词语的上下义关系进行再次的检索，直到找到两者之间的交叉点。这个交叉点对于我们对隐喻的理解能起到很大的帮助。

4 韩国语隐喻知识库的构建

隐喻理解是自然语言理解不可回避的一个难题，如何让计算机可以理解隐喻是计算语言学研究者们急需解决的难题。利用计算机进行隐喻处理，目前通常是依靠隐喻知识库进行处理的。由上面的讲述可以看出，WordNet 作为一个关系清晰的大型语义网络，其中有关名词、动词、形容词等概念的层级描述，以及词汇上下义关系的描述，在隐喻的判断方面发挥着巨大的作用。基于这个优势，我们可以对 WordNet 进行合理的裁剪，加入一定的句法语义知识和隐喻知识，建立源域和目标域的映射关系，构建韩国语隐喻知识库。

本研究主要以名词及名词短语作为研究对象，韩国语隐喻知识库的构建的基本步骤如下：

1. 首先以东亚日报作为语料来源，初期取样抽取以新闻标题为主，通过分析选出具有隐喻意义的句子。

2. 然后利用韩国语句法分析器对所有的取样例句进行句法分析，提取出主、谓、宾上的名词性成分，分别对名词性成分进行分析，把经常作为源域的词语建立为源域词语词表，把经常作为目标域的词语建立为目标域词语词表，对 WordNet 进行合理裁剪，以源域词语为核心建立隐喻知识框架，实现源域词语和目标域词语的对应。

源域词语	谓语词	目标域词语	标记词	隐喻类型	CCD 映射	评价
흐름		세계화/세계질서/경제/시대	의	后喻型	是	
봄		여행/수학/경제/교육/과학 기술		整体隐喻型	是	
막	내리다	새 천년	의	后喻型	是	
지옥		인간	의	后喻型		消极

图 1　源域与目标域对应的知识库

3. 对于加工后的隐喻知识框架继续进行完善，加入句法和语义知识，这样构建完成的韩国语隐喻知识库实际上是一个句法语义知识库，可以兼顾语法与语义范畴。

依靠这样建立起的隐喻知识库，我们不仅可以判断一个句子是否使用了隐喻表达，还可以同时检索出发生隐喻表达的名词性词语的句法语义知识。同时由于目标域概念是以同义词集合的形式进行组织的，这样就便于通过一个隐喻表达检索出一类隐喻表达(如图 1)。

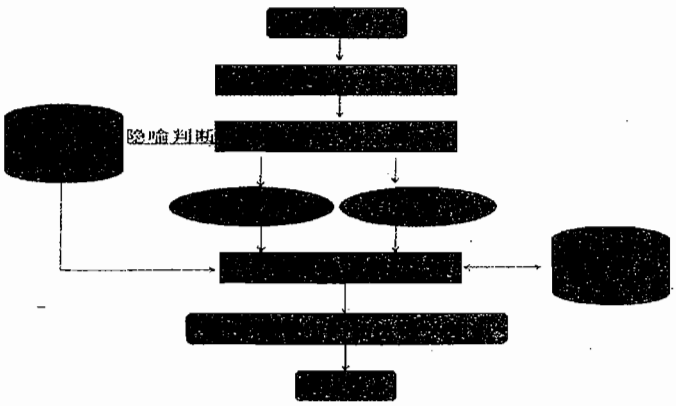

图 2　基于隐喻知识库的隐喻识别过程

参 考 文 献

[1] 胡壮麟. 认知隐喻学, 北京大学出版社, 2004.
[2] 束定芳. 隐喻学研究. 上海: 上海外语教育出版社, 2000.
[3] 程琪龙. 认知语言学概论. 北京: 外语教学与研究出版社, 2001.
[4] 兰纯. 认知语言学与隐喻研究. 外国教学和研究出版社, 2000.
[5] 林杏光. 词汇语义与计算语言学. 北京: 评议出版社, 1999.
[6] 陈群秀. 一个在线义类词典: 词网WordNet. 语言文字应用, 1998.
[7] 俞士汶. 语料库与综合型语言知识库的建设, “自然语言处理若干重要问题”学术研讨会报告, 2002.
[8] 黄孝喜, 周昌乐. 隐喻理解的计算模型综述[J]. 计算机科学, 2006.
[9] 王治敏. 隐喻的计算研究与发展[J]. 中文信息学报, 2006.
[10] 张霄军, 曲维光. 隐喻研究与隐喻知识库建设. 心智与计算, 2008.
[11] 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.
[12] 王金锦, 周昌乐. 面向隐喻计算的实体概念知识库构建方法研究[J]. 计算机科学, 2009.
[13] 戴帅湘, 周昌乐. 隐喻计算模型及其在隐喻分类上的应用. 计算机科学, 2005.
[14] 姜柄圭. 学术语言的隐喻现象与汉韩翻译. 汉韩语言对比研究, 北京语言大学出版社, 2007.
[15] 毕玉德, 崔杞鲜, 刘扬. 多语种词汇语义网构建中的几个问题, 第八届全国计算语言学学术会议论文集, 2005.
[16] 毕玉德, 阎艳萍. 一种基于WordNet的多语种词汇语义网半自动构建方法. 解放军外国语学院学报, 2008, 5.