面向汉韩机器翻译的隐喻研究及隐喻知识库构建设想

徐超

解放军外国语学院 洛阳 471003

E-mail: xcsuper@yahoo.cn

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。本文引介了国内外的几个典型隐喻知识库，从语言学和计算机科学的双重角度，分析了目前机器翻译中的隐喻处理策略。并且针对目前一些面向小语种的机器翻译忽略隐喻处理，导致翻译效率不高的问题，以韩国语为例，提出了一种面向汉韩机器翻译的双语隐喻知识库的建设设想。

关键词：隐喻，隐喻知识库，韩国语，机器翻译

An Idea of Metaphor Study and Metaphor Knowledge Corpus Building for Chinese-Korean Machine Translation

Xu Chao

PLA University of Foreign Languages, Luoyang 471003

E-mail: xcsuper@yahoo.cn

Abstract: Metaphor processing is a difficult problem of Natural language processing, there must be a metaphor knowledge corpus to support.This article introduces several typical metaphor knowledge corpuses of the domestic and international metaphor knowledge corpuses.And this article,from the dual perspective of linguistics and computer science,tries to analyze the current metaphor processing strategies for machine translation.As numbers of machine translation for minority languages ignore the metaphor processing,it resulting in the consequence that translation efficiency is not high. Case in Korean,we try to put forward a bilingual Chinese and Korean Metaphor Knowledge Corpus building idea for Chinese-Korean Machine Translation.

Keywords: metaphor,metaphor knowledge corpus,Korean,machine translation.

1. 引言

机器翻译的关键就是识别与消解自然语言固有的歧义。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。当机器处理自然语言的时候，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，人们现在可以利用各种机器翻译系统实现绝大多数文本的多语言对译。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机器翻译提出了新的挑战。

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）。

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，引介了国内外的几个典型隐喻知识库，从语言学和计算机科学的双重角度，分析目前机器翻译中的隐喻处理策略。并且针对目前一些面向小语种的机器翻译系统忽略隐喻处理，导致翻译效率不高的问题，以韩国语为例，提出了一种面向汉韩机器翻译的双语隐喻知识库的建设设想。

2. 隐喻理解对机器学习提出的挑战

2.1 从汉语隐喻说起

语言学家 W. Taubert 曾说过，“自然语言是一套规则加噪声”。这些噪声就包括隐喻理解。隐喻是修辞学的传统研究内容，但认知语言学认为隐喻是一种思维方式——隐喻概念体系。在计算语言学领域，计算语言学家为了实现自然语言理解，开始关注隐喻的识别和求解。

隐喻分为词汇级、语句级和篇章级隐喻。

词汇级隐喻中，像“桂冠”、“铁公鸡”、“老狐狸”、“蜻蜓点水”、“快刀斩乱麻”“眉毛胡子一把抓”“韩信点兵，多多益善”等这些词语，都是借助隐喻形成的。例如，“蜻蜓点水”有两个义项：①蜻蜓在水面飞行时用尾部轻触水面的动作；②做事肤浅不深入。②显然是隐喻。只要机器中的词汇知识库登录了这些词语的各种义项（包括本义或隐喻义），识别和理解这些词语就不会特别的困难。

语句级的隐喻主要是由词汇级的隐喻延伸而来的，有“人生如旅途”、“知识是海洋”、“论争是战争”、“时间是金钱”等。“旅途”、“海洋”、“战争”、“金钱”都是普通的名词，用在这里使整个语句有了隐喻的意义。

“打起黄莺儿，莫叫枝上啼。啼时惊妾梦，不得到辽西。”是篇章级隐喻的典型例子。其中“辽西”喻指古战场，整首诗则反映妻子对在远方征战的亲人的魂牵梦绕。篇章级的隐喻就是通常所说的“弦外之音”。另外，唐诗还经常把“杨柳”和“离别”、“思念”联系在一起。计算机能不能学到这样的知识呢？对包含“杨柳”的诗篇进行比较、计算、判别，探求诗篇所表达的情感，进而达到对不同语境中的“杨柳”词义的理解。可见，隐喻理解是机器学习急需解决的瓶颈问题。

2.2 韩国语隐喻问题

前面我们已经讨论过汉语的隐喻问题，这里在讨论一下韩国语隐喻问题。汉韩两种语言的隐喻有着许多共同点，但也有不同的表达方式，不管是词汇级的隐喻还是语句级的隐喻，都有所差异。

词汇层级上，表现为两个方面，一方面是韩国语汉字词的隐喻，这显然是受汉语影响的，这类隐喻类似于汉语的隐喻现象，可以参照汉语的隐喻问题进行理解和处理；另一方面韩国语

还有很多带有韩国文化的隐喻词汇，“쥐꼬리”、“가방근”“공주병”、“냄비근성”、“백수”等词语都是隐喻形成的，例如“쥐꼬리”本义是老鼠尾巴，此处隐喻为微不足道的东西。可见，这些隐喻问题已经不同于汉语的隐喻问题，这些显然是受韩国文化影响的，词汇中隐喻的形式和喻体，两种语言明显有着显著的差异，这样在建立双语隐喻知识库的时候就必须进行特殊的标注和处理，然后使之相链接。

语句层级上，也有一些带有韩国特色的隐喻表现。例如“인력을 사다”中，“购买人力”显然是隐喻用法，这里把“人力”隐喻为一种“商品”；而在“벼룩의 간을 빼먹어라”中，字面义是“吃跳蚤的肝脏”，而隐喻为“要钱没有，要命一条”。

语句级隐喻进行翻译时有直译和意译两种方式，一般机器翻译的隐喻处理通常是直译的方式，然而有时候，直译处理会不太符合语言习惯，很多情况下会使人产生误解。汉语中“研究队伍”若直译为“연구부대”，显然不太容易被韩国人接受，然而进行意译“연구팀”，就符合韩国语的习惯了。例如：

在语言信息处理领域打了一场遭遇战。

언어처리 영역에서 한차례 조우전을 펼쳤다.(直译)

언어처리 영역에서 뜻하지 않았던 문제들과 부딪히게 되었다.（意译）

可见，通常直译翻译虽然有时会导致翻译质量不高，但还是便于处理的，而意译翻译完全不出现隐喻词的对应，这就是机器难以理解和处理的，如何实现汉韩两种语言的隐喻词对应，也是机器学习的一个难题。

3. 从计算的角度看隐喻问题

自然语言处理面临的棘手问题之一就是隐喻理解，因为隐喻不具有固定的意义。为此，自然语言处理的研究者或强词夺理地说隐喻并不重要，不必过多考虑；或者宣称他们已经开发了一个令人满意的算法，可以处理各种各样的隐喻问题。（Debatin, 1992）要解释隐喻的深层现象，犹如给洋葱一层一层剥皮，直到构建语言的隐喻表现出相当明显的系统性和规律性，这是隐喻出现的印记。正是基于这个认识，促使有些学者用计算理论处理隐喻用法与“已经存在”的知识的关系。本章节将重点介绍一下国内外的一些隐喻计算模型。

3.1 基于优选限制的隐喻计算模型

Wilks于1975年提出优先语义学。他认为，在词义排歧的过程中，涵义的取舍不要看成是完全的接受或完全的拒绝，而应看成是在各种可能的涵义中进行优选。当单词彼此结合的时候，优选程度最高的那些涵义被确定为可接受的涵义，而优选程度低的涵义则被拒绝。

基于优选限制的方法对一个语言中的多义词的各语义进行描述，并对各语义赋以语义优先特征，当计算机在处理隐喻问题时，就可以根据不同的选择限制，消解歧义。比较有代表性的系统有Fass（1991）提出的可以处理隐喻、转喻、字面义反常表达的隐喻理解模型met5系统，该系统使用隐喻知识库为Sense-frame。

图1左下方显示了名词“animal”在Sense-frame中的描述。“[supertype, organism1]”是一个语义网络体系，“node()”表示该词条是一个名词词条，“[biology1, animal1]”和“[composition1, flesh1]”为该词条的语义优先特征，“[it1, drink1, drink1]”和

“[it1,eat1,food1]”为该词条的句法组合模式。“it1”指该词条，即，“animal1”在“[it1,drink1,drink1]”中能够被“it1”代替，在“[it1,eat1,food1]”中也能够被“it1”代替。这句法组合模式在Sense-frame中被称为“cell”。利用上述形式化描述，Fass用程序Met5实现了对隐喻理解的模型解释。如对“car drinks gasoline”，Met5系统表示如下：

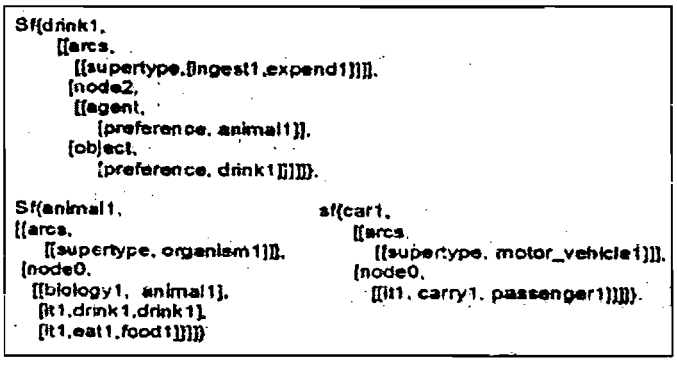


图1 “car drinks gasoline”隐喻的语义解释

首先分析动词“drink”，本身应与动物范畴及其下位范畴结合构成句子，而此处位于主语位置的词是“car”，不属于动物范畴，所以计算机进行隐喻处理，分析此处的“car”喻指动物，，进而分析出上述语义解释。

3.2基于实例的隐喻计算模型

基于实例的隐喻计算模型通过实例对比，解释旧隐喻，识别新隐喻。比较有代表性的系统有Martin（1990）提出的可以实现表达、理解和学习新隐喻的隐喻理解模型MIDAS系统，该系统使用隐喻知识库为MetaBank。

MIDAS系统对隐喻的理解是通过MIS(Metaphor Interpretation System,隐喻理解系统)子系统实现的。MIS首先进行句法切分和初步的语义表达，基于本义。然后处理最后的意义，基于两个推理过程。第一个推理以更为特定的概念替换抽象的概念；第二个推理用已知的喻源概念替换相应的目标概念。该系统可以从常规隐喻的表述中提出本体来获得潜在的隐喻解释，如“INFECTING AS GIVING”（传染是给予）。试以“How can I kill the process?”（我该如何杀掉该程序?）为例。MIDAS会对照已经确定的概念和隐喻图的限制因素，如“killing-as-slaying(a living thing)”，“killing-as-terminating(proceedings)”，“killing-as-dismissing(an arraignment)”，“killing-as-eating(consume whole contents of)”，“killing-as-defeating(an opponent)”等。由于目标概念仅满足“killing-as-terminating”的要求，系统会作出正确的理解，该问题的意思为“我该如何终止这个过程？”MIDAS系统还可以对新隐喻进行识别，由子系统“Mes”(Metaphor Extension System，隐喻延伸系统)来实现。为找寻一个新隐喻的意义，Mes先搜索，然后评估备取的隐喻，最后选取概念上最接近者，整个过程以该隐喻的应用并储存新隐喻图结束。

3.3基于大规模语料库的隐喻计算模型

前述的计算模型大部分依赖于手工构造的语义信息库，在完备性和通用性上具有较大的局限性。随着语料库语言学的飞速发展和各种语料库的构建，语料库也被应用与隐喻计算，比较

代表性的系统是Mason提出的基于语料库识别和分析常规隐喻的CorMet系统。

CorMet首先给定领域关键字，通过搜索引擎从Internet中收集具体领域的语料库，利用apple pie parser解析器分析所搜集的文档，获得语句格框架。然后获取领域的特征谓词，主要是根据语料库中各词干在所有词干中的比率与通用的频率词典进行比较，相对频率高的那些动词词干即为领域特征谓词。最后利用Resnik的选择优先学习算法，获取一个动词的语义优先，给出一个可行的选择优先之间相似度比较的度量，以计算不同领域特征谓词之间的相似度和相关度，谓词的选择优先用一个向量来表示，其中的元素对应于WordNet中相应节点与它之间的选择相关度，而谓词之间的相似度用选择优先向量的点积来表示，然后利用最近邻knn聚类分析算法对节点进行聚类，用于表示给定领域的特征。CorMet利用极性来确定两个概念在隐喻中的成分，极性表示两个概念或领域之间概念转移的方向和数量，当一个概念在某个领域的语言特性被应用与其他领域的另外一个概念时，极性就变为非零，如果一个概念a适用的动词同样适用于概念b，而有些b适用的动词在a中不适用，则称a为喻体概念，b为本体概念。

从上面的隐喻计算模型来看，隐喻的识别与理解往往需要隐喻知识库的支撑。综合上述隐喻计算模型可知，一个隐喻知识系统的建立主要包括以下五个方面的内容：（1）相关语料库或实例库；（2）语义类型分类体系（3）隐喻知识表述体系；（4）隐喻识别算法及隐喻理解模型；（5）新隐喻学习模型。其中，语料库提供一定的语言理解环境，提供一定的背景知识和领域知识，不同语言的隐喻存在方式和隐喻类型是不同的，建立健全的语料库有助于计算机参照理解。隐喻的形式化表示是计算机理解隐喻的前提，必须搞清楚语言的内部关联性特征，不同的语言的隐喻构成是不同的，因此对隐喻的表示也是各有看法，但首要的原则是解释充分性和可操作性。借助隐喻系统可以实现隐喻的处理，但隐喻总是随着社会的发展而层出不穷，因此需要建立隐喻学习机制，不断添加新的隐喻知识表示，不断提高系统的健壮性。

4. 面向汉韩机器翻译的双语隐喻知识库的建设设想

在汉语隐喻知识库的建设方面，北京大学计算语言学研究所已经构建了用于识别的名词性隐喻知识库、用于识别的动词优先选择语义类之知识库、用于一类隐喻句理解的名词显著特征知识库等多个隐喻知识库，已经可以处理绝大多数汉语隐喻问题。

其中下面是用于识别的名词性隐喻知识库的基本构成，如图2，图3：

19980123-11-003-008/m 握/v 着/u 这样/r 的/u 手/n ，/w 能/v 不/d 倍感/v 温暖/n 和/c 鼓舞/n ？/w 像/p 他们/r 这个/r 年龄/n 的/u 小伙子/n ，/w 有/v 多少/r 还/d 依偎/v 在/p 父母/n 的/u 身旁/s ，/w 享受/v 着/u 温馨/a 的/u 抚爱/vn ；/w 有/v 多少/r 正/d 坐/v 在/p 宽敞/a 明亮/a 的/u 教室/n ，/w 遨游/v 于/p 〈知识/n 的/u 海洋/n〉，/w 而/c 他们/r ，/w 却/d 走/v 进/v 了/u 绿色/n 的/u 军营/n ，/w 来到/v 了/u 北大荒/ns ，/w 用/p 青春/n 和/c 血汗/n ，/w 为/p 军队/n 为/p 国家/n 创造/v 巨大/a 的/u 财富/n 。/w

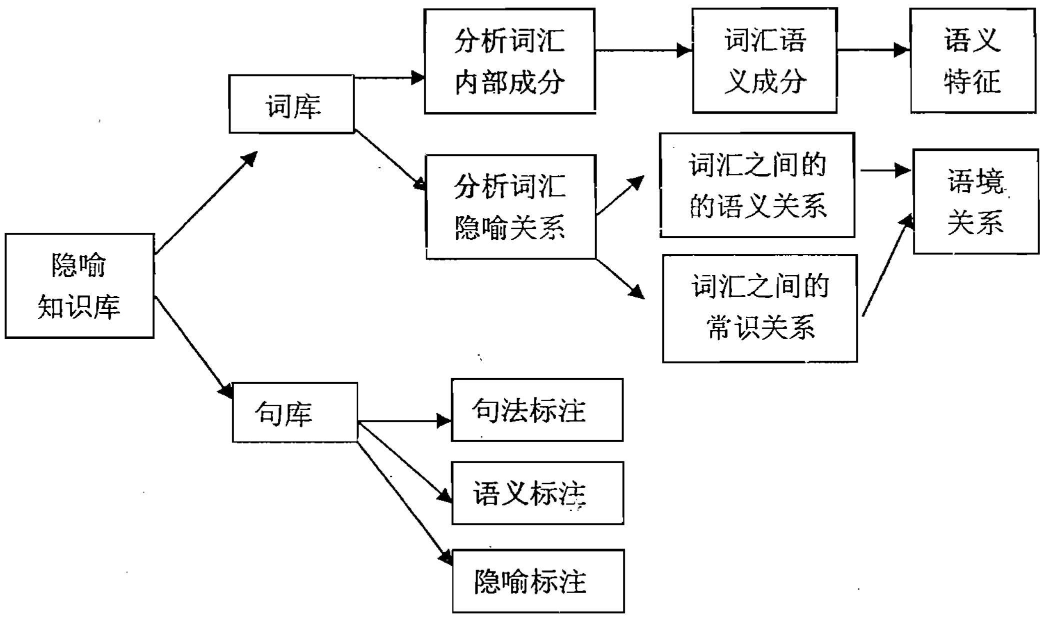
19980303-05-003-003/m 与会/vn 同志/n 认真/ad 回顾/v 1997年/t 全军/n 文艺/n 工作/vn 后/f 认为/v ，/w 军队/n 文艺/n 创作/vn 的/u 根本/a 任务/n 是/v 贡献/v 党/n 的/u 十五大/j 精神/n ，/w 高举/v 邓小平理论/n 伟大/a 旗帜/n ，/w 站/v 在/p 〈时代/n 高度/n〉，/w 积极/ad 弘扬/v 主旋律/n ，/w 为/p 部队/n 和/c 社会/n 提供/v 更/d 多/a 更/d 好/a 的/u 〈精神/n 食粮/n〉；/w 加强/v 文艺/n 创作/vn 要/v 从/p 深入/v 生活/vn 抓起/v ，/w 把握/v 〈时代/n 脉搏/n〉，/w 使/v 作品/n 更/d 具有/v 时代/n 〈生活/vn 气息/n〉；/w 要/v 进一步/d 树立/v 精品/n 意识/n ，/w 把/p 主要/b 精力/n 放/v 到/v 军事/n 题材/n 创作/vn 上/f ，/w 在/p 艺术/n 上/f 进行/v 新/a 的/u 探索/vn 和/c 表现/vn ，/w 开创/v 崭新/b 的/u 多样化/vn 创作/vn 局面/n ；/w 全面/ad 提高/v 文艺/n 队伍/n 的/u 整体/n 素质/n ，/w 使/v 我们/r 的/u 队伍/n 成为/v 一/m 支/q 政治/n 强/a 、/w 业务/n 精/a 、/w 作风/n 硬/a 、/w 特别/d 能/v 战斗/v 的/u 队伍/n 。/w

图2 标注了隐喻短语的语料库

[illegible]	[illegible]	目标域词语	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
海洋		知识/时间/人生/歌/花/舞/笑声/灯/市场/经济/光/科学/知识/影事/大赛/生活	的		是		
大战		伞兵/洪水/经历/羊绒衫/缺苗/储蓄/料车/啤酒/价格/贸易/点球			是		
道路	健康	谈判/社会主义/和平/人生/革命/资本主义/致富	的		是		
堤坝		旧观念/心灵/ 理智/心/感情	的		是		
地狱		人间					消极

图3 源域与目标域对应的知识库

相对于汉语隐喻知识库的完善，韩语的隐喻知识库还是很欠缺的，这也是目前汉韩机器翻译质量不高的重要原因之一，因此建立一个完善的韩语隐喻知识库迫在眉睫。我们将根据国内外的一些隐喻计算模型，以及已经建成的汉语隐喻知识库，进行韩语隐喻知识库的构建。我的设想是：隐喻知识库由词库和句库两部分构成：词库包括动词库、名词库、形容词/副词库，在各词库中要对词语进行隐喻概念分类、隐喻知识描述、隐喻知识体系构建；句库包括韩国语的各主要句型，并且对句型进行句法标注、语义标注和隐喻标注。框架结构如下：



在构建好汉语隐喻知识库和韩语隐喻知识库后，还需要将两隐喻知识库进行对应链接，实际工作就是双语隐喻句获取，可利用检索的方式从汉韩双语平行语料库中抽取带标注的隐喻句，建立汉韩双语隐喻句语料库，这样以汉语隐喻知识库、韩语隐喻知识库、汉韩双语隐喻句语料库为基础，就可构建汉韩双语隐喻知识库了。

以该隐喻知识库为核心，最终可应用于汉韩机器翻译系统以及汉韩隐喻句检索系统等现实的应用系统。利用汉韩机器翻译系统可对汉韩的大部分隐喻句进行处理，输入汉语隐喻句，得到比较理想的韩语隐喻句；利用汉韩隐喻句检索系统可以实现汉韩隐喻句实例的抽取，输入汉

语隐喻句，可从系统中抽取出相对应的韩语隐喻句。当然，在应用系统中，还可以添加统计记忆程序，不断存储新登录的隐喻词句，以不断提高系统的语言处理能力。

参考文献

[1] 胡壮麟. 认知隐喻学, 北京大学出版社, 2004.

[2] 张维鼎. 意义与认知范畴化，四川大学出版社，2007.

[3] 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.

[4] 崔健. 韩汉范畴表达对比，中国大百科全书出版社，2002.

[5] 胡壮麟. 功能主义纵横谈，外语教学与研究出版社，2000.

[6] 王逢鑫. 英汉比较语义学，外文出版社，2001.

[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] 张霄军. 隐喻与换喻计算综述：第四届全国认知语言学研讨会. 2005.

[12] 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.

[13] 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.

[14] 伍铁平. 普通语言学概要，高等教育出版社，2006.

[15] 陆俭明. 现代汉语语法研究教程，北京大学出版社，2004.

[16] 李正栓，孟俊茂. 机器翻译简明教程，上海外语教育出版社，2009.

[17] 冯志伟. 机器翻译研究，中国对外翻译出版公司，2004.

[18] 王克友. 翻译过程与译文的演生，中国社会科学出版社，2008.

[19] 李亚舒，黄忠廉. 科学翻译学，中国对外翻译出版公司，2004.