情感隐喻语料库构建与应用

张冬瑜, 杨亮, 郑朴琪, 徐博, 林鸿飞*

大连理工大学信息检索研究室, 大连 116023

* 通信作者. E-mail: hflin@dlut.edu.cn

收稿日期: 2015-09-24; 接受日期: 2015-10-20; 网络出版日期: 2015-12-07

国家自然科学基金 (批准号: 61572102, 61402075) 和辽宁省自然科学基金 (批准号: 201202031, 2014020003) 资助项目

摘要 隐喻普遍存在于人类语言中, 是对复杂而抽象的情感进行概念化的重要手段, 也是情感计算的重要组成部分. 无论面向计算机科学还是语言学的应用, 情感计算都需要大规模、高质量的语料库作为资源支撑, 才能进行更加深入、广泛的研究. 因此, 本文构建一个大规模汉英双语情感隐喻语料库, 用以满足情感计算对语言资源建设的迫切需求. 首先归纳情感隐喻语料库构建的理论框架, 然后从概念界定、设计原则、语料收集、语料标注、质量监控 5 个方面分别进行说明, 最后论述了该语料库在机器翻译、外语教学及情感分析方面的应用.

关键词 情感计算 情感隐喻语料库 机器翻译 外语教学 情感分析

1 引言

人类具有丰富而复杂的情感. 随着人工智能和互联网的发展, 特别是微博、微信、在线购物、社交网站等网络媒介的迅速崛起, 产生了大量包含多种情感倾向的叙事或评论性信息. 人们越来越迫切地希望计算机像人一样具备感知、理解并生成复杂情感的能力, 从而对日益庞大的海量数据所传递的情感信息进行分析 and 处理. 情感计算就是赋予计算机人类所具有的情感能力, 对来源于情感或与情感相关的信息进行计算^[1]. 它是自然语言处理与人工智能的重要研究内容, 也是计算机语言智能向纵深发展的关键问题之一, 具有重要的理论和实际应用价值.

情感是抽象、模糊、不易描述的, 在这种情况下, 人们常常使用生动形象的隐喻表述内心情感, 例如: “焦头烂额”、“心花怒放”、“吓破胆”、“飘飘然”、“gust and boiling wave”、“I’m six feet off the ground”等. 情感隐喻作为思维方式普遍存在于人类语言中, 使用频率非常高, 日常交流中每三句话中就可能出现一次^[2], 人一生平均使用大约 2100 多万次隐喻表达^[3]. 隐喻是建立在人类生理体验之上的, 是对复杂、抽象的情感进行概念化的重要手段, 是情感表达的主要方式^[4], 因此情感隐喻是情感计算中不可缺少的组成部分, 具有十分重要的意义. 情感隐喻语料库是情感隐喻计算的基础, 只有在大规模、高质量语料库的支撑下, 情感隐喻研究才能更加深入、广泛地进行. 虽然对情感隐喻语料资源建设需求迫切, 但是目前国内外情感隐喻语料库资源比较匮乏, 因此下文从隐喻以及相关情感语义资源这两类资源的建设研究加以回顾.

Master Metaphor List (简称 MML) 是最早在概念隐喻理论指导下建立的常规隐喻语料库^[5]. 许多有影响力的隐喻语料资源都一定程度上借鉴了 MML 的构建理论和方法. MML 语料来源于英文隐

引用格式: 张冬瑜, 杨亮, 郑朴琪, 等. 情感隐喻语料库构建与应用. 中国科学: 信息科学, 2015, 45: 1574–1587, doi: 10.1360/N112015-00242

喻书籍文献、在线论坛和学生作文中的隐喻实例. 以此为基础 George Lakoff 等将隐喻分为情感、时间结构、心理事件和其他 4 大范畴. 在此范畴下的隐喻包含了从源域到目标域映射实现的过程描述和对应例子. 但是 MML 在对概念隐喻映射描述的同时, 并没有兼顾隐喻词汇化方面. Metalude 是一个在线交互隐喻英语语料库^[6], 建设初期收录了 9000 多条隐喻实例及其本体、喻体、字面义、比喻义、词性、分类等信息, 并在逐步扩充和完善. 与 MML 不同的是, Metalude 从词汇角度考察隐喻, 并且按照词汇学标准选择和录入语料. 但是其情感按照概念隐喻进行描述, 并没有对类别进行详细划分, 不利于面向情感计算的隐喻研究. MetaBank 是应用于自然语言处理的隐喻知识库^[7], 其构建体系建立在隐喻的采集、泛化与拓展等方面, 推动了隐喻计算中资源建设方法的研究. 语料采集使用了 MML 的两个语料库: 华尔街日报语料库和 UNIX 下的邮件文本语料库, 由此可以看出, MetaBank 并不关注隐喻的情感因素, 仅仅适合用于专门领域的隐喻研究. Hamburg Metaphor Database 是少数面向多语言的隐喻语料资源代表, 可供在线搜索的德语、法语数据库^[8]. 主要包括例句、词条、隐喻义、字面义、源域、目标域等字段, 也以词汇化隐喻为建库对象. 此外还有 Sense-frame^[9] 和 ATT-Meta^[10] 等隐喻语料资源. 目前国内隐喻资源建设的研究仍处于起步阶段, 以厦门大学研制并公布的汉语隐喻句库为代表^[11], 它包括 1 万个汉语隐喻句, 规模有限, 且不支持情感分析.

另外, 情感语义资源建设也为情感隐喻研究提供了重要基础和有利支撑. 例如 CCDb 和 HPI 情感知识库^[12]; Berardinelli 电影评论语料库; 清华大学标注的景点介绍方面的情感语料库^[13]; 中国科学院计算技术研究所谭松波的数码产品、酒店等情感语料库; 大连理工大学信息检索实验室标注的情感语料库^[14]等.

由此可见, 国外情感隐喻资源建设大致上能够传承和相互借鉴, 在标注体系设计和语料采集加等方面比较全面和完善, 但是对隐喻的情感分类体系和描述并不尽如人意, 并鲜有专门用于情感计算的语料资源, 另外对跨语言的语料建设与对比研究也存在着不足. 国内研究都针对各自不同的构建目的在语料采集、标注中各自为政, 情感的形式化表达和分类体系也没有一套统一的标准. 因此, 为了尝试解决上述问题, 本文目标在于构建一个大规模英汉双语情感隐喻语料库, 该语料库具备以下两个方面的特色.

(1) 面向情感隐喻识别及应用. 鉴于带有情感词汇的隐喻对于情感分析有比较明显的标志, 本文的研究对象更加关注不包含情感词汇的隐喻, 通过领域分类和标注, 计算不同领域语义之间距离, 实现情感隐喻的识别.

(2) 建立以喻底为桥梁的汉语和英语情感隐喻理解. 汉英隐喻的内涵和所传递的情感具有很多差异, 本文对本体、喻体以及喻底进行标注, 并通过喻底建立汉英隐喻语义关联, 获取语料库中具有相同喻底的汉语和英语隐喻实例, 从中可以发现语言文化思维和由此导致的隐喻表达及寓意的差异, 找出语言转换规律.

2 理论基础

2.1 概念隐喻理论

概念隐喻理论框架包括情感隐喻本质、工作机制等方面. 概念隐喻理论是从认知角度对隐喻的研究, 以 Lakoff 和 Johnson 在 1980 年出版的《我们赖以生存的隐喻》为代表性成果^[15]. Lakoff 和 Johnson 认为隐喻是一种认知方式, 用形象具体的事物去理解、认知抽象事物, 其工作机制是从已知概念域(源域)到未知概念域(目标域)的映射过程^[16]. 而情感是复杂抽象、难以描述的, 情感概念需要以隐喻为手段进行, 因此情感隐喻的本质是抽象情感隐喻化的认知过程^[17~19]. 例如英汉两种语言都

有以人体为容器, 人的喜怒哀乐为这个容器中物质的隐喻, 因此就有了“喜出望外”、“他生了一肚子气”、“She is full of sorrow”等隐喻表达. 情感隐喻以人体体验为基础建立起来, 是对复杂抽象情感概念化的重要手段^[20]. 情感隐喻的认知观与以往语言学中的另外两大学派——结构主义和转换生成语法具有明显的不同. 结构主义强调语言形式的重要性, 关注语言系统结构的语义作用, 而忽视了语言要素和语义互动等因素. 转换生成理论实质上是一种“客观主义”, 它重视神经系统和心理机制对语言的作用, 而低估了语境等社会因素和情感、认知等主观性.

具体地说, 需要通过对源域、目标域以及两者之间的映射来观察和描述情感隐喻, 还要重视语言、文化等多维角度. 语料语义标注体系需要以隐喻的本体、喻体以及两者之间语义互动过程中的相似性即喻体凸显的语义特征(称为“喻底”)为主体. 根据以上理论, 本文情感隐喻语料库标注体系中不仅包括本体、喻体, 还把喻底作为主要标注对象, 并尝试以喻底为桥梁, 建立英汉双语的语义关联, 为跨语言情感隐喻的对比研究奠定基础. 另一方面, 概念隐喻理论认为隐喻是发生在两个不同认知领域之间的映射, 本文语料标注中以此为指导, 增加了领域(概念的语义域)的标注.

2.2 自然语义元语言理论 (natural semantic metalanguage)

与概念隐喻理论一样, 自然语义元语言理论(以下简称 NSM 理论)也是从认知语言学角度对情感隐喻的研究, 探讨了存在于各种语言中普遍适用的情感隐喻描写模式. 该理论创建于 20 世纪 70 年代, 由波兰语义学家 Wierzbicka 提出^[21]. NSM 理论构建的语义分析体系对语言具有普遍、广泛的解释和描写能力, 在当今语言学领域具有十分深远的影响. Wierzbicka 的理论起源于理性主义哲学思想^[22], 认为语义在认知过程中承载着不同的文化特征, 因此只有运用那些人类共有的普遍概念, 才能对那些抽象、模糊和不确定的语义特征作出客观地解释. 因此, NSM 理论框架具备了对主观、模糊、思维性的隐喻进行描述的能力, 不仅可以解释隐喻语言内部的语义特征, 还可以用以解决跨语言的隐喻理解, 使用人类共有的普遍概念对蕴藏在不同文化下的隐喻加以诠释^[23]. 具体地, 自然语义元语言理论使用词汇描写概念, 并在此基础上创建了语义描写工具“语义基元”(semantic primitives). 语义基元是词或者短语等简单的语言表达式, 是跨越于文化的, 能够表达语言中最简单、最基本意义的语义核心. 例如“say”, “time”, “good”, “there is”等. 目前经过包括跨语言研究在内的反复验证, 已经形成了 60 多个语义基元^[24], 其中还包括诸如 good, bad, the same 等抽象、模糊的主观性基元, 这也形成了 NSM 理论区别于其他以客观语言为基础的语义分析理论的独特之处.

NSM 理论为语料库建设中情感隐喻表达的特征描述(或称语义形式化描述)提供了理论指导. “化简释义”是 NSM 理论最基本的研究方法, 即在语义分析过程中采用自然语言中基本、简单并具有广泛适应性的词汇解释语义概念. 本文在情感隐喻语料库标注中, 对带有标注者强烈主观判断的喻体的描述中, 遵循 NSM 的理论指导, 尽量使用简单、基础的词汇来描述, 如名词性、动词性或者形容性的单词或者短语, 例如在隐喻句“他的理论站不住脚”中, 使用“不可靠”这个词语对喻底进行解释标注. 而在“Tom is as merry as a cricket”这个英文隐喻句中, 使用单词“merry”作为隐喻义即标注的喻底, 这样的语义特征描写简洁方便, 并且有利于今后的英汉隐喻跨语言语义研究.

3 语料库的构建

3.1 概念界定

根据语言学、语料库构建及情感计算的理论和原则界定情感、情感类别、隐喻及本体、喻体、喻

底等涉及概念的内涵和特征; 通过文献研究、语言学和计算机专家意见征询及情感隐喻实例调查, 从便于情感隐喻形式化表达与计算的角度出发, 对概念进行整合、修改、删减等规范化处理, 确定面向英汉两种语言的情感隐喻语料库的相关概念界定。

以“情感”概念的界定为例, 心理学家们对情感进行了大量的定义和描述 (例如文献 [25,26]) 并探讨了情绪 (emotion)、情感 (affect)、心境 (mood) 和感受 (feeling) 等概念之间的联系和区别。其中文献 [27] 对此进行了细致地划分, 把情绪描述为由刺激引发的, 短暂而强烈的心理反应; 而心境是对情绪的体验和感受; 情感所指范畴较为广泛, 通常涵盖了情绪和心境, 具有持续性、社会性等特点; 感受则是强调情感过程的体验和感受。本文从情感计算角度, 以便于计算机分析、理解和表示为原则, 把“情感”定义为对客观事物的所有主观反应与体验, 对以上概念并不作具体区分。

对“隐喻”的定义, 本文采用前文概念隐喻理论 (见 2.1 小节) 中 Lakoff 和 Johnson 对隐喻最广泛的定义: 隐喻的本质是用一类事物理解和体验另一类事物, 即用源域中的简单/具体的事物去感受理解目标域中的复杂、抽象事物。依据此定义, 本文将所有在思维过程中使用事物 B 范畴的词语来描述事物 A 的性质特征的过程都归于隐喻范畴。按照这种划分, 从修辞角度定义的明喻、暗喻、拟人、通感、借代等都归属到了隐喻的范畴。其中被描述的对象 A 称为本体 (tenor), 用于描述本体事物 B 为喻体 (vehicle), A 与 B 相似性或相邻的特征为喻底 (ground)。连接 A 与 B, 表示隐喻关系的连接词, 如“as”, “like”, “好像”等称为触发词 (indicator)。

3.1.1 情感分类

目前对于情感的分类还没有形成统一的标准, 学者们从不同角度将情感划分为两类至几十类不等。其中影响比较广的几种分类, 例如 Plutchik 提出了 8 种人类基本的情感, 并以相似性、两极性及强度作为划分依据 [25]; Ekman 等 [28] 根据对面部表情的研究, 指出高兴、惊奇、恐惧、悲伤、愤怒以及厌恶 6 类基本情感, 目前这种分类得到了较为广泛的认可。我国对情感种类的分类始于先秦时期, 当时“七情六欲”的说法一直沿用至今。林传鼎 [29] 在心理学研究中把情感划分为安静、喜悦等 18 类; 徐琳宏等 [30] 在 Ekman 等情感划分基础上将情感分为 7 大类, 20 小类。

根据情感隐喻的特点, 并鉴于本文所收集的情感词汇本体及情感语料库都是在文献 [30] 对情感分类基础上建立起来的, 本文情感隐喻分类体系也依此建立, 把情感分为: 乐、好、怒、哀、惧、恶、惊 7 大类和 22 个小类。 (此 7 类情感仅仅是简称, 具体说明见图 1 中的 22 小类)。并在此分类基础上对每类情感的极性加以划分。因此本文从基本分类和情感极性两个维度对情感隐喻的情感类别进行描述, 有助于细粒度情感计算研究的开展。

3.1.2 领域

领域是指概念的语义域, 而隐喻是从已知概念域 (源域) 到未知概念域 (目标域) 的映射过程, 映射发生在两个不同认知领域之间, 本体与喻体之间的语义冲突, 或者称为范畴错位是隐喻的主要语义特征 [31]。因此, 可以通过计算本体与喻体在不同领域之间的语义距离来识别隐喻。语料库中的领域标注可以作为隐喻识别的重要依据, 在情感隐喻计算中具有深远的应用价值。

为了便于情感隐喻计算, 本语料库标注了本体、喻体的领域范围。语义赋码集参考《Roget's International Thesaurus》[32] 和梅家驹先生的《同义词词林》[33]。我们自己设定了领域分类, 共分为 5 个层级, 包括人、动物和无生物 3 个根领域和 27 个语义域。截取的部分领域划分如图 2 所示。

图 1 情感隐喻语料库情感分类树
Figure 1 Classification tree for emotion metaphor corpus

图 2 部分领域划分
Figure 2 Part of the domain division

3.2 设计原则

本文构建了一个面向情感计算, 开源、上线使用的英汉双语情感隐喻语料库, 采用人工标注和机器标注相结合的方式. 为实现此目标需要以下 5 个策略予以支持: (1) 确立知识采集策略及收录原则; (2) 制定语言标注体系的原则和规范, 选择标注集; (3) 设计质量监控体系, 保证情感隐喻知识的正确性和一致性; (4) 实现机器翻译以隐喻结构 (包括本体、喻体、喻底、情感等) 为查询条件的相关隐喻推荐; (5) 提供外语教学中便于检索和使用的语言数据资源. 为此本文依据以下流程构建模型, 如图 3 所示.

图3 (网络版彩图) 情感隐喻语料库构建流程模型

Figure 3 (Color online) The diagram for corpus construction

3.3 语料采集

按照概念界定、构建目标及应用需求,语料采集遵循以下两个原则:(1)历时与共时;(2)含有丰富的情感信息.首先,语料收集考虑到历时性与共时性,才能更加客观、全面、均衡.历时性语料可以帮助考察英汉情感隐喻的形成轨迹和语义变化规律^[34],例如英语的 happy, sad, angry, 汉语的“喜、怒、哀、惧怕”等语义并不明晰的词语,只有从历时的角度才能观察其概念形成的过程,从共时角度很难做到.而那些语义明晰的表达,例如“千疮百孔”、“欲盖弥彰”、“he is boiling with anger”等,从共识的角度就可以理解和描写其映射实现过程.另外,面向情感计算的应用目的,把文本内容含有较丰富的情感作为语料收录的另一项准则.

为了达到历时性要求,本文采集的语料既包括30年多前完成的经典文学作品,也包括最近一年的期刊杂志、网络微博等,其中网络新催生情感隐喻实例保证了语言数据的鲜活性.在共时性收录原则指导下,以互联网、语料库、书籍资料文档为主要来源,收录包括中英文教材、文学作品、网络博客、微博评论、在线购物、社交网站、歌词等覆盖各种领域与题材的语料,并且保持中英文语料的数量基本平衡,且在题材、体裁和时间的选择上基本保持一致.另外,语料集中在小说、散文、杂文、神话、寓言故事、戏剧、网络评论等富含感情色彩的文学作品及评论.表1列出了情感隐喻语料库语料来源的信息.

3.4 语料标注

语料库标注是对原始语料进行加工,选择有利于语料存储的标注附码及表述形式,以便于机读的过程. TEI (text encoding initiative) 是目前国际上采用的主流机读语篇的国际信息编码规范,由多国参与研究与制订,具有便于文本描述、适合多域标注与分析的特性.目前包括英国国家语料库 (The British National Corpus) 在内的很多的大型语料资源都是基于 TEI 的标注准则.

结合情感隐喻的特点及情感标注无歧义、简洁、便于标注等特殊需求及 TEI 的优点,本文采用 TEI 标注与自定义标注相结合的标注体系,以提高情感隐喻标注的准确性、一致性和标注速度.标注

表 1 情感隐喻语料库的语料来源信息
Table 1 Resources of the metaphor corpus

Sources	Details	Words	Sentences	Pages
Books (metaphors)	《隐喻学研究》《认知隐喻翻译研究》《认知隐喻学》 《Metaphors We Live by》《Metaphors of mind: conception of the nature of intelligence》等 11 本书籍	18711	5123	262
Dictionaries	《Collins COBUILD English Dictionary》 《A Dictionary English Metaphors》等 3 部词典	22100	—	—
Books (fairy tales)	《伊索寓言》,《安徒生童话》,《格林童话》	77054	4933	182
Movie scripts	《美丽人生》《飘》《怪物史莱克》等 12 个电影剧本	124176	10875	489
Journals and magazines	《青年文摘》《少年文艺》等 36 本期刊	4139112	226083	3623
Textbooks	大学语文 college English 等 22 册	321058	13109	568
Microblog	对事件及新闻等评论	1250489	104207	52014
Corpora	包含文学作品、网络博客、歌词和时事新闻等多种文体	1586670	137923	52959

粒度为句子级. 情感隐喻语料库标注体系基本框架如下:

MetaphorModel=(tenor, vehicle, ground, [indicator], category, emotion, [note])

各项标注含义如下: 本体 (tenor)、喻体 (vehicle)、喻底 (ground)、触发词 (indicator)、类别 (category)、情感 (emotion)、注释 (note). 其中的触发词 (indicator) 表示“好像”, “仿佛”, “as”, “like”等比喻词; 类别 (category) 表示隐喻的本体和喻体所在领域分类; 注释 (note) 表示此情感隐喻的来源或者习惯用法、搭配等. 其中的触发词 (indicator) 和注释 (note) 可以为空, 其余各项不能为空. 例如录入的句子中如果有本体、喻体等一项或者两项缺省的情况, 要求标注者自行找出, 补充完整, 对应的英文标记方法由图 4 可见. 以下是两个标注示例:

(1) 曹元朗脸上一圈圈的笑痕, 像投了石子水面.

[曹元朗脸上的笑痕, B, 11111]

[投了石子水面, Y, 12312]

[开心, D]

[PA]

[像, C]

(2) Tom is as merry as a cricket.

[Tom, B, 11111]

[cricket, Y, 12315]

[merry, D]

[PA]

3.5 质量监控

情感隐喻语料库建设工作庞大而繁杂, 本文由 7 名人员 (英语教师、英语研究生、计算机专业研究生等组成) 共同完成语料标注工作. 虽然标注人员对标注原则、规范以及标注示例等问题进行多次分析和讨论, 但是由于受到主观因素的影响, 对情感隐喻理解仍旧存在着差异. 7 名标注人员被分成 3 个两人组和 1 个一人组. 采取交叉校验方式.

Name		Corresponding		Note
Tenor		B		被比喻的事物或情境
Vehicle		Y		打比方的事物或情境
Ground		D		上面两者间相似性/共同点
Category		乐	PA	参考之前语料标注工作
		好	PB	
		怒	NA	
		哀	NB	
		惧	NC	
		恶	ND	
		惊	PC	
Indicator		C		“like”，“as”，“是”，“像”等等
Note	颜色	红	Y1	根据语句中出现与否进行标注即可
		黄	Y2	
		蓝	Y3	
		绿	Y4	
		紫	Y5	
		白	Y6	
		黑	Y7	
	方向	上	F1	赞扬
		下	F2	贬低
		左	F3	激进
		右	F4	保守
	容器	R		“怒火中烧”
	温度	W		“冷嘲热讽”
	器官	Q		“撕心裂肺”

图 4 情感隐喻语料库英文标记说明
Figure 4 Introduction to corpus annotation

两人组进行标注, 1 人组参与分歧产生后的决策. 当同组两人标注没有分歧时, 标注完成. 当同组两人标注出现差异时, 由另外一组人员再次标注, 如仍有差异, 由最后一组标注. 对于最后一组标注仍旧不一致的隐语句子, 所有标注人员 (7 人) 将集中讨论作出决定, 从而最大限度保持标注的一致性和正确性. 在标注过程中, 经过统计 93% 组内没有异议; 6.7% 组内异议, 另外一组参与标注通过; 0.3% 由最后一组处理.

设计一套规范化的操作方法, 保证语料在录入、标注和标注完成时的准确性和一致性十分必要. 本文主要从标注准则、录入界面和改错系统 3 个方面完成语料库的质量监控.

(1) 标注准则. 我们制作了详细的情感隐喻标注说明书, 逐条解释了标注的规范、标准以及难点和常见问题 (如喻底的识别和标注), 并提供了大量标注实例作为参考.

(2) 录入界面. 对词条信息采用多重人工检查流程, 并严格控制信息的更新. 在此基础上设计一个录入界面, 可以准确、方便、快速地录入信息. 详细见图 5.

(3) 纠错机制. 大连理工大学情感语料库与情感词汇本体为本文的工作提供了必要的支撑, 因此当

图 5 (网络版彩图) 情感隐喻录入系统界面
Figure 5 (Color online) Interface for submission

表 2 本体及喻体在领域大类中所占的比例
Table 2 Proportion of metaphors in different areas

	Humankind	Living things	Non-living things
Proportion in tenors	0.8862	0.0142	0.0996
Proportion in vehicles	0.4351	0.2221	0.3428

表 3 本体及喻体在特定领域大类中所占的比例
Table 3 Proportion of metaphors in certain areas

	Natural phenomena	Human society	Food/Animals	Others	Natural objects	Man-made objects
Tenor	0.8864	0.1136	0.3333	0.6667	0.4815	0.5175
Vehicle	0.8069	0.1931	0.374	0.626	0.5459	0.4541

标注隐喻情感与标注语料中句子或词汇本体情感不一致时, 则不容许录入。即

$$\text{flag} = \text{WordConsistency}(M_{\text{emo}}, W_{\text{emo}}) \cap \text{SentConsistency}(M_{\text{emo}}, S_{\text{emo}}), \quad (1)$$

其中, 当隐喻标注结果与词汇本体情感一致时, $\text{WordConsistency}(M_{\text{emo}}, W_{\text{emo}})$ 为 1, 否则为 0; 同理于 $\text{SentConsistency}(M_{\text{emo}}, S_{\text{emo}})$. 当且仅当 flag 值为 1 时, 运行直接录入, 否则提示确认检查后录入。

通过概念界定、设计原则、语料收集、语料标注、质量监控 5 个流程, 本文完成了情感隐喻语料库的建设. 表 2 和 3 为标注语料的一部分统计数据.

表 4 百度在线翻译与基于情感隐喻语料库的翻译推荐对比
Table 4 Comparison between Baidu translation and the metaphor corpus

Chinese metaphorical sentences	Baidu online translation	Affective metaphor corpus translation
他像热锅上的蚂蚁.	He's like ants on a hot pot.	He's like a cat on the bricks.
宁为鸡头, 不为凤尾.	Better be the head, not everything.	Better be the head of dog than the tail of a lion.
这个产品已成为我们公司的摇钱树.	This product has become our company's Qian Shu.	This product has become our company's milk cow.
她是父母的掌上明珠.	She is a pearl in the palm of parents.	She is the apple of her parents' eye.
真金不怕火炼.	True gold does not fear fire.	True blue will never stain.

4 语料库的应用

4.1 机器翻译应用

首先, 利用情感隐喻语料库的领域标注, 可以用于情感隐喻句情感类别的识别, 尤其那些没有鲜明情感特征、不包含情感词汇的隐喻, 即隐式情感类型识别对于机器翻译应用, 可以通过汉英喻底关联, 并通过相似度计算进行排序, 推荐翻译结果. 例如汉语隐喻句“那个理论站不住脚”, 百度翻译给出的英语翻译是“*That theory can't stand*”. 从这个错句可以看出, 现有的在线翻译如百度翻译对隐喻句不能有效地翻译. 但如果依赖情感隐喻语料库, 通过检索具有相同喻底的英语隐喻句推荐, 就可以找出地道的翻译“*That theory doesn't hold water*”.

如表 4 所示, 列举了汉语隐喻句百度在线翻译与基于情感隐喻语料库的翻译推荐对比实例. 另外, 情感隐喻语料库方便快捷的检索功能对外语教学有很大的应用前景.

语句级的隐喻进行机器翻译时一般有直译和意译两种, 而通常机器翻译时采用的是直译的方式, 导致翻译结果质量不高, 因此隐喻的机器翻译一直是自然语言处理的一个难点^[35,36]. 情感隐喻语料可以帮助实现识别与理解语言中的隐喻现象. 具体地说, 可实现本体、喻体、喻底、情感的任意组合查询, 根据融合相似性结果对隐喻机器翻译结果进行排序, 并将翻译结果的前 K 项推荐给用户. 设置隐喻识别器, 实现句子的转换算法流程主要分为

- (1) 直接检索, 找到对应的隐喻;
- (2) 对于找不到精确匹配的隐喻, 按照喻底, 计算句子相似度, 找到最为相似的句子;
- (3) 对于找不到对应的喻底, 计算喻底与之相似度, 找到与之最相似的喻底, 再返回相应的句子.

4.2 外语教学应用

隐喻能力始终是外语教学的难点. 英语作为母语和作为第二语言的学习者在隐喻理解和使用上会有很大的区别, 能否熟练地使用目标语中的隐喻是外语学习者语言沟通能力高低的重要标志. 有研究表明, 二语学习者隐喻能力偏低是导致不能地道、规范地使用英语的原因之一^[37]. 因此越来越多的研究者强调隐喻能力在外语教学中的重要地位, 建议把隐喻教学作为外语教学的一项主要内容.

然而隐喻知识数量庞大、分布零散, 很少有系统、清晰的语料资源可以专门用于外语教学, 普通课堂上基于少量单语或双语语料的隐喻学习又带有很强的主观性. 在这种情况下, 本文构建的英汉双语情感隐喻语料库在外语教学中的应用体现在以下 3 个方面:

首先, 英汉双语情感隐喻语料库提供大量真实、直观的参考实例以及方便快捷的检索功能. 例如依据喻底为查询条件的检索中, 学生把中文喻底作为查询条件输入平台检索系统, 系统会给出所有具有相同喻底的英文隐喻句及智能排序, 提供地道的英文隐喻句子以及很多具有相同喻底的实例作为参考, 可以据此开展双语情感隐喻构成 (本体、喻体、喻底)、隐喻类型以及情感等方面的研究.

其次, 教师可以指导学生基于语料库对比英汉情感隐喻的异同, 深刻理解不同的文化风俗及语言习惯. 在分析比较中解决隐喻学习中目标概念知识和喻源概念知识获取问题, 揭示概念隐喻认知机制的本质和隐藏在隐喻背后的意识形态, 了解语言转换规律和文化渊源.

最后, 隐喻能力被看作是母语使用者语言流利性的基本特征之一, 二语学习者需要掌握用以匹配隐喻表层结构和概念内涵的编码知识, 才能向外语学习的高级水平过渡, 达到母语使用者的流利程度. 由此看来, 教师可以把学生在写作中隐喻的使用数量作为写作评价的一项指标. 在情感隐喻语料库应用中, 根据系统的隐喻语料分析比较学生的隐喻使用数量和程度, 建立起更加科学完善的写作评估体系.

4.3 情感分析应用

情感隐喻作为一种特殊的隐式情感表达一直是影响情感分析工作精度的一个重要因素. 为解决上述问题, 构建情感隐喻语料库成为了一个迫切需求. 对于情感隐喻的识别方面, 主要从两部分开展, 一部分是显性情感隐喻, 即将情感隐喻的识别看成一个分类问题, 依赖如“像”等隐喻触发词获取部分样本, 实现对情感隐喻的识别^[37]. 此外, 中西方思维差异也是影响情感分析精度的一个因素, 而因思维差异往往也会造成对同一隐喻存在着不同本体或喻体的表达, 对此在情感隐喻语料库指导下, 结合中西方思维差异的特性, 在一定程度上解决了上述问题^[38]. 用于情感隐喻语料库构建资源均在以上工作中发挥了一定的作用, 为后续的情感分析研究奠定了一定的资源基础.

与目前大部分隐喻资源和情感语义资源相比, 情感隐喻语料库的建设对于情感分析尤为重要. 目前多数隐喻资源建设仅仅从概念和语言的角度对隐喻进行研究, 忽略了隐喻在情感表达中的作用, 因此对于情感分析作用有限; 而大部分情感资源建设中由于要面对隐喻语义标注、情感量化等诸多困难而较少涉及对隐喻的研究, 从而忽略了情感表达中广泛存在的隐喻, 降低了情感的分析效果. 例如几个影响力较大的隐喻数据库 Master Metaphor List, Metalude, Hamburg Metaphor Database 等都仅仅包括了隐喻本体、喻体、字面义、比喻义、例句等信息, 没有情感标注信息. 再如知网 Hownet、台湾大学的情感极性词典 NTUSD 以及大连理工大学的情感常识知识库等语义资源也不涉及隐喻中包含的情感.

我们在电影票房预测的研究中发现, 电影评论论坛上有大量情感隐喻影评的出现. 例如“是个夺人眼球的题材”、“整部电影男主角就是个备胎呀”、“一场羊头狗肉式的执导改编”、“帝王般的享受”、“影院制冷经常偷懒”等情感隐喻表达. 但是目前在线评论情感分析基本上依赖于情感词典, 对字面语言进行分析, 从中挖掘出情感词或属性等信息, 极少对于隐喻所传递的情感进行分析. 这导致大量由情感隐喻传递的情感信息的缺失, 从而影响情感分析的精确度.

因此我们借助情感隐喻语料库, 构建情感隐喻识别引擎, 将评论句子输入到该引擎中, 识别所蕴含的情感, 将其输出. 对于没有隐喻的句子则直接输出到通用的情感分析模块处理, 最后挖掘出评论者的情感倾向性.

5 结束语

在情感隐喻计算对语料资源建设的需求越来越迫切的背景下, 本文研究了大规模汉英情感隐喻语料库的构建方法与应用. 目前已经完成了第一阶段的标注, 包括 2350 个情感隐喻句子, 总计 31550 字. 在情感隐喻语料库构建第一阶段初步标注的基础上, 计划总共完成句子 2 万个, 约 22 万字, 并尝试实现在隐喻情感类别判定、机器翻译以及外语教学中的应用功能. 针对语料库建设中情感隐喻类别分布不均匀、标注体系和标注规范有些粗略的不足, 我们将在下一步的工作中不断完善.

参考文献

- 1 Picard R W. *Affective Computing*. Cambridge: MIT Press, 1997
- 2 Richards I A. *Philosophy of Rhetoric*. London: Oxford University Press, 1936
- 3 Shu D F. *Studies in Metaphor*. Shanghai: Shanghai Foreign Language Education Press, 2000 [束定芳. 隐喻学研究. 上海: 上海外语教育出版社, 2000]
- 4 Huang X X, Zhou C L. Computational models for metaphor comprehension: a survey. *Comput Sci*, 2006, 33: 178–183 [黄孝喜, 周昌乐. 隐喻理解的计算模型综述. 计算机科学, 2006, 33: 178–183]
- 5 Shutova E. Models of metaphor in NLP. In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Stroudsburg: Association for Computational Linguistics, 2010. 688–697
- 6 Shutova E, Teufel S. Metaphor corpus annotated for source-target domain mappings. In: *Proceedings of the International Conference on Language Resources and Evaluation, Malta*, 2010. 3255–3261
- 7 Martin H. Metabank: a knowledge-base of metaphoric language conventions. *Comput Intel*, 1994, 10: 134–149
- 8 Lonneker-Rodman B. The Hamburg metaphor database project: issues in resource creation. *Language Resour Eval*, 2008, 42: 293–318
- 9 Fass D. Met*: a method for discriminating metonymy and metaphor by computer. *Comput Linguist*, 1991, 17: 49–90
- 10 Barnden J, Glasbey S, Lee M, et al. Reasoning in metaphor understanding. *Assoc Computat Linguist*, 2002, 2: 1–5
- 11 Li J F, Yang Y, Zhou C L. Corpus designing and constructing for metaphor computation. *Mind Comput*, 2007, 1: 142–146 [李剑锋, 杨芸, 周昌乐. 面向隐喻计算的语料库研究和建设. 心智与计算, 2007, 1: 142–146]
- 12 Dhall A, Goecke R, Lucey S, et al. Collecting large, richly annotated facial-expression databases from movies. *IEEE MultiMedia*, 2012, 19: 34–41
- 13 Yang H W, Meng H M, Wu Z Y, et al. Modeling the global acoustic correlates of expressivity for Chinese text-to-speech synthesis. In: *IEEE/ACL Spoken Language Technology Workshop, Palm Beach*, 2006. 138–141
- 14 Xu L H, Lin H F, Zhao J. Construction and analysis of emotional corpus. *J Chinese Inf Process*, 2008, 22: 116–122 [徐琳宏, 林鸿飞, 赵晶. 情感语料库的构建和分析. 中文信息学报, 2008, 22: 116–122]
- 15 Hu Z L. *Metaphor and Cognition*. Beijing: Peking University Press, 2004 [胡壮麟. 认知隐喻学. 北京: 北京大学出版社, 2004]
- 16 Lakoff G, Johnson M. *Metaphors we Live by*. Chicago: the University of Chicago Press, 1980
- 17 Kovecses Z. *Metaphor and Emotion*. Cambridge: Cambridge University Press, 2000
- 18 Kovecses Z. *Metaphor in Culture: Universality and Variation*. Cambridge: Cambridge University Press, 2005
- 19 Kovecses Z. *Metaphors of Anger, Pride and Love: a Lexical Approach to the Structure of Concepts*. Amsterdam: John Benjamins Publishing, 1986
- 20 Lakoff G, Kovecses Z. *The Cognitive Model of Anger Inherent in American English*. Cambridge: Cambridge University Press, 1987. 195–221
- 21 Wierzbicka A. *Semantic Primitives*. Frankfurt: Athenum, 1972
- 22 LI J Y, LI B J. An examination of the objectives, principles and methodology of natural semantic metalanguage approach. *Contemp Linguist*, 2007, 9: 68–77 [李炯英, 李葆嘉. NSM 理论的研究目标、原则和方法. 当代语言学, 2007, 9: 68–77]
- 23 Wierzbicka A. *Semantics, Culture, and Cognition*. Oxford: Oxford University Press, 1992
- 24 Wierzbicka A. Conceptual primes in human languages and their analogues in animal communication and cognition. *Lang Sci*, 2004, 26: 413–441

- 25 Plutchik R. A general psychoevolutionary theory of emotion. In: Plutchik R, Kellerman H, eds. *Emotion: Theory, Research, and Experience. Theories of Emotion*. New York: Academic Press, 1980, 1: 3–33
- 26 Izard C E. *The Face of Emotion*. New York: Appleton-Century-Crofts, 1971
- 27 Kelly J R. Mood and emotion in groups. In: Hogg M A, Tindale R S, eds. *Blackwell Handbook of Social Psychology: Group Processes*. Maiden: Blackwell Publishers, 2002
- 28 Ekman P, Power M J. *Handbook of Cognition and Emotion*. Sussex: John Wiley & Sons, 1999
- 29 Lin C D. Emotions in Social Psychology. *Sci Soc Psychol*, 2006, 21: 37–62 [林传鼎. 社会心理学中的情绪问题. *社会心理学*, 2006, 21: 37–62]
- 30 Xu L H, Lin H F, Yang Z H. Text orientation identification based on semantic comprehension. *J Chinese Inf Process*, 2007, 21: 91–95 [徐琳宏, 林鸿飞, 杨志豪. 基于语义理解的文本倾向性识别机制. *中文信息学报*, 2007, 21: 91–95]
- 31 Jackendoff R. *Semantics and Cognition*. Cambridge: MIT Press, 1983
- 32 Roget P M, Mawson C O S. *Roget's International Thesaurus*. New York: Crowell, 1962
- 33 Mei J J. *Synonymy Thesaurus*. Shanghai: Shanghai Dictionary Press, 1996 [梅家驹. 同义词词林. 上海: 上海辞书出版社, 1996]
- 34 Zhang H. Conceptualization and expressions of emotions in Chinese and English. *J Foreign Lang*, 2000, 5: 27–32 [张辉. 汉英情感概念形成和表达的对比研究. *外国语*, 2000, 5: 27–32]
- 35 Huang X X, Zhou C L. Computational models for metaphor comprehension: a survey. *Comput Sci*, 2006, 33: 178–183 [黄孝喜, 周昌乐. 隐喻理解的计算模型综述. *计算机科学*, 2006, 33: 178–183]
- 36 Wang Z M. Recent developments in computational approach to metaphor research. *J Chinese Inf Process*, 2006, 20: 16–24 [王治敏. 隐喻的计算研究与进展. *中文信息学报*, 2006, 20: 16–24]
- 37 Danesi M. The role of metaphor in second language pedagogy. *Rassegna Italianna di Linguistica Appl*, 1986, 18: 1–10
- 38 Lin H F, Xu K, Ren H. Mechanism of dominant sentimental metaphor identification based on lexical domain and semantic similarity. *J Dalian Univ Tech*, 2012, 52: 761–766 [林鸿飞, 许侃, 任惠. 基于词汇范畴和语义相似的显性情感隐喻识别机制. *大连理工大学学报*, 2012, 52: 761–766]

Construction and application of affective metaphor corpus

ZHANG DongYu, YANG Liang, ZHENG PuQi, XU Bo & LIN HongFei*

Information Retrieval Laboratory, Dalian University of Technology, Dalian 116023, China

*E-mail: hflin@dlut.edu.cn

Abstract Metaphors are commonly used in human languages, and are an important way of conceptualizing complicated and abstract emotions for humans. Meanwhile, metaphors also play an important role in affective computing. Affective computing requires a large amount of high-quality corpus for application to both computer science and linguistics in order to perform additional extensive research. Based on this idea, we construct a large-scale Chinese-English bilingual emotional metaphor corpus to meet the urgent demand for affective computing. First, we introduce the theoretical framework for the construction of emotional metaphor corpora. Then, we introduce the framework from five perspectives: concept definitions, design principles, data collection, data annotations, and quality control. Finally, we analyze the application of the constructed corpus on the area of machine translation and English teaching.

Keywords affective computing, affective metaphor corpus, machine translation, foreign language teaching, sentiment analysis

ZHANG DongYu is an associate professor at the Software School of the Dalian University of Technology. Her research interests include affective computing and metaphor recognition.

YANG Liang was born in 1986. He received his Bachelor's degree in Software Engineering from the Dalian University of Technology, Dalian. He is now studying for his Ph.D. degree. His research interests include sentiment analysis, text mining, and machine learning.

XU Bo was born in 1988. He received his Bachelor's degree in Software Engineering from the Dalian University of Technology, Dalian. He is now studying for his Ph.D. degree. His research interests include information retrieval and learning to rank.

LIN HongFei was born in 1962. He received the M.S. degree from Dalian University of Technology, China, in 1992, and the Ph.D. degree from Northeastern University, China, in 2000. He is currently a Professor in the College of Computer Science and Technology at the Dalian University of Technology. His research interests include information retrieval and sentiment analysis.