重庆大学学报(社会科学版) 2009 年第 15 卷第 6 期
JOURNAL OF CHONGQING UNIVERSITY(Social Science Edition) Vol. 15 No. 6 2009

基于 CBR 的双语语境知识表征系统研究

牛书杰,李 红
(重庆大学 语言认知及语言应用研究基地,重庆 400044)

摘要:文章简要回顾了基于案例①的推理(case-based reasoning,简称 CBR)机制和第二语言习得理论中"补缺假设"理论的语境观,分析了 CBR 在双语语境知识表征中应用的可行性,并在此基础上提出了双语语境知识表征 CBR 系统的计算模型、算法和该系统的一般实现过程,是机器学习和人工智能理论在第二语言习得理论研究中应用的初步探讨,旨在给第二语言习得研究提供新的研究方法和视角。

关键词:案例(CBR);推理;补缺假设;认知模拟;语境
中图分类号:H319 **文献标志码**:A **文章编号**:1008-5831(2009)06-0144-05

对语言习得的研究是人类认知过程研究的重要组成部分,也是近年人工智能领域非常重视的一个研究课题[1]。国外相关文献显示,利用计算机严密的逻辑性和准确性来模拟语言习得的认知过程已经成为重要的研究方式之一,不少研究者利用这一手段对语言的习得进行了系统的研究[1-5]。然而几乎所有的研究都是以儿童母语习得为基础进行的。中国是世界上外语学习者最多的国家,将计算机科学相关理论引入第二语言习得研究,不但可以给这一领域带来工具性的革命,而且对外语教和学也有指导意义。可是,由于学科设置等原因,国内二语习得研究领域中,利用计算机模拟为手段的研究文献目前还没有看到。笔者基于"补缺假设"的语境观,分析基于案例的推理(CBR)技术在模拟双语语境知识中应用的可行性,并在此基础上提出双语语境知识 CBR 系统的算法和一般实现过程。该双语语境知识学习系统具有增量和自适应性特征,具有一定的现实意义。

一、基于案例的推理

(一)CBR 的产生

CBR 是受认知科学领域中对人类解决问题策略研究的启发而产生的[6]。它类似于人类解决问题方法中的启发法,即凭借经验解决新的、类似的问题。CBR 的基本概念最早由美国耶鲁大学的 Schank 教授于 20 世纪 80 年代初提出,后来由他的学生 Kolodner 完善并开发出了第一个基于该概念的系统[6-8]。

收稿日期:2009-05-05
基金项目:2007 年重庆大学人文社科青年教师科研基金项目"基于计算机模拟的第二语言习得研究"(CDSK2006-20)
作者简介:牛书杰(1976-),男,河南禹州人,重庆大学外国语学院讲师,重庆大学语言认知及语言应用基地研究员,重庆大学计算机学院博士研究生,主要从事第二语言习得及计算机应用技术研究;李红(1962-),女,重庆人,重庆大学外国语学院教授,博士,重庆大学语言认知及语言应用基地专职研究员,硕士生导师,主要从事第二语习得及心理语言学研究。
①CBR 在文献中翻译为"案例"、"实例"、"事例"、"范例"等,本文从"案例"。

(二)CBR 的基本结构

基于案例推理系统主要由检索系统、案例库、案例改写等核心部分构成[9]。其中,案例库是过去问题求解经验的总和,为新的问题求解提供支持,而新的求解结果也可以作为案例存储在库中,作为知识的积累。

(三)CBR 的特点

跟其他人工智能的学习和推理机制不同的是,CBR 依赖的不是某一领域泛化的世界知识,而是将知识具体化、案例化,然后加以提取,并服务于新的情形,同时产生出新的知识片段(案例)[10]。这样以来,不但提取和检索方便,而且有利于知识的增量,克服了基于规则推理机制的知识获取瓶颈。系统的准确性也会随着使用而提高,不会出现基于规则推理机制的规则冲突等现象。

二、"补缺假设"理论的语境观

"补缺假设"是由王初明教授首次提出,并进行了深入探讨的一个全新的第二语言习得理论。该理论尝试从语境的角度来廓清中国人学习外语的认知机理。该假设认为:"语言形式与语境知识的有机结合是语言正确流利使用的前提。由于外语环境缺少与外语表达方式匹配的真实语境,在外语理解、习得和使用的过程中,母语语境知识介入补缺,进而激活与母语语境知识配套的母语表达式,母语迁移因此而发生。"[11]

该假设区分了"内部语境(internal context)"和"外部语境(external context)"②。外部语境是说话发生的语言环境,包括物理环境和社会环境,比如,说话的参与者、说话的时间、地点等。内部语境是外部语境在大脑中的表征。因为母语(L1)的习得是内部语境和外部语境匹配的过程,所以二者有机结合,习得母语语言结构的同时也习得了与之配套的语境知识。

但是,外语(L2)的学习则完全不同于母语的习得过程。外语学习多是在课堂上完成的,外语的外部语境几乎为零。所以,外语的内部语境和外部语境的匹配无法完成,从而造成断裂。在使用外语交际时,由于外语内部语境知识的缺乏,引起母语语境知识的补缺,致使外语(英语)的语言结构和母语的语境知识结合,产生所谓的"汉式英语"。倘若连母语的语境知识也没有得到激活,则会产生所谓的"哑巴英语"[11]。

由于母语语境知识没有被激活,加上英语语境知识的缺省,产生"哑巴英语"是显而易见的。笔者试图开发一个 CBR 系统来模拟双语语境知识在大脑中的表征和"汉式英语"的产生过程,并以此来说明"补缺假设"的解释力,以期对该假设进行相应的评介。

三、基于 CBR 的双语语境知识表征系统的可行性

(一)CBR 是对基于规则推理的反动

基于案例的推理是对基于规则的推理(rule - based reasoning,简称 RBR)的反动,它强调的是案例,而不是规则。它试图从案例库中检索到可以应用的相关案例,重新使用,或者做出适当修改后加以应用,同时产生出新的案例。基于案例的推理对于规则难于提取的研究领域很有帮助。例如,在社会科学的一些研究中,把研究对象规则化、数学模型化几乎是无法做到的,而应用 CBR 就比应用 RBR 显得要恰当,而且易于操作。

(二)双语语境知识难以规则化

"补缺假设"的语境理论涉及的认知过程是无法单纯使用规则来描述的[11]。语境本身就是一个动态的过程[12]。例如,外部语境就包含人物、地点、事件、话题、谈话的正式程度、社交活动等。这些因素又有各自不同的属性,任何一个因素都会给系统带来影响,而且内部语境也涉及各种因素,比如说话者和听话者的意图、文化背景知识等。它们与语言结构相互作用、影响,使整个系统变得非常复杂。此外,外语学习的过程也是一个不断变化的过程,学习者通过学习,增进语言结构和语境知识,从而提高外语水平。所以,试图使用规则来描述语境知识的获取和表征的方法很难达到预期的效果。

(三)语境知识案例化的优势

如前所述,CBR 是对基于规则的推理的反动,语境知识在大脑中的表征是难以用规则来描述的,所以使用 CBR 思想来描述语境知识在大脑中的表征是恰当的。案例化语境知识对于研究人类的内隐记忆和语感也很有帮助。通过案例化,一个成功的 CBR 系统便可相对准确地模拟人类的认知过程,对于打开内隐记忆和语感的黑箱将会起到重要的作用。

四、基于 CBR 的双语语境知识表征系统

(一)系统概貌

通过建立一个 CBR 计算机模拟系统,使该系统模拟人类内部语境的认知机理,将"补缺假设"的语境知识理论付诸实施,然后将该系统产生出的语言行为与外语学习者的真实语言产出进行比较,从而反过来对模拟系统和理论本身进行调整、评估。

为了将研究范围具体化,笔者暂时将模拟系统中的内部语境限定为母语语境知识的内部表征。排

②参见:WANG C. The compensation hypothesis—— the role of context in language Transfer [P],2004.

除外语语境知识的原因是:“补缺假设”假定了外语内部语境知识的缺省,在外语内部语境知识被激活的情况下,如果外语语境知识案例库中有匹配的案例,则系统就不必到母语语境案例库中检索,也就无从补缺。

图 1 是基于“补缺假设”语境知识理论的 CBR 系统的概貌。在真实的交际场景中,由于外语语境知识的完全或者部分缺失,外语学习者调用母语(汉语)内部语境知识补缺,结合英语表达式,产生出语言输出,即“汉式英语”。如果该系统经过训练的语言输出与英语内部语境知识缺失的学习者的语言输出表现出显著的相关,则可以说明“补缺假设”理论和该系统的有效性,反之,该理论或模拟系统被证伪。

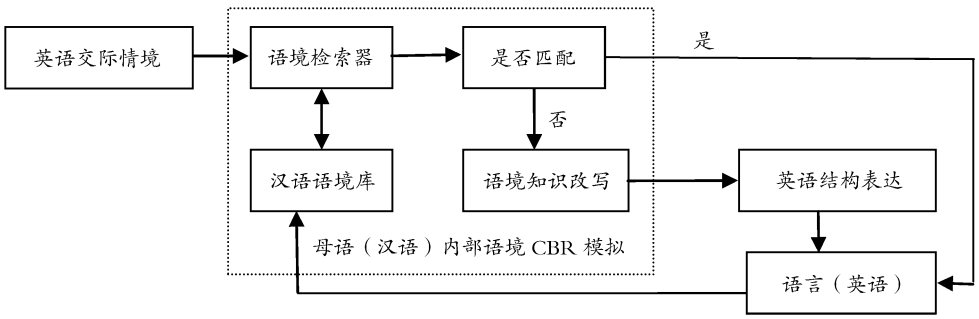

图 1 基于 CBR 的双语境知识表征系统

(二)案例表示

该模拟系统中的一个案例实际上是母语内部语境知识的单个片段。内部语境是外部语境在大脑中的表征,涉及的因素有:话题(topical,以下简称 TOPi)、空间(spatial,以下简称 SP)、时间(temporal,以下简称 TP)、关系(relational,以下简称 RT)、参与者(participational,以下简称 PP)、文化规则、肢体语言以及参与者的性格、心情、文化水平等。这里只对前五种特征因素进行案例表示,其余的因素在抽象系统中的影响相对较弱,所以暂不涉及。但是,需要说明的是,当系统随着案例的增加变得庞大时,就必须提高案例表征的颗粒度(granularity),考虑更多的特征因素。

一个有效的案例表示一般应包括三个部分的内容:案例发生的背景、案例的特点、解决方法或者结果[13]。这里以汉语中一个典型的告别语境为例来说明,为便于检索,可以采用英语代码来描述。案例的背景(话题)为“道别”,涉及四个方面:Default 普适于任何场景;该语境表达一个交际的结束,可以应用于正式(fml)或非正式(infml)语境;牵涉到的交际者超过两个人;结果是激活汉语表达式“再见”。

“再见”的案例表示举例:

[CASE 1
TOPi: Biding farewell;
FEATURES
SP: Default;
TP: Ending a communication;
RT: fml * infml;
PP: ≥ 2 people;
SOLUTION: “再见”]

需要注意的是,当案例库增大时,案例(语境)涉及的特征和因素就必须更加具体、更加复杂,否则很难区分两个比较类似的语境,并进而跟英语结构表达式结合后产生的语言(英语)输出就不能对类似的语境进行区分。这也和观察到的现象相吻合:英语语境知识缺省的学习者在类似的语境下经常重复使用同一个英语的表达式来交流,而这个表达式从汉语的意义角度看没有问题,但是从英语语境的角度看就不地道,甚至会引起误解。比如,英语常见的告别是“Byebye”,如果使用案例的特征表示,则如下所示。

[CASE #
TOPi: Biding farewell;
FEATURES
SP: Default;
TP: Ending a communication;
RT: infml;
PP: ≥ 2 people;
SOLUTION: “Byebye”]

它跟汉语“再见”的案例表示的唯一区别在于:Byebye 只能用于非正式的(infml)语境中,而“再见”在正式(fml)或非正式的语境中都可以使用。外语(英语)语境知识缺省的情况导致了母语(汉语)语境知识的补缺。所以,语境知识不完全的学习者即使在一个极其正式的场合道别时,也会使用 Byebye 来结束交际任务,从而造成不地道的表达。

(三)案例库和案例修改

高频的母语语境知识表示首先是案例库中必须包括的内容。例如,招呼(greeting)、抱怨(complaining)、道歉(apologizing)、命令(directing)等。如前文所述,案例库超过一定规模时,就必须更细化案例特征,以便增加区分度。根据频率的高低还要对案例增加权重,并据此对案例库中的案例进行排序,以便以后检索。

案例库中除了单个的案例之外,还存储有一些汉语语境知识的对应英语语言产出。跟其他的 CBR 系统不同的是:其他系统把不匹配的案例经过修改后直接存入案例库,而"补缺假设"的模拟系统则需要把最终的语言(英语)输出存入案例库,这样一来,案例库中不仅有汉语语境知识的案例,还存储了常用的跟汉语语境知识对应的英语输出(output)。这样做的原因是:对中国的英语学习者来说,母语语境知识是一个相对稳定的系统,它的更新主要通过母语来实现;这个系统模拟的是外语语境知识缺省的情况下外语的输出情况,所以在母语语境知识 CBR 系统之外,必须有个英语结构表达系统,把汉语语境知识跟英语的表达式结合起来。语言输出存入案例库后,在以后的案例调用时,就可以在案例库中检索匹配,直接进行语言输出。

"再见"案例修改后的表示:

[CASE 1 #
　　TOPi: Biding farewell;
FEATURES
　　SP: Default;
　　TP: Ending a communication;
　　RT: fml * infml;
　　PP: ≥ 2 people
;SOLUTION: "再见";
PRODUCTION:"Byebye"]

把母语语境知识跟外语语言结构形式结合后的语言输出存入案例库是有心理现实(psychological plausibility)意义的,它能够解释外语学习过程中的几个现象。首先,如果新的案例可以在案例库中直接匹配到对应的母语语境知识和英语输出,那么英语产出就非常快,这有助于解释为什么大量的外语课堂练习,尽管不一定在真实的语境下完成,对于学习者的流利度也是有帮助的。其次,它能够解释外语学习过程中的"石化(fossilization)"现象。该现象表现为:尽管外语学习者在语言应用方面非常流利,但是他们可能在语音、句法、语用等方面表现出持续的错误,而这些错误是很难消除的,甚至是"永久性的"[14]。从"补缺假设"理论模拟的 CBR 系统可以看出,母语语境知识跟外语语言结构表达结合后的外语输出一旦存入案例库,该表达的调用和激活就非常方便、快捷,而且很难从案例库中消除,即表现为学习过程中的"石化"现象。

在前文特征提取的基础上,案例的修改采用权值设定的方法,通过考虑特征频率等因素来实现。这里可以采用神经网络中经常用到的 S 曲线函数(sigmoid function)来设定案例权值,从而对案例库中的案例进行排序。

$$\text{Ranking} = \frac{2}{1 + e^{-0.1\mu}} - 1$$

其中,Ranking 为案例排序的的权值;u 代表案例被调用的次数,所以,其取值范围为[0, + ∞]。因为,S 曲线函数为连续升函数,所以,Ranking 的对应取值范围也就是[0,1)。可以看出,案例被调用的次数越多,对应的排序值也越高;调用次数为 0 时,对应的排序值也是 0。

对案例的修改也可以适当采用人机结合的方式来进行,让人类介入案例的修改。这更类似学习者在教师指导下学习。虽然在母语习得的过程中,反面证据(negative evidence)的作用不大,但是,成人第二语言习得跟儿童母语习得是有区别的,一定的教师指导很有必要。所以,对该系统的案例修改进行一定的人类干预有其理论依据。

(四)案例库的检索

由于该模拟系统的案例相对简单并且数量相对有限,所以常用的检索算法就可以满足需要。这里简述最相邻算法在本 CBR 系统案例检索时的工作机制。最相邻算法通过累加目标案例与案例库中案例的每个域的相似度值来确定总的相似度,然后把超过相似度阈值的案例返还[13]。当新的问题输入系统时,系统通过计算各个案例与新输入问题的特征(TP, SP, RT, PP 等),然后输出相关案例和权重,并按照权重的大小顺序进行排序。超过设定权值为匹配,否则需要建立新的母语语境知识案例,通过与外语(英语)语言结构产生式结合,把语言(英语)作为新的案例存储到案例库中,以便以后的检索。

基于 CBR 的双语语境知识表征系统算法伪代码:

```
for each [context representation Ci]
if [Ci is retrieved in L2 corpus]
then
reuse Ci
increase Ci ranking
else
if [Ci is not retrieved in L2 corpus]
then
search for Ci in L1 corpus
if [Ci is retrieved in L1 corpus]
then
reuse Ci with L2 structure
revise and update L1 corpus
increase Ci ranking in L1
else
create new Ci representation in L2 corpus
if [multiple representations retrieved in L2 corpus]
if [Ranking (Ci = Cmax)]
then
reuse Cmax
increase Cmax ranking
```

end

五、结论

笔者在简要回顾 CBR 和"补缺假设"语境理论的基础上,提出了基于 CBR 的双语语境知识表征系统的基本实现过程。将机器学习的理论与语言习得理论相结合,利用计算机的逻辑性、准确性对人类的认知过程进行模拟和研究是认知科学的一个重要走向。若能成功将计算机科学和第二语言习得研究交叉将会给后者带来工具性革新,也会获取相应的研究成果。笔者在此只是对该领域进行了初步的、理论性的探讨,具体实现该系统还需要辅助大量的实验和实地语料的获取。另外,文中所提及的英语结构表达式实际上是语言的产生式系统,虽然语言的产生不是文章研究的重点,但是这也是本模拟系统正常工作的一个必要环节。尽管语言的产生已经成为了人工智能、自然语言处理、认知科学等领域的研究热点,但是研究者对此研究时假设仍然多于结论。通过对该模拟系统的更进一步探讨,希望对语言产出机制也能带来一定的启示。

参考文献:

[1]WALCZAK S. A context-based computational model of language acquisition by infants and children [J]. Foundations of Science 2002,7: 393 - 411.

[2]ALBRIGHT A, HAYES B. Rules vs analogy in English past tenses: a computational / experimental study [J]. Cognition 2003,90: 119 - 161.

[3]COOPER R. Modeling high-level cognitive processes [M]. Mahwah NJ: Lawrence Erlbaum Associates, 2002.

[4]KAZMAN R. Simulating the child's acquisition of the lexicon and syntax——experiences with babel [J]. Machine Learning ,1994,16: 87 - 120.

[5]HANNEMAN R. Computer-assisted theory building [M]. Newbury Park, California: Sage Publications, 1988.

[6]张光前,邓贵仕,李朝晖. 基于事例推理的技术及其应用前景[J]. 计算机工程与应用,2002(20):52 - 55.

[7]KLODONER J. An introduction to case-based Reasoning [J]. Artificial Intelligence Review,1992,6. (1): 3 - 33.

[8]KLODONER J. Case-based reasoning [M]. Morgan Kaufmann,1993.

[9]朱福喜,汤怡群,傅建名. 人工智能原理[M]. 武汉:武汉大学出版社,2002.

[10]AAMODT A,PLAZA E. 1994. Case-based reasoning: fundamental issues, methodological variations, and system approaches [J]. AI Communications,1994,7(1):39 - 59.

[11]王初明. 补缺假设与外语学习[J]. 外语学刊,2003(1): 1 - 5.

[12]何兆熊,蒋艳梅. 语境的动态研究[J]. 外国语,1997(6): 16 - 22.

[13]郭艳红,邓贵仕. 基于事例的推理(CBR)研究综述[J]. 计算机工程与应用,2004(21),1 - 5.

[14]BROWN H D. Principles of language learning and teaching [M]. (4th Edition) Beijing: FLTRP, 2005:217.

Simulating Bilingual Contextual Knowledge of Chinese Learners of English

NIU Shu-jie, LI Hong

(*Research Center of Language, Cognition and Langeage Application, Chongqing University, Chongqing 400044, China*)

Abstract: This paper, by constructing a CBR-based computational model, attempts to evaluate the validity of the Compensation Hypothesis, an SLA theory proposed by Professor Wang Chuming for the purpose of elucidating the cognitive mechanisms of Chinese learners of English. Instead of simulating the processes of phonological, lexical, semantic or syntactical acquisition, as is often done in first language acquisition research, the paper, based on the Compensation Hypothesis, examines the feasibility of modeling the cognitive mechanisms of bilingual contextual knowledge acquisition. This model holds that bilingual contextual knowledge can be best represented by the use of case-based reasoning, not rule-based reasoning. The architecture of the CBR-based model is detailed, related algorithm for the system proposed and a case-based corpus depicted. The computational model can account for some of the issues in SLA research, such as Chinglish, fossilization, etc. This methodology can also serve as an alternative research paradigm for SLA studies.

Key words: CBR (Case-based reasoning); Compensation Hypothesis; cognition simulation; SLA (Second-language acquisition)

(责任编辑 胡志平)