

基于词向量的中国学习者英语词汇深度知识习得研究

邓海龙

赣南师范大学

摘要: 词汇知识是二语学习者语言能力的重要方面,但其测量与描写却较为费时费力,相关成果比较零星,缺乏系统性。本研究借助自然语言处理中的词向量技术,通过大规模计算和分析学习者语料库中的词汇使用情况,系统描述中国学习者英语词汇深度知识习得现状,为外语词汇教学提供参考。研究表明,当代人工智能技术有助于揭示学习者二语词汇深度知识习得规律,在外语教学研究中具有广泛的应用潜力。

关键词: 词向量; 词汇知识; 中国学习者

【DOI】10.12252/j.issn.2096-6288.2022.10.171

引言

词汇知识是二语学习者语言能力的重要基础。Wilkins (1972) 早就指出“不懂语法,能说的非常有限;没有词汇,却什么也说不” (Without grammar very little can be conveyed; Without vocabulary nothing can be conveyed.) 二语学习者通常花费大量时间用于学习词汇。因此,大学校园随处可见手捧一本词汇记忆手册的情形也就可以理解了。然而,词汇知识是多维度的,其测量与评估却并非易事。已有学习者词汇知识测量大多采用问卷答题或心理实验方法,费时费力,试验范围较窄,难以获取大规模数据。时至今日,受限于测量工具,学习者词汇知识相关探讨还相对比较零星,研究结果众说纷纭,缺乏系统性。当代人工智能技术飞速发展,为大规模学习者词汇知识测量和描述提供了可能。为进一步推动学习者词汇知识研究,本文引入自然语言处理技术中的词向量模型,探讨词向量在学习者词汇深度知识习得研究中的应用途径。

一、词汇深度知识

词汇知识具有多维特征。Nation (2013: 49) 将二语词汇知识概括分为词形 (Form)、词义 (Meaning) 以及使用 (Use) 三大类。其中,每个大类下面又分为三个小类,每个小类包括接受性 (receptive) 与产出性 (productive) 两种类别。

中国学者马广惠 (2007) 在前人的基础上构建了二语词汇知识理论框架,提出十二种词汇知识。该词汇知识框架的主要特点在于增加了元词汇知识即词汇策略知识。词汇策略知识涉及词汇学习与交际策略,前者是指学习者为掌握词汇知识而采取的方法与行动,后者是指学习者在词汇使用过程中,为了克服词汇知识不足,

在保证实现既定交际目的前提下,而采取的各种应对策略。

词形辨识知识可以通过简单的答题测试得以检验,但词义及使用相关知识的测量就相对困难。从词义来看,大多数单词具有多义性,学习者掌握其中的单个或多个词义大为不同。就词汇使用而言,涉及多种语境和搭配,更是极其复杂,使得一般的测量手段很难实施,效果有限。

词汇水平测试 (Vocabulary Levels Test, 简称 VLT) 就属于单词辨识 (recognition) 类测试,由 Nation (1990) 设计,后经 Schmitt et al. (2001) 修订完善。词汇量测试 (Vocabulary Size Test, 简称 VST) 则是专门用于评估学习者词汇量 (Nation & Beglar, 2007)。VLT 和 VST 主要用于测量接受性词汇广度知识。二者都是以单词的词形与意义匹配为测试目标,无法考虑单词的实际运用情况,基本忽略词语搭配等短语学知识,仅适用于考察词汇知识中相对浅层次内容。

词汇联想测试 (Word Associates Test, 简称 WAT) 将词汇深度知识理解为单词之间的相互关联网络,考察学习者对目标词两种主要语义关系的掌握程度:纵聚合 (paradigmatic) 与横组合 (syntagmatic) 联系 (Read, 1993; Read, 2000)。WAT 是接受性词汇广度知识测量工具,能够考察学习者词汇语义以及搭配知识。它已经不再局限于词形,相对较好地反映了词汇深度知识。但是,WAT 仍然存在一个重要问题,它并非检测学习者的真实语言运用,得到的是类似问卷调查式的诱导性数据。

词汇知识量表 (Vocabulary Knowledge Scale,

简称VKS)最初用于词汇知识发展研究,它把词汇知识分成5个级别,采用被试自我报告(self-report)的方式,考察学习者词汇知识掌握程度(Paribakht & Wesche, 1993)。VKS属于混合型词汇深度知识测量工具。它存在一些缺陷,包括对词汇知识发展过程的单向线性预设,测试结果主观性较强等问题(Schmitt, 2010: 220-221)。

词汇频率概貌(Lexical Frequency Profile)(Laufer & Nation, 1995)以及词汇多样性(Lexical Diversity)(Malvern et al., 2004)等测量工具主要基于学习者自由产出数据,而不是通过访谈、问卷调查等方式得到的诱导式数据,为产出性词汇深度知识测量提供了重要技术。因此,这类测量工具可以计算分析由学生写作文本构成的学习者语料,并结合语料库方法开展二语词汇习得研究。不过,现有词汇测评工具大多仍然以词形频数或词汇覆盖度为基础,没有真正反映学习者词汇深度知识或者隐性知识。

二、词向量技术

随着深度学习技术蓬勃兴起,人工智能领域的自然语言处理研究迎来了新的发展阶段。2013年,谷歌公司发布Word2vec词向量工具(Mikolov et al., 2013a; Mikolov et al., 2013b)。由于Word2vec在语义相似度以及类比推理(analogical reasoning)方面表现出众,立即引起了学界和业内的极大兴趣。

词向量技术的具体表现形式多种多样,除了Word2vec,还有Glove、FastText等类似工具。词向量与当前的生成式人工智能(如ChatGPT)等一脉相承,都属于神经网络语言模型(袁毓林, 2023)。当然,二者的规模差异极大,故后者又常称作“大语言模型”(Large Language Models)。语言模型(Language Model),通常又称概率语言模型(Probabilistic Language Model),指关于特定语言的概率分布。具体来说,就是给定一个字符串s,该模型能够计算出所给字符串出现在特定语言中出现的概率 $P(s)$ 。在实际应用中,语言模型主要基于N-Grams进行统计与计算。

经过大规模语料训练得到的Word2vec词向量模型能够计算词汇语义相似度,甚至还能够进行类比推理,其结果往往与人类直觉相当吻合。例如,Word2vec

词向量能够识别国家与首都关系,使用Paris的向量减去France向量,再加上Italy的向量($\text{Paris} - \text{France} + \text{Italy}$),可以得到Rome(Mikolov et al., 2013a)。这是一种类比推理关系:巴黎之于法国相当于罗马之于意大利。另一项值得关注的表现是Word2vec词向量对性别语义关系的强大挖掘能力。一个经常引用的经典例子是:用king的词向量减去man词向量,在加上woman词向量,可以得到queen,即 $\text{king} - \text{man} + \text{woman} \approx \text{queen}$ (Mikolov et al., 2013c)。上例表明,国王减去男人的属性,加上女人的属性就是王后。有科学家在国际顶尖期刊《科学》(Science)上发文指出,词向量模型学习到了人类的性别和职业偏见(Caliskan et al., 2017)。

由此可见,词向量在文本语义挖掘方面表现优越。Word2vec词向量训练及相似度计算的Python实现具体技术细节可参考(邓海龙, 2019)。

三、基于词向量的词汇深度知识习得

运用词向量自动分析学习者词汇知识涉及跨学科兼容性论证、具体技术实现以及消除词汇语义量化表示与语言学描述之间的沟通壁垒等诸多内容。基于词向量的学习者词汇深度知识习得研究需要构建基于词向量的学习者英语词汇知识自动分析方法。基于词向量的学习者词汇知识自动分析方法面临两个关键技术问题:(1)构建基于词向量的词汇知识表示系统;(2)制定适宜的学习者词汇知识运用对比测量方法。

(一) 基于词向量的词汇知识表示系统

词向量是词汇语义表示技术,即每一个单词的意义用若干个维度的实数进行表示。例如,dog这个单词可以由“-0.3, 0.4, 0.15, -0.2”等一组数表示,而cat则由“-0.1, 0.2, 0.35, -0.4”等一组数表示。当然,实际应用中的词向量一般在100维以上,此处以4个维度为例仅为示意方便。数字本身较为抽象,很难为人类所理解。这是神经网络常受到诟病的主要问题:可解释性差。

显然,揭示和刻画词向量词汇知识与人类日常理解词汇知识之间的相互对应关系是极为关键。解决词向量技术与语言学研究之间的相互对接问题有助于减少当代自然语言处理技术与二语习得研究之间的联通壁垒。

还是以上述两个单词的向量表示为例。假设通过某种方法,发现了该词向量的四个维度分别表示“动物属性、家养属性、宠物属性、攻击属性”,那就很容易对两个单词的不同维度词义进行对比分析了。

(二) 学习者词汇知识运用对比测量方法

基于学习者语料库进行词向量训练,可以得到学习者词汇知识特征;而基于本族语语料库进行词向量训练,则可以得到本族语词汇知识特征。对比两者,我们可以发现学习者与本族语者之间的词汇知识运用差异。

由于两个向量空间坐标系不同,学习者与本族语词向量模型之间无法直接对比。以目前技术来看,可以从三个方面制定词向量对比测量方法,为后续研究奠定技术基础。第一,定性比较方法,即人工对比分析目标单词在两个词向量模型中的邻近词分布差异。第二,网络图比较方法,即通过构建两个词向量模型的网络图,比较以目标单词为中心的网络结构差异。第三,数学量化比较法,即通过数学方法进行向量空间对齐,然后比较目标单词在两个对齐向量空间中的向量差异。三种方法有不同特点,具有不同的适应场景。

参考文献

- [1] Caliskan, A., Bryson J. J., Narayanan A. Semantics Derived Automatically from Language Corpora Contain Human-Like Biases[J]. *Science*, 2017 (6334): 183-186.
- [2] Laufer, B., Nation P. Vocabulary Size and Use: Lexical Richness in L2 Written Production[J]. *Applied Linguistics*, 1995 (3): 307-322.
- [3] Malvern, D., Richards B., Chipere N., et al. Lexical Diversity and Language Development: Quantification and Assessment[M]. New York: Palgrave Macmillan, 2004.
- [4] Mikolov, T., Chen K., Corrado G., et al. 2013a. Efficient Estimation of Word Representations in Vector Space[M], Workshop at International Conference on Learning Representations (ICLR 2013); Scottsdale, Arizona.
- [5] Mikolov, T., Sutskever I., Chen K., et al. 2013b. Distributed Representations of Words and Phrases and Their Compositionality[M], International Conference on Neural Information Processing Systems (NIPS 2013); Lake Tahoe, Nevada: 3111-3119.
- [6] Mikolov, T., Yih W.-T., Zweig G. 2013c. Linguistic Regularities in Continuous Space Word Representations[C]//; City. 746-751.
- [7] Nation, I. S. P. Teaching and Learning Vocabulary[M]. New York: Newbury House, 1990.
- [8] Nation, I. S. P. Learning Vocabulary in Another Language[M], 2nd ed. Cambridge: Cambridge University Press, 2013.
- [9] Nation, I. S. P., Beglar D. A Vocabulary Size Test[J]. *The Language Teacher*, 2007 (7): 9-13.
- [10] Paribakht, T. S., Wesche M. B. Reading Comprehension and Second Language Development in a Comprehension-Based Esl Program[J]. *TESL Canada Journal*, 1993 (1): 9-29.
- [11] Read, J. The Development of a New Measure of L2 Vocabulary Knowledge[J]. *Language Testing*, 1993 (3): 355-371.
- [12] Read, J. Assessing Vocabulary[M]. Cambridge: Cambridge University Press, 2000.
- [13] Schmitt, N. Researching Vocabulary: A Vocabulary Research Manual[M]. Hampshire: Palgrave Macmillan, 2010.
- [14] Schmitt, N., Schmitt D., Clapham C. Developing and Exploring the Behaviour of Two New Versions of the Vocabulary Levels Test[J]. *Language Testing*, 2001 (1): 55-88.
- [15] 邓海龙. Python词向量训练与应用技术解析[J]. *语料库语言学*, 2019 (2): 88-109.
- [16] 马广惠. 二语词汇知识理论框架[J]. *外语与外语教学*, 2007 (04): 22-24.

基金项目: 本文系江西省高校人文社会科学研究项目“基于词向量的中国学习者英语词汇知识研究”(编号YY20104)的阶段性研究成果