信号与认知系统

词向量与创造力

专业 ： 心理学

班级 ： 心理 2202

学号 ： 3220102692

姓名 ： 毛沛炫

性别 ： 男

词向量与创造力

毛沛炫 [1]

([1] 浙江大学心理与行为科学系，浙江杭州，310058)

摘　要 本研究旨在探索词向量语义距离计算方法在创造力评估中的应用，并考察联想发散任务(Divergent Association Task, DAT)作为一种新型创造力测量工具的特点。本研究首先利用通过 Global Vectors for Word Representation (GloVe) 算法得到的词向量计算词语间的语义距离，验证词向量模型对词语语义关系的表征能力。随后，笔者作为被试完成 DAT 任务，即列出 10 个尽可能互不相关的词语，并通过计算词向量的平均语义距离得到 DAT 分数，以此评估个体的发散性思维能力。实验结果表明，GloVe 词向量模型能够有效捕捉词语间的语义距离，DAT 任务所得分数可反映个体的发散性思维水平，并具备客观、简便、可跨文化比较等优势。然而，DAT 分数的有效性亦受词向量模型质量和文化背景等因素的影响，人们也应谨慎对待使用 DAT 任务对人工智能创造力的评估结果。未来研究可进一步完善 DAT 任务，并结合其他评估方法，以期更全面、深入地理解人类与人工智能的创造力。

关键词 创造力；发散性思维；语义距离；词向量；联想发散任务

分类号 B842.3

1 实验背景

创造力是指个体产生兼具新颖性和适宜性的产品或观念的能力 (Runco & Jaeger, 2012)。简单来说，创造力不仅要求想法独特，还要在特定情境下具有实际意义。关于创造力的理论，主要有两种主流观点：联想理论和控制注意理论。

联想理论认为，创造性思维是一种自发的、低认知控制的联想加工过程。根据这一理论，高创造力的人能够更容易地检索到语义网络中较为“遥远”的概念，并将这些概念联结起来，形成新颖的想法(Mednick, 1962)。例如，当一个人看到“苹果”这个词时，普通人可能会联想到“水果”或“红色”，而高创造力的人可能会联想到“牛顿”或“科技公司”。这种远距离联想的能力被认为是创造力的核心(Kenett & Faust, 2019)。控制注意理论则强调，创造性思维不仅仅依赖于自发的联想，还需要有意识的、自上而下的认知控制。根据这一理论，创造性的解决方案需要个体在生成想法的同时，能够抑制常见的、过于普通的联想，并选择那些既新颖又合适的想法 (Beaty et al., 2014)。例如，在设计一个新产品时，设计师不仅需要生成大量的创意，还需要筛选出那些既独特又可行的方案。

而有关创造性思维的心理成分，经典的理论认为，包含两大重要成分：发散性思维 (divergent thinking) 和聚合性思维 (convergent thinking) (Guilford,1950)。其中，发散性思维指的是从一个目标出发，通过不同的途径和方法，寻求多种可能的答案，进行问题解决的思维方式，对个体创造性潜能有很好的预测 (Acar & Runco, 2014)。目前常用的发散性思维测验任务有用途创意任务 (Alternative Uses Task, AUT)、开放结局任务 (Consequences Task, CT)、现实情境问题 (Realistic Presented Problem, RPP) 等。而聚合性思维是指从某个目标出发，按照一定逻辑和思考方向找出问题正确答案或最优解的思维方式，其目标任务的答案是特定的或有限的。创造性问题中经典的聚合性思维任务是顿悟问题(Insight)，经典的图形顿悟任务有火柴棍问题(Match Stick Arithmetic Problems) (Knoblich et al., 1999)、言语性顿悟问题有远距离联想测验 (Remote Associate Test, RAT) (Mednick, 1962) 等。

然而，这些经典的方法存在一些局限性。首先，手动评分过程繁琐且耗时，评分者的主观性可能影响结果的可靠性 (Olson et al., 2021)。例如，在 AUT 任务中，评分者需要根据参与者的回答判断其“原创性”和“灵活性”，这一过程容易受到评分者个人

偏见的影响。其次，评分结果依赖于样本，难以进行跨文化或跨时间的比较 (Acar & Runco, 2014)。某些物品的用途在不同文化中可能有很大差异，导致原创性评分的标准不一致。此外，传统任务通常需要较长的文本回答，难以进行自动化评分。

为了更加客观、便捷的评估创造力水平，Olson 等人 (2021) 提出了一种新的测量方法——联想发散任务 (Divergent Association Task, DAT)。该任务要求参与者列出 10 个在意义和用法上尽可能不相关的词语，并通过计算这些词语之间的语义距离来评估个体的创造力。DAT 任务的优点在于其评分过程完全自动化，避免了主观偏差，且任务简短，适合大规模应用。研究表明，DAT 任务与其他传统创造力测量方法具有较高的相关性，显示出良好的收敛效度 (Olson et al., 2021)。

本次实验将使用通过Global Vectors for Word Representation (GloVe) 算法得到的词向量，计算不同词汇间的内积，以此了解不同词语之间的相近性。随后进行DAT任务来探究DAT任务对创造力(具体来说是发散性思维) 的评估特点。最后笔者也会讨论该计算方法是否可以测量机器智能的创造力这一问题。

2 实验方法

2.1 被试

仅笔者一人。

2.2 仪器与材料

GloVe (Global Vectors for Word Representation) 是一种无监督学习算法，用于生成词语的向量表示。该算法通过分析语料库中词语的全局共现统计信息来训练词向量，生成的词向量能够捕捉词语之间的语义关系，并在词向量空间中展现出特定的线性子结构。

GloVe 的核心思想是通过词语共现概率的比率来捕捉语义信息。具体来说，GloVe 模型的目标是学习词向量，使得两个词向量的点积等于它们共现概率的对数。由于概率比率的对数可以表示为对数概率的差值，GloVe 通过词向量的差值来编码语义信息。例如，词语“ice”和“steam”与不同探测词(如“solid”和“gas”) 的共现概率比率可以反映它们的热力学相态特性 (Pennington et al., 2014)。

2.3 实验流程

本实验包括比较词语之间距离与评估创造力两部分。其中，两部分中用于表征自然语言的词汇与计算语义距离的词向量一共有 400000 个，均来自 GloVe 模型。

在比较词语之间的语义距离之前，首先对任意的词向量$\boldsymbol{w}$进行均方根归一化，归一化后的词向量为：

$$\bar{w}_i = \frac{\boldsymbol{w_i}}{|\boldsymbol{w}|}$$

其中，w_i是$\boldsymbol{w}$的第i维，$|\boldsymbol{w}|$为原词向量的均方根。$|\boldsymbol{w}|$的计算公式为：

$$|\boldsymbol{w}| = \sqrt{\frac{1}{n}\sum_1^n w_i^2}$$

将词向量进行归一化可以消除向量长度对词语之间距离计算的影响、简化语义距离计算并提高语义距离的可比性。在比较词语之间距离的过程中，实验者将读取“dog”“bone”和“bird”的词向量，经均方根归一化后计算“dog”与“bone”“dog”与“bird”之间的内积，并比较“dog”与“bone”“bird”之中何者更为相近。随后，以相同的方式计算“dog”与 GloVe 中任一词的内积，找出内积最大的词。

在接下来的创造力评估任务中，实验者将想出 10 个尽可能互不相关的词语，然后计算归一化之后的词向量两两之间的语义距离，并取平均得到创造力指数 DAT (Olson et al., 2021)。

其中两个词向量$\boldsymbol{w}_1$、$\boldsymbol{w}_2$之间的语义距离通过余弦距离来定义：

$$semanticdistance = \left(1 - \frac{\boldsymbol{w}_1 \cdot \boldsymbol{w}_2}{|\boldsymbol{w}_1||\boldsymbol{w}_2|}\right) \times 100$$

基于余弦距离的计算方法，DAT 的计算公式为：

$$DAT = \frac{100}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \left(1 - \frac{\mathbf{w}_i \cdot \mathbf{w}_j}{|\mathbf{w}_i||\mathbf{w}_j|}\right)$$

然后实验者将从词语库中随机取 10 个词计算 DAT，并重复 100 遍，然后比较自己想的词组和随机词组的 DAT 大小。

3 结果分析

3.1 比较词语之间的距离

“dog”在 GloVe.6B.50d.txt 原始文件中的词向量见附录 1。将所有词语的词向量进行均方根归一化后，得到“dog”词向量和“bird”词向量的内积为 33.45，“dog”词向量和“bone”词向量的内积为 19.40。因此相较于“bone”，“dog”和“bird”在大量文本中的共现概率更大。

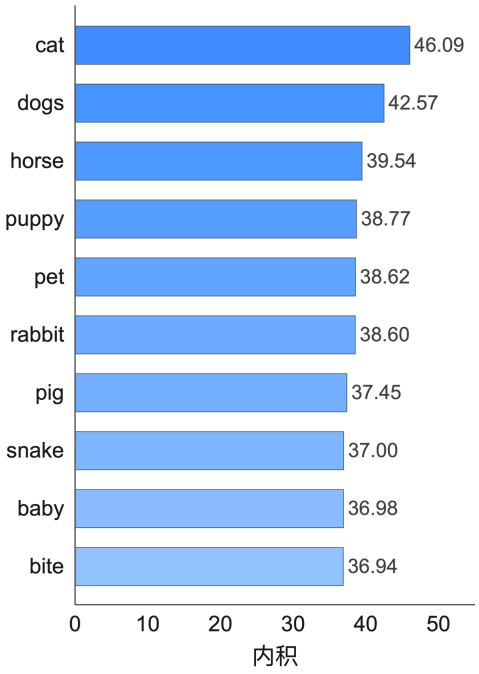

图 1 与“dog”内积最大的 10 个词

注：计算内积前所有向量都经过了均方根归一化。

同样的，计算其他所有词语与“dog”的内积，得到和“dog”内积最大的 10 个词语 (见图 1)。结果表明，在 399999 个词语中，和“dog”内积最大的是“cat”，接着是“dogs”。可以发现这 10 个词语中，前 8 个都是直接或间接指代动物 (少部分情况下“pet”也会指代非动物界的客体或智能体)，而“bite”这一动词排到了第十位，在“baby”的后面。

3.2 创造力

笔者所想的 10 个词汇以及词语之间的语义距离如图 2A 所示，计算得到这 10 个词语的 DAT 分数为 90.78。为了计算随机抽取的 10 个词语的 DAT 分数，首先对 400000 万个词语词语进行筛选，去除了含有非英文字符的词语，最后保留 335725 个词语。在 335725 个词语当中随机 10 个词语计算 DAT 分数并重复 100 次后，得到 DAT 分数平均值为 89.90，中位数为 90.10。笔者的 DAT 分数超过了 56%的随机情况，大致在平均水平。

笔者又将上述随机取样过程重复了 100 次，并记录每次随机取样下笔者的 DAT 分数超过的随机词组数量。结果表明，笔者的 DAT 分数最多一次超过了 72%的随机情况，最少一次超过了 47%的随机情况，平均超过了 61%的随机情况；在 99 次情况中，笔者的 DAT 分数超过了 50%的随机十元词组。100 次随机取样 (即一共随机进行了 10000 次随机抽取 10 个词汇) 的 DAT 分数平均值为 88.92，标准差为 0.55，基于 100 次随机取样的 DAT 分数分布显示出较高的一致性。

4 讨论

本次实验旨在探索基于词向量的语义距离计算方法在创造力评估中的应用价值，并考察联想发散任务 (DAT) 作为一种创造力测量工具的特点。在词语距离比较部分，笔者观察到“dog”和“cat”的

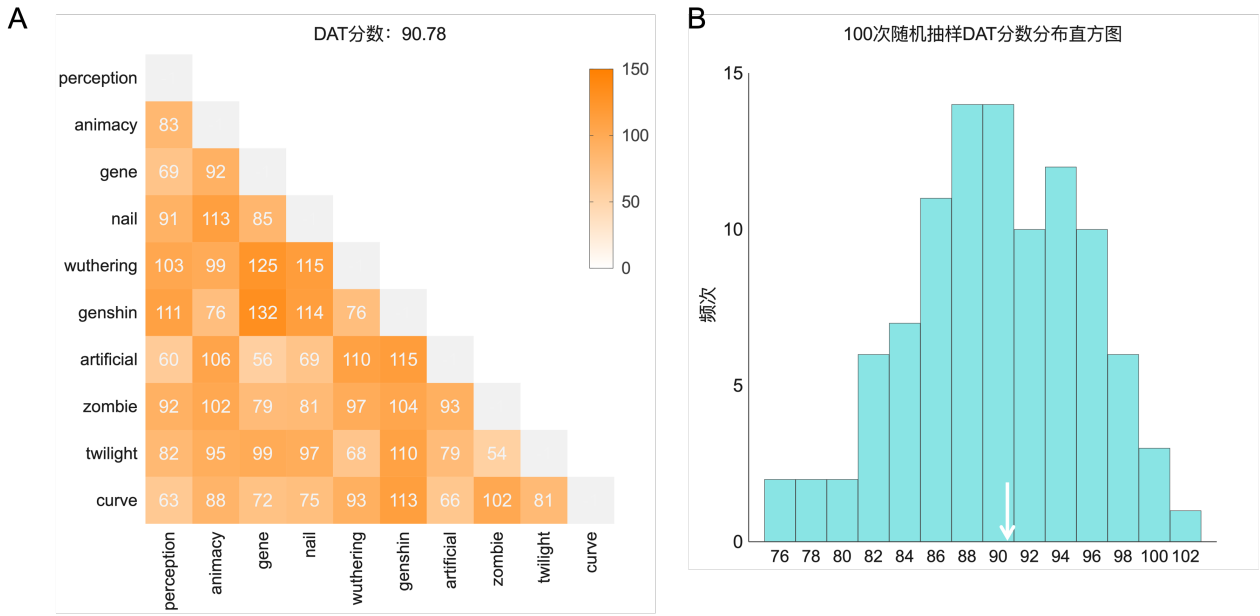

图 2 笔者的 DAT 分数以及 100 次随机取样得到的 DAT 分数分布

注：计算 DAT 分数前所有向量都经过了均方根归一化。A) 笔者的 DAT 分数以及各词语之间的语义距离。B) 100 次随机取样的 DAT 分数分布直方图，图中白色箭头为笔者的 DAT 分数。

语义距离小于“dog”和“bone”的语义距离，这一结果与笔者的日常认知存在一定偏差。实验还发现与“dog”词向量内积最大的词语大多与动物相关，而与“dog”语义联系紧密的动作或行为词汇 (如“bite”) 排名相对靠后。笔者通过个人的常识判断，“bite”“bone”似乎是与“dog”语义关联更紧密的词语；然而，基于 GloVe 词向量模型的计算结果表明，在大规模文本语料库中，“dog”与“cat”等其他动物的共现概率更高。这反映出在 Common Crawl 等语料库中，“dog”可能和“cat”“horse”等其他动物词语更频繁地在描述自然、动物等宏观概念的语境下共同出现。笔者认为，这一结果与笔者自身认知的偏差，或可部分归因于东西方思维模式的差异。西方思维模式倾向于范畴化和规则化，强调客体的属性，因而在类别判断任务中，更易将猩猩和熊猫归为同一类，因二者均属于“动物”范畴；而东方思维模式则更倾向于关联性和情境性，强调事物之间的相互关系，在类别判断任务重，会将猩猩和香蕉归为一类 (Nisbett, 2003)。采用大规模英文文本训练的词向量模型，可能在一定程度上内化了西方思维模式的特点，从而在语义表征上更倾向于将同属类别的词语聚拢

在创造力评估部分，笔者的 DAT 分数为 90.78，超过了约 56%的随机十元词组，初步表明笔者的发散性思维能力处于中等水平。然而，DAT 分数的绝对值和相对排名易受多种因素的影响。首先，DAT 分数显著依赖于训练词向量的数据集。本次实验采用的 GloVe 模型基于大规模英文语料库训练，其构建的词向量空间可能更契合英语语言的使用习惯和知识表征模式。对于母语非英语的被试而言，其 DAT 分数可能受到语言和文化背景差异的潜在影响。以笔者为例，笔者的母语为中文，在进行英语词语联想时，可能受到中式思维模式的干扰，进而影响 DAT 得分。但这种影响的具体方向尚不明朗，既可能因思维方式差异而导致分数偏高，也可能因不熟悉西方文化语境下的词语关联而导致分数偏低。其次，DAT 分数亦与 GloVe 算法的参数量级密切相关。理论上，参数量级越大，模型能够捕捉的语义信息越丰富、越精细，DAT 分数的区分度亦可能随之提高。本次实验所用 GloVe 模型参数量级为 6B，词向量空间为 50 维，这可能在一定程度上限制了 DAT 分数的区分效力。此外，随机抽取的十元词组的 DAT 分数亦受词语库质量的影响。尽管实验已剔除所有含非英文字符的词语，但剩余约 33 万个英文单词中，仍可能包含非常规用词 (如“vvv”“dys”“mpx”等) 或生僻词。这些词汇在日常语境中较少出现甚至只在某些，可能导致随机词组的 DAT 分数整体偏高，进而相对拉低笔者的 DAT 分数排名。为更精确地评估个体创造力水平，未来研究可考虑采用更大规模、且更贴近研究对象文化背景的语料库训练词向量模型，并结合更多同文化背景被试的数据进行参照。

结合实验内容和实验过程，笔者认为 DAT 任务作为一种新兴的创造力测量手段，在本次实验中展现出独特的优势与局限。相较于传统发散性思维任务 (如 AUT、CT) ，DAT 任务在评估过程、评估指标和评分标准等方面均有所优化。DAT 任务的评分立足于词语在词向量空间中的绝对距离，而非基于样本的相对排名，这赋予 DAT 分数更强的同文化跨语言和跨时间可比性，为未来进行大规模、跨地域的创造力研究提供了可能。

然而，DAT 任务的局限性亦不容忽视。DAT 分数的有效性高度依赖于词向量模型的质量与适用性。词向量模型本质上是在特定语料库上训练的，其语义空间可能存在固有的偏差或局限，进而影响 DAT 分数对个体创造力的精确表征。若语料库对某些词语的语义关系建模失真，或语料库本身即存在文化或地域偏见，则可能从根本上影响 DAT 分数的有效性。此外，DAT 任务得到的创造力指标可能过于笼统。尽管 DAT 任务与传统发散性思维任务存在一定程度的相关性 (Olson et al., 2021)，但其对创造力构成要素的评估可能不够全面，例如灵活性 (flexibility) 、独特性 (originality) 和精细度 (elaboration) 等维度在 DAT 任务中均难以有效测量。更重要的是，DAT 任务的跨语言文化适用性仍待进一步考量。词向量模型是基于特定语言的语料库训练的，DAT 任务在不同语言文化背景下的适用性和解释力可能存在差异。未来研究亟需深入探讨 DAT 任务的跨文化有效性，并开发针对不同语言文化的 DAT 任务版本。

近年来，有研究者尝试采用 AUT、CT、DAT 等任务评估人工智能的创造力水平，并将其与人类被试的表现进行比较研究。部分研究发现，在特定任务和评估指标下，人工智能甚至展现出超越人类的创造力潜能 (Cropley, 2023; Haase & Hanel, 2023; Hubert et al., 2024)。对此，笔者认为有必要对人工智能的“创造力”与人类创造力的本质差异保持清醒的认识。DAT 任务的理论基础是联想主义，其核心假

设是创造力与语义记忆网络的结构和联想能力存在内在关联。人类的语义记忆网络是在长期生活经验、文化濡染和社会互动的复杂过程中逐步构建的，天然蕴含着丰富的情感、意图和情境信息。反观当前的人工智能，特别是基于深度学习的语言模型，其“知识”和“联想”能力主要来源于对海量文本数据的统计学习和模式识别。尽管这些模型在特定发散性思维任务上表现出彩，但其内在的认知机制和创造性过程与人类迥异。目前大部分大语言模型都是基于联结主义 (connectionism) 的人工智能，其语义空间的构建是纯粹数据驱动的，缺乏人类经验的具身性 (embodiment) 和情境性 (situatedness)，难以真正理解词语的深层语义和语用内涵，更遑论产生人类那样具有复杂意图和价值导向的创造性构想。此外，已有研究表明，尽管 GPT 等大型语言模型在 DAT 任务上可获得较高得分，但在某些方面，人类的创造性回答仍保有其独特优势，如总词语数量更多等，更能体现人类个体的原创风格和独特视角 (Hubert et al., 2024)。还有研究显示，人工智能在辅助写作等创意任务中，反而可能导致作品间的差异性降低，造成创意同质化现象 (Doshi & Hauser, 2024)。因此，在评估人工智能的“创造力”时，不应仅关注其在特定任务上的量化得分，更需深入剖析其创造性过程的内在特点与局限，并警惕人工智能对人类创造性思维可能产生的潜在影响。

进一步而言，创造力本身即是一个多维度、多层级的复杂概念，发散性思维仅是其重要构成维度之一。DAT 任务作为一种发散性思维的测量工具，在创造力评估领域具有一定的应用价值，但无法完全捕捉创造力的全貌。除发散性思维外，聚合性思维、直觉思维以及情感、动机等因素在创造性活动中同样扮演着不可或缺的角色。科学领域的突破性进展往往离不开顿悟和直觉的驱动，而艺术领域的创新则更仰赖于个体情感的表达和价值的。未来研究亟待开发更多元、更全面的评估方法，以期更深入地理解人类与人工智能的创造力。

总结而言，本次实验探讨了基于词向量的语义距离计算方法和 DAT 任务的特性。实验结果表明，DAT 任务作为一种简便、客观的发散性思维测量工具，在创造力评估领域展现出一定的应用潜力，但其局限性亦不容忽视。针对 DAT 任务及人工智能创造力的评估，未来研究应秉持审慎和批判的态度，持续探索和完善评估方法，以期更深刻地理解创造力的本质，并更好地利用人工智能服务于人类的创造性发展。

参 考 文 献

Acar, S., & Runco, M. A. (2014). Assessing Associative Distance Among Ideas Elicited by Tests of Divergent Thinking. *Creativity Research Journal*, *26*(2), 229–238. https://doi.org/10.1080/10400419.2014.901095

Beaty, R. E., Silvia, P. J., Nusbaum, E. C., Jauk, E., & Benedek, M. (2014). The roles of associative and executive processes in creative cognition. *Memory & Cognition*, *42*(7), 1186–1197. https://doi.org/10.3758/s13421-014-0428-8

Cropley, D. (2023). Is artificial intelligence more creative than humans?: ChatGPT and the Divergent Association Task. *Learning Letters*, *2*, 13. https://doi.org/10.59453/ll.v2.13

Doshi, A. R., & Hauser, O. P. (2024). Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, *10*(28), eadn5290. https://doi.org/10.1126/sciadv.adn5290

Haase, J., & Hanel, P. H. P. (2023). Artificial muses: Generative artificial intelligence chatbots have risen to human-level creativity. *Journal of Creativity*, *33*(3), 100066. https://doi.org/10.1016/j.yjoc.2023.100066

Hubert, K. F., Awa, K. N., & Zabelina, D. L. (2024). The current state of artificial intelligence generative language models is more creative than humans on divergent thinking tasks. *Scientific Reports*, *14*(1), 3440. https://doi.org/10.1038/s41598-024-53303-w

Kenett, Y. N., & Faust, M. (2019). A Semantic Network Cartography of the Creative Mind. *Trends in Cognitive Sciences*, *23*(4), 271–274. https://doi.org/10.1016/j.tics.2019.01.007

Knoblich, G., Ohlsson, S., Haider, H., & Rhenius, D. (1999). Constraint relaxation and chunk decomposition in insight problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(6), 1534–1555. https://doi.org/10.1037/0278-7393.25.6.1534

Mednick, S. (1962). The associative basis of the creative process. *Psychological Review*, *69*(3), 220–232. https://doi.org/10.1037/h0048850

Nisbett, R. E. (2003). *The geography of thought: How Asians and Westerners think differently ... And why.* (pp. xxiii, 263). Free Press.

Olson, J. A., Nahas, J., Chmoulevitch, D., Cropper, S. J., & Webb, M. E. (2021). Naming unrelated words predicts creativity. *Proceedings of the National Academy of Sciences*, *118*(25), e2022340118. https://doi.org/10.1073/pnas.2022340118

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://doi.org/10.3115/v1/D14-1162

Runco, M. A., & Jaeger, G. J. (2012). The Standard Definition of Creativity. *Creativity Research Journal*, *24*(1), 92–96. https://doi.org/10.1080/10400419.2012.650092

附录 1

“dog”的词向量

[0.1101 -0.3878 -0.5762 -0.2771 0.7052 0.5399 -1.0786 -0.4015 1.1504 -0.5678 0.0039
0.5288 0.6456 0.4726 0.4855 -0.1841 0.1801 0.9140 -1.1979 -0.5778 -0.3799 0.3361
0.7720 0.7556 0.4551 -1.7671 -1.0503 0.4257 0.4189 -0.6833 1.5673 0.2769 -0.6171
0.6464 -0.0770 0.3712 0.1308 -0.4514 0.2540 -0.7439 -0.0862 0.2407 -0.6482 0.8355
1.2502 -0.5138 0.0422 -0.8812 0.7158 0.3852]

Word Vectors and Creativity

MAO Pei-Xuan[1]

([1]*Department of Psychology and Behavioral Sciences, Zhejiang University, Hangzhou, Zhejiang 310058*)

Abstract

This research explores using word vector calculations of semantic distance to assess creativity, focusing on the Divergent Association Task (DAT) as a new way to measure creative thinking. First, researcher used the GloVe algorithm to measure how related words are, checking if this method could accurately show word meanings. Then, researcher took the DAT, listing 10 unrelated words and getting a DAT score based on how different these words were in meaning. This score aimed to measure divergent thinking, or the ability to think creatively. The results showed GloVe effectively measured word meaning differences, and DAT scores reflected divergent thinking. DAT is also objective, easy to use, and can compare creativity across cultures. However, DAT's accuracy depends on the quality of the word model and cultural context. We should also be careful when using DAT to judge AI creativity. Future studies can improve DAT and use other methods to better understand creativity in both humans and AI.

Key words: Creativity, Divergent Thinking, Semantic Distance, Word Vector, Divergent Association Task