arXiv:2601.01488v1 [cs.CL] 4 Jan 2026

难度的四个象限：一种简单的分类及其局限

Vanessa Toborek[1,2],Sebastian M¨uller[1,2] 和 Christian Bauckhage[1,2,3]

[1]德国波恩大学 [2]德国 Lamarr研究所[3]德国弗劳恩霍夫智能分析与信息系统研究所

摘要。课程学习（CL）旨在通过估计样本难度并据此安排样本顺序，改善模型训练效果。在自然语言处理中，通常使用与任务无关的语言学启发式方法或人类直觉来近似衡量难度，这隐含地假设这些信号与神经模型认为难以学习的内容相关。我们提出一种难度信号的四象限分类——人类与模型、与任务无关与任务相关——并在一个自然语言理解数据集上系统分析它们之间的相互作用。我们发现，与任务无关的特征大多相互独立，只有与任务相关的特征彼此一致。这些发现对课程学习中常见的直觉提出了质疑，并凸显出需要轻量级、与任务相关的难度估计器，以更准确地反映模型学习行为。

1 引言

课程学习（CL）基于这样一种直觉：神经网络训练应以模仿人类学习的方式进行组织，即从较容易的概念开始，再逐步过渡到较难的概念 [1, 2]。因此，任何 CL 策略都包含两个组成部分：一个为每个训练实例分配难度分数的函数，以及一个决定每个实例在训练过程中何时可用的调度器。在自然语言处理（NLP）中，难度估计尤其具有挑战性，因为语言学难度具有多个维度，难以用单一指标衡量 [3]。因此，课程学习研究采用了多种与任务无关的衡量方式，例如句子长度、句法复杂度或可读性分数，同时也采用直接借助人类直觉的方法 [4, 5]。

这些策略隐含地假设，人类感知到的语言学难度与神经网络实际面临的难度相符。然而，心理语言学、标注分歧 [6], 和训练动态分析 [7] 等多个领域所体现的难度概念各不相同，分别基于对人类处理过程、任务歧义或模型学习行为的不同假设。尽管相关证据不断累积，该领域仍缺乏对语言学难度、人类分歧与模型学习难度之间关系的系统分析。现有研究有时会区分特定于任务的难度与无关任务的难度 [8, 9], ，但这种区分仍纯粹是操作层面的，并非旨在捕捉更广泛的概念空间。

为弥补这一空缺，我们提出一种有理论依据的难度四象限分类，区分（i）难度来源是人类还是模型，以及（ii）信息是任务无关还是任务相关。这一分类综合了以下领域的见解：

任务无关的人类 长度、词语稀有度、SLE、多样性、复杂度、FRE、习得年龄、具体性、普及度 任务无关的模型 困惑度 任务相关的人类 标注者间分歧 任务相关的模型 置信度、变异性、正确性、损失

表 1：所有样本难度代理指标概览。我们提出将其划分为四个不同的组别。

可读性研究、人工标签变异和训练动态，并使我们能够提出并通过实证检验关于不同难度信号如何相互作用的预期。我们针对以下四个象限开展了系统的跨象限分析：（1）任务无关的人类语言学特征；（2）任务相关的人类难度；（3）任务无关的模型信号；以及（4）任务相关的模型难度。结果表明，任务无关难度与任务相关难度彼此正交。语言复杂度无法预测标注分歧或模型学习难度；只有任务相关的人类信号与任务相关的模型信号表现出显著的一致性。这些发现对许多启发式 CL策略背后的核心假设提出了质疑：即任务无关的语言学难度能够单方面反映模型在训练过程中所经历的难度。

其影响有两方面。首先，任务无关的课程学习启发式方法之所以成功，必定源于准确难度估计以外的机制，例如课程学习调度器在训练期间进行的分布重塑。其次，标注熵虽然信息量很高，但成本依然不低。因此，我们的结果强调，有必要开发新的低成本方法，在模型训练开始之前的预处理阶段近似估计任务相关的难度。

2 难度信号的分类

难度的四个象限我们根据两个维度将难度信号划分为四个象限：其来源（人类或模型）和范围（任务无关或任务相关）。表 1 概述了这一分类。

任务无关的人类难度（*TA-H*）这一组指标旨在衡量输入本身的语言学难度，而不受任何下游任务影响。其中一些代理指标是简单的表层启发式指标，例如平均句长或词汇稀有度，它们是根据训练语料库中的频率分布计算得出的。弗莱施阅读易读度（FRE）最初是为较长文本设计的，但也常用于句子级可读性评估 [3]。我们纳入了习得年龄（AOA）、具体性和普及度这几项心理语言学测量指标；研究表明，它们是词汇复杂性的有效预测指标 [10]。此外，我们还纳入了两项句法测量指标：多样性，定义为输入中词性标签的集合大小；以及复杂度，定义为句子依存句法树的平均深度。最后，我们还测试了学习得到的无参考指标 SLE；研究表明，该指标与人类对难度的感知高度相关 [11]。

任务无关的模型难度（*TA-M*）在大多数机器学习场景中，若不进行特定任务训练，就无法获得基于模型的难度信号。然而，在自然语言处理中，我们可以通过考察大型预训练语言模型在微调之前的行为，来近似衡量任务无关的模型难度。困惑度反映了预训练模型依据其习得的语言分布预测输入的能力，因此能够捕捉预训练过程中形成的流畅性和词汇预期等方面。我们使用输入中掩码词元的平均困惑度，作为任务无关的模型信号。

任务相关的人类难度（*TD-H*）对于这一类别，我们考察标注者之间的分歧。对于每个实例的多条标注，分歧反映了人们对正确标签的不确定性。这种不确定性可能源于多种因素，包括词汇或句法复杂度、固有的语义歧义、输入中的欠指明，或标注者的主观变异 [6]。我们使用标注熵来量化分歧，这是一种基于标签分布的不确定性度量。

任务相关的模型难度（*TD-M*）在特定任务上训练模型时，可以根据其训练动态推导出任务相关的难度代理指标[7]。我们跟踪几项统计数据：模型对正确标签的平均置信度、模型在各训练轮次中的正确性，以及模型的变异性（定义为训练过程中置信度的标准差）。此外，我们还监测损失的均值和标准差。这些指标提供了事后视角，展现每个实例对模型而言有多难学习。

四象限之间的预期交互四象限分类使我们能够对不同难度信号之间的关系提出预期：

H1：*TA-H*的内部一致性。众所周知，语言学难度具有多维性，词汇、句法和概念复杂度分别体现了不同方面。因此，我们预计TA-H各特征之间的内部相关性较低。

H2：*TA-H* $\leftrightarrow$ *TA-M*。困惑度反映了预训练语言模型对输入序列的意外度，这种意外度由模型习得的词汇和句法预期所决定。由于若干TA-H特征与词汇可预测性存在一定关联，我们预计困惑度与人类语言学特征之间存在中度相关性。

H3：*TA-H* $\leftrightarrow$ *TD-H*。基于语言复杂度可能导致人类分歧这一观察 [6], ，我们预计语言学难度与标注熵之间存在中度相关性。

H4：*TA-H* $\leftrightarrow$ *TD-M*。鉴于在某些情境下，采用表层语言学特征的CL策略已被证明有效，我们预计语言学难度与模型学习难度之间存在交互作用。

H5：*TD-H* $\leftrightarrow$ *TD-M*。关于“数据集制图”的先前研究表明，标签歧义较高的实例往往更难供模型学习 [7]。我们预计人类分歧与模型训练难度之间存在正相关关系。

尽管这些假设有既有研究作为依据，但目前该领域尚缺乏系统评估，无法确定这些理论上有依据的交互关系在实践中是否成立。这一分类为我们的实证研究奠定了概念基础。

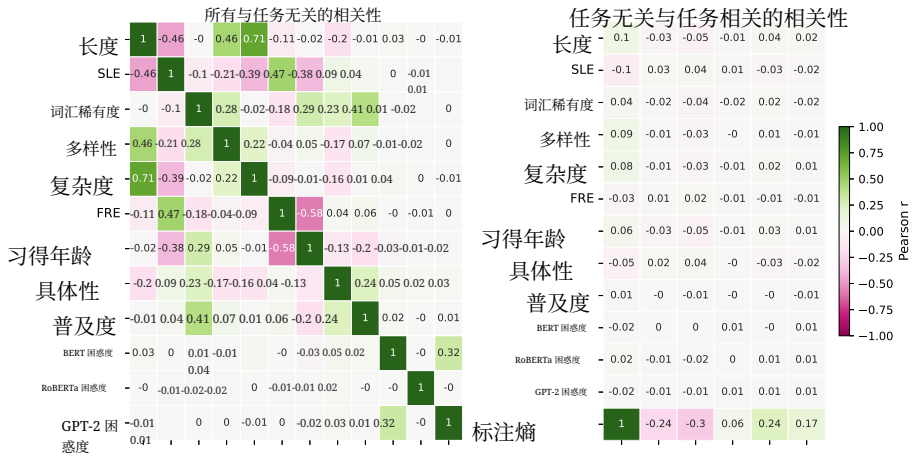

图 1：四个难度象限的皮尔逊相关系数。（左）任务无关的人类难度信号与任务无关的模型难度信号之间的相关性。（右）所有任务无关的难度信号与所有任务相关的难度信号之间的相关性。所有任务相关模型均取三个模型和各自十个随机种子的平均值。

3 实证研究：检验象限之间的交互作用

我们使用 SNLI 数据集 [12]，对难度象限之间交互关系的假设进行实证评估。该数据集的训练集划分中，每个样本都有四名独立标注者给出的标签，因此可以计算标注熵，作为任务相关人类难度代理指标。任务无关人类特征在预处理步骤中计算。对于任务无关模型特征，我们使用预训练的 BERT-base [13], RoBERTa-base [14], 和 GPT-2-base [15] 模型计算困惑度。在微调期间，我们在 12 个间隔均匀的检查点收集 3 647 个数据点的任务相关模型特征。BERT/RoBERTa 的训练批次大小为 b=64，学习率为 lr=3×10^{-5}；GPT-2 的训练批次大小为 b=16，学习率为 lr=10^{-5}。所有模型均训练 5 个轮次，使用 AdamW（权重衰减 0.01），并采用线性 lr 调度，预热比例为 6%。

我们通过四种方式评估各象限之间预期的关系：(1) 计算皮尔逊相关系数：(a) TA-H 象限内部的相关性，(b) TA-H 与 TD-H 之间的相关性，以及 (c) TA-H 与 TD-M 之间的相关性。(2) 为评估所有 TA 特征能否共同预测任一 TD 难度，我们分别以 TD-H 或 TD-M 特征为目标进行多元回归。(3) 遵循数据集制图的诊断方法 [7], ，我们找出最容易学习和最模糊的数据点中排名最高与最低的 25%。随后，我们比较这些数据点在任务无关特征上的分布，以检验这些特征能否有效地区分它们。(4) 作为合理性检查，我们针对不同的模型架构，复现标注熵与模型模糊性之间已确立的关系 [7] 。

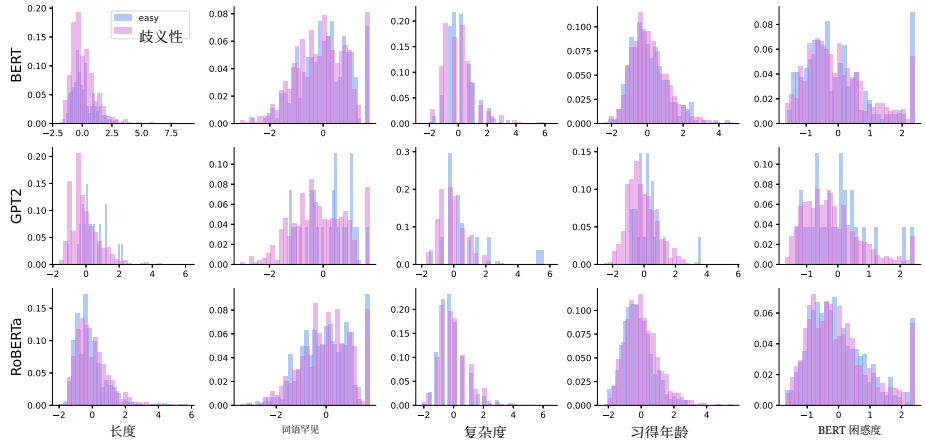

图 2：针对选定的任务无关代理指标——长度、词汇稀有度、复杂度、习得年龄和BERT困惑度——绘制的简单样本与模糊样本直方图（依据相应模型的置信度和变异性）。

综合来看，这些分析全面评估了四个象限中的难度信号是否呈现出假设中的交互作用。

结果图 1（左）展示了与任务无关的难度信号之间的皮尔逊相关系数。结果支持*H1*：任务无关人类难度信号相关性大多处于中等（r <0.5）到较低（$r < 0.2$）水平，只有长度—复杂度（$r = 0.71$）和习得年龄—弗莱施易读性（$r = -0.58$）较为突出。这表明，与任务无关的人类难度信号捕捉的是语言学难度的不同方面，而非某个单一的潜在因素。再来看*H2*，图 1（左）底部显示，困惑度与任务无关的人类难度信号并不相关，因此没有证据支持假设中的弱对齐关系。图 1（右）进一步呈现了这一情况：所有与任务无关的特征几乎都与任务相关的人类难度（*H3*）或任何任务相关的模型难度信号（*H4*）毫无相关性。这表明：（a）SNLI 标签分歧很可能源自固有歧义，而非语言复杂度；（b）表层语言属性难以预测模型认为哪些内容难以学习。为通过多变量分析检验这些关系，我们使用线性模型和基于树的模型，将每种任务相关信号回归到所有与任务无关的特征上。预测能力仍微乎其微（标注熵为 $R^2 < 0.05$，所有任务相关模型指标为 $R^2 < 0.1$），进一步印证了与任务无关的人类难度信号和任务相关难度之间联系薄弱。通过数据集制图定义“容易”和“有歧义”样本后，对其进行分布分析，提供了更多证据。如图 2 所示，对于所有模型，两组样本的与任务无关特征分布几乎完全重叠，这表明语言学难度无法区分易于学习的样本和有歧义的样本。最后，图 1（右）底行证实了预期中的任务相关信号间相关结构（*H5*）：标注熵与正确性、置信度和平均损失相一致，反映了它们共同受标签不确定性影响。

课程学习的4条经验

我们对 SNLI 数据集的分析表明，与任务无关和任务相关的难度信号在很大程度上相互独立，这对一种常见但隐含的假设提出了质疑：由浅层代理指标捕捉到的语言学难度，能够直接近似模型学习难度。然而，据报道，这两类难度度量在实践中都能产生有效的课程，这说明课程学习（CL）的效果不能仅归因于难度估计。相反，课程学习的效果源于难度度量、调度决策与任务特征之间的相互作用，其中调度器设计，尤其是如何施加难度约束以及随时间推移如何调整样本呈现，起着核心作用。我们进一步确认，任务相关的模型难度与标注者熵相一致，支持将人类分歧视为一种有意义的难度信号。然而，其高昂的获取成本限制了实际应用。未来研究的一个重要方向是开发轻量级的、预先计算的任务相关难度近似指标，使其更能反映模型的学习行为，并支持更有原则的课程设计。

参考文献

[1] J. L. Elman。《神经网络中的学习与发展：从小处着手的重要性》。《 *Cognition*》，48(1)，1993年。[2] Y. Bengio、J. Louradour、R. Collobert和J. Weston。《课程学习》。载于 *ICML*，2009年。[3] A. Battisti、D. Pf¨utze、A. S¨auberli、M. Kostrzewa和S. Ebling。《用于德语自动可读性评估和文本简化的语料库》。载于 *LREC*，2020年。[4] M. Elgaar和H. Amiri。《HuCurl：人类诱导的课程发现》。载于 *ACL*，2023年。[5] V. Toborek、S. M¨uller、T. Selbach、T. Horv´ath和C. Bauckhage。《超越浅层启发式方法：利用人类直觉进行课程学习》。载于 *ICNLSP*，2025年。[6] B. Plank。《人工标签变异的“问题”：论数据、建模和评估中的真实标签》。载于 *EMNLP*，2022年。[7] S. Swayamdipta、R. Schwartz、N. Lourie、Y. Wang、H. Hajishirzi、N. A. Smith和Y. Choi。《数据集制图：利用训练动态对数据集进行映射和诊断》。载于*EMNLP*，2020年。[8] F. Christopoulou、G. Lampouras和I. Iacobacci。《用于课程学习的训练动态：单语和跨语言自然语言理解研究》。载于 *EMNLP*，2022年。[9] V. Toborek、F. Seiffarth、S. M¨uller、T. Horv´ath和C. Bauckhage。《探索语言的课程学习：来自正规语言任务的经验》。载于 *Discovery Science*，2025年。[10] A. T. Desai、K. North、M. Zampieri和C. Homan。《SemEval-2021任务1中的LCP-RIT：探索词汇复杂度预测的语言学特征》。载于 *SemEval-2021*，2021年。[11] L. Cripwell、J. Legrand和C. Gardent。《简易程度估计（SLE）：一种用于句子简化的学习型无参考指标》。载于 *EMNLP*，2023年。[12] S. R. Bowman、G. Angeli、C. Potts和C. D. Manning。《用于学习自然语言推理的大型标注语料库》。载于 *EMNLP*，2015年。[13] J. Devlin、M.-W. Chang、K. Lee和K. Toutanova。《BERT：用于语言理解的深层双向Transformer预训练》。载于 *ACL*，2019年。[14] Y. Liu、M. Ott、N. Goyal、J. Du、M. Joshi、D. Chen、O. Levy、M. Lewis、L. Zettlemoyer和V. Stoyanov。《RoBERTa：一种经过稳健优化的BERT预训练方法》。《 *CoRR*》，2019年。[15] A. Radford、J. Wu、R. Child、D. Luan、D. Amodei和I. Sutskever。《语言模型是无监督的多任务学习者》。《 *OpenAI blog*》，2019年。