阿尔茨海默病早期诊断的语义连贯性标记

戴维德 科拉[1], 马泰奥 德尔桑托[1], 马尔科 阿戈斯托[2], 贝内代托维蒂耶洛[2] 和达尼埃莱·P. 拉迪乔尼[1]

[1]都灵大学计算机系[2]都灵大学公共卫生与儿科系

arXiv:2302.01025v1 [cs.CL] 2 Feb 2023

摘要

本文探讨如何利用语言模型分析语言，并通过困惑度指标区分认知障碍受试者与健康受试者。困惑度最初被视为一种信息论度量，用于评估给定语言模型预测某个文本序列的能力，或者说，评估某个词序列与特定语言模型的契合程度。我们使用公开数据开展了广泛的实验，并采用了多种语言模型，包括从2元语法到五元语法的N元语法模型，以及基于Transformer的语言模型GPT-2。我们研究了困惑度分数能否用于区分健康受试者与阿尔茨海默病（AD）受试者的转录文本。表现最佳的模型在对阿尔茨海默病组和对照组受试者进行分类时，准确率和F分数均达到满分（精确率/特异度和召回率/敏感度均为1.00）。这些结果表明，困惑度可以成为一种有价值的分析度量，有望用于辅助精神障碍症状的早期诊断。

关键词

痴呆诊断、困惑度、自动语言分析、语言模型、早期诊断、精神与认知障碍

1. 引言

本文是文章《语义连贯性标记：困惑度指标的贡献》的（大幅）删节版（*https://doi.org/10.1016/j.artmed.2022.102393, [1]*），原文还包含对所用数据及本文所述工作的实现方式的引用。

在经济发达的社会中，精神障碍带来的负担日益显现，对人们的日常生活造成负面影响，也给医疗卫生系统带来巨大成本。尽管许多精神病性障碍尚无治愈方法，但已有共识认为，对有较高风险罹患精神分裂症或相关精神病性障碍的人进行治疗，可从早期检测和干预中获益 [2]。为此，旨在分析思维和交流模式、以识别精神障碍早期症状的方法可能发挥核心作用 [3]。

人类语言分析最近成为一个新兴研究领域，有望帮助分析精神疾病的诊断与治疗。自然语言处理技术的最新进展

引用格式：*Colla D., Delsanto M., Agosto M., Vitiello B., Radicioni D. P.*（*2022*年）。语义连贯性标记：困惑度指标的贡献。《医学人工智能》，*134*，*102393*。*https://doi.org/10.1016/j.artmed.2022.102393, [1]*

davide.colla@unito.it (D. Colla); matteo.delsanto@unito.it (M. Delsanto); marco.agosto@unito.it (M. Agosto);benedetto.vitiello@unito.it (B. Vitiello); daniele.radicioni@unito.it (D.P. Radicioni)

使得开发精准的语言模型成为可能。这些模型可以视为文本序列上的概率分布，可用于估计文本与此类语言模型的一致程度（更准确地说，即文本在多大程度上可由这类模型预测）。为了衡量实际词元序列与概率分布之间的距离，我们提出使用困惑度这一度量；在已有文献中，它常用于语言模型的内在评估。在这项工作中，我们报告了一项简单实验的结果，旨在评估困惑度是否有助于区分健康受试者与精神障碍患者。

尽管在已有文献中，困惑度作为比较健康受试者与确诊受试者语言的工具并不新鲜，但我们报告的实验结果与已有文献中的结果相比表现良好。此外，据我们所知，此前没有研究比较过通过 GPT-2 和 N元语法等差异显著的语言模型计算出的困惑度分数，以区分健康受试者和患有阿尔茨海默病的受试者。这种差异具有实际影响，主要是因为训练和使用此类模型所需的计算成本不同，以及所学模型的描述能力各异。

2. 相关工作

过去十年间，自然语言处理技术的进步使人们得以构建自动处理语言分析和语言生成等任务的方法，其中也包括上述许多语言层面。这些方法发现了一些标记，可帮助区分精神障碍患者与健康对照组，并在个体患者层面预测高风险组中精神障碍的发作。

尽管困惑度最初旨在评估语言模型对先前未见数据的建模能力，但它也可用于比较（并区分）健康受试者或患有语言相关障碍者所生成的文本序列。简而言之，给定一个语言模型和一个词序列，困惑度是一个正数，用于表示该模型生成该序列的可能性有多低。第3节将更详细地介绍困惑度。在 [4] 本研究采用词性（POS）标签的N元语法来识别句法层面的模式。随后，研究人员训练了两个语言模型（一个基于患者数据，另一个基于健康对照组数据），再通过两个语言模型分别对新样本（即不属于任一训练数据集的样本）计算困惑度，据此对该样本进行分类。如果基于健康受试者数据训练的语言模型所得困惑度低于另一个语言模型，则将该样本归类为由健康受试者生成；如果基于患者数据训练的语言模型所得困惑度较低，则将其归类为由患者生成。近期研究提出，困惑度可作为认知衰退的指标 [5]；具体而言，研究人员根据访谈转录文本，记录了生理性衰老以及阿尔茨海默病（AD）和轻度认知障碍（MCI）发病初期口语内容的复杂程度。本研究使用的语言模型通过利用1元语法和2元语法信息构建；如下一节所示（参见公式2），这些模型所采用的上下文信息量有所不同。困惑度分数通过十折交叉验证计算：将受试者的转录文本分为十份；用其中九份构建模型，并用剩余一份进行测试。该过程重复十次，以确保每份文本都恰好作为测试集使用一次。研究进行了四轮检查

各轮检查之间的观察间隔超过20年；研究发现，实验开始时转录文本的困惑度分数与后来出现轻度认知障碍/阿尔茨海默病的受试者在痴呆筛查工具上的得分存在相关性。

困惑度已被用于预测阿尔茨海默病（AD），研究对象是 DementiaBank 的 Pitt 语料库中的转录文本，该语料库包含健康对照组和阿尔茨海默病患者的数据 [6]。更具体地说，研究者在 [7] 中构建了两个基于 LSTM 模型的神经语言模型，一个基于健康对照组的数据构建，另一个则使用痴呆组患者的数据训练。他们设计了留一说话人交叉验证；按照这一设置，为每位说话人s创建一个语言模型 $\mathcal{M}_{-s}$，使用该说话人所属组中除s本人以外所有说话人的转录文本。随后，使用两个语言模型 $\mathcal{M}_{-s}$对说话人s 的数据进行测试，从而得到一个困惑度分数 p_{own}；同时，还使用基于该说话人不属于的组的全部转录文本构建的语言模型进行测试，从而得到困惑度分数p_{other}。研究者计算了困惑度分数之间的差值 $\Delta_s = p_{own} - p_{other}$，以此描述说话人$s$。随后，他们通过设定一个阈值对每位说话人进行分类，以确保两组达到相同的等错误率。作者在 499 份转录文本上取得了 85.6% 的准确率，并表明困惑度还可用于预测患者的简易精神状态检查（ MMSE ）得分。我们在已有文献中发现，这项研究采用的方法与我们的工作最为接近；不过，两者在一些方面也有所不同。首先，我们研究了困惑度在评估健康受试者的语言时有多可靠。也就是说，我们分析了同一个体的困惑度分数如何变化，作为初步步骤，以评估困惑度是否适用于检查文本片段/转录文本；这些材料（与 Pitt 语料库的情况类似）通过多次访谈和检查收集，时间跨度长达数年。此外，我们关注的是评估同一个体的所有文本片段，以在受试者层面预测阿尔茨海默病诊断，而非预测每份转录文本各自的类别。为了评估困惑度作为辅助诊断工具的作用，我们仅分析了至少有两份转录文本的受试者的数据。

按照 [7], 中提出的方法，研究人员进一步探讨了困惑度在健康受试者和阿尔茨海默病患者分类中的应用 [8]。具体而言，研究人员使用 Pitt 语料库中对照组和阿尔茨海默病受试者的转录文本训练了不同的语言模型 [6]。这些语言模型被用于评估困惑度分数在多大程度上反映了语言使用方面的缺陷。我们的方法与之不同。首先，我们探索了两种不同类型的语言模型（N元语法和 GPT-2 模型，并分别使用 5、10、20 和 30 个训练轮次进行微调），以收集实验依据，考察用于计算困惑度分数的不同语言模型所达到的准确率。其次，我们比较了四种不同的决策规则，这些规则基于对照组和语言受损受试者的平均困惑度分数及其相应的标准差。此外， [8] 中是在转录文本层级进行分类，而我们关注的是受试者的分类。

3. 困惑度背景

大多数方法都依赖一种简单而强大的描述性（及预测性）理论框架，即分布假说。分布假说指出，

在相似语境中出现的词往往表达相似的含义 [9]。人们可以采用多种技术来获取词语的分布特征，通常将其表示为连续高维欧几里得空间中的实数稠密单位向量。在这种框架下，每个词都可以用一个向量来描述，而每个向量都可以映射到多维空间中，其中距离（例如向量之间的欧几里得距离）可作为相似度的替代指标，而相似度可以解释为一种度量。因此，语义内容相似的词预期会比语义上不相似的词彼此更接近。本文可以采用不同的度量来估算词语和词义的语义邻近度/距离 [10]。

语言模型（LMs）是一种统计推断工具，可用于估计词序列 $W = \{w_1, \ldots, w_k\}$ [11, 12] 的概率。该概率可以计算为

$$p(W) = \prod_{i=1}^{k} p(w_i | w_1, \ldots, w_{i-1}), \tag{1}$$

通常近似为

$$p(W) \approx \prod_{i=1}^{k} p(w_i | w_{i-N+1}, w_{i-N+2}, \ldots, w_{i-1}). \tag{2}$$

在后一种情况下，只考虑由少数几个（准确地说是 N 个）单词组成的块来预测整个 W：因此，我们可以根据 N 元语法来预测词序列；这些 N 元语法是由两个、三个或四个前置元素组成的块（分别称为二元语法、三元语法和四元语法）。一般而言，随着 N 增大，N 元语法模型往往能取得更好的性能，但代价是 P 的估算变得更加困难（$w_N | W_{1,N-1}$）。这类模型面临的另一个问题是，随着上下文长度增加，在训练语料库中找到相同长度序列的可能性会降低。为了处理训练语料库中未出现的 N 元语法（称为词汇表外N元语法），语言模型必须增加一个正则化步骤，使此前未见过的N元语法也能获得非零概率 [13, 14]。语言模型所赋予的概率是学习过程的结果，在这一过程中，模型会接触特定类型的文本数据。学习过程的目标是训练模型，使其能够预测与训练期间见到的句子非常相似的词序列。

如前所述，语言模型本质上是词序列的概率分布：困惑度最初被构想为一种用于语言模型的内在评估工具，因为它可以用来衡量给定语言模型 [12] 时，某个输入序列出现的可能性。这一度量定义如下。考虑一个由 k 个元素组成的词序列 $W = \{w_1, \ldots, w_k\}$；由于我们关注的是评估模型在未见过的数据上的表现，测试序列 W 必须是新的，且不能属于训练集。给定语言模型 LM，我们可以计算句子 W 的概率，即 LM(W)。这样的概率自然可以用来衡量语言模型本身的质量：概率越高，模型越好。根据模型计算出的平均对数概率定义为

$$\frac{1}{k} \log \prod_{i=1}^{k} \mathrm{LM}(W) = \frac{1}{k} \sum_{i=1}^{k} \log \mathrm{LM}(W),$$

这相当于整个测试序列 W 的对数概率除以序列中的词元数量。给定语言模型 LM，序列 W 的困惑度计算如下

$$\text{PPL(LM,W)} = \exp\{-\frac{1}{k}\sum_{i=1}^{k}\log \text{LM}(w_i|w_{1:i-1})\}. \tag{3}$$

现在可以清楚地理解，为什么较低的 PPL 值（对应较高的概率值）意味着词序列与模型拟合良好，或者说，模型能够预测该序列。

神经语言模型是基于神经网络的语言模型。这类模型利用神经网络处理较长历史信息的能力，提升了 N元语法的语言建模能力。此外，神经模型无需针对未见过的 N元语法进行正则化步骤，并通过处理分布式表示来应对 N元语法的数据稀疏性难题。在使用相同训练集的情况下，神经语言模型的预测能力高于 N元语法语言模型。尽管神经语言模型在自然语言处理任务上取得了显著进步，但其训练时间长于 N元语法语言模型。

4. 实验

本节所述实验旨在回答一个核心问题：对于某一特定类别的受试者，他们被诊断为患有影响常见语言能力的障碍，能否仅根据困惑度指标，自动区分其语言与健康对照组的语言？本实验使用了 Pitt 语料库，并从中选取了对偷饼干刺激图片作出回应的转录文本 [15], ，其中包括被诊断为患有痴呆的患者的转录文本（n = 194）以及健康对照组的转录文本（n = 99）。[1]

4.1. 比较的语言模型

我们设计了不同的实验方案，以比较通过两种不同架构（N元语法和 GPT-2）训练得到的语言模型所计算出的困惑度。

4.1.1. N元语法

由于 N元语法实现了最简单的带上下文语言模型，其中每个词仅以其前面的 N-1 个词元为条件，因此我们在第一个实验设置中采用了 N元语法。为清晰起见，我们介绍二元语法的形式化定义；这种表述还可以进一步推广到任意 N。

[1]复现实验的代码可在 https://github.com/davidecolla/semantic_连贯性_标记，[16]。

我们将词序列 $W_{1,\ n} = \{w_1, w_2, \ldots, w_n\}$ 的概率定义为：

$$P(W_{1,n}) = \prod_{i=1}^{n} P(w_i|w_{i-1}),$$

其中，每个二元语法的概率均通过最大似然估计（MLE） [17, 第 3] 章进行估算。[2] 根据最大似然估计，我们可以按如下方式估算二元语法（$w_{i-1},\ w_i$）的概率：

$$P(w_i|w_{i-1}) = \frac{C(w_i|w_{i-1})}{C(w_{i-1})} \tag{4}$$

其中，C（$w_i|w_{i-1}$）表示二元语法（$w_{i-1},\ w_i$）在训练集中的出现次数，而 C（w_{i-1}）仅统计单词 w_{i-1} 的出现次数。值得一提的是，在有限词汇表上训练二元语法可能会遇到未登录词，即训练过程中未出现过的词。未登录词会导致难以计算其所在句子的概率：在这种情况下，我们无法计算包含该未知词的二元语法的概率，从而影响整个序列的概率。我们通过插值 Kneser-Ney 平滑技术处理未见过的N元语法问题。该技术属于插值策略，基于绝对折扣技术 [14]。在本研究中，我们试验了从二元语法到五元语法的 N 元语法；Kneser-Ney 折扣因子 d 设为 0。1。[3] 每项实验均采用封闭词汇表：也就是说，各项实验所用的 N 元语法模型，均基于本文转录文本拼接后得到的词汇表进行训练。由于困惑度受词汇表大小限制，固定词汇表的规模有助于比较在不同语料库上训练的 N 元语法模型的困惑度分数。

4.1.2. GPT-2

我们设计的第二种实验设置采用了 GPT-2 神经模型，具体来说，我们使用了通过 Hugging Face Transformers 库提供的 GPT-2 预训练模型。[4]在这种设置下，输入文本经过预训练分词器处理，并分组成长度为1024 个词元的文本块。该预训练模型针对输入文本被专门训练为因果语言模型（CLM），即根据左侧上下文预测一个词。由于模型将每个词元的平均对数似然作为损失返回，因此文本的困惑度根据公式3计算。

4.2. 基于 PPL 的分类评估

尽管此前已广泛研究 PPL 的可靠性 [1], ，我们目前要探究语音文本转录稿的困惑度分数能否区分患者与健康对照组。研究使用了公开可用的 Pitt 语料库数据。[5] 这些数据

[2]在这种设置下，通常不会过滤停用词，因为它们能提供有用的序列信息。[3]为计算 N元语法，我们使用了 NLTK 版本 *lm*包中的语言建模模块（3.6.1，https: //www.nltk.org/api/nltk.lm.html。[4] https://huggingface.co/gpt2 [5]https://dementia.talkbank.org/access/English/Pitt.html。

表1。描述实验3所采用转录文本的统计数据。对于每个类别，我们报告每次访谈的平均词元数、每次访谈的平均唯一词元数、参与者人数、转录文本总数以及形符-词元比（TTR）。

类别	平均词元数	平均唯一词元数	参与者	转录文本	TTR
对照组	437	26	74	218	0.07
阿尔茨海默病	409	25	77	192	0.08

是匹兹堡大学医学院阿尔茨海默病及相关痴呆研究项目开展的一项更大规模研究方案的一部分 [6]。具体而言，我们选取了受试者对偷饼干图片的描述。偷饼干图片是一种常用测试，言语语言病理学家借此评估患有痴呆等疾病的受试者的说明性话语。

4.2.1. 材料

该数据集由552个文件组成，分别存放在“对照组”（243项）和“痴呆”（309项）目录中。这些文件对应于对99名对照组受试者以及219名确诊痴呆的受试者进行的多次访谈。本文档中的文本均按照CHAT格式转录，[6] 因此我们对这些文档进行了预处理以提取文本。在此过程中，原始文本在一定程度上得到了简化：例如，停顿被忽略了，犹豫现象等未得到一致标注的内容也被忽略 [18, 19]。

为了收集足够多的文本进行分析，我们剔除了只参加过一次访谈的受试者的访谈记录。最终，我们获得了74名对照组受试者的相关材料（共收集到218份转录文本），以及77名确诊痴呆的受试者的相关材料（共192份转录文本）。

实验3所用转录文本的词元数量、独特词元数量及形符-词元比的统计数据见表1。

4.2.2. 程序

本实验旨在检验困惑度分数的区分特征：具体而言，我们测试了一种简单的分类算法，用于区分认知障碍者和健康受试者。我们采用了文献 [7] 中的实验设置：分别利用对照组和阿尔茨海默病组的所有转录文本，建立了两个语言模型 LM_C 和 LM_{AD}。这些模型旨在掌握两组言语的主要语言特征，从而体现对照组和阿尔茨海默病组受试者所采用的典型语言。对于两组，我们均采用留一受试者法：使用同组中除一名受试者以外所有其他受试者的文件来优化语言模型，并将这名受试者的数据用于测试。对于每位受试者 s，我们使用同组中除该受试者 s 的转录文本，建立模型 LM_s s。随后，用两个困惑度分数 P_C 和 P_{AD} 来表征语料库中的每份转录文本，分别表示通过基于对照组和阿尔茨海默病组所建立的语言模型得到的分数。更

[6] https://talkbank.org/manuals/CHAT.pdf。

具体而言，如果受试者 s 属于阿尔茨海默病组，则其转录文本的分数 P_C 通过 LM_C 得到，而分数 P_{AD} 则利用 LM_s 计算。反之，如果受试者 s 来自对照组，则其转录文本的分数 P_C 通过 LM_s 得到，而分数 P_{AD} 则利用 LM_{AD}计算。此外，由于我们希望研究表征每位受试者的分数，因此将每位受试者各份转录文本的分数取平均值，汇总其困惑度分数 P_C 和 P_{AD}，从而得到P_C和 P_{AD}。

为了将阿尔茨海默病患者与健康受试者区分开来，我们采用了基于阈值的分类策略。我们探索了三种不同的方法来估计该阈值：

（i）在第一种设置中，我们使用训练过程中所有对照组受试者的平均困惑度分数；（ii）在第二种设置中，我们将阈值计算为属于阿尔茨海默病组的所有受试者的平均困惑度分数；（iii）在第三种设置中，我们利用差异$P_{AD} - P_C$估计两个不同的阈值，首先采用 [7] 和 [8]中报告的方法。

对于每位受试者，阈值估计过程均通过留一受试者设置进行，并针对（i）至（iii）中的三种方法分别重复。在第一种设置中，阈值根据对照组中除测试受试者 s之外的所有受试者进行估计：对于每位受试者 s，我们将阈值计算为对照组中除 s 之外的所有受试者的 P_C 分数的平均值——如果 s 来自健康对照组——。如果受试者 s 的困惑度分数P_C 高于健康对照组阈值，则将该受试者标记为患有阿尔茨海默病；否则标记为健康受试者。类似地，在第二种设置中，我们将阈值计算为阿尔茨海默病组中除 s之外的所有受试者的 P_{AD} 分数的平均值。如果受试者 s 的困惑度分数 P_{AD} 高于阿尔茨海默病组阈值的平均值，则将该受试者标记为健康受试者；否则标记为患有阿尔茨海默病。前两种设置的依据在于，每位受试者可能会被从同一组转录文本中获得的语言模型更准确地刻画：换言之，我们预期，对照组（阿尔茨海默病组）受试者使用在对照组（阿尔茨海默病组）受试者转录文本上训练或微调的语言模型时，其困惑度分数会低于属于另一组的受试者。

依据已有文献，在第三种设置中，我们用差异$D = P_{AD} - P_C$来表征每位受试者。我们定义了两个阈值：D_{AD}，即 AD 组患者所有差异得分的平均值；以及 D_C，即健康对照组所有差异得分的平均值。在这两种情况下，我们都纳入了该组中除测试受试者 s 之外的所有患者（将 s 排除在外，唯一目的是从 D_{AD} 或 D_C 中剔除其贡献）。不同于已有文献使用等错误率的做法，我们采用 D_{AD} 和 D_C 作为类别 AD 和 C 的紧凑描述符。该分类方案的基本原理是，将受试者归入其困惑度分数与自身最相似的类别。我们通过选择与受试者 s 对应的D 值之间差距最小的阈值（D_{AD} 或 D_C），来对受试者 s 进行分类，具体公式如下：

$$\text{class}(s) = \underset{x \in \{C, AD\}}{\text{argmin}} \left| D - \overline{D}_x \right|. \tag{5}$$

这种设置（涉及 D_{AD} 和 D_C）将称为 D。

此外，我们改进了决策规则 D，使其纳入标准差信息。除了平均值 D_{AD} 和 D_C，我们还计算了 σ_{AD} 和 σ_C，即受损组和对照组差异得分 D 的标准差。我们探讨了 3σ 规则，这是一种在经验科学中广受欢迎的启发式方法：它指出，对于假定由正态分布随机变量描述的总体，超过 99.7% 的数值落在均值三个标准差范围内，95.5% 落在两个标准差范围内，68.3% 落在一个标准差范围内 [20]。据此，我们尝试了三种方案，分别在平均得分上加上 1、2 和 3 个标准差；采用 2 个标准差时结果最佳。随后，我们对阈值作了如下调整：

$$\begin{aligned} \overline{D}^*_{AD} &= \overline{D}_{AD} + 2 \cdot \sigma_{AD}, \text{ and} \\ \overline{D}^*_{C} &= \overline{D}_{C} - 2 \cdot \sigma_{C}. \end{aligned}$$

分类所用的更新后决策规则随后调整为

$$\text{class}(s) = \underset{x \in \{C, AD\}}{\text{argmin}} \left| D - \overline{D}^*_x \right|. \tag{6}$$

这一设置包含 D^*_{AD} 和 D^*_C，将称为 D^*。

我们设计了双重实验设置，包括使用 N元语法和 GPT-2 的实验，并将窗口大小设为 20，以处理较短的文本样本（训练数据中最短的文本仅包含 23 个词元）。对于 N元语法，我们构建了从 2元语法到五元语法的模型；GPT-2 模型则采用 5、10、20 和 30 个训练轮次进行微调。

4.2.3. 评估指标

为评估结果，我们采用了精确率和召回率指标（特异度和敏感度），以及它们的调和平均数——F1分数和准确率。精确率（特异度）定义为$P = \frac{TP}{TP+FP}$，而召回率（敏感度）定义为 $R = \frac{TP}{TP+FN}$。精确率用于估计分类系统的精确程度，而召回率则表示在所有可能的结果中，有多少结果被识别出来。随后使用 F_1 度量来提供精确率和召回率的综合值，其中通过它们的调和平均数对这两个度量进行等权处理：$F_1 = 2 \cdot \frac{P \cdot R}{P+R}$

准确率计算为 $ACC = \frac{TP+TN}{P+N}$，即正确预测的数量（TP 与 TN 之和）占所检查记录总数（正例与负例，即 P 和 N 之和）的比例。

最后，为了记录一个用于评估相关两组的准确率和 F1分数的综合指标，我们采用了这三个值的调和平均数。其计算方式为

$$\text{HM}(\text{Acc.}, \text{F1}_{AD}, \text{F1}_C) = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} = \left(\frac{\sum_{i=1}^{n} x_i^{-1}}{n} \right)^{-1}.$$

其中，n 设为 x_i 个待平均值的数量。

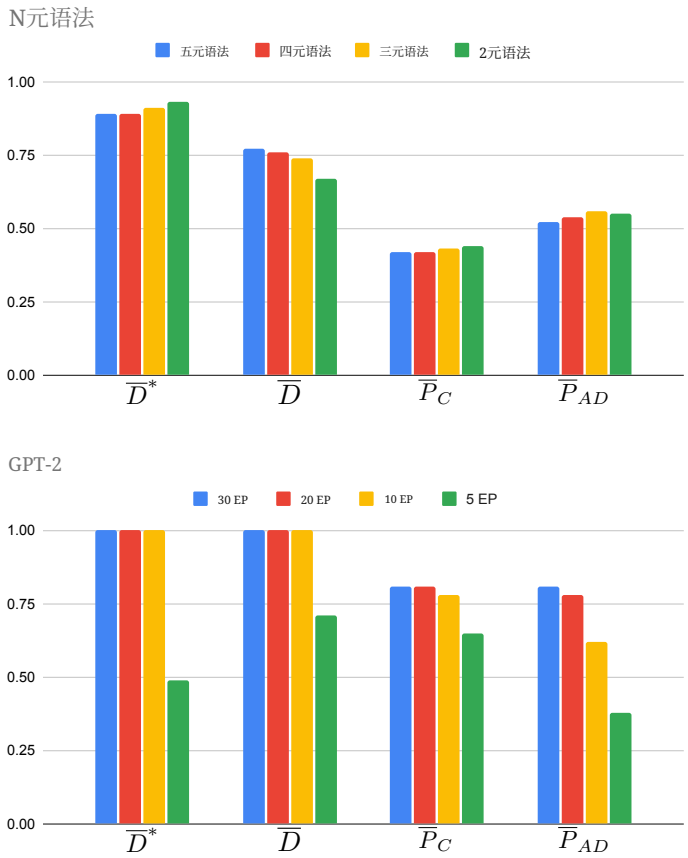

图 1： AD/对照组受试者分类第三项实验的准确率得分图。上方子图中的直方图显示了N元语法的准确率，下方直方图则报告了通过 GPT-2 模型获得的结果。不同颜色分别对应不同阶数的N元语法和不同的微调轮次。这些直方图分别展示了通过 $\overline{D}^*$, $\overline{D}$, $\overline{P}_C$ 和 $\overline{P}_{AD}$ 决策规则获得的得分。

4.2.4. 结果

总体准确率得分见图 1，不同实验条件下的详细数据见表3，载于 附录。

首先报告 N 元语法模型的结果。总体而言，最有效的策略是$\overline{D}^*$ (公式 6)，该策略基于阿尔茨海默病患者与健康对照组之间差异设定阈值，并结合 3σ 规则。表现最佳的模型基于二元语法，取得了 。93 的准确率、阿尔茨海默病组的 。92 F1分数，以及 C 组的 。93 F1分数。采用对照组困惑度分数的模型（图 $\overline{P}_C$ 中标示为1 ，见表 3）在所有条件下的准确率均最低，远低于随机猜测；而 $\overline{P}_{AD}$ 策略的准确率始终高于 。5。总体而言，我们观察到，随着马尔可夫假设长度增加，所有决策规则下 N 元语法模型的准确率都会降低（采用更多上下文似乎会对这类模型略有不利影响），但

表2.比较每个语言模型中阈值策略与分类策略有效性的研究。报告各模型得分最高的策略。

N 元语法模型	分类策略	平均 HM 分数
2元语法	$\overline{D}^*$	0.93
三元语法	$\overline{D}^*$	0.91
四元语法	$\overline{D}^*$	0.89
五元语法	$\overline{D}^*$	0.89
GPT-2 模型：训练轮次	分类策略	平均 HM 分数
5 个训练轮次	$\overline{D}$	0.71
10 个训练轮次	$\overline{D}, \overline{D}^*$	1.00
20 个训练轮次	$\overline{D}, \overline{D}^*$	1.00
30 个训练轮次	$\overline{D}, \overline{D}^*$	1.00

的 D 策略除外。¯

GPT-2 模型的结果显示，整体准确率较高，范围为。71其中，采用5 个训练轮次微调获得的最佳模型准确率为1。00 ，其余所有微调步骤均达到这一准确率。同样的趋势也体现在阿尔茨海默病（AD）和对照组受试者相关子任务的F1分数上，分别从约0。69 （采用5个训练轮次微调获得的最佳模型（D 阿尔茨海默病类别上的策略）到1。00 ，其余所有模型和子任务均达到这一分数。若考察阈值策略及其相关决策规则的效果，精细差异规则 D 是基于GTP-2的模型中表现最佳的策略，表3最右列对此进行了说明。这些分数是准确率、阿尔茨海默病受试者分类F1分数和对照组受试者分类F1分数的调和平均数。表2简要展示了同一列中的数据，并列出各模型对应的最佳策略。

为了结合已有文献来理解我们的结果，我们先从 Becker及其同事 [6]在研究第一版中获得的基线临床诊断准确率谈起：该准确率为 86%；考虑随访临床数据后，这一数值升至 91.4%，灵敏度为 0.988，特异度为 0.983。后续已有文献将其视为比较实验结果的金标准。需要指出的是，这些数据尤为重要，因为人工评估包含多个分析步骤，例如内科及神经科病史和检查、半结构化精神科访谈以及神经心理评估。后续研究中的实验结果则仅使用对一幅相当简单图像的描述转录文本，来接近上述评定。一项重要研究通过基于长短期记忆网络的模型 [7] 对单份转录文本进行分类，达到了 85.6% 的准确率。随后， [8], 的研究复现并改进了这一结果，其中报告的最佳模型在实验中取得了 0.872 的准确率。

总讨论

鉴于我们的实验结果似乎优于已有文献所报告的准确率分数，我们意识到，简短且受控的引导任务在某些情况下可能优于从说话者那里获取的自然语言数据。我们的结果质量还需要在不同

的设置下进行检验（包括更多语言和不同的实验条件：因此仍有大量实验工作有待开展），但这一事实表明，采用基于语言模型和困惑度分数技术的系统，可能会有效地为专业人员提供辅助。此外，通过比较差异如此之大的语言模型（如N元语法模型和基于较新GPT-2的模型），我们观察到，二元语法模型优于经过5个训练轮次微调的GPT-2模型。这一事实或许有助于深入了解结果准确率与计算时间及成本之间可能存在的权衡。

虽然困惑度被证明总体上是研究人类语言的一种可行工具，但我们发现，相关模型的输出存在一致差异，这些差异主要源于语言模型的内在属性、模型所考虑的上下文量、可用训练数据的规模，以及用于改进模型本身的训练量。首先，即便 N元语法难以与基于 GPT-2 的模型相提并论，尝试辨别这类模型在哪些情境下能取得更好的结果，仍可能有所裨益。令人有些意外的是，在我们的实验中，表现最佳的 N元语法模型（2元语法）达到的准确率为0.93，其调和平均数高于表现最佳的基于 GPT-2 的模型（HM =0.73；请参阅表 3），该模型经过5 轮训练微调，并采用了 D 决策规则。

—

这一结果或许可以从偷饼干图片描述所使用的语言相当规范这一点来理解，因此这项任务对 N元语法语言模型而言难度较低。从这些方面来看，我们得到的启示是：对于词汇和语言层面难度较低的任务，可以采用 N元语法；在某些此类问题中，与更全面、计算成本更高的 Transformer模型相比，采用 N 元语法模型可能更为便利（兼顾训练和测试所需的工作量）。关于准确率与计算开销之间的权衡，补充一些数据或许有助于说明。我们的实验使用了科学计算能力中心 [21] 提供的设备。具体而言，我们使用了配备 2 个 Intel Xeon E5-2680 v3处理器和 128GB 内存的节点。报告中的实验每种 GPT-2 配置耗时约 8 小时，所有 N 元语法模型的实验总共耗时约 12 分钟。

5. 结论

本研究探讨了困惑度是否适用于支持临床诊断中的自动语言分析，以及其适用程度。痴呆诊断是一个复杂、耗时且劳动密集的过程，涉及神经精神评估，包括内科及神经科病史和检查、半结构化精神科访谈以及神经心理测验 [22, 23]。因此，若能定义一种用于检测精神障碍症状的语言标记，便可为临床医生提供自动化语言分析程序，以高效、无创的方式助力精神疾病的早期诊断和治疗。为此，我们探讨了一个基本研究问题：困惑度分数能否以及在多大程度上将健康受试者的转录文本与阿尔茨海默病（AD）患者的转录文本分类开来。本实验使用了公开数据集 Pitt 语料库。我们设计了多种实验设置，以考察困惑度分数的预测能力和判别能力，并评估所得分类准确率如何随训练

／微调用于获取语言模型的训练量而变化。我们还比较了（2、3、4 和 5 元）N 元语法模型、0 至 30 个 GPT-2 微调轮次，以及四种不同的阈值策略。我们提出了新的阈值，并将其与已有文献报道的阈值进行了比较：新提出的分类策略确保结果持续优于最先进的结果。

最后，我们展望一下未来的工作。不同的语言模型只需付出一小部分训练/微调成本，便可能取得准确率相近的结果。例如，我们还对长短期记忆网络进行了初步测试（为简洁起见，本文未报告），结果显示其性能较差，且计算负载高于 GPT-2 架构。[7]此外，还可以采用不同的分类算法来区分患者和对照组受试者；改进所用的语言模型和整体分类策略，或许能带来显著提升。然而，仍需开展更多实验，以评估更大样本中的困惑度，并考察不同类型的口语。正如前文所述，用于描述偷饼干图片的语言相当规整。要更全面地刻画困惑度分数的判别能力，还需在不同语言及其相应的语言模型上开展实验。

然而，这项概念验证研究的发现具有若干启示：在预测转录文本的作者是痴呆患者还是健康受试者时，我们取得了有价值的结果。尤其值得注意的是，我们的预测仅依据困惑度分数，而所使用的信息量远少于人类专家在执行同一分类任务时，于各个诊断步骤中收集的临床证据。 [6]

参考文献

[1] D. Colla、M. Delsanto、M. Agosto、B. Vitiello、D. P. Radicioni，语义连贯性标记：困惑度指标的贡献，《医学人工智能》（2022年）。URL：https://www.sciencedirect.com/science/article/pii/S0933365722001440。doi：https://doi.org/10.1016/j.artmed.2022.102393。[2] M. Marshall、S. Lewis、A. Lockwood、R. Drake、P. Jones、T. Croudace，未治疗的精神病持续时间与首次发作患者队列结局之间的关联：系统综述，《普通精神病学档案》62（2005年）975–983。[3] M. K. Larson、E. F. Walker、M. T. Compton，精神分裂症及相关精神病性障碍前驱期的早期体征、诊断和治疗，《神经治疗学专家评论》10（2010年）1347–1359。[4] A. Stolcke、E. Shriberg，用于言语不流畅现象的统计语言建模，载于：1996年 IEEE 声学、语音与信号处理国际会议论文集，第1卷，IEEE，1996年，第405–408页。[5] C. Frankenberg、J. Weiner、T. Schultz、M. Knebel、C. Degen、H.-W. Wahl、J. Schroeder，困惑度——口语中认知变化的新预测指标？——以下研究的结果

[7]更具体地说，通过长短期记忆网络计算的困惑度分数波动极大（标准差值通常超过平均困惑度值）；即使增加训练轮次，情况也没有改善，而且所需时间几乎是训练 GPT-2 基础模型的两倍。

成人发展与衰老跨学科纵向研究（ILSE），Linguistics Vanguard 5（2019年）1–10。[6] J. T. Becker、F. Boiler、O. L. Lopez、J. Saxton、K. L. McGonigle，阿尔茨海默病的自然病程：研究队列描述及诊断准确率，Archives of Neurology 51（1994年）585–594。[7] J. Fritsch、S. Wankerl、E. Nöth，使用神经网络语言模型自动诊断阿尔茨海默病，载于：ICASSP 2019-2019 IEEE 声学、语音与信号处理国际会议（ICASSP），IEEE，2019年，第5841–5845页。[8] T. Cohen、S. Pakhomov，两种困惑度的故事：神经语言模型对阿尔茨海默型痴呆词汇提取缺陷的敏感度，载于：计算语言学协会第58届年会论文集，计算语言学协会，在线，2020年，第1946–1957页。URL：https://aclanthology.org/2 020.acl-main.176。doi：10.18653/v1/2020.acl-main.176。[9] Z. S. Harris，分布结构，Word 10（1954年）146–162。[10] D. Colla、E. Mensa、D. P. Radicioni，使用词义嵌入计算语义相似度的新型度量，Knowledge-Based Systems 206（2020年）106346。[11] C. Manning、H. Schutze，统计自然语言处理基础，麻省理工学院出版社，1999年。[12] Y. Goldberg，自然语言处理的神经网络方法，Synthesis Lectures on Human Language Technologies 10（2017年）1–309。[13] W. A. Gale、K. W. Church，加一法有什么问题，基于语料库的语言研究：献给 Jan Aarts（1994年）189–200。[14] R. Kneser、H. Ney，m 元语法语言建模的改进回退方法，载于：1995年声学、语音与信号处理国际会议，第1卷，IEEE，1995年，第181–184页。[15] H. Goodglass、E. Kaplan，波士顿失语症诊断检查手册，Lea & Febiger，1983年。[16] D. Colla，语义连贯性标记：源代码，2022年。URL：https://doi.org/10.5281/ze nodo.7118402。doi：10.5281/zenodo.7118402。[17] D. Jurafsky、J. H. Martin，语音与语言处理，第3卷，Prentice Hall，2014年。[18] B. MacWhinney，CHILDES 项目：分析言语的工具，第二卷：数据库，Psychology Press，2014年。[19] B. MacWhinney，分析言语的工具第1部分：CHAT 转录格式，Carnegie（2017年）1–115。[20] J. R. Helms，健康科学数学：综合方法，Cengage Learning，2009年。[21] M. Aldinucci、S. Bagnasco、S. Lusso、P. Pasteris、S. Rabellino、S. Vallero，OCCAM：灵活、多用途且可扩展的高性能计算集群，Journal of Physics: Conference Series 898（2017年）082039。[22] F. J. Huff、J. Becker、S. Belle、R. Nebes、A. Holland、F. Boller，阿尔茨海默病的认知缺陷与临床诊断，Neurology 37（1987年）1119–1119。[23] O. L. Lopez、A. Swihart、J. T. Becker、O. Reinmuth、C. Reynolds、D. Rezek、F. Daly，用于诊断阿尔茨海默病的 NINCDS-ADRDA 临床标准的可靠性，Neurology 40（1990年）1517–1517。

A. 详细结果

表3。
详细结果。该表报告了两项任务的准确率（Acc.）、精确率（P）、召回率（R）和F1分数这两项任务旨在识别阿尔茨海默病和对照组受试者。最右列报告调和平均数（HM）即准确率、阿尔茨海默病类别和C类的F1分数的调和平均数。最佳结果以粗体标示。

模型		Acc.	痴呆（阿尔茨海默病）			对照组（C）			HM(acc,F1$_{AD}$,F1$_{C}$)
			P	R	F1	P	R	F1	
2元语法	$\overline{P}_C$	0.44	0.41	0.21	0.28	0.46	0.69	0.55	0.39
	$\overline{P}_{AD}$	0.55	0.54	0.75	0.63	0.57	0.34	0.42	0.52
	$\overline{D}$	0.67	0.68	0.66	0.67	0.66	0.68	0.67	0.67
	$\overline{D}^*$	**0.93**	0.99	0.87	**0.92**	0.88	0.99	**0.93**	**0.93**
三元语法	$\overline{P}_C$	0.43	0.40	0.22	0.28	0.44	0.65	0.53	0.39
	$\overline{P}_{AD}$	0.56	0.55	0.70	0.62	0.57	0.41	0.47	0.54
	$\overline{D}$	0.74	0.76	0.71	0.74	0.72	0.77	0.75	0.74
	$\overline{D}^*$	**0.91**	1.00	0.83	**0.91**	0.85	1.00	**0.92**	**0.91**
四元语法	$\overline{P}_C$	0.42	0.38	0.23	0.29	0.43	0.61	0.51	0.38
	$\overline{P}_{AD}$	0.54	0.54	0.65	0.59	0.54	0.43	0.48	0.53
	$\overline{D}$	0.76	0.81	0.70	0.75	0.73	0.82	0.77	0.76
	$\overline{D}^*$	**0.89**	1.00	0.78	**0.88**	0.81	1.00	**0.90**	**0.89**
五元语法	$\overline{P}_C$	0.42	0.38	0.23	0.29	0.43	0.61	0.51	0.38
	$\overline{P}_{AD}$	0.52	0.53	0.62	0.57	0.52	0.42	0.46	0.52
	$\overline{D}$	0.77	0.86	0.66	0.75	0.72	0.89	0.80	0.77
	$\overline{D}^*$	**0.89**	1.00	0.79	**0.88**	0.82	1.00	**0.90**	**0.89**
GPT-2 训练轮次：5	$\overline{P}_C$	0.65	0.64	0.70	0.67	0.66	0.59	0.62	0.65
	$\overline{P}_{AD}$	0.38	0.42	0.58	0.49	0.29	0.18	0.22	0.33
	$\overline{D}$	**0.71**	0.76	0.62	**0.69**	0.67	0.80	**0.73**	**0.71**
	$\overline{D}^*$	0.49	0.50	0.09	0.15	0.49	0.91	0.64	0.30
GPT-2 训练轮次：10	$\overline{P}_C$	0.78	0.70	0.99	0.82	0.98	0.57	0.72	0.77
	$\overline{P}_{AD}$	0.62	0.63	0.58	0.61	0.60	0.65	0.62	0.62
	$\overline{D}$	**1.00**	1.00	1.00	**1.00**	1.00	1.00	**1.00**	**1.00**
	$\overline{D}^*$	**1.00**	1.00	1.00	**1.00**	1.00	1.00	**1.00**	**1.00**
GPT-2 训练轮次：20	$\overline{P}_C$	0.81	0.73	1.00	0.84	1.00	0.61	0.76	0.80
	$\overline{P}_{AD}$	0.78	0.91	0.64	0.75	0.71	0.93	0.81	0.78
	$\overline{D}$	**1.00**	1.00	1.00	**1.00**	1.00	1.00	**1.00**	**1.00**
	$\overline{D}^*$	**1.00**	1.00	1.00	**1.00**	1.00	1.00	**1.00**	**1.00**
GPT-2 训练轮次：30	$\overline{P}_C$	0.81	0.73	1.00	0.84	1.00	0.61	0.76	0.80
	$\overline{P}_{AD}$	0.81	0.96	0.65	0.78	0.73	0.97	0.83	0.80
	$\overline{D}$	**1.00**	1.00	1.00	**1.00**	1.00	1.00	**1.00**	**1.00**
	$\overline{D}^*$	**1.00**	1.00	1.00	**1.00**	1.00	1.00	**1.00**	**1.00**