Logion：面向希腊语文献学的机器学习

Charlie Cowen-Breen [*] **Creston Brooks** [†] **Johannes Haubold** [‡] **Barbara Graziosi** [‡]

arXiv:2305.01099v1 [cs.CL] 1 May 2023

摘要

本文介绍了利用机器学习解决希腊语文献学中各种问题的方法。我们在迄今用于此目的的最大前现代希腊语数据集上训练 BERT 模型后，识别并纠正了抄写员在文本传承过程中造成的、此前未被发现的错误；据我们所知，这是首次成功通过机器学习识别此类错误。此外，我们展示了该模型填补前现代手稿因材料劣化而造成的缺文的能力，并将模型的表现与领域专家的表现进行比较。我们发现，向领域专家提供模型建议以供启发时，效果最佳。着眼于这种人机协作，我们探究了模型的可解释性，并发现某些注意力头似乎能够编码前现代希腊语的特定语法特征。

1 引言

前现代希腊文本保存在手稿中，而这些手稿往往存在因材料（如纸莎草纸、羊皮纸、纸）的劣化以及抄写员引入的错误所造成的缺漏。对于一些前现代文本，抄写员在漫长的文本传承过程中，抄录早期范本中留存下来的内容，而这些范本如今已经失传。例如，荷马的《伊利亚特》成书于公元前8世纪，并通过数个世纪间接续进行的手抄传抄流传至今：现存最早的纸莎草纸残片可追溯至公元前3世纪，而整部诗篇最早的手稿则可追溯至公元10世纪。为了编纂前现代希腊文本的版本，语文学家会识别文本传承过程中逐渐累积的抄写错误，尝试对其进行校勘，并填补由材料劣化造成的缺漏。

在一项相关研究中，Assael 等人训练了一个多任务的基于 Transformer 的模型，用于确定古希腊铭文的年代、地点并填补缺漏 [1]。铭文呈现的是原始文本，而古代流传至今的大部分内容，都是通过从早期范本进行长期手抄传承而保存下来的。因此，Assael 等人专注于由物理损坏造成的缺漏，而不处理抄写错误。

这里概述的方法旨在服务于研究所有经由手稿传承保存下来的希腊文文本的语文学家。具体而言，我们介绍 Logion 背后的智力工作。Logion 是一个帮助语文学家开展工作的框架。[1]这个名字在希腊语中意为“神谕”，我们选择它是为了强调解读机器生成结果的必要性。在一篇概念验证论文中，我们已经使用 Logion 找到了前现代希腊文本中先前未被发现的错误 [2]。在本文中，我们描述并研究了得出这些结果所采用的方法。更广泛地说，我们表明，与语文学研究相关的若干任务适合使用上下文语言模型。

1.1 论文结构

在 第 1 节中，我们概述了前现代希腊文 BERT 模型的训练流程 [3], , 该模型使用的数据集据我们所知是迄今为止用于此目的的最大数据集。

在第 2 节中，我们陈述了检测抄写错误的问题，并概述了我们在[2]中所使用的方法，以发现米海尔·普塞洛斯这位 11 世纪拜占庭作者作品中此前未被检测到的错误。为验证该方法，我们随机生成模拟抄写员所犯错误的错误，并研究模型发现这些错误的有效性。

在第 3 节中，我们随机生成了人工缺文，并比较领域专家的建议与模型为填补这些缺文所给出的建议

[*] Department of Pure Mathematics and Mathematical Statistics,剑桥大学
[†] Department of Computer Science，普林斯顿大学
[‡] Department of Classics，普林斯顿大学

[1]https://github.com/charliecb/Logion.git。

缺失部分。当领域专家能够利用模型的建议来指导自己的判断时，缺失部分的填补准确率最高，这也为希腊语文献学领域未来的人机协作提供了动力。

基于此类合作，第 4 节对 BERT 模型的可解释性进行了初步研究，分析了似乎能够编码前现代希腊语中特定语法特征的注意力头。

在第 5 节中，我们概述了未来的工作步骤，以探索错误检测方法，并更广泛地开发支持语文学家编辑前现代希腊文本的方法。

1.2 训练流程

我们最初使用一个包含640万个前现代希腊语单词的数据集训练了一个BERT 模型，该数据集由 Pranaydeep Singh 慷慨提供。这是 [2] 中使用的基础模型。Singh等人在训练 [4] 中介绍的古代和中世纪希腊语 BERT 模型期间，从 Perseus Digital Library 和 First1KGreek 语料库等开源数据库中汇集了这些数据。随后，我们汇集了一个规模大得多的数据集，约包含7000万个单词。[2] 我们将这些数据按照90:10的比例划分为训练集和测试集，并使用两块 NVIDIA A100 GPU 训练 BERT 模型，共进行200个训练轮次，直至验证损失趋于稳定。为了准备分词输入，我们尽可能增加每个输入中的标点符号分隔文本量，上限为512个输入词元。我们采用了16的批大小和0.15的掩码比例。[3]

为了评估Singh等人的预训练对现代希腊语的影响，我们训练了两个模型，一个采用随机初始化，另一个则从Singh等人的古希腊语BERT [4]初始化。两次训练中，我们都使用了Singh等人为现代希腊语子词创建的分词器，因为他们自己微调过一个现代希腊语BERT。我们发现，两种训练最终都收敛到相同的在少量训练轮次后，验证损失表明，在现代希腊语上进行预训练并未带来明显收益。未来的模型使用针对古希腊语优化的分词器后，效果可能会更好：参见第 5 节。

2 错误检测

前现代文本经由复杂的文本传承过程流传至今，有时会因此产生抄写错误。这些错误包括简化难懂的表达、遗漏内容、修改古代拼写以符合后来的发音，以及将某个单词误认为另一个发音、形状或功能相似的单词，等等。传统上，发现此类错误的唯一方法，是缓慢而仔细地阅读手稿。本节将展示机器学习如何帮助检测和校勘抄写错误。列出一份潜在抄写错误的“候选清单”，可以显著提升语文学家的工作速度和准确率。Logion 已经帮助发现了一些此前未被研究同一批文本的语文学家注意到的错误 [2]。

2.1 相关文献

在英语和其他现代语言中，先前关于文本错误检测的研究通常着重于拼写和语法检查 [6] [7] [8], ，而抄写员引入的文本错误往往更为复杂（例如，图2）。因此，检测抄写错误与分布外检测更加接近：这项任务旨在辨别样本——在我们的案例中是单词——是否可能由给定的分布——在我们的案例中是作者的全部作品——生成，还是属于分布外，即传抄过程中产生错误的结果。 [9] 提出使用似然比来确定分布外样本，这是一种我们略作修改的指标。[4] 有时，错误检测会由语文学专家验证；有时，则会由手稿加以证实。这些手稿曾在编纂第一版或后续印刷版的过程中被先前的学者搁置一旁或误读。

2.2 方法

我们提出了一种用于发现文本错误的两阶段方法。在第一阶段，我们在我们正在研究的作品上微调前现代希腊语 BERT 模型

[2]部分文本由 Thesaurus Linguae Graecae 主任慷慨提供给我们，用于微调模型，但不得进一步传播；另一些文本则由同事提供，他们乐于让我们与任何感兴趣的人分享。因此，我们可以应要求分享部分训练数据，但无法分享全部数据：目前，限制访问 Thesaurus Linguae Graecae 的许可证对我们有所约束。[3]在测试集的单词元预测任务中，模型的 Top-1 准确率为 84.4%，前五准确率为 95.2%，伪困惑度也较低，为 [5], 2.162。请注意，这些指标取决于具体的分词方式，不应与采用不同分词方式的模型进行比较。

[4][9] 我们使用似然比，在基因组学数据集的分布外检测以及图像深度生成模型方面实现最先进的性能。

Psellus 的作品。给定一个词元序列 $w_1, \ldots, w_n$，考虑一个词元 w_i，并将其上下文记为 $w_{-i} = (w_1, \ldots, w_{i-1}, w_{i+1}, \ldots, w_n)$。通过掩码语言模型（MLM）训练任务 [3]，模型学习到分布

$$p(w|w_{-i}) \tag{1}$$

词元的 w，这些词元出现在句子中的 i^{th} 位置，且周围有上下文 w_{-i}。对于由多个词元组成的单词，我们通过束搜索将 p 扩展为词元序列上的分布，具体方法见第 3 节。因此，下文中，当$(w_1, \ldots\ldots, w_k)$ 表示单词序列而非词元序列时，我们将用 $p()$表示相应的单词分布 w，该分布由表达式1经束搜索得出。

在第二步（下一节将对此进行说明）中，我们应用现有的统计理论处理学习得到的分布 p，以确定最可能包含错误的词元。

2.3 候选清单生成

给定一个语料库，目标是生成**标记**（特别可能有误的单词），为领域专家提供一份潜在错误候选清单。如果某个单词满足根据下文定义的指标设定的特定条件，就会被标记。

2.3.1 指标

我们提出了三个用于标记潜在错误的指标。未来，其他指标或许能进一步提高错误检测的准确率。[5] 话虽如此，如命题所示，这里介绍的指标具有可解释性这一优点1。

1. 给定单词 w_i 及其上下文 w_{-i}，单词 i 的 **偶然性** 定义为

   $$p(w_i|w_{-i})$$

   也就是说，由模型确定该单词在给定上下文中存在的概率。

2. 模型在单词 i 处的**置信度** 定义为

   $$\max_{\text{word } w} p(w|w_{-i})$$

   也就是说，由模型确定的在位置 i周围的给定上下文中，排名最高的建议替换词的概率。

3. 在词元 i 处的**抄写距离** 定义为

   $$d\left(w_i, \quad \operatorname*{argmax}_{\text{word } w} p(w|w_{-i})\right)$$

   其中，$d(x, y)$ 表示字符串 x 和 y之间的改进的莱文斯坦距离，并考虑了抄写员可能会犯的错误。[6]

2.3.2 罕见词

虽然偶然性较低似乎是最直观的错误指标，但我们发现，另外两个指标有助于避免假阳性。如果只考虑偶然性，那么真实但罕见的单词就会被错误地标记为错误。[7] 此外，抄写错误有时在字形或语音上与正确文本相似。因此，与仅随机猜测相比，通过标记偶然性较低、且在读音或字形上与高置信度模型建议相近的文本片段，我们能够大幅提高错误检测的准确率，正如第 2.5.3 小节所示。

2.3.3 组合指标

根据需要，可以通过多种方式组合指标来生成错误标记。下文将介绍两种排序方案，它们似乎能有效找出真实的抄写错误，或为测试我们的方法是否有效而引入的人工错误。

机会度-置信度比排序

作为衡量每个单词出错可能性的指标，我们提出如下量：

$$\frac{\textbf{chance}}{\textbf{confidence}} \quad \Big| \quad \textbf{distance} \leq k \tag{2}$$

其中，右侧表示用于计算偶然性和置信度的分布 $p(\cdot|w_{-i})$ 应进一步以

[5] 这些指标当然并非人类语文学家判定某个单词可疑的唯一依据，但目前它们仍是有用的工具，[2]便证明了这一点。未来，我们预计，更多端到端的方法（例如直接针对错误检测进行训练），以及纳入更多指标的回归，将会优于此处展示的方法。

[6] 具体而言，带下标点的元音之间的替换ι，υ，η 比其他元音组合之间的替换受到的惩罚更小，这一选择基于 [10] 中列出的显而易见的错误。这种编辑距离确实是一种度量 [11]。[7]

[7] 这是因为偶然性只考虑单词 w_i 在上下文 w_{-i}中的绝对概率，而不考虑与合理的替代项相比时的相对概率。机会-置信度比可以得出这样的相对概率，我们将在下一节介绍它。

事件 $d(\cdot, w_i) \leq k$，其中 k 为某个固定值。实际上，这将建议限制为与 w_i 的抄写距离不超过 k 的那些项。

形式上，设

$$\mathcal{W}_k(w_i) = \{w : d(w, w_i) \leq k\}$$

给定一个句子 $s = (w_1, \ldots, w_n)$，我们将表达式 2 记为

$$\rho_i(s) := \frac{p(w_i|w_{-i})}{\max_{w \in \mathcal{W}_k(w_i)} p(w|w_{-i})}$$

并将其称为位置 i 处的**机会-置信度比**。

机会-置信度比的一种自然动机如下：假设我们只能更改一个句子中的一个字符，并希望以此尽可能贴近某位给定作者写下的内容。那么，我们应当更改的字符，恰好是会使受影响单词的机会-置信度比最小的那个字符。以下命题对此作了形式化表述，我们将在附录中证明该命题。[8]

命题1。（机会-置信度比与句子相对概率之间的对应关系）设 $p(s)$ 是句子上的一个联合分布 s，且满足贝叶斯定理。给定一个句子 s，假设

$$s^* = \operatorname*{argmax}_{s' \in \mathcal{W}_1(s)} p(s')$$

那么 $s^* = s$ 当且仅当 $\rho_i(s) > 1$ 对所有 i 都成立。此外，若 $s^* \neq s$，且 i^* 是 s^* 与 s 不同之处的词索引，则

$$i^* = \operatorname*{argmin}_i \rho_i(s)$$

此外，s^* 是通过将 w_{i^*}替换为 i^* 处受限于 $\mathcal{W}_1(w_{i^*})$的模型的首选建议而获得的。

换言之，该命题指出，假设联合概率分布存在，[9] 机会-置信度比表示单字符修改 s 模型认为作者最有可能写下的内容。[10] 这促使我们按偶然性-置信度比值对单词进行排序 $\rho_i(s)$，以检测可能的错误。在第 2.5 小节中，我们人为生成错误，并发现索引为

$$\operatorname*{argmin}_i \rho_i(s)$$

确实有 90% 的概率包含错误，这表明此类排名对于检测人为生成的错误十分有效（见 表 1和 图 3）。此外，在这类实例中的 98% 里，错误词 w_i 处的模型首选建议 $\operatorname{argmax}_{w \in W_1(w_i)} p(w|w_{-i})$，能够恢复正确的真实词。

ρ_i 的另一种解释是：假设关于 w的先验分布在 $\mathcal{W}_k$ 上均匀分布、在其他地方为零，它就是似然比统计量。从这个意义上说，机会-置信度比建立在[9] [12], 之上，而后者利用似然比统计量成功检测了分布外样本。这种解释相当于将位置 i 处的真实词视为一个未知参数，其取值决定了周围单词的条件分布 $p(w_{-i}|\cdot)$。在这种情况下——再次假设抄写员只会犯抄写距离不超过 k的错误——我们可以将错误检测表述为假设检验问题

H_0 : 单词 w_i 书写正确。H_1 : 原单词已被修改，位于 $\mathcal{W}_k \setminus \{w_i\}$。

在图1(i)中，我们绘制了假设 H_0和 H_1下似然比统计量的分布。每个假设下的分布各不相同，因此可以通过似然比检验进行正式的假设检验。

阈值设定

在某些应用中，分别为每项指标设定阈值，可以更灵活地生成错误候选清单。在 [2], 中，结果是通过筛选置信度至少为 50%、抄写距离至多为 3 的单词生成的，并将剩余单词按偶然性递增的顺序排列。在

[8]对于 $k > 1$ 字符的修改，该命题可推广为相应的陈述，但假设改为：s' 属于这样一个句子集合，其中的每个句子与s 仅有一个单词不同，且最多相差 k 个字符。[9]

在现有文学中，似乎没有关于这一点的理论保证。在 [5], 作者试图通过证明 BERT 是一个马尔可夫随机场，直接构造一个联合分布 $p(s)$，但这篇论文后来已被撤回。

[10]尽管如此，仍须谨慎，不能据此断定 s^* 就是作者最初的表述。抄写错误可能更倾向于文本中较易读的异文，因此可能会增加 p。我们将在第 5 节进一步探讨这一影响。

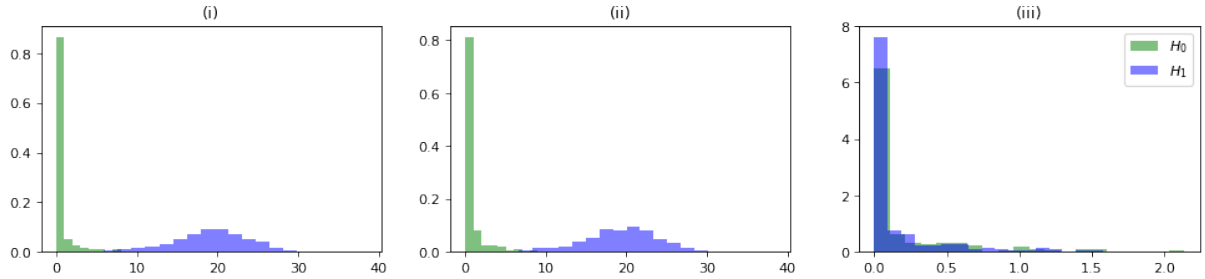

图1：假设下指标的分布 H_0 和 H_1。图中所示的指标为：（i）表达式 2中的机会-置信度比，（ii）偶然性，（iii）置信度。每个横轴均为 $-\log T$，其中 T是与该图对应的指标。这里，H_1 根据第 2.5 小节所述的人工错误生成方案建模；为高效生成样本，我们仅限于单词替换。每张图包含大约 500,000 个来自 H_0 的样本，以及 1,000 个来自 H_1的样本。

（2）的记法中，此方案使用以下量对单词进行排序

$$\textbf{偶然性} \,\Big|\, \textbf{距离} \;\le 3\; \textbf{置信度} \;\ge 1/2$$

采用该方案得到的一些标记见 第 2.4 小节。将置信度阈值设为 50% 很方便，因为它满足这样一个性质：在通过阈值筛选的单词中，按抄写距离进行阈值设定前后，模型给出的首选建议相同。

阈值决定模型用于预测讹误词时的精确率和召回率。对于希望找到一份错误词的有力候选列表的应用（即期望较高的精确率），可以将置信度阈值设得较高（例如 90%），并将偶然性和抄写距离阈值设得较低（例如分别为 10^{-6} 和 2）。对于希望找到更多讹误词、且能够接受筛选较弱候选项的应用（即期望较高的召回率），可以将置信度阈值设得较低（例如 50%），并将偶然性和抄写距离阈值设得较高（例如分别为 10^{-4} 和 4）。

2.4 具有语文学意义的结果

本文所列指标成功识别出此前未被发现的错误，包括手稿中的抄写错误、印刷版中的排印错误，以及由数字化造成的错误。这些发现经过语文学同行评审，并已获准发表于古典研究学会的研究期刊 *TAPA*。

在 [2], 中，我们通过概念验证阶段的示例展示此处介绍的方法如何通过识别并有时解决若干不同的语文学问题，增进我们对前现代希腊文本的认识。详细示例及进一步讨论请见 [2]。下面我们提供一个示例，说明可能检测到的一类错误（在本例中，是对手稿的误读，而非手稿本身存在的实际错误）。

在 Psellus 的《历史概要》第 81.89–90 行中，Aerts 的版本记载：

> οὗτος δὶς βασιλεύσας ηὔχετο καὶ τρὶ ϛκαὶ τετράκις· ἦ δὲ γάρ, φησι, μετὰ νέ-ϕοϛὁ ἥλιος. 此人曾两度担任国王，祈求第三任和第四任。†而且确实†，他说：“云后见日。”

在对置信度和抄写距离设置阈值时，词元 δε 是测试集中偶然性最低的词元之一。图中所示的算法输出图2以及随后对本版本所依据的手稿进行的检查，使我们意识到，手稿实际记载的是“ἡδὺ”，而不是“ἦ δὲ”。现在可以将句子翻译如下：“这个人曾两度为王，祈求再任第三、第四个任期。因为他说：‘雨过天晴，令人心旷神怡。’”

2.5 有效性

本节中，我们通过引入人工生成的错误，研究所提出的错误检测方法的有效性。

2.5.1 人工生成的错误

由于数据生成机制本身复杂且难以复现，人工生成抄写错误颇为困难。

ουτος δις βασιλευσας ηυχετο και τρις και τετρακις · η δε γαρ, φησι, μετα νεφος ο ηλιος.

ηκε	1.3%
ηδει	0.06%
ηδυ	0.05%

图2：导致发现单词中抄写错误的算法输出 η δε。 上方一行： 单词根据其似然比进行颜色编码，如 公式 2所示；单词 δε被算法标记出来，因为在其给定的上下文中（周围的512个词元，此处未全部展示），它的机会-置信度比是所有单词中最低的。 上方一行下方： 算法生成的建议，根据其偶然性进行颜色编码。在每种情况下，算法都建议通过删除 δε前面的空格来合并两个单词。第三条建议 ηδυ实际上见于相关手稿中，因此必定是作者最初写下的内容 [2]。Logion 给出的较低概率反映了抄写错误的复杂性。有些错误很简单，包括我们人为生成的那些错误；另一些则更难校勘，包括这一处错误。

这种机制本身复杂且难以复现。这类错误往往取决于具体的抄写员、他们工作的上下文，以及他们对所抄内容的兴趣：抄写错误可能千差万别，也相当复杂。

尽管如此，有些错误相当平常，例如，由伊塔主义等现象引起的发音变化可能导致拼写错误。[11] 为了进行这项模拟，我们生成此类抄写错误：在每个段落中，将随机选取的一个字符替换为另一个随机字符，使修改后的单词出现在作者所使用的单词词典中，且出现次数至少为 10 次。如果修改后的单词不符合这一条件，我们就继续替换字符，直到满足条件为止。这一过程确保我们生成的错误无法通过简单的词典检查来发现。

2.5.2 结果

在每个段落中，我们按偶然性-置信度比率对单词进行排序（公式 2），如小节 2.3.3所述，采用 $k = 1$。在随机生成的 615 个实例中，错误词有 556 次排名第一，Top-1 准确率为 90.5%。在错误词排名第一的实例中，真实词有 98.1% 的情况下是错误词最可能的替换词。结果总结见表 1。

2.5.3 消融研究

为证明综合考虑第 2.3 节介绍的全部三项指标能够提高检测人工错误的准确率，我们在此将按机会-置信度比排序，与两种未涵盖全部三项指标的替代排序方案进行比较：

[11]伊塔主义一词描述的是不同元音和双元音之间的混淆，这些元音和双元音后来都被读作 /i/。

（1）仅限抄写距离为 1 时按置信度排序；（2）按仅随机猜测排序。

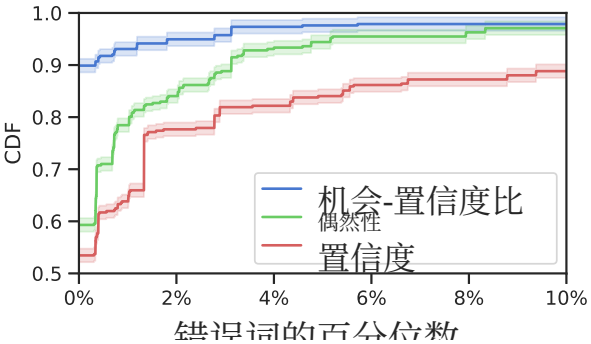

图 3：在每个段落中随机选取一个单词插入人工错误（平均每段 230 个单词）。计算段落中每个单词的指标（机会-置信度比、仅随机猜测、仅置信度），并测量错误词在段落内按指标排序后的百分位数。此图展示了这些百分位数的累积分布函数。按机会-置信度比排序时，90.5% 的单词在其段落中排名第一（即 0th 百分位数）；按偶然性排序时，59.7% 的单词排名第一，这与 表 1一致。误差带通过 DKW不等式计算，覆盖概率为99% 。

图 3展示了按所提出的排序方案排序后，人工植入错误的单词的分布。每种排序方案的分布都明显左偏：按任一指标排序时，超过 85% 的错误词位于单词排序结果的后 10%。这表明，所提出的每种排序方法都与人工错误相关。

然而，在不到 60% 的情况下，按照任一替代排序方案排在末位的单词实际上是错误的（见 表 1）。因此，我们得出结论，仅按偶然性或置信度对单词进行排序，在识别人工制品方面效果较差。

准确率	机会-置信度比	仅随机猜测	仅置信度
前1名	90.5%	59.7%	54.2%
前5名	95.9%	88.2%	81.1%
前10名	97.6%	93.1%	83.5%

表 1：根据不同的指标组合方案，检测 230 个单词中单个人工错误的准确率。采用机会-置信度比能够取得最佳性能，尽管偶然性也是一种可行的指标。由于该任务旨在生成潜在错误的候选列表以供审查（而且领域专家通常能够迅速核实被标记的单词是否确实存在错误），因此，前十准确率在此处是一项重要指标。

比按机会-置信度比排序更容易产生错误。

3 缺文补全

3.1 背景

除了在文本传承过程中累积的抄写错误，前现代希腊文本还经常因机械损坏或其他形式的劣化而出现缺文。制造类似真实缺文的人工缺文，比引入人工错误更为直接，因为真实缺文实际上是随机缺失的文本片段。

机器学习已用于推测缺文中可能适合填入什么文本这一任务。Bamman 和 Burns 使用大型拉丁文本语料库训练了一个 BERT 模型，并评估了该模型在文本补全方面的准确率 citeburns。他们遮蔽单个词元，因此可以直接利用 BERT 的掩码语言模型目标，但这限制了该方法在实际语文学研究中的实用性：面对真实缺文时，即便在许多情况下能够以一定准确度估算缺失字符的数量，也无法得知缺失文本片段由多少个词元构成。其他作者则提出了能够根据缺文大小调整的缺文补全模型。

针对古希腊铭文，Assael 等人开发了 Pythia-Word，这是一种字符级序列到序列方法，结合字符和单词表征，通过字符级束搜索补全缺失字符 [13]。在之后开发的模型 Ithaca 中，输入文本被编码为字符、单词和位置嵌入序列，并输入采用 Transformer 架构的模型，该模型设有三个任务头，专门用于区域归属判定、年代归属判定和缺文补全[1]。Ithaca 特别注重人机协作，并设计了一种评估模式，我们也采用了这种模式，但做了一处调整：Assael 等人将机器生成的结果与真实值进行比较；随后，他们将古典学研究生进行缺文补全的尝试与真实值进行比较；最后，他们评估通过人机协作获得的结果。下文中，我们采用相同的方法，但在我们的研究中，人类的能力由一位专业语文学家来代表；这位语文学家此前已用传统方法研究过相关文本 [14]。我们仍然发现，模型的建议能够提升人的表现。

3.1.1 模型设置

根据构成文本缺口的缺失词元数量，可以使用针对掩码的迭代式束搜索启发式方法生成预测。虽然推理时并不知道缺口中缺失词元的数量，但通常可以根据物理缺口的大小推断缺失字符的数量；下文将尝试仅根据缺失字符的数量和上下文来估算缺失词元的数量。考虑一个人工生成的缺口，即真实值已知的缺口。令 S_n 为所有可能的、由 n 个字符组成的跨度的集合。令 toks(s) 表示构成跨度 s 的词元数量，令 gap_toks 表示构成缺口所覆盖的真实文本的词元数量（推理时未知的量）。

标准答案	Οὐκ οἶδα πότερον
分词结果	[ουκ, οι, -δα, ποτε, -ρον]
插入空缺	Οὐκ???? πότερον

空缺_标记 $= 2$

给定围绕一个由 c 字符组成的空缺的上下文 n，我们寻找能够最大化 $s \in S_n$ 的对象

$$\mathbb{P}(s|c) = \mathbb{P}(s|\mathsf{gap_toks} = \mathsf{toks}(s), c) \cdot \mathbb{P}(\mathsf{gap_toks} = \mathsf{toks}(s)|c)$$

我们将 $\mathbb{P}(s|\mathsf{gap_toks} = \mathsf{toks}(s)\ c)$ 识别为恰好 $p(s|c)$，其可通过束搜索计算得到

搜索 toks(s) 词元。[12]

要计算 $\mathbb{P}(s|c)$，现在只需估计$\mathbb{P}$（间隔_toks $=$ toks(s)$|c$）。为此，我们将全连接神经网络（FCNN）作为间隔中字符数 n 以及每次（$n//2+2$）束搜索所得前十个概率的函数进行训练，并将这些作为输入。输入维度为 78，具体方法是将维度为（7×10）的概率矩阵展平并连接起来，再加上对字符数进行编码的独热向量。该网络包含一个隐藏层，并使用交叉熵损失进行训练；它有 7 个输出节点，分别代表间隔_toks 的潜在取值。

训练数据的生成方式是：在 Psellus 训练集中随机遮蔽连续的完整单词，直到被遮蔽的字符数超过随机选定的 3 至 10 个字符。我们不将标点计入字符数，尽管某些间隔包含标点；例如，ἐστηκότα· αἱ 是一个包含 10 个字符的间隔。对于每个经过人工遮蔽的跨度 s，我们生成每次束搜索的前十个概率，并使用 toks(s) 作为训练标签。

由此得到的 FCNN 输出概率分布 $\mathbb{P}$（间隙_词元 $=k|c$），针对的是 $k\in[1,7]$。因此，在实践中，要对长度为n 的间隙进行预测，我们可以执行 $(n//2+2)$ 次束搜索，生成概率分布$\mathbb{P}$（间隙_词元 $=k|c$），并返回长度恰为 n、且 $\mathbb{P}(s|c)$ 最大的跨度 s。

3.1.2 连续的掩码词元

在上述推理流程中，可能会执行最多包含 7 个连续掩码词元的束搜索。但包含许多连续掩码词元的输入与 BERT 的 MLM 训练任务并不十分匹配：例如，当 15% 的输入词元被掩码时，在一次训练迭代中，任意五个连续词元同时被掩码的偶然性约为 10^{-5}。

为缩小训练与推理之间的差异，我们对原始 BERT 模型进行微调，使其能够处理特定长度的连续掩码词元序列：分别为包含 2、3、4 和 5 个掩码词元的序列训练四个新模型。在束搜索中进行推理时，根据输入中连续掩码词元的数量选择合适的模型，

[12]词元的实际数量 gap_toks 显然有上界 n，但在实践中，超过 99% 的情况下，该数量不超过 $n/2$。因此，为降低计算成本，我们将掩码词元的数量限制在 $[1,n//2+2]$ 范围内进行束搜索。

输入中连续掩码词元的数量为≥5 时，使用 5 掩码模型。

3.2 实验

我们进行了一项实验，以比较模型填补普塞洛斯验证集中空缺的准确率与领域专家的表现。我们选择 Johannes Haubold 担任专家，因为他此前已使用传统方法校勘过 Psellus 的文本 [14]。我们通过随机遮蔽连续的完整单词片段，创建了 48 个长度不一（3 至 10 个字符）的空缺示例。[13] 我们测试专家的空缺数量受到一项目标的限制：模拟真实的语文学工作环境。在这种环境中，领域专家通常会花大量时间考量校勘方案。

每个示例平均包含 234 个单词，为处理空缺提供了充足的上下文。此外，参与者还可以看到每个示例所选文本的标题，并可搜索整部希腊文学语料库，但不包括该示例的来源文本。

对于 48 个空缺中的每一个，专家在看到模型前十预测及其可能性之前，先给出了“专家预先查看前预测”。随后，他可以选择坚持原先的预测，也可以给出新的预测；我们将后者称为“专家查看后预测”。只有与标准答案完全一致的预测才被视为正确。Haubold 总共花了 7 小时填补这 48 个空缺；对于文献学家而言，这被认为是较快的工作速度。

3.2.1 结果与讨论

	准确率
模型（前1名）	58.3%
专家预览	64.6%
模型（前2名）	70.8%
专家查看后	75.0%
模型（前10名）	85.4%

研究发现，专家的初始预测比模型的首选更准确。然而，允许模型提供多条建议后，其准确率大幅提升，而专家也能借助模型排名前十的建议，改进自己的初始猜测。由于生成这些建议的目的是启发语文学家，

[13]尽管真实间隙并不遵循词边界，但考虑到子词分词，为方便起见，我们屏蔽了整个单词。

而非取代语文学家，因此我们认为前十准确率是衡量成功与否的良好指标。

4 可解释性

与语文学家不同，他们能够为工作中的每个步骤提供有理有据的论证，而 BERT 产生输出的过程却不易理解。本节将展示，前现代希腊语学者（尤其是 Denniston 对希腊语虚词的研究）的语法观察 [15]与 BERT 架构中的某些注意力机制之间存在对应关系。具体而言，我们发现，若将某些注意力头视为无需训练的分类器，它们在特定语法任务上的表现出奇地好。

4.1 背景与文献综述

研究发现，BERT 模型在基于语言的任务上表现出色，但究竟是什么使其能够做到这一点，目前尚不完全清楚。Transformer 并非通过单一且易于理解的步骤处理语言信息，而是通过一系列堆叠的注意力头 [3],组件来处理信息，这些组件显然难以解释。

文献 [16], 提出，若将 BERT 的某些注意力头视为用于特定英语语法任务的无需训练的分类器，它们的表现似乎出奇地好。这促使研究者在其他几种现代语言中开展类似研究，例如 [17] [18]。

在本节中，我们尝试针对前现代希腊语 BERT 模型复现这些结果。

4.1.1 背景：注意力

关于注意力的完整解释，请参阅 [19]。就我们的研究目的而言，我们注意到，BERT 每层包含12个注意力头，共有12层，因此总计包含144个注意力头。

给定一个输入序列，每个注意力头都会计算一组**注意力权重**，这些权重取决于输入序列，记作 α_{ij}，每个输入词元 i 和输出词元 j各对应一个。注意力权重的取值范围为 $[0, 1]$, , 且每个词元对所有输出词元的注意力之和为 1。每个注意力头通过对值向量 v_i（即输入的线性变换）按注意力加权求和来生成输出：

$$o_i = \sum_{j=1}^{n} \alpha_{ij} v_j \tag{3}$$

给定一个指定的输入序列和注意力头，如果词元 i **最关注**词元 j，则称

$$j = \underset{j'}{\operatorname{argmax}}\, \alpha_{ij'}$$

随后，我们可以按照以下方式将这一界定扩展到由多个子词词元组成的单词：对于来自多词元单词的注意力，我们对注意力权重取平均值；对于指向多词元单词的注意力，我们则对注意力权重求和。正如 [16] 的作者所指出的，这些转换保留了这样一个性质：来自每个单词的注意力权重之和为 1。

[16] 的作者还观察到，对于 BERT 中的特定注意力头和英语中的特定依存关系，单词 i 对单词 j 的依存关系与单词 i 最关注单词 j 相关。由于训练目标中没有任何部分明确鼓励注意力头专门学习单个语法特征，因此这种专门化现象或许令人惊讶。

4.2 案例研究： μέν/δέ 注意力头

作为第一个案例研究，我们聚焦于结构μέν...δέ...，它由两个不发生屈折变化的小品词组成，因此便于研究。在前现代希腊语中，这些小品词标记平行分句，相当于英语中的“on the one hand……, on the other hand……”。在少数罕见情况下，μέν 单独出现，充当副词，意为“因此”或“所以”。更罕见的是， μέν 的应答词并非 δέ，而是其他小品词，例如 ἔπειτα或 μήν。

我们假设，网络中索引为 5-0 的注意力头可充当一种无需训练的分类器，用于识别应答 μέν 的单词。如果没有单词作应答，且 μέν 用作副词，它往往会指向自身。为检验这一假设，我们首先考察 5-0 注意力头所关注的 μέν 的分布。在 99% 的情况下， μέν 关注的是 δέ[14] 或自身。

为进行进一步分析，我们区分两类情况：一类是存在 δέ ，它似乎回应了前面的 μέν；另一类是不存在 δέ。在后一种情况下， μέν 有92.2%的时候照应自身；有5.8%的时候照应除 δέ之外的合适小品词，例如ἔπειτα或 μήν。

ἔπειτα或 εἶτα 作为 δέ的替代词

在 μέν 之后的对应分句不包含 δέ 的情况下，通常会出现 ἔπειτα or

[14]这里我们还包括省略形式 δ'。

εἶτα，而最常见的类型是：当 μέν 从句包含 πρῶτον，意为“首先”（Denniston[15]，第376页）。如果第一个从句包含 πρῶτονμέν，意为“首先，一方面”，第二个从句很可能包含ἔπειτα δέ，意为“然后，另一方面”，此处的 δέ 变得多余，很容易被省略。

与这些事实一致， μέν 在前现代希腊语 BERT 中会关注 ἔπειτα or εἶτα，这涵盖了测试样本中 πρῶτον μέν 出现的全部五种情况。

μήν 作为 δέ的替代词

在没有 δέ 跟在 μέν之后的实例中，有 3 次 μέν 转而关注 μήν，这再次符合希腊语虚词用法的标准描述。事实上， μήν 用于与前面的μέν [15]相呼应，尤其常与否定词 οὐ（“不”）连用。5-0 注意力头与这一语法观察相符：每当 μέν关注 μήν时， οὐ μήν 都会出现。

4.3 古希腊语依存关系与 BERT

在本节中，我们将展示网络的特定层如何充当分类器，以执行若干特定的语法任务。我们沿用 [16]中的设置，并为考虑到希腊语是一种屈折语这一事实，从珀耳修斯数字图书馆维护的古希腊语和拉丁语依存树库中提取依存关系 [20]。现在，我们将检验所有注意力头再现特定依存关系的准确率（表 2），并以古典时期两位研究充分的作者柏拉图和修昔底德为例。

我们发现，虽然没有任何注意力头在 [20], 中使用的依存关系分类任务上表现特别出色，但若聚焦于某些词类，表 2所示的某些层似乎表现更好。与 [16], 一样，我们使用固定偏移基线作为对照。[15]

如 表 2所示，与固定偏移量基线相比，若干注意力头在所示语法任务上表现出色。虽然注意力头 5-0 似乎专门处理 μέν/δέ 关系，而注意力头 1-3 似乎专门负责将辅助虚词与支配动词进行匹配，注意力头 9-1 和 9-2 则似乎专门处理一般依存关系，例如定语结构中的依存关系。注意力头在语法任务上的表现优于固定偏移基线的一个可能原因是，前现代希腊语的语序具有高度灵活性。有关各个固定偏移基线下注意力头的分析，请参见图4。

5 未来工作

未来工作的一个重要方向，是开发一款由领域专家采用并用于辅助其工作的应用程序。对于任意文本，这款应用程序都能够生成疑似错误清单供人审阅，并就如何复原缺失或受损的文本部分提出建议。该领域未来的研究方向包括开发高效且符合语言学原理的子词分词方案、提升固定字符间隙填充的表现，以及能够在推理时选择将数据集中的某些部分纳入或排除在考量之外；当人们希望生成符合特定作者作品风格的间隙填充建议时，这一点尤其重要。为实现后一目标，一个有前景的架构是 DEMix[21], ，它支持在推理时进行动态专家混合。

未来研究的另一个思路，也是使抄写错误检测有别于传统错误检测的一个方面，是将抄写修改视为扩散过程。由于抄写错误往往受上下文驱动，抄写员修改后的文本在评估时，反而可能比原始文本具有更高的概率。[16] 因此，依照这一观点，文本会随时间推移，作为一个扩散过程不断演变，其转移核由 p 推导而来（例如，可以根据条件分布$p(w_i|w_{-i})$，通过吉布斯采样对文本的演变轨迹进行建模）。扩散模型 [23] 旨在从扩散后的数据中恢复原始数据，因此，将此类模型用于从经抄写修改的文本中恢复原始文本，或许会很有成效。

虽然其本身并非扩散模型，但 ELECTRA[24] 是一种很有前景的架构，可用于此类未来研究。

[15]对于一个注意力头，我们按以下方式定义其固定偏移基线：对于每个固定整数 $-10 \leq k \leq 10$，我们考察该注意力头在关注距离为 k 个位置的单词这一任务上的表现。注意力头的固定偏移基线定义为其在上述任务中对所有 k所取得的最高表现。使用这一基线表现来检验以下假设：给定的注意力头并非真正学会了语法，而只是学会关注附近的单词。

[16]在语文学中，这一原则称为 难读优先原则。由于“通常的倾向是简化、使之平庸化，并消除陌生的单词或结构”，通常应将较难的读法（即 难读优先原则）视为真实读法 [22]。

语法任务	准确率	专用注意力头	固定偏移基线
μέν ⟶ 应答小品词，例如 δέ	95%	5-0	8% (1)
感叹词 ⟶ 呼格	99%	9-1	30% (1)
冠词 ⟶ 带冠词的不定式	90%	9-1	30% (1)
对应小品词 ⟶ 愿望语气动词	94%	1-3	43%（1）
定语冠词 ⟶ 名词	86%	9-2	54%（1）
定语冠词 ⟶ 名词性形容词	86%	9-2	54%（1）
定语形容词 ⟶ 名词	86%	9-2	54% (1)
作定语的属格名词 ⟶ 名词	84%	9-2	54% (1)

表 2：选定注意力头作为无需训练的分类器对古希腊语依存关系的性能。右侧列中的括号表示偏移量大小 k，该偏移量在关注相距 k 个位置的单词时可使性能最大化。

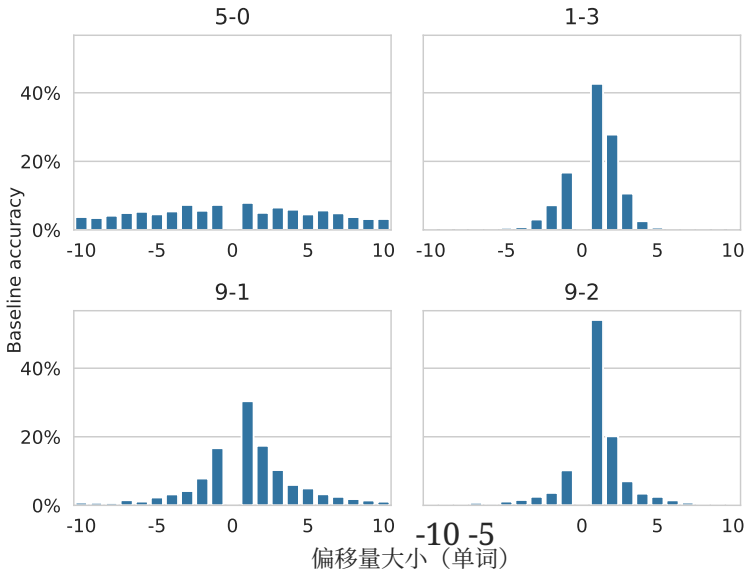

图4：比较在基准偏移量下发现的专用注意力头在表2 中不同偏移量大小下的性能。专门处理语法任务的注意力头倾向于寻找附近的单词，例如感叹词与呼格之间的关系，其最佳性能集中在邻近单词范围内（即偏移量较小）。相反，专门处理 μέν/δέ 关系的注意力头 5-0，在任何偏移量大小下的性能都相对较低，这可能是因为 μέν 与回答的 δέ 之间的距离可能非常大。

工作。ELECTRA 通过同时训练生成器和判别器来学习表征：生成器旨在使用合理的替换词元破坏输入文本，而判别器则将输入序列中的词元分类为已替换或原始词元。因此，未来的研究工作可以考虑将 ELECTRA 扩展到这样一种情形：输入词元并非仅由生成器随机替换一次，而是连续多次被替换，以模拟抄写错误生成的扩散视角。此外，如果进一步要求替换词元不仅依据上下文进行选择，还要遵循视觉和语音相似性，那么这种架构或许能够同时用于对抄写错误进行建模和检测。

6 结论

在本研究中，我们训练了一个 BERT 模型，以支持对前现代希腊文本的文献学工作。我们采用统计方法和基于机器学习的方法来识别错误，并修复由物质损坏造成的缺失。我们开展了一项实验，展示了人机协作在校勘前现代文本方面的潜力。最后，本文对 BERT 模型的可解释性进行了初步研究，揭示了特定注意力头如何编码前现代希腊语的某些语法特征。从更广泛的意义上说，本研究旨在为未来的

语文学作出贡献。语文学是一门致力于保存、阐释并向公众开放全球前现代文本档案的学科。本文所述的部分内容，对于我们未曾考察的作者和语言也具有一定的参考价值。

致谢

我们感谢 Kasia Kobalczyk、Simon Babb、Suma Bhat、David Cox、Justin Curl、Bernhard Haubold、Max Haubold、Peter Heslin、Mika Hyman、Karthik Narasimhan、Maria Pantelia、Stratis Papaioannou 和 Pranaydeep Singh 提供的宝贵交流与建议。我们也感谢的里雅斯特 Bar Libreria Knulp 的店主、员工和其他顾客。

参考文献

[1] Y. Assael, T. Sommerschield, B. Shilling-ford 等。“利用深度神经网络修复并归属古代文本”。载于：《自然》603（2022），第 280–283 页。[2] B. Graziosi、J. Haubold、C. Cowen-Breen 和 C. Brooks。“机器学习与语文学的未来：一项案例研究”。载于：*TAPA* 153（2023）。即将出版。[3] J. Devlin、M.-W. Chang、K. Lee 和 K. Toutanova。“用于语言理解的深度双向 Transformer 预训练”。载于：*2019* 年北美计算语言学协会会议论文集：人类语言技术。第 1 卷（长篇与短篇论文）。计算语言学协会，2019 年，第 4171–4186 页。[4] P. Singh、G. Rutten 和 E. Lefever。“BERT 语言建模与古代和中世纪希腊语形态分析的试点研究”。载于：第五届文化遗产、社会科学、人文学科与文学计算语言学联合 *SIGHUM* 研讨会论文集。多米尼加共和国蓬塔卡纳：计算语言学协会，2021 年，第 128–137 页。[5] A. Wang 和 K. Cho。“BERT 有一张嘴，它必须说话：作为马尔可夫随机场语言模型的 BERT”。载于：神经语言优化与评估方法研讨会论文集生成。明尼阿波利斯：计算语言学协会，2019，第 30–36 页。[6] P. Etoori、C. Manoj 和 M. Radhika。“利用深度学习为资源匮乏语言自动纠正拼写”。载于：第 56 届计算语言学协会年会论文集。学生研究研讨会。计算语言学协会，2018，第 146–152 页。[7] M. Can Ganiz 等。《土耳其语语法与拼写检查》。CSE4197 – 分析与设计文档。土耳其马尔马拉大学工程学院计算机工程系，2020。网址：http://cse.eng.marmara.edu.tr/dosya/eng/cse/documents/grad_proj/ add _3.pdf。[8] D. Naber。《基于规则的风格与语法检查器》。比勒费尔德大学技术学院毕业论文。2003。网址：http:// www . danielnaber . de / languagetool / download/style_and_grammar_checker. pdf。[9] J. Ren 等。“用于分布外检测的似然比”。载于：《神经信息处理系统进展》第 32 卷（NeurIPS 2019）。H. Wallach 等编。第 19 卷。纽约州雷德胡克：Curran Associates Inc.，2019，第 14640–14651 页。[10] D. R. Reinsch。《Michaelis Pselli Chronographia》。柏林：德格鲁伊特，2014，第 XXXVI–XXXVII 页。[11] R. A. Wagner 和 J. M. Fischer。“字符串到字符串的纠错问题”。载于：《计算语言学协会期刊》21.1（1974），第 168–173 页。[12] V. Gangal、A. Arora、A. Einolghozati 和 S. Gupta。“面向任务的对话中用于无监督域外检测的似然比与生成式分类器”。载于：第三十四届 AAAI 人工智能大会论文集（AAAI-20）。人工智能促进协会，2020，第 7764–7771 页。[13] Y. Assael、T. Sommerschield 和 J. Prag。“利用深度学习复原古代文本：希腊铭文学案例研究”。载于：2019 年自然语言实证方法会议论文集。

处理与第九届自然语言处理国际联合会议（*EMNLP-IJCNLP*）：系统演示。编者：S. Padó 和 R. Huang。计算语言学协会，2019 年，第6368–6375页。[14] J. Haubold。“关于米海尔·普塞洛斯《论哲学》的校勘猜测”。载于：《论哲学》。《拜占庭》93（2023）。即将出版。[15] J. D. Denniston 和 K. J. Dover。《希腊语小品词》。伦敦：Bristol Classical Press，2005年。[16] K. Clark 等。“BERT关注什么？对BERT注意力机制的分析”。载于：*2019*年*ACL*研讨会论文集：*BlackboxNLP*：自然语言处理神经网络的分析与解释。佛罗伦萨：计算语言学协会，2019年，第276–286页。

[17] R. Guarasci、S. Silvestri、G. De Pietro 等。“评估 BERT 学习意大利语句法的能力：关于零主语和一致现象的研究”。载于：《环境智能与人性化计算杂志》14（2023），第 289–303 页。

[18] V. Ravishankar 等人：《注意力可以反映句法结构（只要你允许它）》。载于：欧洲计算语言学协会第十六届会议论文集：主卷。计算语言学协会，2021 年，第 3031–3045 页。

[19] A. Vaswani 等人。《注意力就是你所需要的一切》。载于：第 *31* 届神经信息处理系统国际会议论文集（*NeurIPS'17*）。纽约州雷德胡克：Curran Associates Inc.，2017 年，第 6000–6010 页。

[20] D. Bamman 和 G. Crane。《古希腊语和拉丁语依存树库》。载于：文化遗产语言技术：自然语言处理理论与应用。C. Sporleder、A. van den Bosch 和 K. Zervanou 编。柏林和海德堡：Springer，2011 年，第 79–98 页。[21] S. Gururangan 等。*Demix*层：为模块化语言建模解耦领域。2021 年。arXiv：2108.05036。

[22] M. L. West。《适用于希腊语和拉丁语文本的文本批评与编校技术》。斯图加特：Teubner，1973 年，第 51 页。[23] J. Sohl-Dickstein 等。“使用非平衡热力学进行深度无监督学习”。载于：第 *32* 届国际机器学习会议论文集。F. Bach 和 D. Blei 编。第 37 卷。《机器学习研究论文集》。PMLR，2015 年，第 2256–2265 页。

[24] K. Clark 等人：《ELECTRA：作为判别器而非生成器的预训练文本编码器》。载于第八届国际学习表征会议论文集。国际学习表征会议。埃塞俄比亚亚的斯亚贝巴，2020年。网址：https://openreview.net/group?id=ICLR.cc/2020/Conference。

7 附录

命题 *1* 的证明。

$$\begin{aligned}
\max_{s'\in\mathcal{W}_1(s)} p(s') &= \max_{1\le i\le n}\max_{w\in\mathcal{W}_1(w_i)} p(w_1,\dots,w,\dots w_n)\\
&= \max_{1\le i\le n}\max_{w\in\mathcal{W}_1(w_i)} p(w|w_{-i})p(w_{-i})\\
&= \max_{1\le i\le n}\max_{w\in\mathcal{W}_1(w_i)} \frac{p(w|w_{-i})p(w_{-i})}{p(w_i|w_{-i})p(w_{-i})}p(s)\\
&= p(s)\max_{1\le i\le n}\max_{w\in\mathcal{W}_1(w_i)} \frac{p(w|w_{-i})}{p(w_i|w_{-i})}\\
&= p(s)\max_{1\le i\le n} \frac{1}{\rho_i(s)}
\end{aligned}$$

这表明，对于 $s^*\in\mathcal{W}_1(s)$，

$$p(s^*) = \max_{s'\in\mathcal{W}_1(s)} p(s')$$

当且仅当 s^* 与 s 在单词 i^* 上不同，并且

$$\rho_{i^*}(s) = \min_i \rho_i(s).$$

另一方面，我们有

$$\begin{aligned}
\rho_i(s) &= \frac{p(w_i|w_{-i})}{\max_{w\in\mathcal{W}_1(w_i)} p(w|w_{-i})}\\
&= \min_{s'\in\mathcal{W}_1(s):s' \text{ differs from } s \text{ at word } i} \frac{p(s)}{p(s')} > 1
\end{aligned}$$

当且仅当 $\forall s'$ 使得 s' 与 s 仅在单词 i上有所不同，且仅相差一个字符时，我们有$p(s) > p(s')$。如果这对所有 i都成立，那么 $s = s^*$

根据定义。如果不是，则存在某个 i，使得$p(s)$ $\leq p(s')$。在这种情况下，根据最大值的唯一性，对于满足这一条件的某个i，我们必须有 $p(s) < p(s')$。因此 $s \neq s^*$。 □