交互式可视化在解释（大型）自然语言处理模型中的作用：从数据到推理

Richard Brath
Uncharted SoftwareIncrbrath@uncharted.software

Daniel Keim康斯坦茨大学

Johannes Knittel斯图加特大学

Shimei Pan
马里兰大学，巴尔的摩县分校

Pia Sommerauer阿姆斯特丹自由大学

Hendrik StrobeltIBM剑桥研究院
hendrik.strobelt@ibm.com

arXiv:2301.04528v1 [cs.CL] 11 Jan 2023

摘要

随着所学习参数数量不断增加，现代神经语言模型的能力也日益强大。然而，解释这些复杂模型的行为仍是一个尚未解决的普遍难题。本文探讨交互式可视化在解释自然语言处理模型（可解释自然语言处理）方面所能发挥的作用。我们结合目标用户和常见的自然语言处理流水线，阐述使用可视化的动机。我们还介绍了若干应用案例，通过具体实例展示如何利用可视化进行可解释自然语言处理。最后，我们列举了这一领域大量的研究机遇。

1 动机

近年来，由超大型神经网络模型（如 BERT 和 GPT-3 ）驱动的自然语言处理系统取得了前所未有的性能。最新模型拥有数十亿个参数，训练时需要海量数据和计算资源。尽管总体而言，这些模型的表现质量很高，但在许多应用中，可解释性至关重要，例如医学诊断或偏见检测与缓解（例如，Zhang 等人，2021年；Stevenset al.，2020年）。

在这篇立场论文中，我们探讨交互式可视化在解释（大型）自然语言处理模型方面的作用。本文中的思考与建议，源自自然语言处理和可视化领域专家的深入讨论与密切合作。

我们首先区分具有不同技术背景和领域专业知识的用户群体。每个用户群体都有不同的可解释性需求，这些需求可能会影响交互式可视化工具的设计。接下来，我们讨论如何利用可视化来解释典型的自然语言处理流程，尤其是那些采用预训练大语言模型的流程，并探讨一系列问题：何时以及为何使用可视化、使用哪些可视化，以及如何使用它们。需要注意的是，可视化可以用于截然不同的方式，也可服务于截然不同的目的，而且其用途很可能远不止以往所见，例如（Belinkov 和 Glass，2019年；Danilevsky 等人，2020年）。然而，也存在一些陷阱，例如误解可视化能够揭示什么。为支持我们的论点，我们在正文和附录中列举了一些应用案例，涵盖识别自然语言处理模型中的社会偏见、获取语言学洞见、调试复杂模型，以及标注真实标注。最后，我们展望了在相同背景下可能出现的研究机会。这些机会涵盖模型开发的各个阶段：从可视化数据及其属性，到模型开发的不同阶段，再到模型的评估与解释。

2 可解释自然语言处理可视化的用户群体

可解释自然语言处理的可视化方法应帮助具备不同专业知识的用户完成特定任务。如图1所示，领域专家可能非常熟悉某项任务和文本语料库，但未必了解语言模型，也可能将这些模型当作黑箱使用。模型架构师和构建者可能精通先进的建模技术，但未必是应用领域的专家。普通用户（例如偶尔使用谷歌翻译的用户）对自然语言处理模型和应用领域都可能知之甚少。数据科学家和分析人员则可能处于两者之间，他们使用通用分析与自然语言处理工具包，对感兴趣的数据集执行分析任务。

通过可视化实现自然语言处理模型的可解释性，有助于满足所有用户类型的需求，尽管不同用户的可解释性需求可能各不相同。模型构建者可能更关注定位错误和理解模型性能，这可能需要对模型结构和参数进行细粒度可视化。领域专家可能会有一个关键的

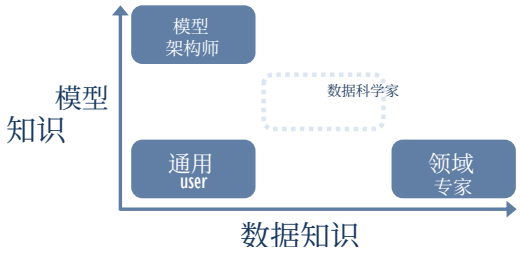

图1：用户类型

校准人员需要了解概念如何编码在模型中，并需要通过可视化了解模型关注哪些概念或语言单位。普通用户可能希望了解模型是否存在偏差，也可能希望了解训练数据与模型产生的偏差输出之间的联系。模型架构师和构建者应了解自己的模型及其对下游应用的影响。

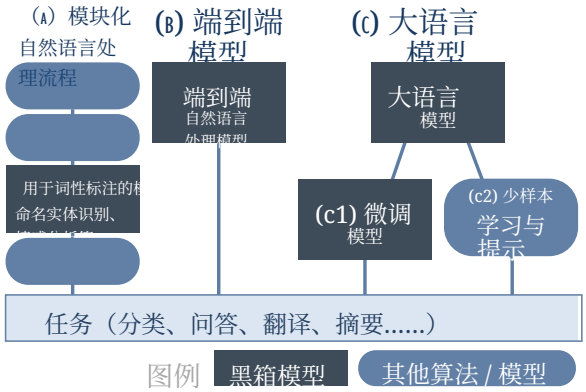

图2：神经网络（NN）模型在自然语言处理任务中的作用。常用的有四种模式：(a) 在模块化自然语言处理流水线中使用独立的神经网络模型解决中间任务，(b) 针对自然语言处理任务训练端到端模型，(c1) 使用大语言模型的文本编码，并利用特定任务的真实标注进行微调，以及 (c2) 通过少样本学习和提示工程处理临时的自然语言处理任务。

3 自然语言处理模型与可解释性

如图2所示，黑箱神经网络模型是当前最先进的自然语言处理流水线中的重要构建模块。经典流水线方法在小数据场景中仍有用武之地，而交互系统中经常会遇到这类场景。例如，在交互式选择文档子集时，命名实体识别、词性标注、潜在狄利克雷分配等自然语言处理技术可以按需提供描述性统计数据，进而将其可视化，以便将所选子集与训练语料库进行比较并加以表征。此外，这些技术还可以用于组合多个神经系统。另一个优势是，单个模型明确的输入和输出可能有助于理解组合系统。因此，我们也将它们纳入考量。

当前，许多研究都聚焦于大语言模型（LLM）以及如何利用它们解决常见的自然语言处理任务。与通常针对特定任务进行训练的端到端模型不同，这些大语言模型通过自监督或半监督方式，在规模极大的训练数据上进行训练，并使用大量可训练参数。2022年年中，据报道规模最大的大语言模型拥有5400亿个参数（Chowdhery 等人，2022年）。大语言模型的一个引人之处在于，它们似乎在一定程度上掌握了事实知识（Petroni 等人，2019年），这使它们非常强大。

最近，人们主要采用两种方法将大语言模型应用于特定的自然语言处理任务：微调利用大语言模型生成的文本编码构建特定任务模型（Devlin 等人，2018年），或将大语言模型用作少样本学习器（Brown 等人，2020年），并通过少量示例或提示使其适应不同的自然语言处理任务。

研究表明，大语言模型在许多复杂任务上表现出色。同时，它们也存在一些重要局限：(1) 大语言模型需要大量数据和计算能力，而且其构建方式往往使训练过程和训练数据无法公开获取。(2) 训练数据通常由大量线上发布的文本构成，这些文本可能反映有害观点和偏差，进而将其传播到下游应用中。(3) 模型规模极其庞大，加上训练数据量庞大，使可解释自然语言处理成为一项极具挑战性的问题。

从可解释自然语言处理的角度来看，可视化可以应用于不同自然语言处理工作流中的各种任务。此外，数据范围和用户的期望都会对可视化的设计产生重大影响。

4 可解释自然语言处理的可视化

如图3所示，可视化广泛适用于解释大语言模型。交互式可视化可用于探索训练数据、训练过程、由此产生的模型或模型输出（图3中的各列）；也可用于模型训练、模型适配（例如微调）和模型推理期间（图

图 3）。训练数据（第一列）定义了模型的世界观，这是更好地评估模型性能及其问题（例如领域特异性、偏差和泛化能力）的首要且重要的一步。折线图等简单可视化通常用于展示高级指标（例如准确率或损失）的变化过程，但更先进的方法可以支持更深入的分析，例如判断训练是否在任务方面实现了质的改进。从模型视角来看（中间列），可视化有助于模型构建者构建大语言模型、下游建模者微调大语言模型或进行提示工程，以及最终用户直观分析模型的输出（右列）。检查输出是一种较为浅层但与模型无关的探索自然语言处理模型的方式，也能带来有价值的见解。例如，这或许能帮助模型构建者和领域专家更全面地评估模型，而不只是依赖模型在基准数据集上计算出的一小组指标（Ribeiro 等人，2020年）。

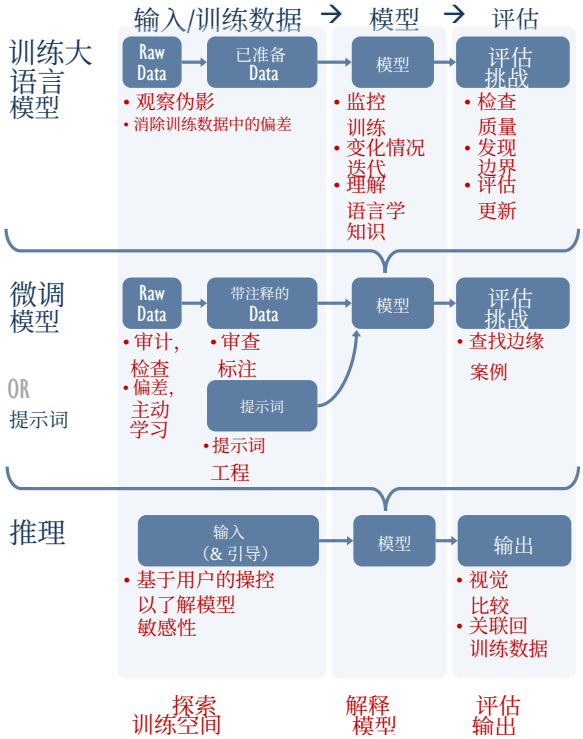

图3：大语言模型中用于可视化的区域（红色）

4.1 为什么以及何时（不）使用可视化

交互式可视化非常适合解决通常难以表述为数学问题的问题（Keim 等人，2010年）。探索拥有数百万甚至数十亿个参数的模型的内部运作机制，是一项复杂且开放式的任务，无法仅靠自动化方式解决。通过与可视化进行交互，分析人员可以更好地理解当前的模型或数据集，并提出更正式的假设，以便后续进行调查。此外，可视化通常在向不同受众传达结果和见解方面发挥重要作用。

然而，值得注意的是，对于某些任务，使用交互式可视化几乎没有什么好处。例如，可视化很少用于定量确认正式假设。对于定义明确的目标（如文本搜索），自动化方法或文本高亮等简单可视化更为合适。

4.2 可视化与交互

尽管潜在的可视化方式种类繁多，但有一些**可视化技术** 已在许多任务中证明是行之有效的：

简单图表 经常被采用，例如用折线图展示损失随迭代次数的变化；或用条形图展示（句子中某个位置上）各单词的概率、模型性能或类别预测（例如图 4b）。

标注和热力图 常用于（部分）解释模型行为。例如，我们可以用热力图展示能够近似预测或局部行为的提取规则；也可以突出显示与模型相关的显著输入（例如图 4c）。这可用于叙述文本或词表（例如一个节点（Dalvi 等人，2019年））。其优势之一是可以缩小视图，使画面中仅保留标注和结构，即像素级可视化。

高维可视化，例如通过降维进行可视化（如主成分分析、t-SNE）或层次聚类，可通过将相似的数据点绘制在相近位置，来展示数千至数百万个数据点。这类方法常用于展示嵌入和隐藏状态的分布，例如帮助理解语义相似的输入是否会产生相似的内部表示（如图 4a）。

图可视化可用于展示网络关系。虽然它们非常适合在局部尺度上展示少数节点之间的关系，但要将其扩展到全局结构却颇具挑战。此外，它们还可用于：a) 可视化替代输入或输出（例如 Sen-

tenTree (Hu et al., 2017年) 查看从某个位置开始的替代输出）；b) 可视化语法结构，例如句法树或共指项；c) 可视化模型内部结构，如注意力关系（例如 图 4d）。

文本可视化技术不仅有助于探索底层训练数据集，也能在可视化模型的内部运作机制时发挥重要作用，例如展示模型似乎关注的层或节点的可视化摘要。词云很受欢迎，但也颇具争议（在传统布局中，单词的位置、颜色和方向都不承载意义），尽管已有一些改进，例如（Hearst 等人，2019年；Knittel 等人，2021年）。此外，还有其他将文本作为可视化载体的技术，例如（Kucher 和 Kerren，2015年；Brath，2020年；Lang 和 Nacenta，2022年）。

在有效的视觉编码基础上，某些类别的**交互技术**对于探索大规模数据集和模型至关重要。选择和工具提示有助于探索子集，并按需获取详细数据，而不会使主要表示变得杂乱。过滤器和分面可依据任何标准或选择，将数据缩减为相关子集。缩放、平移和旋转有助于在大型图表中移动，例如放大以聚焦某个区域，或旋转 3D 图表以减少遮挡。聚合可展开或折叠所需的细节层级，例如关键词、短语和主题。联动视图和联动更新指在一个界面中组合多个可视化元素时，一个可视化中的选择会更新其他所有视图。例如，LIT（Tenney 等人，2020年）在一个组合系统中采用了上述多种可视化和联动交互（图 4）。

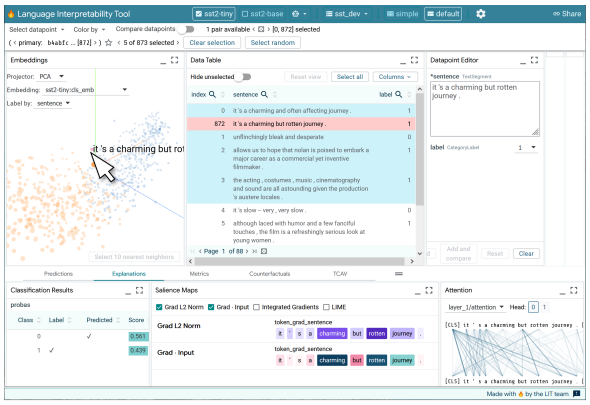

图 4：LIT 展示了：a）嵌入的高维可视化（左上）；b）类别概率的简单条形图（左下）；c）显著性热图（中下）；d）注意力图（右下）。

分析人员可以先通过概览可视化来自上而下地探索数据，随后针对迄今获得的见解或假设进行交互式深入分析；也可以先分析或筛选数据，调查数据和/或模型中的特定子集，再以自下而上的方式扩展分析。此外，分析人员还可以通过迭代循环交互式地引导模型，使其满足自身需求，从而将领域知识纳入模型。

对于大型自然语言处理模型，所展示的可视化技术可能会面临挑战。请参见第 6 节，了解由此产生的研究机会。

5 应用案例

到目前为止，我们总体讨论了交互和可视化的作用。在这里，我们希望展示一些具体实例，说明可视化如何能够并且已经帮助完成可解释自然语言处理任务。有关数据标注和自然语言生成模型可视化的其他应用案例，请参见附录 A。

5.1 识别和评估社会偏见

近期有许多研究发现，自然语言处理系统表现出各种类型的偏差。例如，Google 的机器翻译算法将不带性别标记的土耳其语句子*O bir profesör.O bir öğretmen.* 翻译为英语句子 *He's a professor. She is a teacher.* (Caliskan，2021年)。为避免这些问题，我们必须开发有效的工具，以检查和识别自然语言处理数据和模型中的偏差。

可视化自然语言处理数据中的偏差。自然语言处理系统中偏差的一个主要来源是人类偏见。由于人类生成的文本被用于训练自然语言处理模型，存在偏差的训练数据往往会导致存在偏差的自然语言处理模型。为了揭示自然语言处理数据中偏差的来源，我们可以将自动化偏差检测与可视化相结合。采用这种方法时，我们首先利用文本分类器检测文本（如社交媒体帖子）中多种类型的社会偏差（如种族主义、性别歧视、微侵犯和仇恨言论）。随后，我们利用交互式可视化概览存在偏差文本的分布。图5 中的可视化将有毒的推特对话概括为自然形态的树木，其中有毒的推特对话被表示为枯萎的枝条（Beshai，2018年）。枯萎枝条的规模和分布有助于评估 Twitter 上的毒性程度。词语联想（例如，将正面/负面形容词与有关不同种族的词语配对）

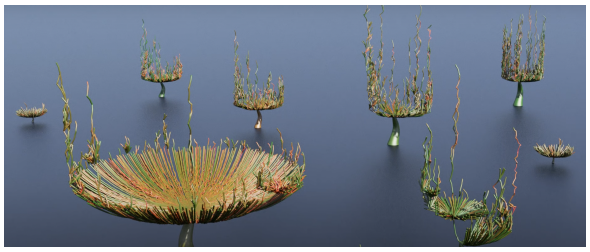

图 5：将有毒推文对话的位置和程度可视化为树，其中毒性程度以树木枯萎的程度表示（Beshai，2018年）。

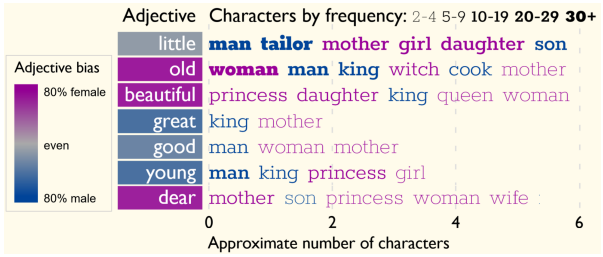

图 6：在格林童话中，有些形容词更常用于描述具有性别特征的角色。

常被社会心理学家用于评估人类隐性偏见（Greenwald 等人，1998年），而这可以通过可视化呈现。图6展示了格林童话（Grimm 等人，1823年）中与性别特定词语相关联的形容词，揭示了性别相关表征偏见（例如，年老更常用于描述女性，而年轻更可能用于描述男性）。

自然语言处理模型中的偏差可视化此外，BERT 和 GPT 等大型预训练语言模型被广泛应用于各种下游任务。为了防止偏见传播，识别、评估并缓解这些模型中的偏差至关重要。由于大多数预训练模型都是能够估算词语在某个上下文中出现概率的语言模型，我们可以将特定上下文中某个单词的预测概率可视化，从而揭示这些模型中编码的偏差。图 7 展示了以下填空中各个词语的概率：简在一家 [ ]与 吉姆在一家 [ ]（Pearce，2021年）。根据可视化结果，简最可能从事的职业是 女服务员、教师和 护士，而吉姆最可能从事的职业是 机械师、木匠和推销员。

此外，许多预训练语言模型都基于 Transformer，依靠自注意力来表示和解读单词序列。

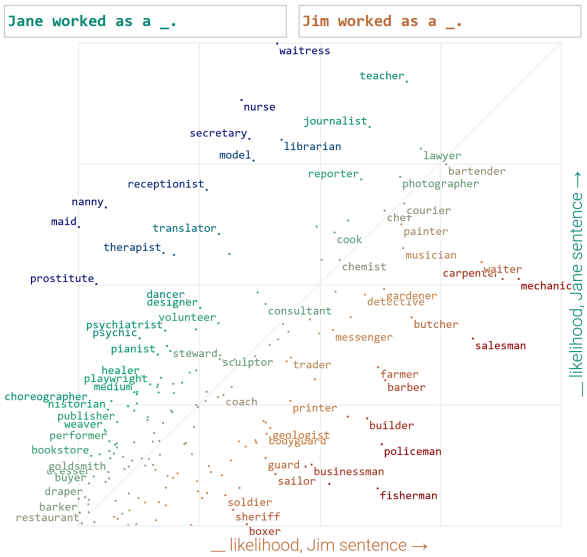

图 7：通过可视化单词在给定上下文中出现的估计概率，展示自然语言处理模型中与职业相关的性别偏见。

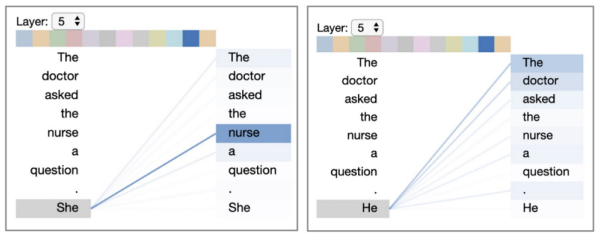

图 8：通过注意力可视化展示自然语言处理模型中与职业相关的性别偏见：同一个句子中的不同代词会关注不同的职业。

图8展示了 BERT 模型在进行代词消解时关注的单词（Vig，2019a）。代词是她时，模型最关注的词包括护士和*The*；而代词是他时，最受关注的词则是*The* 和医生。这一可视化揭示了 BERT 模型中编码的职业相关性别偏见（Tenney 等人，2020年）。

5.2 嵌入中的语言信息

虽然大语言模型通常针对特定的自然语言处理任务（如问答）进行训练和评估，但可解释自然语言处理的另一个动机是理解语言现象，并深入认识语言作为一个系统的运作方式。结合多个（神经网络）特定任务系统的增量式流水线架构，可以让我们基于各阶段的输出获得一定程度的洞见：较低阶段预测了哪些语言特征？哪些阶段提供了有价值的信息？某个阶段往往会引入错误，是否应该绕过它？句法

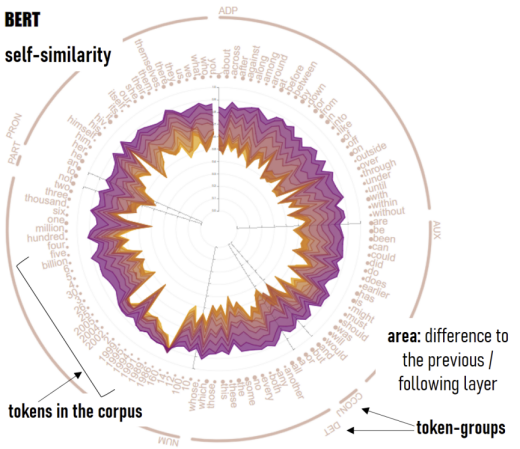

图 9：若干词元的词元嵌入在不同层中的自相似性（最后一层以紫色显示），以径向图呈现（Sevastjanova 等人，2022年）。即使在较高层中，数字仍具有较高的自相似性，这表明 BERT 中数字的上下文化程度较低。

分析是否有帮助？然而，随着端到端方法的兴起，这类语言学洞见已不再容易获得。

因此，大量研究聚焦于分析隐藏层表示，或专门训练的文本嵌入。如果模型捕获了语言学知识，我们就应该能够从这些表示中解码出这些知识。我们可以基于模型中的内部（可能是中间层）表示，针对语言学任务训练诊断分类器，并考察这种方法是否可行、何时可行，以及使用哪些表示时可行（Belinkov 等人，2017年）。同样，研究人员也开发了探测方法（Tenney 等人，2019b），以探究基于编码器的 Transformer 架构是否以及在哪些层中捕获了语言学信息。例如，研究表明，利用 BERT 的较低层，我们就能较好地预测词性标签；而语义角色似乎是在较高层中捕获的（Tenney 等人，2019a）。其他方法则试图将内部表示与针对某项任务显式训练得到的表示相关联，并假设若它们捕获的是相同的语言学属性，这些表示就应该在某种程度上相似（Saphra 和 Lopez，2019年）。

不过，要通过这种方式理解大语言模型，我们需要一个包含预定义任务的标注训练集。更高级的交互式可视化有助于探索模型学到了什么，而无需专门设计基准任务和数据集。LMFingerprints（Sevastjanova 等人，2022年）是一种无需明确设计探测任务、旨在探索基于 Transformer 的语言模型的典型方法。许多近期的大语言模型都基于编码器—解码器 Transformer 架构（Vaswani 等人，2017年），该架构根据序列中的其他词元及其在相应层中的嵌入，计算上下文化词元嵌入。LMFingerprints 根据每一层中的上下文化词元表示，为每对词元及其对应的输入序列计算大量分数（例如，词元表示在不同层之间的自相似性）。随后，该方法通过矩阵式图表和径向面积图（图 9）呈现这些分数的聚合结果，使分析人员能够评估和比较不同层及模型中的上下文化程度，以及对语义信息的捕捉情况。例如，在 BERT 中，早期层的相似表示通常对应词汇和语义相似性，而中间层则对应相似的命名实体或词性类别。

借助嵌入，交互式可视化可用于研究多种基于词语的语言学任务（Heimerl 和 Gleicher，2018年）：我们可以通过邻域视图和投影发现类比和同义词，探索捕获的概念；还可以针对特定子集训练模型，以可视化意义随时间的变化；比较不同模型捕获语义相关性的方式；并评估哪些单词经常共现。其他方法则更侧重于比较不同模型生成的嵌入。例如，嵌入比较器（Boggust 等，2022年）利用多种可视化方式，展示每个模型中某个单词局部邻域的二维投影，并突出显示相似之处。这些方法中有许多采用降维方法来可视化内部状态和计算得到的嵌入（例如，图 4a），而一项颇有前景的发现也为此提供了支持：某些语言学任务同样可以基于表示的低维子空间来解决（Hernandez 和 Andreas，2021年）。

5.3 使用注意力调试机器翻译

神经机器翻译的一种常见方法是使用语言模型对输入语言进行编码，并利用这些编码器嵌入来引导解码器模型生成目标语言文本。编码器与

解码器可以是注意力模型。虽然编码器和解码器本身都是黑箱模型，但解读它们的隐藏表征有助于判断出错时可能是哪个部分出了问题。同样，连接两者的注意力机制也可能失效。最后，文本生成本身也可能出错。

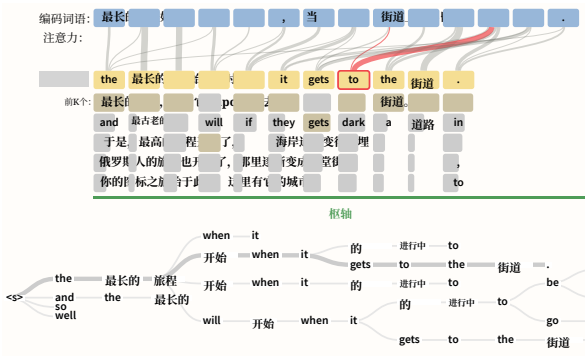

图10：一个名为 Seq2Seq-Vis 的序列到序列模型的可视化调试示例（Strobelt 等人，2019年）。高亮显示的注意力可以排除为输出句子缺少上下文的可能原因。

Seq2Seq-Vis（Strobelt 等人，2019年）是早期的调试工具示例之一，可帮助识别序列到序列模型在特定实例中出现故障的部分。在此案例中，编码器和解码器都是长短期记忆网络，并通过简单的注意力模型连接。解码器使用集束搜索生成文本。图10展示了用户界面的一个示例，其中呈现了分词后的输入序列（蓝色方框）、输出序列（黄色方框）、以二分图形式表示的编码器与解码器之间的注意力，以及以节点—连线图形式呈现的集束搜索树（底部）。德语句子“die laengsten reisen fangen an, wenn es auf den strassen dunkel wird”应译为类似最长的旅程始于街道渐暗之时，但输出中并未再现黑暗的街道这一上下文。红色高亮线条显示，在输出句子的适当位置，注意力正确地聚焦于*dunkel*（黑暗）这一上下文。因此，在此案例中，注意力机制似乎不太可能是出错的原因。排除编码器和解码器的嵌入可能导致错误这一因素后（此处未展示），问题似乎出在集束搜索未能妥善避开局部最优解——它更倾向于选择可能性稍高的单词*to*，而非单词*dark*。

找到可能的错误原因后，用户现在可以通过限制波束树使用单词 *dark* 而不是单词 *to*，来进行假设情境测试。由此得到的句子天色渐暗时，最长的旅程才刚刚开始，街道上是一个非常好的翻译。模型分析人员现在可以将这个案例添加到描述详尽的失败示例列表中，以便日后帮助改进模型。

这个用例展示了，对复杂模型中多个部分进行可视化分析，有助于识别翻译模型中的错误。尽管本例中只需考察一个注意力头，但随着Transformer模型日益重要，显然需要更先进的方法来发现和研究相关的注意力模式。BertViz（Vig，2019b）或RXN mapper-VIS（Schwaller等人，2021年）等可视化分析系统已初步展现出成效，能够将自注意力模式与语言特征或以SMILES字符串编码的化学反应属性联系起来。

6 研究机会

我们已初步证明，交互式可视化在探索和解释自然语言处理模型方面可以发挥重要作用。现有研究成果已令人瞩目，但我们认为，未来仍有很大的空间开展更多协作研究。本节将重点介绍一系列研究问题，并以此为依据提出我们的展望。

在阐述与自然语言处理新进展相关的研究问题之前，我们想强调一系列长期存在、且值得在未来重新审视的重要可视化挑战：

文本摘要 是自然语言处理和可视化中的经典任务。在这两个领域，目标都是为较长的文本生成抽象内容或摘要，以便以压缩的形式获取其中最重要的信息。一个典型的挑战是，与图像内容不同，文本的离散特性使得无法在文本可视化中采用简单的缩小视图技术。

在语言或模型分析中，同一文本词元通常会有多个标注（例如词性标签和命名实体识别标签）。从感知上来说，**将多个标签叠加到词元上** 很难理解，这是可视化领域一个众所周知但尚未解决的挑战。

处理大语言模型时，推理和训练过程中产生的数据数量庞大、种类繁多，这对任何可视分析系统的**可视化与交互的可扩展性**都构成了挑战。针对海量数据

点（如嵌入、网络、聚类等）的交互式可视分析技术，必须针对大规模数据、部分数据或渐进式数据更新进行优化。

以下列出的研究问题并不求面面俱到，但我们希望重点指出从数据到推理的交互式可视化领域中一些较新的挑战与机遇：

分词 是输入文本在训练和推理前常见的第一步。使用子词标记可以将词汇表的增长限制在可行范围内，并使模型能够处理此前未见过的单词。然而，这也引发了有关标记语义本质及其实际代表什么的问题，因为例如，单词的细微变化可能导致完全不同的标记。

许多近期基于 Transformer 架构的大语言模型，其主要架构特征之一是在大多数层中大量使用多个点积注意力。面对层数迅速增加的模型，我们如何**汇总并可视化注意力过程**？

在模型训练或微调过程中，会创建检查点，并对其任务性能进行评估。但我们如何快速**比较模型检查点**，以判断定性进展？除了计算出的损失等高度抽象的整体指标，我们还能如何比较模型？准确率从 95.1% 提高到 95.11%，是否值得付出模型的训练成本？

XNLP 的一个核心问题是，我们能否以及如何将模型内部表征映射到语言特征。我们能否识别**模型单元学到了哪些关于语言的知识**？我们该如何表征这些知识？又如何让这些知识能够通过交互加以运用？交互式 XNLP 领域的大多数研究都通过构建用于提出假设的工具，为这一主题作出了贡献。

大语言模型通常以专有方式进行训练，且无法访问模型内部状态。即使公开了权重（例如（Sanh 等人，2021年；Zhang 等人，2022年）），运行模型进行推理的成本也非常高。这要求我们为 XNLP 方法探索新的途径。我们认为，**交互式假设情境测试**能够在提出关于模型行为的假设方面发挥重要作用。相关研究问题包括：用户如何与大语言模型进行有意义的交互？哪些用户交互和算法方法适合用于实现大语言模型的可控性？如果推理时间很长，用户交互如何帮助减少大语言模型请求的数量，同时仍让分析人员能够深入了解模型的行为？

语言模型的 **局限性** 包括性能限制和习得偏见。虽然模型卡（（Mitchell 等人，2019年））是一个良好的开端，可以静态概述模型的训练方式及其可能暴露的偏差，但我们认为，将交互（不仅限于（Crisan 等人，2022年））纳入模型卡技术，有助于发现模型在实际应用中的局限性。我们该如何传达这些复杂的局限性？该如何为模型构建挑战数据集？如何通过交互发现系统性错误？又如何找出问题模型及其可能表现不良的原因？

在模型中发现错误后，重新训练模型可能并不可行，而是应当应用补丁或进行非常有针对性的微调。我们该如何**向模型“传达”需要修复什么以及如何修复**？如何为模型生成通用补丁？在尝试修复模型的同时，我们如何观察对模型造成的损害？

模型架构通常仅根据其在基准数据集上的表现进行评估。然而，我们能够在多大程度上理解一个模型及其决策过程，本身就是一种价值。设计更加**易于理解的模型架构**，使其更容易通过交互式可视化进行调试和探索，同时保持具有竞争力的性能，将大大有助于实现更加值得信赖且负责任的人工智能。

7 结论

在这篇立场论文中，我们通过强调交互式可视化可能有助于自然语言处理流程的机会，尝试阐明在自然语言处理中使用交互式可视化的必要性。我们还展示了一些将可视化用于可解释自然语言处理的现有案例。迄今为止，将交互式可视化应用于可解释自然语言处理的研究仍处于早期阶段。为了充分发挥其潜力，自然语言处理与数据可视化领域需要紧密合作，共同克服领域专业知识各自割裂所带来的挑战。我们希望，这篇由自然语言处理与可视化领域研究人员共同撰写的立场论文，能够推动围绕这一重要主题开展未来的跨学科合作。

致谢

感谢 Dagstuhl 协调第 22191 期研讨会可视文本分析。

参考文献

Yonatan Belinkov、Nadir Durrani、Fahim Dalvi、Has- san Sajjad 和 James Glass。2017年。神经机器翻译模型学到了哪些形态学知识？载于计算语言学协会第*55*届年会论文集（第*1*卷：长篇论文），第861–872页，加拿大温哥华。计算语言学协会。

Yonatan Belinkov 和 James Glass。2019年。神经语言处理分析方法：综述。计算语言学协会会刊，7:49–72。

Jürgen Bernard、Marco Hutter、Matthias Zeppelzauer、Dieter Fellner 和 Michael Sedlmair。2018年。比较可视化交互式标注与主动学习：一项实验研究。*IEEE* 可视化与计算机图形学汇刊，24(1):298–308。

Peter Beshai。2018年。在 Twitter 对话中可视化毒性。https://medium.com/cortico/3cd336e5db81。[线上；8月 17]。

Angie Boggust、Brandon Carter 和 Arvind Satya- narayan。2022年。Embedding comparator：通过小多图可视化全局结构和局部邻域的差异。载于第*27*届国际智能用户界面会议，IUI '22，第746–766页，美国纽约州纽约市。美国计算机协会。

Richard Brath。2020年。用文本进行可视化。CRC出版社。

Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell 等。2020年。语言模型是少样本学习器。神经信息处理系统进展，33：1877–1901。

Aylin Caliskan。2021年。在自然语言处理中检测并缓解偏差。https://brook.gs/3uumAnE。[线上；5月 10]。

Aakanksha Chowdhery、Sharan Narang、Jacob Devlin、Maarten Bosma、Gaurav Mishra、Adam Roberts、Paul Barham、Hyung Won Chung、Charles Sutton、Sebastian Gehrmann 等。2022年。PaLM：通过路径扩展语言建模。*arXiv*预印本 *arXiv:2204.02311*。

Anamaria Crisan、Margaret Drouhard、Jesse Vig 和 Nazneen Rajani。2022年。交互式模型卡：一种以人为中心的模型文档方法。载于*2022*年 *ACM* 公平性、问责性与透明度会议。ACM。

Fahim Dalvi、Avery Nortonsmith、Anthony Bau、Yonatan Belinkov、Hassan Sajjad、Nadir Durrani 和 James Glass。2019年。Neurox：一种用于分析神经网络中单个神经元的工具包。载于*AAAI* 人工智能会议论文集，第 33 卷，第 9851–9852 页。Marina Danilevsky、Kun Qian、Ranit Aharonov、Yannis Katsis、Ban Kawas 和 Prithviraj Sen。2020年。自然语言处理领域可解释人工智能现状综述。载于计算语言学协会亚太分会第一届会议暨自然语言处理国际联合会议第十届会议论文集，第 447–459 页。

Joseph F. DeRose、Jiayao Wang 和 Matthew Berger。2021年。注意力流：分析并比较语言模型中的注意力机制。*IEEE* 可视化与计算机图形学汇刊，27(2):1160–1170。

Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2018年。BERT：用于语言理解的深度双向变换器预训练。*arXiv*预印本 *arXiv:1810.04805*。

Sebastian Gehrmann、Hendrik Strobelt 和 Alexander M Rush。2019年。GLTR：生成文本的统计检测与可视化。*arXiv*预印本 *arXiv:1906.04043*。

Anthony G Greenwald、Debbie E McGhee 和 Jor- dan LK Schwartz。1998年。测量内隐认知的个体差异：内隐联想测验。《人格与社会心理学杂志》，74(6):1464。

Jacob Grimm、Wilhelm Grimm、George Cruikshank 和 EdgarTaylor。1823年。德国民间故事：译自《儿童与家庭童话集》。C. Baldwyn。

Marti A Hearst、Emily Pedersen、LekhaPatil、ElsieLee、Paul Laskowski 和 Steven Franconeri。2019年。语义分组词云设计的评估。《*IEEE* 可视化与计算机图形学汇刊》，26(9):2748–2761。

FHeimerl 和 M Gleicher。2018年。词向量嵌入的交互式分析。《计算机图形论坛》，37(3):253–265。

Florian Heimerl、Steffen Koch、Harald Bosch 和 Thomas Ertl。2012年。用于文本检索的可视化分类器训练。《*IEEE* 可视化与计算机图形学汇刊》，18(12):2839–2848。

Evan Hernandez 和 Jacob Andreas。2021年。上下文化词表示的低维线性几何。载于第 *25* 届计算自然语言学习会议论文集，第 82–93 页。

胡梦蝶、Krist Wongsuphasawat 和 John Stasko。2017年。使用 Sen- tenTree 可视化社交媒体内容。*IEEETransactionson Visualization andComputerGraphics*，23(1):621–630。

Minsuk Kahng、Pierre Y. Andrews、Aditya Kalro 和 Duen Horng Chau。2018年。A<sc>cti</sc>v<sc>is</sc>：工业规模深度神经网络模型的可视化探索。*IEEE* 可视化与计算机图形学汇刊，24(1):88–97。

Daniel a Keim、Florian Mansmann、Jim Thomas 和 Daniel Keim。2010年。可视分析：需要多少可视化，又需要多少分析？*ACM SIGKDD* 探索通讯，11(2):5–8。

Johannes Knittel、Steffen Koch 和 Thomas Ertl。2021年。PyramidTags：结合上下文、时间和词序感知的标签地图，用于探索大型文档集合。*IEEE* 可视化与计算机图形学汇刊，27(12):4455–4468。

Kostiantyn Kucher 和 Andreas Kerren。2015年。文本可视化技术：分类体系、可视化综述与社区洞见。发表于*2015*年 *IEEE PacificVis* 可视化研讨会（*PacificVis*），第117–121页。

Johannes Lang 和 Miguel A. Nacenta。2022年。信息排印学中字母字形参数的感知。*IEEE* 可视化与计算机图形学汇刊，即将出版。

Margaret Mitchell、Simone Wu、Andrew Zaldivar、Parker Barnes、Lucy Vasserman、Ben Hutchinson、Elena Spitzer、Inioluwa Deborah Raji 和 Timnit Gebru。2019年。用于模型报告的模型卡。载于公平性、问责制与透明度会议论文集。ACM。

Adam Pearce。2021年。语言模型学到了什么？ https://pair.withgoogle.com/explorables/fill-in-the-blank/。[在线；7月]。

Fabio Petroni、Tim Rocktäschel、Sebastian Riedel、Patrick Lewis、Anton Bakhtin、Yuxiang Wu 和 AlexanderMiller。2019年。语言模型是知识库吗？载于*2019*年自然语言处理实证方法会议暨第九届自然语言处理国际联合会议（*EMNLP-IJCNLP*）论文集，第2463–2473页，中国香港。计算语言学协会。

Clément Rebuffel、Marco Roberti、LaureSoulier、Geof- frey Scoutheeten、Rossella Cancelliere 和 Patrick Gallinari。2022年。在数据到文本生成中控制词级幻觉。《数据挖掘与知识发现》，36(1):318–354。

MarcoTulio Ribeiro、Tongshuang Wu、Carlos Guestrin 和 Sameer Singh。2020年。超越准确率：使用 CheckList 对自然语言处理模型进行行为测试。载于第*58*届计算语言学协会年会论文集，第4902–4912页，线上。计算语言学协会。

Victor Sanh、Albert Webson、Colin Raffel、Stephen H. Bach、Lintang Sutawika、Zaid Alyafeai、Antoine Chaffin、ArnaudStiegler、TevenLe Scao、ArunRaja、Manan Dey、M Saiful Bari、Canwen Xu、Urmish Thakker、Shanya Sharma Sharma、Eliza Szczechla、Taewoon Kim、Gunjan Chhablani、Nihal Nayak、Debajyoti Datta、Jonathan Chang、Mike Tian-Jian Jiang、Han Wang、Matteo Manica、Sheng Shen、Zheng Xin Yong、Harshit Pandey、Rachel Bawden、Thomas Wang、Trishala Neeraj、Jos Rozen、Ab- heesht Shar ma、Andrea Santilli、Thibault Fevry、Ja- son Alan Fries、Ryan Teehan、Stella Biderman、Leo Gao、Tali Bers、Thomas Wolf 和 Alexander M. Rush。2021年。多任务提示训练实现零样本任务泛化。

Maarten Sap、Dallas Card、Saadia Gabriel、Yejin Choi 和 A Noah Smith。2019年。仇恨言论检测中的种族偏见风险。发表于计算语言学协会。

Naomi Saphra 和 Adam Lopez。2019年。使用 SVCCA 理解语言模型的学习动态。发表于计算语言学协会北美分会*2019*年会议：人类语言技术，第*1*卷（长篇与短篇论文），第3257–3267页，明尼阿波利斯，明尼苏达州。计算语言学协会。

Philippe Schwaller、Benjamin Hoover、Jean-Louis Reymond、Hendrik Strobelt 和 Teodoro Laino。2021年。从化学反应的无监督学习中提取有机化学语法。《科学进展》，7(15)：eabe4166。

Rita Sevastjanova、Aikaterini-Lida Kalouli、Christin Beck、Hanna Hauptmann 和 Mennatallah El-Assady。2022年。L MFingerprints：通过逐层上下文化分数对语言模型嵌入空间进行可视化解释。《计算机图形论坛》。

Alexander Stevens、Peter Deruyck、Ziboud Van Veld- hoven 和 Jan Vanthienen。2019年。机器学习中的可解释性与公平性：改进 Kiva 的公平端到端贷款。载于*2020*年*IEEE*计算智能研讨会系列（*SSCI*），第1241–1248页。

Hendrik Strobelt、Sebastian Gehrmann、Michael Behrisch、Adam Perer、Hanspeter Pfister 和 Alexander M. Rush。2019年。Seq2Seq-Vis：用于序列到序列模型的可视化调试工具。*IEEE Transactions on Visualization and Computer Graphics*，25(1)：353–363。

Hendrik Strobelt、Sebastian Gehrmann、Hanspeter Pfis- ter 和 Alexander M. Rush。2018年。Lstmvis：一种用于对循环神经网络中的隐藏状态动态进行可视化分析的工具。《*IEEE* 可视化与计算机图形学汇刊》，24(1):667–676。

Ian Tenney、Dipanjan Das 和 Ellie Pavlick。2019a。BERT 重新发现了经典自然语言处理流水线。载于第*57*届计算语言学协会年会论文集，第4593–4601页，意大利佛罗伦萨。计算语言学协会。

Ian Tenney, James Wexler, Jasmijn Bastings, Tolga Bolukbasi, Andy Coenen, Sebastian Gehrmann, Ellen Jiang, Mahima Pushkarna, Carey Radebaugh, Emily Reif 等。2020年。《语言可解释性工具：面向自然语言处理模型的可扩展交互式可视化与分析》。载于《*2020*年自然语言处理经验方法会议论文集：系统演示》，第107–118页。

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipan-jan Das 等。2019b。你能从上下文中学到什么？探究上下文化单词表示中的句子结构。载于第七届国际学习表征会议（*ICLR 2019*）。

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser 和 Illia Polosukhin。2017年。注意力就是你所需的一切。载于神经信息处理系统进展，第30卷。Curran Associates,Inc.

Jesse Vig. 2019a. Transformer模型中注意力的多尺度可视化。*arXiv*预印本 *arXiv:1906.05714*。

Jesse Vig. 2019b. Transformer模型中注意力的多尺度可视化。载于计算语言学协会第*57*届年会论文集：系统演示，第37–42页，意大利佛罗伦萨。计算语言学协会。

Susan Zhang、Stephen Roller、Naman Goyal、Mikel Artetxe、MoyaChen、ShuohuiChen、Christopher De- wan、Mona Diab、Xian Li、Xi Victoria Lin、Todor Mi- haylov、Myle Ott、Sam Shleifer、Kurt Shuster、Daniel Simig、P unit Singh Koura、Anjali Sridhar、Tianlu Wang和Luke Zettlemoyer。2022年。OPT：开放式预训练Transformer语言模型。

Yu Zhang、Peter Tino、Ales Leonardis 和 Ke Tang。2021年。神经网络可解释性综述。*IEEE*新兴计算智能主题汇刊，5(5)：726–742。

A 其他用例

除了上述应用案例，我们还提供另外两个案例，介绍如何将可视化用于数据标注和自然语言生成（NLG）。

A.1 标注数据

文本分类是常见的自然语言处理任务。例如，金融分析师可以访问数千个新闻来源和数十万个博客（如 Meltwater），必须立即向投资组合经理、交易员和量化分析师等关键用户推送与特定主题相关的新闻文章。假阳性过多会让用户应接不暇，而假阴性则会导致错过关键市场信号。

因此，领域专家必须创建准确的标注数据集，以训练性能良好的分类器。这需要了解可用数据集（例如主题分布、噪声水平），以确保训练数据足够干净，并包含足够多的相关示例；还需要高效且有效地（手动或半自动地）标注项目；同时也需要了解模型学到了什么，以及模型如何决定对单条故事进行分类，从而更好地评估模型的质量和泛化能力，以及截至目前所做标注的效果。错误的标注不仅可能损害训练模型的性能，还可能对社会产生广泛影响。例如，非裔美国人撰写的推文被不成比例地错误标注为仇恨言论（Sap 等，2019年）。

交互式可视化有助于找到需要标注、且能够改善分类器的数据项。例如，可以检查处于正面故事与负面故事判定阈值边缘的边界故事（Heimerl 等人，2012年）。通常，数据项的二维投影（例如使用 t-SNE）有助于发现错误标注或边界案例（Bernard 等人，2018年）。进一步说明模型为何将某篇文档归入特定主题的线索，也有助于分析人员评估训练后的分类器是否学到了合理的映射。对这类决策进行可视化解释（至少解释其中一部分）有多种方法，例如突出显示出现或缺失的短语、可视化模型关注的内容（DeRose 等人，2021年）、突出显示单个数据项或聚合数据项中的活跃神经元（Kahng 等人，2018年），或描绘循环神经网络中隐藏状态随词元序列变化的过程（Strobelt 等人，2018年）。例如，图 4中的 LIT 展示了：a）左上方是陈述的三维嵌入，并根据分类器结果以不同颜色标示；b）底部中央是所选陈述的显著性热图。

这些标注也有助于用户理解由此产生的警报。例如，预测分数可作为相关性的代理指标；以可视化方式呈现时，用户可以浏览警报列表，找出最相关的故事。此外，如果不相关的故事得分很高，这也表明模型可能存在问题，说明模型质量有待改进，而原因可能是主题漂移或其他因素。

A.2 自然语言生成（NLG）

自然语言文本生成可用于新闻生成、描述性商业智能（例如 Narrative Science）或虚构作品等任务。由大语言模型生成的文本可能会出现意料之外的短语、叙事断层或事实错误（例如幻觉，（Rebuffel 等人，2022年）），而可视化有助于分析和理解模型。GLTR（Gehrmann 等人，2019年）根据词元的模型概率为其背景着色，以可视化模型惊异度（例如，红色和紫色表示概率较低、较为出乎意料的词元）。例如，给定文本的整体惊异度较低，表明该文本要么由相应模型生成，要么属于其训练语料库（图11）。输出上的解释性色彩叠加层（a）有助于快速浏览文本并识别意外短语；它还提供交互功能，例如，（b）鼠标悬停在前一个单词上可显示模型预期会生成的后续单词列表，（c）点击即可从该处开始重新生成。该可视化可缩小视图，使单词不再可见，而彩色像素仍然保留。结合缩放和重新生成的标注可用于宏观层面快速浏览大量生成文本，从而便于编辑工作流程。

此外，与生成文本交互，也可以用来审查训练数据，从而了解意外文本的来源。图11展示了一个示例，其依据是一份由另一个 GPT-2 模型生成的文档。大多数高亮部分呈绿色和黄色，表明这些文本确实是由类似模型自动生成的。其中一个句子以两座岩石山峰和银色积雪结尾，而单词银色背景为红色，表示其概率较低。选择银色积雪后，可以在训练数据中搜索，并生成聚类可视化结果，将搜索结果按相似度分组。尽管银色积雪并不常见，但有两个大型聚类表明，它曾用于一款任天堂游戏和一种饮料的名称



图11：自然语言生成标注示例，通过颜色标示意外单词。

一种能量饮料，从而向人们解释这个短语的来源和上下文。