一双新的 GloVes

Riley Carlson, **John Bauer** and **Christopher D. Manning**
Stanford NLP Group
Stanford University
353 Jane Stanford Way, Stanford CA 94305-9035, U.S.A.
{ rileydc, horatio, manning } @stanford.edu

Abstract

本报告记录、描述并评估了新的 2024 年英文 GloVe（词表示的全局向量）模型。虽然 2014 年构建的原始 GloVe 模型被广泛使用并被认为有用，但语言和世界不断发展，我们认为当前的使用可能会从更新的模型中获益。此外，2014 年的模型在使用的确切数据版本和预处理方面没有得到认真记录，而我们通过记录这些新模型来纠正这一问题。我们使用 Wikipedia、Gigaword 和 Dolma 的一个子集训练了两组词嵌入。通过词汇比较、直接测试和命名实体识别任务的评估表明，2024 年的向量包含了新的文化和语言相关的词汇，在类比和相似性等结构性任务上表现相当，并在最近与时间相关的命名实体识别数据集（如非西方新闻数据）上展示了改进的性能。

1 介绍

神经语义词向量空间模型用一个称为词嵌入的实值向量表示每个词。一种广泛使用的词嵌入是 Pennington et al. (2014) 引入的 GloVe 词嵌入。该算法利用了一个以局部上下文为焦点构建的全局共现矩阵。GloVe 模型在这个全局矩阵上进行训练，所得的权重就是词嵌入，这样相似的词在向量空间中被分得更近。结果得到的词嵌入在密集的向量空间中对语义关系进行编码，使它们对各种 NLP 任务有用。尽管基于变压器的模型兴起，预训练的静态嵌入如 GloVe 在资源有限的环境中、计算高效的模型和注重可解释性应用中仍然很有价值。

我们更新词嵌入以纳入更近期数据的动机在于，自 2014 年最初训练以来，出现了新词，而现有词汇的语义意义也发生了变化。例如，“covid” 在 2014 年的嵌入中是没有表示的。使用这种更新的词汇表的嵌入在下游任务中有许多好处，例如减少词汇不在词汇表中的问题。为了反映当前英语词汇的使用情况，需要训练基于近期语言的新嵌入。

在这项工作中，我们将最小频率阈值（MFT）纳入词汇选择过程，以训练更新的词向量，该过程建立在 GloVe-V (Vallebueno et al., 2024) 工作的基础之上。使用 MFT 使我们能够在过滤掉过于稀有和噪声的词语的同时，保留不太频繁但在上下文中重要的术语。GloVe-V 框架通过引入统计不确定性估计来扩展这种方法，这考虑了由于数据稀疏性导致的嵌入位置的可变性。这使得训练的词向量不仅具有鲁棒性和表现力，而其更适合下游任务，其中稀有词通常具有关键重要性，确保适应现代语言的使用 (Vallebueno et al., 2024) 。

通过本文，我们详细说明了确切的训练程序，并证明 2024 年的词嵌入具有更新的词典，反映了当今的语言使用和文化趋势。在词类比和相似性任务中，它们与 2014 年的嵌入表现相当，表明相似的结构和核心语义表达能力。此外，2024 年的嵌入在时间依赖的命名实体识别（NER）数据集中表现出更好的性能。

在这份报告中，我们首先描述用于创建两组不同词嵌入的训练数据，包括选择的语料库和预处理步骤。然后，我们概述训练过程以便易于重复。最后，我们展示了嵌入的评估指标，包括新的词汇覆盖、直接评估以及在下游任务中的表现。

2 数据：2014 年 vs. 2024 年

2014 Data	2024 Data
Wikipedia & Gigaword – 6 billion tokens)	Wikipedia (2024) & Gigaword (5th edition) – 11.9 billion
Common Crawl – 42 and 840 billion	Dolma subset – 220 billion
Twitter – 27 billion	

Table 1: 用于训练单词嵌入的 2014 年和 2024 年数据源的比较，包括语料库的大小（以十亿标记为单位）。

为了生成 2024 年的嵌入，我们使用了 3 个语料库来训练两组嵌入：Wikipedia、Gigaword 和 Dolma。为与 2014 年的 Wikipedia 和 Gigaword 嵌入进行充分比较，我们采用了相同的语料库，但更新了 Wikipedia 数据转储。

Wikipedia 语料库是一个由 Wikipedia 文章组成的数据集，它比字典等提供了一个更自然的环境中的单词定义来源。除了 Wikipedia 之外，我们还使用了 Gigaword (Parker et al., 2011) 。具体来说，对于 2024 年的向量，我们使用了 Gigaword 的第 5 版。此语料库包含了来自 1994 年至 2010 年间 4 到 7 家（视年份而定）不同国际新闻机构的英语新闻电讯。自 2014 年以来，Wikipedia 的数据转储在标记数量上大约翻了一倍。为平衡这种增长，我们在训练语料中放入了两份 Gigaword 的副本。

Dataset	Percent Taken	Tokens (Billions)
Common Crawl	5 %	87.2
C4	40 %	60
Reddit	100 %	68.9
Project Gutenberg	100 %	2.3

Table 2: Dolma 训练子集

除了前述的语料库外，我们还使用了 Dolma v1.6 (Soldaini et al., 2024) 。该语料库于 2024 年 1 月发布，由来自书籍、编程脚本、参考资料、学术文章和在线内容的 3 万亿个令牌组成。我们从 Dolma 中获取了超过 1TB 的子集。表格 2 展示了从 Dolma 中获取的子集及所使用的令牌数。具体来说，我们有来自 Common Crawl 和 C4 (Raffel et al., 2020) 的网页、来自 Project Gutenberg 的书籍、以及来自 Reddit 的社交媒体。C4 的数据截止到 2019 年，而其他的截至到 2023 年。

首先，我们将更详细地描述用于训练不同嵌入的三个语料库，以及采取的任何预处理步骤。接下来，我们将描述用于所有嵌入的训练过程。随后，将展示不同的评估实验。

2.1 语料库 # 1：维基百科与 Gigaword

2024 向量训练语料库的维基百科部分从维基百科转储 https://dumps.wikimedia.org/enwiki/20240720/enwiki-20240720-pages-meta-current.xml.bz2 下载。然后使用 Wiki-Extractor (Attardi, 2015) 提取数据。通过去除诸如 <doc> 和 <unk> 标记的数据对维基百科数据进行了清理。

数据使用斯坦福的 CoreNLP 分词器（版本 4.4.1）[1] 进行小写化预处理。然后将 Wikipedia 和 Gigaword 语料库合并，其中 Gigaword 被包括了两次。合并后的语料库大约为 60GB，其中 Gigaword 约占 74%，其余由 Wikipedia 占据。

Wiki/Giga 向量的词汇量是根据 Vallebueno et al. (2024) 所述的方法选择的。具体来说，词汇量的大小是通过设置一个用于包含单词到语料库中的 MFT 来确定的。通过对使用不同 MFT 训练的向量进行实验，观察到 MFT 为 20 时，训练向量与其加权最小二乘 (WLS) 向量之间的平均余弦相似度最高。与 WLS 向量具有高余弦相似度表明训练得到的嵌入与从共现矩阵推导出的统计最优解非常一致，反映了稳健且准确的词表示 (Vallebueno et al., 2024) 。对于这个语料库，使用 MFT 为 20 结果是词汇量为 1,291,146 个单词。

像其他语料一样，我们以相同的方式使用斯坦福的 CoreNLP 分词器进行预处理。在预处理中，我们移除了 <unk> 个标记。使用了最大词汇量为 120 万的限制。词汇构建过程在 Dolma 的不同子集上独立进行，并在最后合并以达到 120 万的词汇量。此外，通过将合并词汇上的共现矩阵合并，创建了共现矩阵。

2.2 训练

培训过程在所有嵌入和语料库中是一致的。对于每个嵌入，首先构建词汇表和共现矩阵。维基百科和 Gigaword 语料库的词汇量设置为超过 120 万，Dolma 语料库的词汇量为 120 万。使用大小为 10 的对称上下文窗口来定义共现。一旦构建了共现矩阵，就使用固定的种子 123 对 Wiki/Giga 矩阵进行洗牌，使用固定的种子 2024 对 Dolma 矩阵进行洗牌。

对于维度为 50、100、200 和 300 的嵌入模型，分别在 Wikipedia 和 Gigaword 语料库上进行了训练，并在 Dolma 语料库上训练了 300 维的嵌入。这些嵌入使用 GloVe 的原始优化器 AdaGrad 进行优化。训练过程使用 demo.sh 脚本执行，该脚本在 GloVe 存储库 中提供，并且在 Training_README.md 文件中有更多文档。训练过程中使用的超参数汇总于表 3 。

2.3 评估：更新的词典

为了评估嵌入的质量，我们检查 2024 年嵌入中存在但 2014 年嵌入中不存在的单词，以确定更新后的嵌入是否反映了新的常用词。比较了来自 Wikipedia 和 Gigaword 语料库的 2014 年和 2024 年嵌入词汇表，以及 2024 年 Dolma 嵌入词汇表与 2014 年在 Common Crawl 上训练的 840B 向量的词汇表。通过将词汇表表示为集合，我们通过从 2024 年集合中减去 2014 年集合来计算差异。从这个结果集合中，我们

[1] https://nlp.stanford.edu/software/tokenizer.html

Hyperparameter	Value
Learning Rate	0.05 *
Alpha	0.75
XMax	100
Seed	2024 *
Epochs: (50d, 100d)	50
Epochs: (200d, 300d)	100

Table 3: 在 Wiki/Giga 上训练的 50 维、100 维、200 维和 300 维词嵌入，以及在 Dolma 上训练的 300 维词嵌入的训练超参数摘要。对于 50 维的 Wiki/Giga 向量，使用了学习率 0.075 和种子 123。

为每个训练语料库选择 39 个代表性示例来说明我们的发现。

2.4 评价：直接评价

我们对嵌入进行了直接评估任务，将其与 2014 年的嵌入进行比较。评估的重点是两个主要任务：词类比和词相似性。

对于词类比任务，目标是在类比格式 `word_1 : word_2 :: word_3 : ?` 中预测第四个词，并将其与黄金标准标记词进行比较，以计算准确率。我们使用两个基准数据集：

- 谷歌类比数据集 (Mikolov et al., 2013a)，包括 8,869 个语义词对和 10,675 个句法词对。
- MSR 类比数据集 (Mikolov et al., 2013b)，包含 8,000 个句法词对。

对于词语相似度任务，词嵌入通过给词对分配相似度评分并将这些评分与人工标注的基准进行比较来进行评估。我们使用三个基准数据集：

- WordSim353 (Finkelstein et al., 2001)，其中包含 353 个词对，这些词对被分类为高度相似、较少相似但相关或不相关。
- SimLex999 (Hill et al., 2015)，包含 999 个词对，并标注了语义相似度分数。
- MEN (Bruni et al., 2014)，包括 3,000 对人类评判相关性得分的词对。

为了进行这些评估，我们使用了由 Jastrzebski et al. (2017) 开发的嵌入评估软件包，该软件包支持对词汇类比和词汇相似性数据集的分析。

2.5 评估：命名实体识别

为了进一步评估新嵌入的表现，我们在下游任务命名实体识别（NER）上进行评估。在这个任务中，句子中的标记被标注为预定义的实体类别，如人名、地名、组织名和其他专有名词，从而实现从文本中抽取结构化信息。为了这个评估，我们使用斯坦福的 Stanza NER [2] 模型 (Qi et al., 2020)，并对其进行修改，以用我们训练的词嵌入替换默认的词嵌入。我们为每个词嵌入和数据集训练一个模型。

对于 NER 评估，我们使用三个数据集分别训练和测试模型，分别针对 2014 和 2024 的嵌入：

- CoNLL-03：该数据集发布于 2003 年，包括人物、地点、组织以及杂项类别的实体（不属于其他类别的实体）(Tjong Kim Sang and De Meulder, 2003)。
- CoNLL-PP：由 Liu and Ritter (2023) 引入的 CoNLL-03 的改进版本，具有更新和现代化的数据。我们在 CoNLL-03 上训练模型，并在 CoNLL-PP 测试集上进行评估。
- 英语全球新闻专线：该数据集在 (Shan et al., 2023) 中引入，由 2023 年发布的超过 1000 篇英语新闻专线文章组成。这些文章来自 47 个国家，不包括美国的新闻媒体，以确保重点不在西方并关注最近的语言使用。数据集包括对诸如 COVID-19 大流行等重大事件的参考，为评估在 2020 年之前数据（2014 嵌入）训练的嵌入如何推广到近期背景提供了独特的机会。
- 新兴和稀有实体识别（WNUT 17）：这个由 Derczynski et al. (2017) 提供的 6 类数据集包括来自 Youtube、Twitter 和 Reddit 等平台上的用户生成文本中的实体。这些实体更为稀有并且经常是未见过的，这使得它们即使对人类来说在嘈杂文本中进行标注也具有挑战性。此数据旨在提高在动态、真实世界文本场景中检测和分类不常见实体的能力。

这些数据集的训练子集被用于以默认参数训练模型，且没有进行嵌入微调或字符语言建模。开发和测试子集被用于评估。我们报告每个实体和每个标记的 F1 分数。

[2] https://stanfordnlp.github.io/stanza/ner.html

Table 4:

直接评估：类比 XMATHX 类似性

	Analogy		Word Similarity		
Embedding	Google	MSR	WordSim353	SimLex999	MEN
2014 50d Wiki/Giga	0.462	0.355	0.448	0.265	0.652
2024 50d Wiki/Giga	0.455	0.329	0.431	0.256	0.637
2014 100d Wiki/Giga	0.631	0.550	0.477	0.298	0.681
2024 100d Wiki/Giga	0.601	0.486	0.455	0.291	0.672
2014 200d Wiki/Giga	0.698	0.595	0.515	0.340	0.710
2024 200d Wiki/Giga	0.696	0.574	0.480	0.326	0.688
2014 300d Wiki/Giga	0.717	0.614	0.544	0.371	0.737
2024 300d Wiki/Giga	0.718	0.594	0.486	0.338	0.690
2024 300d Dolma	0.708	0.623	0.470	0.270	0.651

Table 5: 2014 年和 2024 年嵌入在单词类比数据集 Google 和 MSR 上的准确性，以及在单词相似性数据集 WordSim353、SimLex999 和 MEN 上的斯皮尔曼等级相关系数。

3 结果

我们展示了三个评估指标的结果：更新词典、直接评估和下游评估。

3.1 更新词典

在这里，我们定性地展示了在 2024 年新的嵌入中出现而在 2014 年嵌入中未出现的词汇。新的维基百科和 Gigaword 向量提供了一个更大的词汇表。在 2024 年维基百科和双倍 Gigaword 嵌入与 2014 年维基百科和 Gigaword 嵌入之间，有超过 70 万个新词（不包括数字和包含非拉丁字母的词）。在 2024 年 Dolma 嵌入与 2014 年 Common Crawl 840B 嵌入之间，有超过 50 万个新词（不包括数字和包含非拉丁字母的词）。在表格 7 和 9 中，我们报告了由作者选择的 39 个具有文化、政治和技术性质的新出现的词汇。

3.2 词嵌入评估

我们在表格 5 中报告了 2014 年和 2024 年在词类比和相似性数据集上的嵌入结果。类比任务使用 Google 和 MSR 数据集进行评估，以准确率作为度量。词相似性任务在 WordSim353、SimLex999 和 MEN 上使用斯皮尔曼等级相关系数（ρ）进行评估。

对于类比任务，2024 年的嵌入在 Google 数据集上的表现大致与 2014 年的嵌入相似，但在 MSR 数据集上的表现略微逊色。此外，在这两个数据集中，随着维度大小的增加，无论是 2014 年还是 2024 年的嵌入，准确性都在持续提高。

对于词汇相似性任务，2024 年的嵌入在大多数数据集和维度上表现得与 2014 年的嵌入竞争力相当。两个 2024 年 300 维度的嵌入与 2014 年的嵌入相比，特别是在 SimLex999 上的等级关联性出现下降。这个下降将在讨论部分中进行说明。

Table 6:

2024 年更新的词典（维基百科）		
afrobeats	antiracism	asmr
binance	bipoc	blockchain
brexit	chatbot	clickbait
covid	fyp	cryptocurrency
deepfake	docuseries	doja
doordash	draftkings	rizz
nonbinary	fintech	fortnite
skibidi *	idk	jungkook
latinx	lgbtqia	microaggression
lstm	metoo	microplastic
pickleball	retweet	zelenskyy
teladoc	web3	tiktok
transwoman	girlboss	viserys

Table 7: 与 2014 年 Wiki/Giga 向量相比，Wiki/Giga 嵌入中包含的新词的 39 个词样本。
* 不在 Dolma 向量中

3.3 命名实体识别

我们使用四个数据集（CoNLL-03、CoNLL-PP、Worldwide 和 WNUT 17）对 2014 年和 2024 年的嵌入进行了 NER 任务评估。对于这些数据集，我们报告了每个实体和每个标记的测试 F1 分数。结果汇总在表格 11 、13 和 15 中。

在这四个 NER 数据集上的结果表明，2024 年的嵌入总体上优于 2014 年的嵌入，尤其在时间相关的数据集上有显著提升。对于表格 11 中的 CoNLL 数据集，2024 年的嵌入在

Table 8:

2024 更新词典 (Dolma)		
dinkies	profeminist †	theranos †
chatgpt †	adagrad	databricks †
huggingface	tarboosh †	gamestonk
badbunny	yeet †	patreon †
brainrot	xgboost †	bytedance †
fakenews	periodt	duolingo †
mansplains	pytorch †	absofreakinglutely
squidgame	trumpism †	clapback
highkey	bffr	situationship
cybertruck †	boujee †	alphafold †
glowup	openai †	scikit
bingewatch	tensorflow †	kubernetes †
aapi †	airpods †	deeplearning

Table 9: 39 个 Dolma 嵌入中包含的新词样本与 2014 年 840B 普通抓取向量中的新词样本进行比较。
† 也在 Wiki/Giga 向量中

CoNLL-03 上表现相当，但在现代化的 CoNLL-PP 版本上显示出明显优势，其中 2024 年的 50d Wiki/Giga 嵌入在每个实体得分上取得了最高的相对改进（83.64 对比 81.58）。在表格 13 中的 Worldwide 数据集上，2024 年的嵌入在每个实体和每个标记的 F1 分数上均表现出一致的改进，2024 年 50d Wiki/Giga 嵌入在每个实体 F1 分数上达到 84.64，显著优于其 2014 年版本的 82.1 分。在表格 15 中的具有挑战性的 WNUT17 数据集上，与 CoNLL 和 Worldwide 相比整体性能显著下降（F1 分数在 30 到 40 之间），但 2024 年的嵌入始终优于其 2014 年的版本，2024 年 200d Wiki/Giga 嵌入分别在每个实体和每个标记 F1 分数上达到了 37.46 和 35.63，这是其中的最高得分。

在所有三个数据集中，2024 嵌入在较低维度上表现出最显著的提升，特别是在 50 维度时，2024 和 2014 嵌入之间的差异最大。虽然较高维度（200 维和 300 维）获得了最佳绝对 F1 分数，但在这些维度上，2024 和 2014 嵌入之间的相对提升不太明显。这些趋势表明，2024 嵌入在与其训练数据时间上对齐的数据集上表现与或优于 2014 嵌入，同时在更能反映当代语言使用和文化趋势的现代、语言多样性较大的数据集上表现出显著提升。

4 讨论

2024 年词嵌入引入了多样化的新词汇，反映了过去十年间的文化、技术和语言变化，包括与全球重大事件相关的词汇（“疫情”和“脱欧”），现代俚语（“脑残”和“有理”），新兴技术（“chatgpt”和“区块链”），以及流行产品（“airpods”）。虽然许多俚语词汇在 Wikipedia & Gigaword 训练数据中缺失，因为现代俚语从社交媒体转移到 Wikipedia 需要时间，但多样化的 Dolma 数据集通过捕捉非正式和对话语言进行了补偿。缩略词（“idk”）、语言融合词（“absofreakinglutely”）和派生词（“转推”）的加入揭示了在数字通信和社交媒体推动下词汇形成模式的演变。通过捕捉这些转变，2024 年的嵌入与当前的日常用语保持一致，为下游任务提供了实际好处，例如减少现代数据集中词汇缺失的问题，同时对这些新词的进一步探索为语言学和社会学研究提供了机会。

在表 5 中报告的类比任务中，2014 年的嵌入表现稍好，与谷歌数据集的差异很小（准确率在 0.01 到 0.03 之间），但在 MSR 数据集上的差距较大，尤其在维度较低的情况下，100d 时的差异达到 0.07，而 50d 时为 0.03。2014 年的 50d 正确预测了大约 1100 个实例，而 2024 年的 50d 预测错误。这些错误中大约有一半与地理相关（例如，开罗，埃及：伯尔尼，_）。另一半主要是 2024 年的嵌入预测了金答案的同义词。例如在 simple, simpler: cold, _ 中，2024 年的嵌入预测的是“cooler”而不是“colder”。尽管 2024 年的 50d 嵌入犯了这些错误，2014 年的嵌入也犯了一些常见错误。大约有 900 个实例是 2024 年的 50d 正确预测，而 2014 年的 50d 预测错误。这些错误大约一半是地理相关，另一半是使用同义词（在 write, writes: think, _ 中使用“knows”而不是“thinks”）。在地理相关的错误中，2024 年的嵌入确实“知道”正确答案，但最近的邻居不是正确答案。例如，尽管 2024 年的嵌入知道伯尔尼在瑞士，但有时最近的邻居并不是瑞士。这可以在 2024 年的嵌入中看到，喀布尔，阿富汗：伯尔尼，_ 不正确，但开罗，埃及：伯尔尼，_ 正确。在 2014 年的嵌入中也看到了相同的现象。MSR 数据集的错误中更多的实例使用了在句法单词类比中的同义词。

因此，2024 和 2014 年的嵌入在类比任务上的表现大致相同。这样的相同表现是预期之中的，因为这些任务主要依赖于句法结构和常用词汇，而这些在过去十年中相对稳定。

对于词相似性任务，不同词向量在斯皮尔曼等级相关系数中表现出差异。为了将这些差异置于背景中，我们考察了在词向量预测与人类评估（缩放至 −1 – 1，而不是 0–10）之间存在 0.3 差异的词对。对于 MEN 数据集，有 70 个词对的 2024 年 300d 词向量预测在阈值内，而 2014 年 300d 词向量则偏离阈值。相反地，

Table 10:

命名实体识别分数:CoNLL

Embedding	Per Entity (2003)	Per Entity (PP)	Per Token (2003)	Per Token (PP)
2014 50d Wiki/Giga	89.52	81.58	89.30	80.77
2024 50d Wiki/Giga	89.74	83.64	89.43	82.23
2014 100d Wiki/Giga	90.62	84.40	90.41	82.73
2024 100d Wiki/Giga	90.34	83.53	90.16	82.04
2014 200d Wiki/Giga	90.88	84.21	90.91	82.89
2024 200d Wiki/Giga	90.69	84.36	90.46	82.75
2014 300d Wiki/Giga	90.60	84.25	90.43	82.70
2024 300d Wiki/Giga	90.72	84.06	90.50	82.74
2024 300d Dolma	90.05	85.14	90.12	83.69

Table 11: 2014 年和 2024 年嵌入在 CoNLL-03 (Tjong Kim Sang and De Meulder, 2003) 和 CoNLL-PP (Liu and Ritter, 2023) 上的每个实体和每个标记的平均测试 F1 分数。我们使用了在 CoNLL-03 上训练的未进行嵌入微调的 Stanford 的 Stanza NER 模型。

Table 12:

命名实体识别分数:全球

Embedding	Per Entity	Per Token
2014 50d Wiki/Giga	82.1	81.04
2024 50d Wiki/Giga	84.64	83.88
2014 100d Wiki/Giga	85.29	84.58
2024 100d Wiki/Giga	85.55	84.25
2014 200d Wiki/Giga	84.41	83.53
2024 200d Wiki/Giga	85.68	84.92
2014 300d Wiki/Giga	84.53	83.89
2024 300d Wiki/Giga	84.89	84.11
2024 300d Dolma	86.23	85.27

Table 13: 对 2014 年和 2024 年嵌入在 Worldwide 数据集 (Shan et al., 2023) 上的每个实体和每个标记的平均测试 F1 分数。我们使用了未进行嵌入微调的 Stanford Stanza NER 模型，该模型在 Worldwide 数据集上训练。

Table 14:

命名实体识别得分:WNUT17

Embedding	Per Entity	Per Token
2014 50d Wiki/Giga	32.95	31.05
2024 50d Wiki/Giga	35.65	33.10
2014 100d Wiki/Giga	36.48	33.39
2024 100d Wiki/Giga	36.33	34.23
2014 200d Wiki/Giga	35.68	33.31
2024 200d Wiki/Giga	37.46	35.63
2014 300d Wiki/Giga	36.64	33.73
2024 300d Wiki/Giga	37.17	33.33
2024 300d Dolma	39.44	34.22

Table 15: 在 WNUT17 数据集上，2014 和 2024 嵌入的每个实体和每个标记的平均测试 F1 分数 (Derczynski et al., 2017) 。我们使用了斯坦福的 Stanza NER 模型，该模型在 WNUT17 数据集上训练时没有进行嵌入微调。

有 197 个词对的 2024 年词向量偏离，而 2014 年词向量保持在阈值内。对于 SimLex999 数据集，这些数字分别是 14 和 43，对于 WS353 数据集，则是 10 和 24。

比较两种在 MEN 上的嵌入模型，2024 嵌入在捕捉近义词和上下位关系（例如：cemetery – graveyard, stair – staircase, ice – snow, sea – water）方面表现优异。在这些情况下，2024 模型与数据集中高相似度评分紧密一致，而 2014 嵌入则倾向于低估它们的相似性。这表明 2024 嵌入可能在聚类共享核心、近似等价含义或清晰的部分-整体关系的术语方面做得更加精确。

然而，也存在 2014 年模型表现优于 2024 年模型的情况。这些通常涉及较松散的主题或分布关联，比如颜色词（蓝色 – 红色，紫色 – 黄色）或数据集因类别或上下文的缘故将其视为相似的常见日常物品（例如，水仙花 – 郁金香，水坑 – 飞溅，鸡肉 – 羔羊，土豆 – 西红柿）。在这些情况下，2024 年模型要么高估要么低估了相似性，而 2014 年模型更准确地捕捉到了这些广泛的关系。

在 WS353 数据集上，我们观察到与先前比较中出现的相似模式。2024 的嵌入更倾向于更精确地捕捉近义词和明显相关的类别对。例如，像 gem – jewel 或者 coast – shore 这样的配对有较高的真实相似性分数，而 2024 与之密切对齐，而 2014 的嵌入则低估了这些紧密连接。其他近义词或上位词-下位词的例子也是如此，比如 magician – wizard 或者 lobster – food，这表明 2024 更可靠地编码了强烈、直接的词汇关系。

相反，2014 年的模型在较松散或更具情境的关联方面表现出色，其中概念是功能上或主题上相关的，而不是严格同义的。例如，在 WS353 中，飞机–汽车或能源–实验室获得了中等偏高的相似度评分，反映出它们共享一个总体类别或背景（运输、科学研究）。2024 年的模型低估了这些关系，暗示它没有捕捉到某些广泛或较松散的概念连接。

在 SimLex-999 上，我们再次看到两个模型在不同领域表现出色。2024 的嵌入在处理一些共享领域或有适度概念重叠的词对时表现

Table 16:

实体识别 WNUT-17 混淆矩阵

t\p	O	CORP	CREATIVE-WRK	GROUP	LOC	PE
O	58	8	−61	−1	−18	8
CORP	0	5	0	2	−6	−
CREATIVE-WRK	26	−3	−39	7	8	−
GROUP	1	1	−2	4	1	−
LOC	−1	1	−7	−4	16	−
PER	2	5	−7	15	−5	−1
PROD	186	10	1	7	3	1

Table 17: 2024/2014 300d Wiki/Giga 混淆矩阵在 WNUT-17 测试集上的差异

Table 18:

实例命名实体识别标注

Sentence	2024	2
His repeated questioning of the system has prompted the Supreme Court to open an investigation into Bolsonaro . (Worldwide)	PER	I
Nationwide, COVID-19 infections in United States are at their peak with an average of 193,863 new cases reported each day over the past week … (CoNLL-PP)	MISC	C
Man Finna bring me a 🍕 up to my job. 😎 [3](WNUT-17)	O	I

Table 19: 来自 NER 数据集的示例句子中，从 2024 模型对加粗词语的正确标记显示在与 2014 模型的标记进行比较。

Table 20:

命名实体识别 CoNLL-PP 混淆矩阵

t \ p	O	LOC	MISC	ORG	PER
O	−17	−3	5	18	−3
LOC	−3	−19	2	21	−1
MISC	12	−8	8	1	−12
ORG	13	−21	13	2	−7
PER	22	−3	−4	−12	−3

Table 21: 在 CoNLL-PP 测试集上 2024/2014 的 300d Wiki/Giga 混淆矩阵差异

困难，例如疾病–感染或河流–海。虽然这些词对不是同义词，但 SimLex-999 对它们的相似性评分相当高，而 2024 对它们的低估程度比 2014 更大。类似地，2024 有时低估了近义动词（例如，应得–赚取，保持–保留，替换–恢复），这表明即使这些词对在语义上有显著接近度，它也可能在微妙的词汇重叠方面无法跟踪。

此外，在 2024 年 Dolma 300 维嵌入和人工注释的十大偏差中，所有十个例子都是模型在反义词之间具有高相似性（例如，agree – argue）。对于 2024 年 Wiki/Giga 300 维向量，10 个例子中有 6 个是高度相似的反义词，另外 4 个是同义词，但没有给出足够高的相似性（例如，creator – maker）。

总体来看，在 MEN、WS353 和 SimLex-999 上，出现了一种一致的模式：2024 年的嵌入特别擅长捕捉强烈的、直接的关系（近义词、上下位链接和部分-整体关系），但有时低估了更具上下文的联系。相比之下，2014 年的嵌入往往在较松散的主题性或分布性关系上表现较好（例如，颜色词、共享领域、功能重叠），但偶尔会在一些最明显的、高相似度的配对上失误，例如严格的同义词。

与直接评估结果对比，其中不清楚哪个嵌入表现得更好，命名实体识别（NER）评估揭示了 2024 年嵌入在更新和领域外的数据集上的表现优于 2014 年嵌入。在经典的 2003 CoNLL 数据集上，结果几乎相同，这表明 2014 年嵌入依然足以应对与其原始训练领域和时代相似的语料库。然而，在 CoNLL-PP 以及更新的 Worldwide 和 WNUT-17 数据集上，2024 年嵌入表现出持续的改进。

对混淆矩阵的检查帮助揭示了为什么 2024 嵌入在更近期的数据上表现更好。例如，表格 21 突显了 2024 模型如何减少将新出现的实体误分类到 O（无实体）类别或错误类型（例如，将 “COVID-19” 标记为 O）的情况。相反，更新后的嵌入捕捉当代术语，因此更有可能为它们分配正确的标签，通常是 MISC，而

Table 22:

NER 全球混淆矩阵 t \ p	O	LOC	MISC	ORG	PER
O	−19	24	68	−52	−21
LOC	14	−29	24	−2	−7
MISC	6	−48	138	−70	−26
ORG	32	−36	63	−30	−29
PER	23	1	−6	−2	−16

Table 23: 2024/2014 300d Wiki/Giga 混淆矩阵在全球测试集上的差异

不是将它们留为未标记状态或与 LOC、ORG 或 PER 混淆。同样地，在表格 23 中，我们看到 2024 模型不太可能将非西方人名与地名混淆，这种变化反映了对这些实体在较新语料库中有更多的曝光或更好的表示。

这些模式在表格 19 中的例句中显而易见。第一句来自 Worldwide，2024 年的嵌入能够将“Bolsonaro”标记为一个人，而 2014 年的嵌入将其标记为一个地点。第二个例子来自 CoNLL-PP，显示 2024 年正确标记了 COVID-19，而 2014 年的嵌入没有标记。这是一个应该正确标记出最近使用的单词的例子。最后一个例子来自 WNUT-17，展示了这个数据集对标记来说是多么困难。尽管如此，2024 年的模型能够识别“finna”是俚语并且没有标记它，而 2014 年的模型则将其标记为一个人。这是 2024 年的嵌入更好地捕捉到口语词汇的例子。

总体而言，我们发现 2024 年的嵌入比 2014 年的嵌入更好地表示了当前的语言使用。在存在时间依赖性的情况下（例如，聊天机器人、NER 标注器等），应使用 2024 年的嵌入。

5 结论

我们提出了新的 GloVe 词向量，介绍了两个在更新的维基百科、Gigaword 和 Dolma 子集语料库上训练的向量集。这些词向量提供了有价值的新词汇，反映了过去十年中的文化和技术变化，并为语言学家提供了丰富的数据以分析英语的发展，特别是在社交媒体影响力不断上升的背景下。

在词类比任务中，新的嵌入表现与 2014 年的嵌入相当，展示了相同的结构和核心语义表现力。对于词相似度任务，维度更高的 2024 年嵌入偶尔会高估反义词之间的语义相似性。尽管如此，2024 年的嵌入在时效性依赖的 NER 数据集上表现出明显优势，例如非西方导向的新闻稿数据。

[3]Thank you Twemoji for the emojis!

这些发现强调了更新词向量以跟上语言和文化变化的重要性。2024 年的词向量代表了现代语言建模方面的一项重要进展，提供了与当代使用更好对齐的工具。它们能够捕捉近期的文化、技术和语言变化，这使得它们在以人为中心的自然语言处理应用中特别有价值，例如改进聊天机器人的互动和设计能够适应多样化和不断发展的用户需求的系统。

References

Giuseppe Attardi. 2015. WikiExtractor. https://github.com/attardi/wikiextractor.

Elia Bruni, Nam Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. J. Artif. Int. Res. , 49(1):1–47.

Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. 2017. Results of the WNUT2017 shared task on novel and emerging entity recognition. In Proceedings of the 3rd Workshop on Noisy User-generated Text , pages 140–147, Copenhagen, Denmark. Association for Computational Linguistics.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. ACM Transactions on Information Systems - TOIS , 20:406–414.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. Computational Linguistics , 41(4):665–695.

Stanisław Jastrzebski, Damian Leśniak, and Wojciech Marian Czarnecki. 2017. How to evaluate word embeddings? On importance of data efficiency and simple supervised tasks. ArXiv preprint arXiv:1702.02170.

Shuheng Liu and Alan Ritter. 2023. Do CoNLL-2003 named entity taggers still work well in 2023? In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) , pages 8254–8271, Toronto, Canada. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. ArXiv preprint arXiv:1301.3781.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies , pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2011. English Gigaword Fifth Edition . Linguistic Data Consortium, Philadelphia, PA. LDC2011T07.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) , pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations , pages 101–108, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research , 21(140):1–67.

Alexander Shan, John Bauer, Riley Carlson, and Christopher Manning. 2023. Do "English" named entity recognizers work well on global Englishes? In Findings of the Association for Computational Linguistics: EMNLP 2023 , pages 11778–11791, Singapore. Association for Computational Linguistics.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Harsh Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Pete Walsh, Luke Zettlemoyer, Noah A. Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. 2024. Dolma: An Open Corpus of Three Trillion Tokens for Language Model Pretraining Research. arXiv preprint arXiv:2402.00159 .

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 , pages 142–147.

Andrea Vallebueno, Cassandra Handan-Nader, Christopher D Manning, and Daniel E. Ho. 2024. Statistical uncertainty in word embeddings: GloVe-V. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing , pages 9032–9047, Miami, Florida, USA. Association for Computational Linguistics.