arXiv:2301.01269v1 [cs.CL] 3 Jan 2023

平均值还不够：多语言评估的注意事项

MatúšPikuliak 和 **Marián Šimko**Kempelen
智能技术研究所 firstname.surname@kinit.sk

摘要

本文是一篇探讨多语言评估问题的立场论文。使用简单的统计数据（如语言平均性能）可能会在评估方法中引入语言偏差，使其偏向占主导地位的语系。我们认为，要检测多语言结果中的此类偏差，需要借助比较语言学进行定性分析。我们的案例研究表明，已发表的研究结果确实可能存在语言偏差；我们还展示了基于URIEL类型学数据库的可视化方法能够检测出这种偏差。

1 引言

自然语言处理研究中的语言多样性正在不断增长（Joshi等人，2020年；Pikuliak等人，2021年），这得益于各种多语言技术的进步，例如机器翻译（Arivazhagan等人，2019年）、多语言语言模型（Devlin等人，2019年；Conneau和Lample，2019年）、跨语言迁移学习（Pikuliak等人，2021年）或语言无关表征（Ruder等人，2019年）。如今，我们可以为许多任务创建性能良好的多语言方法。在处理多语言方法时，我们需要能够评估它们的实际表现，也就是说，它们在广泛且类型多样的语言上的效果如何。请看图1(a)所示的两种方法。如果不考虑具体语言，方法 A 似乎更好。它在大多数语言上的结果更好，平均性能也更优。然而，方法 A 表现更好的三种语言实际上全是非常相似的伊比利亚语言，而第四种语言则属于印度-伊朗语族。那么，方法 A 真的更好吗？还是只在伊比利亚语言上更好？实践中，人们往往会直接采用简单平均值，而不考虑所选语言的语言多样性，尽管许多语料库和数据集都偏向于历史上占主导地位的语言及语系。

此外，随着语言数量不断增加，要注意到此类现象会变得越来越困难。请看表 1中两组结果的比较。面对41种语言，即使具备必要的语言学知识，要从认知上发现语言与其结果之间的各种关系也十分困难。

在这篇立场论文中，我们认为，仅使用平均值等简单统计量来比较多语言方法并非最佳实践。常用的简单评估规程可能会使研究偏向占主导地位的语言，进而损害历来处于边缘地位的语言。相反，我们提议将纳入语言类型学（Ponti 等人，2019年）和比较语言学考量的定性结果分析作为额外的合理性检查。我们认为，这是研究工具箱中经常被忽视的一种工具，应当更广泛地加以运用，以确保我们能够恰当地解读多语言评估的结果，并发现各种语言偏差和问题。除了这番讨论（我们认为其本身就是一项贡献）之外，我们还提出了一种基于URIEL类型学数据库（Littell等人，2017年）的可视化方法，作为此类定性分析的示例，并展示了它能够发现已发表结果中的语言偏差。

2 相关工作

自然语言处理中的语言偏见。 Bender(2009年)曾提出，主要由英语评估驱动的研究会偏向这种语言，而且未必具有充分的语言独立性。即使近年来，诸如 *word2vec*和 字节对编码等流行技术，也被证明在形态丰富的语言上表现较差（Bojanowski等人，2017年；Park等人，2020年）。同样，跨语言词嵌入通常以英语

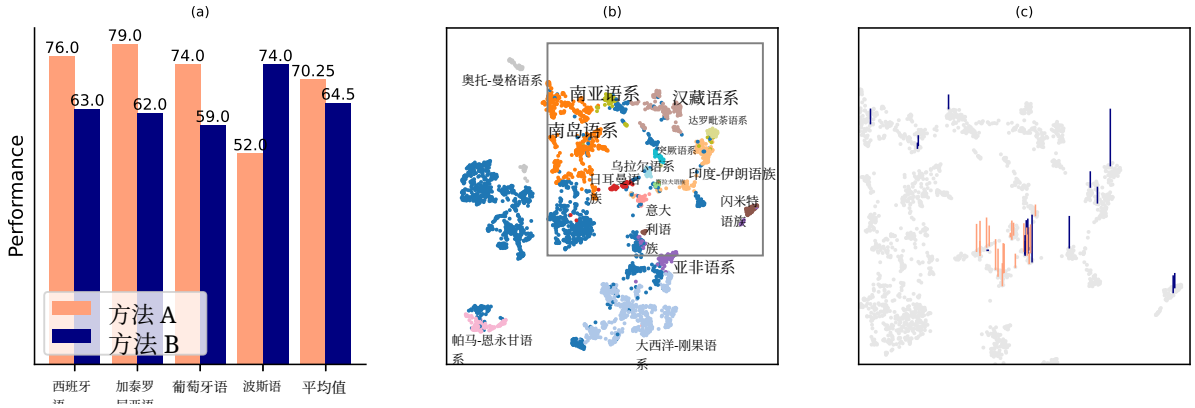

图 1：*(a)* 对比两种方法在语言不平衡集合上的表现。*(b)* 对采用不同语言家族颜色编码的 URIEL 语言进行可视化。*(c)* 对比 Rahimi 等人 提出的两种方法。这里使用与*b*相同的语言地图，但视图经过了缩放。

语言	bul	ben	bos	cat	ces	dan	deu	ell	eng	spa	est	pes	fin	fra	heb	hin	hrv	hun	ind
方法 A	54	60	77	79	72	79	64	34	57	76	71	52	69	73	46	58	77	69	61
方法 B	61	70	63	62	62	62	58	61	47	63	64	74	67	57	53	68	61	59	67

语言	ita	lit	lav	mkd	zlm	nld	nor	pol	rus	slk	slv	alb	tgl	tur	ukr	vie	**AVG**
方法 A	76	75	67	48	63	78	77	77	36	76	76	76	57	67	49	48	64.5
方法 B	60	62	68	67	66	59	65	61	53	62	64	69	66	61	60	55	62.1

表1：两种方法的比较，出自 Rahimi 等人（2019年）。

作为默认枢纽语言，尽管这可能会损害许多语言的表现（Anastasopoulos and Neubig，2020年）。或许，如果研究实践不那么以英语为中心，其他方法和技术就会流行起来。我们的工作与这类问题密切相关。我们表明，语言选择不均衡的多语言评估可能会造成类似的现象。

基准测试。近年来，在自然语言处理社区中，使用基准测试的做法受到了诸多审视。人们认为，基准评估会助长虚假数据过拟合（Kavumba 等，2019年）、鼓励指标操纵（Thomas 和 Uminsky，2020年），或使研究偏离通用的类人语言智能（Linzen，2020年）。同样，基准测试也因主要关注性能、却忽略了其他几项重要属性（例如预测成本或模型鲁棒性）而受到批评（Ethayarajh 和 Jurafsky，2020年）。研究尤其表明，平均值在鲁棒性方面存在若干问题，而采用成对实例评估可以解决这些问题（Peyrard 等，2021年）。为了解决这些问题，一些基准测试不再使用聚合分数，而是同时报告多项指标，将结果的解读留给读者。Gehrmann 等（2021年）就是这样一种基准测试，它提出使用可视化来帮助解读——释结果。在这项工作中，我们也使用可视化来达到类似的效果。

3 多语言评估策略

比较涉及大量语言的多语言方法时，要跟踪语系、书写系统、类型学属性等各种语言学方面的信息，在认知上颇为困难。研究人员通常会采用各种简化策略：

聚合指标。聚合指标（例如平均性能，或某种方法取得最佳结果的语言数量）能够提供一些信息，但正如我们在图1(a)中所示，它们可能无法呈现全貌。聚合结果会使我们失去有关单个语言和语系的重要信息。常用统计数据通常不会考虑底层的语言多样性。这可能导致不良现象，例如偏袒优势语系。聚合指标所编码的价值观，可能与我们想表达的价值观不一致。平均值体现了一种功利主义世界观，而采用最低性能则可视为一种优先主义方法（Choudhury 和 Deshpande，2021年）。尽管分析

指标所编码的价值观是迈向更公平评估的一步，但仍未能呈现结果中更细致的差异。

不同群组的聚合指标。另一种选择是针对特定语言家族或群组计算统计数据。这些做法朝着正确的方向迈进了一步，因为它们能提供更细致的图景，但仍然存在一些问题。目前尚不清楚应选择哪些语言家族。例如，我们应该对所有印欧语系语言取平均值，还是按斯拉夫语族、日耳曼语族等语族分别取平均值？最终，这种选择带有主观性，而不同的选择可能会呈现不同的结果。此外，跨语言家族进行聚合，仍可能掩盖这些家族内部的差异。按照可用数据集的规模对语言分组（例如低资源与高资源）可以揭示模型如何应对数据匮乏，但这些群组在语言学方面仍可能不均衡。

平衡语言抽样。另一种选择是构建一个语言学上均衡的多语言数据集。这一过程称为语言抽样(Rijkhoff 等人，1993年；Miestamo 等人，2016年）。实际上，这意味着为每个语系选取少量代表性语言。通过控制每个语系所纳入的语言数量，就能解决优势语系占比过高的问题。然而，决定纳入哪些语系，再从这些语系中挑选语言，仍然是一个带有主观性的过程。不同语系及其语族的多样性程度可能各不相同。不同的选择可能偏向不同的语言属性，所得结果也可能有所差异。此外，目前也不清楚给定的选择有多全面，即涵盖了多少语言多样性。一些现有研究提到了各自的选择标准：Longpre 等人(2020年统计所选语言覆盖的使用者人数，Clark 等人(2020年采用一组经过选择的类型学属性，Ponti 等人(2020年)采用所谓的多样性语言抽样。公布这些标准，有助于我们在未来开展事后分析，评估这些标准的效果。

定性分析 在本文中，我们认为，定性分析是一种经常被忽视却不可替代的评估技术。在接下来的部分中，我们将介绍一个案例研究，展示如何进行定性分析。

4 案例研究：通过可视化进行定性分析

本节将展示如何利用基于 URIEL 类型学数据库的可视化技术，对多语言结果进行定性分析。我们将说明，借助这一技术，我们可以（1）揭示结果中的语言偏差；（2）理解涉及数量不少、且差异显著的语言的结果。作为案例研究，我们考察Rahimi 等人（2019年）的研究结果。我们的目的并非评估该论文中的特定方法，而是展示基于语言学知识的分析如何帮助研究人员深入理解结果。我们分析这篇论文的结果，并非要对其提出批评，而是因为这是一篇写得很好的论文，确实尝试使用差异显著的多种方法，对数量不少的语言进行多语言评估。我们发现的语言偏差在论文中已有部分讨论。这里，我们仅展示如何借助恰当的可视化，有效地进行定性分析并揭示这些偏差。附录A还展示了对另一篇论文（Heinzerling 和 Strube，2019年）进行的类似分析，其中可以看出语言偏差。

我们使用 URIEL，这是一个语言类型学数据库，包含针对 3718 种语言的 289 项句法和音系二元特征。我们采用 UMAP 特征降维算法（McInnes 和 Healy，2018）创建二维语言类型学空间。该地图如图1 (b)所示。该地图具有交互性，支持动态筛选语言和语系，并可查看单个语言及其属性。[1] 每个点代表一种语言，图中选定的语系以不同颜色标示。尽管用于降维的 URIEL 特征不包含有关语系的信息，但亲缘关系接近的语言自然会在我们的可视化中形成簇。图中也呈现了某些地理关系，例如，苏丹语和乍得语虽属于不同语系，却形成相邻的簇。这体现了按照区域和宏区域对语言进行分类的语言学传统。这表明，我们的可视化能够捕捉语言的语系内相似性和语系间相似性，因此适用于我们的应用场景。

我们在这张语言地图(2019年) 上展示了Rahimi 等人 使用维基百科——

[1]代码可在 GitHub 上获取

语料库进行命名实体识别，并比较了针对 41 种语言的各种跨语言迁移学习算法。他们使用的语言集合并不均衡，其中三大优势语系——日耳曼语族、意大利语族和斯拉夫语族——占所有语言的 55%。关于这篇论文的更多详情，请参阅附录 A 。我们使用 URIEL 地图，展示表 1中涵盖的全部 41 种语言上两种方法的比较。在图 1 (c) 中，我们比较了两种方法——方法 *A* ——使用多种源语言的跨语言迁移学习方法（平均性能 64。5），以及看似表现较差的 方法 *B*——不采用任何形式的跨语言监督进行低资源训练（平均性能 62。1）。我们使用同一张 URIEL 地图，并以彩色柱状图叠加显示两种方法的相对性能。这张地图上的橙色柱表示方法 *A* 表现更好，蓝色柱则表示方法 *B*表现更好。每根柱子的高度表示两种方法之间相对性能差距的大小。也就是说，较高的橙色柱表示 *A*占优，较高的蓝色柱表示 *B*占优。

现在我们可以清楚地看到，彩色列的位置呈现出某种规律。以平均值作为评估指标时，方法 *A* 总体上似乎更好。但我们可以看到，它只在一个特定的语言簇中表现更好——即橙色列所在的语言簇。这些语言都属于彼此相关的欧洲语言。它们大多属于日耳曼语族、意大利语族或斯拉夫语族，少数例外则是不属于印欧语系、但地理位置相邻的语言，例如匈牙利语。另一方面，所有非欧洲语言实际上都更倾向于方法 *B*。这些是散布在其余空间中的蓝色列，代表阿拉伯语（闪米特语族）、汉语（汉藏语系）或泰米尔语（达罗毗荼语系）等语言。

这揭示了平均值掩盖的关于这两种方法的一个重要事实。跨语言监督似乎表现更好，但它仅在占主导地位的相似语言簇中表现更好，因为跨语言监督在该语言簇中更可行。其他语言实际上更倾向于采用单语低资源学习，因为它们无法轻易地从其他语言中学习。在这种情况下，平均值高估了跨语言学习对非欧洲语言的价值。这种高估可能会对这些语言造成损害。

我们还可以看到，也存在一些例外——橙色簇中的蓝色柱。这些例外包括希腊语、俄语、马其顿语、保加利亚语和乌克兰语——它们都是使用非拉丁文字的印欧语系语言。在这种情况下，不同的书写系统可能是造成额外语言偏差的原因。仅仅查看结果表可能很难注意到这一模式，但在这里，我们可以快速识别出这些语言是离群项，随后便很容易意识到它们的共同点。

请注意，我们并不认为大多数论文中都会出现这种程度的语言偏差。我们特意从这篇论文中挑选了这些特定方法，因为它们展示了结果中的语言偏差最为明显的情况。这主要是由于维基百科上的语言选择不均衡，而且从某种意义上说，将跨语言监督与低资源学习进行比较并不公平。

5 结论

自然语言处理中的多语言现象日益普遍，但方法实践有时却未能跟上（Artetxe 等人，2020年；Keung 等人，2020年；Bender，2011年）。缺乏恰当的评估方法，进步必然举步维艰。在这项工作中，我们主张有必要进行定性结果分析，并展示了如何利用此类分析和交互式可视化来改进评估。在我们的案例研究中，我们成功揭示了已发表结果中的语言偏差。

考虑到机器学习和自然语言处理中的实践，将多语言方法的性能简化为一个数字或许很诱人。然而，我们认为，多语言评估的复杂性并不能如此轻易地简化。需要考虑的维度太多，自然语言处理研究者应当理解这些维度。我们认为，适当掌握各种语言学领域的知识（如类型学或比较语言学），对于正确理解多语言结果并进行恰当的定性分析至关重要。我们主张，定性分析是一种常被忽视的结果分析方法，应当更加广泛地采用，以避免我们对研究结果的语言学意义产生各种误解。

6 伦理考量

当前自然语言处理研究的大部分内容都只聚焦于少数几种语言。某些

语言使用者群体因数据匮乏而被落在后面。我们相信，本文可能会产生积极的社会影响。本文聚焦于这些边缘化语言及其社群所面临的问题。遵循我们的建议，可能会使研究界和产业界的多语言评估更加多元、公平。这或许进而能够带来更好的模型和应用，并最终改善一些人的生活质量。

致谢

本研究部分得到了 DisAi 项目的资助，该项目由欧盟 Horizon Europe 计划根据资助协议编号 101079164提供资金支持。

参考文献

Antonios Anastasopoulos 和 Graham Neubig。2020年。所有跨语言嵌入都应该说英语吗？载于计算语言学协会第*58*届年会论文集，第8658–8679页，线上。计算语言学协会。

Naveen Arivazhagan、Ankur Bapna、Orhan Firat、Dmitry Lepikhin、Melvin Johnson、Maxim Krikun、Mia Xu Chen、Yuan Cao、George F. Foster、Colin Cherry、Wolfgang Macherey、Zhifeng Chen 和 Yonghui Wu。2019年。真实场景下的大规模多语种神经机器翻译：发现与挑战。*CoRR*，abs/1907.05019。

Mikel Artetxe、Sebastian Ruder、Dani Yogatama、Gorka Labaka 和 Eneko Agirre。2020年。呼吁提高无监督跨语言学习的严谨性。载于计算语言学协会第*58*届年会论文集，第7375–7388页，线上。计算语言学协会。

Emily M. Bender. 2009年。语言学上的天真！= 语言类型学为何是自然语言处理不可或缺的一环：语言无关性。载于《*EACL 2009* 语言学与计算语言学互动研讨会论文集：相辅相成、相互冲突还是毫无意义？》第 26–32 页，希腊雅典。计算语言学协会。

Emily M Bender. 2011年。论如何在自然语言处理中实现并评估语言无关性。《语言技术中的语言学问题》，6(3):1–26。

Piotr Bojanowski、EdouardGrave、Armand Joulin 和 Tomas Mikolov. 2017年。利用子词信息丰富词向量。《计算语言学协会汇刊》，5:135–146。

Monojit Choudhury 和 Amit Deshpande。2021年。多语言预训练语言模型在语言公平性方面表现如何？载于第三十五届 *AAAI* 人工智能会议人工智能，*AAAI 2021*，第三十三届创新人工智能应用会议（*IAAI 2021*），第十一届人工智能教育进展研讨会（*EAAI 2021*），线上举行，*2021*年*2*月*2*日至*9*日，第12710–12718页。AAAI出版社。

Jonathan H. Clark、Eunsol Choi、Michael Collins、Dan Garrette、Tom Kwiatkowski、Vitaly Nikolaev 和 Jennimaria Palomaki. 2020年。TyDi QA：面向类型学多样语言的信息寻求型问答基准。《计算语言学协会汇刊》，8:454–470。

Alexis Conneau 和 Guillaume Lample。2019年。跨语言语言模型预训练。载于《神经信息处理系统进展》第*32*卷：*2019*年神经信息处理系统年会，*NeurIPS 2019*，*2019*年*12*月*8*日至*14*日，加拿大不列颠哥伦比亚省温哥华，第7057–7067页。

Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2019年。BERT：用于语言理解的深度双向Transformer预训练。载于计算语言学协会北美分会*2019*年会议论文集：人类语言技术，第*1*卷（长篇与短篇论文），第4171–4186页，明尼苏达州明尼阿波利斯。计算语言学协会。

Kawin Ethayarajh 和 Dan Jurafsky。2020年。效用由用户评判：对自然语言处理排行榜的批评。载于*2020*年自然语言处理实证方法会议论文集（*EMNLP*），第4846–4853页，线上。计算语言学协会。

Sebastian Gehrmann、Tosin P. Adewumi、Karmanya Aggarwal、Pawan Sasanka Ammanamanchi、Aremu Anuoluwapo、Antoine Bosselut、Khyathi Raghavi Chandu、Miruna-Adriana Clinciu、Dipanjan Das、Kaustubh D. Dhole、Wanyu Du、Esin Durmus、Ondrej Dusek、Chris Emezue、Varun Gangal、Cristina Garbacea、Tatsunori Hashimoto、Yufang Hou、Yacine Jernite、Harsh Jhamtani、Yangfeng Ji、Shailza Jolly、Dhruv Kumar、Faisal Ladhak、Aman Madaan、Mounica Maddela、Khyati Mahajan、Saad Mahamood、Bodhisattwa Prasad Majumder、Pedro Henrique Martins、Angelina McMillan-Major、Simon Mille、Emiel van Miltenburg、Moin Nadeem、Shashi Narayan、Vitaly Nikolaev、Rubungo An- dre Niyongabo、Salo mey Osei、Ankur P. Parikh、Laura Perez-Beltrachini、Niranjan Ramesh Rao、Vikas Raunak、Juan Diego Rodriguez、Sashank Santhanam、João Sedoc、Thibault Sellam、Samira Shaikh、Anastasia Shimorina、Marco Antonio So- brevilla Cabezudo、Hendrik Strobelt、Nishant Sub- ramani、Wei Xu、Diyi Yang、Akhila Yerukola 和 Jiawei Zhou。2021年。GEM基准：自然语言生成、其评估与指标。*CoRR*，abs/2102.01672。

Benjamin Heinzerling 和 Michael Strube。2019年。使用上下文和非上下文子词表示进行序列标注：多语言评估。载于 第*57*届计算语言学协会年会论文集，第273–291页，意大利佛罗伦萨。计算语言学协会。

Pratik Joshi、Sebastin Santy、Amar Budhiraja、Kalika Bali 和 Monojit Choudhury。2020年。自然语言处理领域语言多样性与包容性的现状与未来。载于计算语言学协会第*58*届年会论文集，第6282–6293页，线上。计算语言学协会。

Pride Kavumba、Naoya Inoue、Benjamin Heinzerling、Keshav Singh、Paul Reisert 和 Kentaro Inui。2019年。选择合理的备选项时，聪明汉斯可能很聪明。载于自然语言处理中的常识推理第一届研讨会论文集，第33–42页，中国香港。计算语言学协会。

Phillip Keung、Yichao Lu、Julian Salazar 和 Vikas Bhardwaj。2020年。不要使用英语开发集：关于上下文嵌入的零样本跨语言评估。载于*2020*年自然语言处理实证方法会议（*EMNLP*）论文集，第549–554页，线上。计算语言学协会。

Tal Linzen。2020年。我们如何加快实现类人语言泛化的进程？载于计算语言学协会第*58*届年会论文集，第5210–5217页，线上。计算语言学协会。

Patrick Littell、David R. Mortensen、Ke Lin、Katherine Kairis、Carlisle Turner 和 Lori Levin。2017年。URIEL 和 lang2vec：将语言表示为类型学、地理和系统发育向量。载于计算语言学协会欧洲分会第*15*届会议论文集：第*2*卷，短篇论文，第8–14页，西班牙瓦伦西亚。计算语言学协会。

Shayne Longpre、Yi Lu 和 Joachim Daiber。2020年。MKQA：用于多语言开放域问答的语言多样化基准。*CoRR*，abs/2007.15207。

Leland McInnes 和 John Healy。2018年。UMAP：用于降维的统一流形近似与投影。*CoRR*，abs/1802.03426。

Matti Miestamo、Dik Bakker 和 Antti Arppe。2016年。抽样以体现多样性。语言类型学，20(2)：233–296。

Hyunji Hayley Park、Katherine J. Zhang、Coleman Haley、Kenneth Steimel、Han Liu 和 Lane Schwartz。2020年。形态学很重要：多语言语言建模分析。*CoRR*，abs/2012.06262。

Maxime Peyrard、Wei Zhao、Steffen Eger 和 Robert West。2021年。优于平均值：自然语言处理系统的配对评估。载于第*59*届计算语言学协会年会暨第*11*届自然语言处理国际联合会议论文集（第*1*卷：长篇论文），第2301–2315页，线上。计算语言学协会。

Matúš Pikuliak、Marián Šimko 和 Mária Bieliková。2021年。文本处理的跨语言学习：一篇综述。《专家系统与应用》，165：113765。

Edoardo Maria Ponti、Goran Glavaš、Olga Majewska、Qianchu Liu、Ivan Vuli´c 和 Anna Korhonen。2020年。XCOPA：用于因果常识推理的多语言数据集。载于*2020*年自然语言处理实证方法会议论文集（*EMNLP*），第2362–2376页，线上。计算语言学协会。

Edoardo Maria Ponti、Helen O'Horan、Yevgeni Berzak、Ivan Vuli´c、Roi Reichart、Thierry Poibeau、Ekaterina Shutova 和 Anna Korhonen。2019年。语言变异与语言普遍性的建模：面向自然语言处理的类型语言学综述。《计算语言学》，45(3):559–601。

AfshinRahimi、Yuan Li 和 TrevorCohn。2019年。面向命名实体识别的大规模多语言迁移。载于第*57*届计算语言学协会年会论文集，第151–164页，意大利佛罗伦萨。计算语言学协会。

Jan Rijkhoff、Dik Bakker、Kees Hengeveld 和 Peter Kahrel。1993年。一种语言采样方法。《语言研究》。由"语言基金会"赞助的国际期刊，17(1):169–203。

Sebastian Ruder、Ivan Vulic 和 Anders Søgaard。2019年。跨语言词嵌入模型综述。人工智能研究期刊，65：569–631。

Rachel Thomas 和 David Uminsky。2020年。指标的问题是人工智能面临的根本问题。*CoRR*，abs/2002.08512。

A 分析论文的详细信息

在本附录中，我们提供有关所分析论文的补充信息。

A.1 Rahimi 等人

这是我们在正文第4节中用于演示的论文。我们使用了该论文表 4 中报告的结果。该论文使用的语言列于此处的表 2中。可以看出，印欧语系语言明显占据主导地位。该论文列出了 14 种不同的方法。我们在图 2中比较了这些方法的结果。在图中可以看到，各方法的平均结果与非 GIS（德语族—意大利语族—斯拉夫语族）语言的平均结果之间的差异。数字对应于原论文中列出的方法顺序。图 1(c) 中比较的两种方法分别以蓝色和橙色表示。橙色的 *方法 A* 在原论文[tok] 中称为 BEA。蓝色的 *方法 B* 称为LSup。通过这种简单的视角，我们也能看出语言偏差。所有基于跨语言学习的方法，其非 GIS 语言的结果都不如不使用跨语言学习的方法（方法 1 和 2）。然而，这项分析无法取代我们在第 4节提出的可视化。它提供了以 GIS 为中心的视角，但无法捕捉其他偏差来源。例如，它没有显示可视化中观察到的各种离群现象，例如表现得与 GIS 语言相似的乌拉尔语系语言，或表现得与其他斯拉夫语族语言不同的、使用西里尔字母的斯拉夫语族语言。

ISO	语言	亚语族	语系
bul	保加利亚语	斯拉夫语族	印欧语系
bos	波斯尼亚语		
ces	捷克语		
hrv	克罗地亚语		
mkd	马其顿语		
pol	波兰语		
rus	俄语		
slk	斯洛伐克语		
slv	斯洛文尼亚语		
ukr	乌克兰语		
afr	南非语	日耳曼语族	
dan	丹麦语		
deu	德语		
nld	荷兰语		
nor	挪威语		
swe	瑞典语		
cat	加泰罗尼亚语	意大利语族	
fra	法语		
ita	意大利语		
por	葡萄牙语		
rom	罗马尼亚语		
spa	西班牙语		
ben	孟加拉语	印度-伊朗语族	
hin	印地语		
pes	伊朗波斯语		
lit	立陶宛语	波罗的语族	
lav	拉脱维亚语		
ell	希腊语		
alb	阿尔巴尼亚语		
est	爱沙尼亚语		乌拉尔语系
fin	芬兰语		
hun	匈牙利语		
ind	印度尼西亚语		南岛语系
tgl	他加禄语		
zlm	马来语		
arb	标准阿拉伯语		亚非语系
heb	希伯来语		
vie	越南语		南亚语系
tam	泰米尔语		达罗毗荼语系
tur	土耳其语		突厥语系

表 2：Rahimi 等人使用的语言Rahimi 等人。

A.2 Heinzerling 和 Strube

在 Heinzer- ling 和 Strube 的研究中也能看到类似的语言偏差。他们评估了各种表示在词性标注和命名实体识别任务上的表现。在图 3 中，我们比较了采用共享嵌入词汇表的多语言模型的词性标注准确率（平均性能96.6，MultiBPEmb+字符 +微调，原论文中的结果）与不使用迁移监督的简单 BiLSTM 基线（平均性能为 96.4，BiLSTM，原论文中的结果）。橙色柱表示更适合多语言模型的语言，蓝色柱表示更适合基线的语言。在这种情况下，几乎所有橙色柱实际上都对应 GIS 语言。其他语言使用这种方法的结果明显更差，其中大多数实际上更适合不使用跨语言监督的简单基线。这表明了所提出的多语言

ISO	语言	亚语族	语系
dan	丹麦语	日耳曼语族	印欧语系
deu	德语		
eng	英语		
nld	荷兰语		
nor	挪威语		
swe	瑞典语		
bul	保加利亚语	斯拉夫语族	
ces	捷克语		
hrv	克罗地亚语		
pol	波兰语		
slv	斯洛文尼亚语		
fra	法语	意大利语族	
ita	意大利语		
por	葡萄牙语		
spa	西班牙语		
hin	印地语	印度-伊朗语族	
pes	伊朗波斯语		
eus	巴斯克语		孤立语言
fin	芬兰语		乌拉尔语系
heb	希伯来语		亚非语系
ind	印度尼西亚语		南岛语系

表 3：使用的语言Heinzerling 和 Strube。

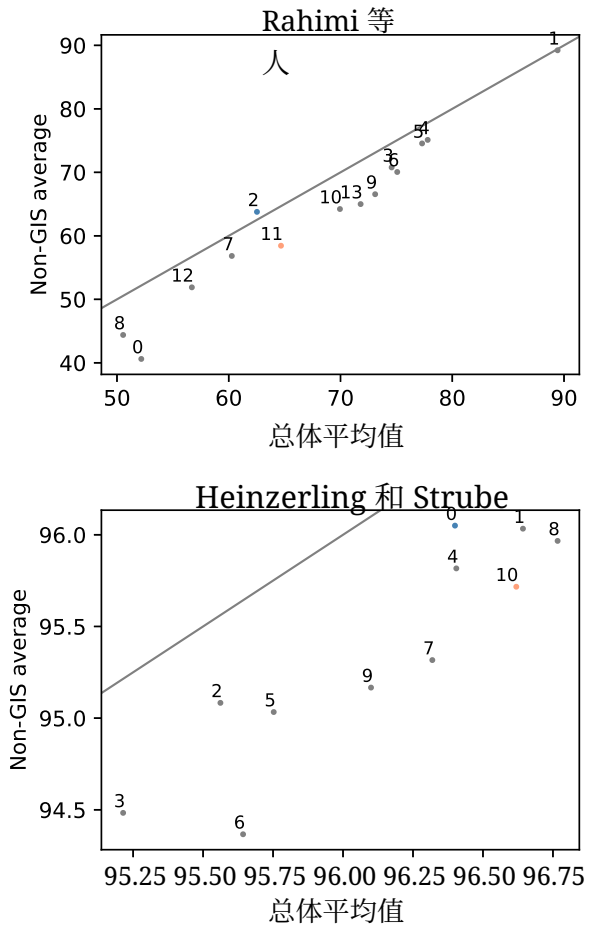

图 2：方法性能比较。图中展示了全局平均值与非 GIS 语言平均值之间的关系。每个点代表论文中的一种方法。

监督方法在离群语言上的局限性。

我们采用其论文中表 5 报告的结果。其使用的语言列于此处的表 3中。同样，我们可以看到 GIS 语言明显占据主导地位。其论文中列出了 11 种不同的方法。我们省略了表 7 中报告的另外 6 种低资源语言的结果，因为在这些语言上仅使用了 11 种方法中的 4 种。我们在图 2中比较了这些方法的结果，与前一篇论文类似。橙色点代表多语言模型，蓝色点代表基线。现在可以看到，BiLSTM 基线实际上是非 GIS 语言上表现最佳的方法。

B 超参数

我们使用 UMAP Python 库[2]，并设置以下超参数：

[2]umap-learn.readthedocs.io

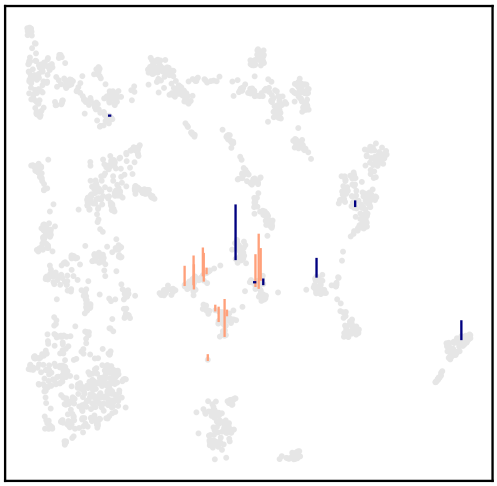

图 3：比较 Heinzerling 和 Strube 提出的两种方法。

- 邻居数量 (n_邻居)：15
- 距离度量（metric）：余弦
- 最小距离（min_dist）：0.5
- 随机种子（随机_状态）：1

C 其他可视化

本节展示了使用 URIEL 语言地图将多语言评估结果可视化的几种其他方式。我们的目标是提出一些额外的技术，供定性分析使用；除了比较本文正文图 1 中使用的两种方法之外，还可以采用这些技术。这里列出的可视化方式并不全面。我们认为，基于用户的需求和要求，利用这种定性分析还可以实现许多其他类型的可视化。

在图 4中，我们展示了如何通过分别呈现每种方法的性能来比较两种以上的方法。我们为三种方法分别绘制了图表，并可以直观地比较它们的性能。可以看出，HSup方法总体表现稳定且出色。LSup的表现较差，但仍相当均衡。最后，BWET 的表现与LSup相近，但可以看出它在某些区域表现不佳，例如图最右侧的语言，其性能明显较差。

在图5 中，我们展示了另一种可视化方式。在这种情况下，我们只需展示每种语言表现最佳的方法。

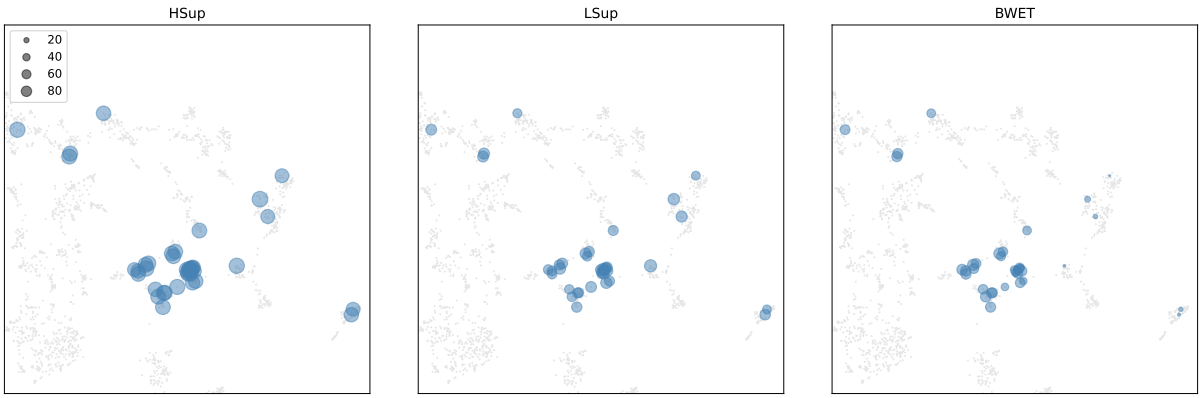

图 4：采用尺寸标示各个语言的方法性能比较。HSup，LSup 和BWET 是（Rahimi 等人，2019年）中报告的方法。

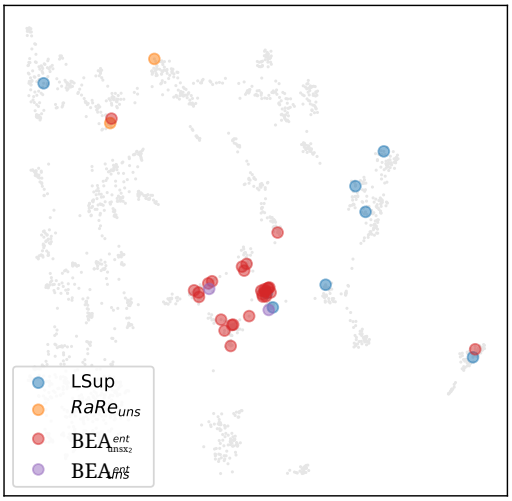

图 5：各种语言中表现最佳的方法。

我们比较了采用跨语言监督和低资源训练（LSup）的方法。在七种方法中，只有四种在至少一种语言上取得了最佳表现，这些方法展示在图中。再次，我们可以看到与之前类似的情况。考虑平均值后，有一种方法（$BEA^{ent}_{uns\times 2}$）表现最佳。然而，从这种可视化中可以看出，它实际上只在欧洲语言这一占主导地位的集群中表现最佳。在其他地方，表现更好的是其他方法。