Scand ScandEval：斯堪的纳维亚自然语言处理基准测试 g

Dan SaattrupNielsen The
Alexandra Institute 哥本哈根，丹
麦 dan.nielsen@alexandra.dk

arXiv:2304.00906v1 [cs.CL] 3 Apr 2023

摘要

本文介绍了一个斯堪的纳维亚基准测试平台 ScandEval，可在斯堪的纳维亚语言的四项不同任务上对任意预训练模型进行基准测试。其中两项任务（语言可接受性和问答）所使用的数据集为全新数据集。我们开发并发布了一个 Python 软件包及命令行界面 scandeval，可对上传至 Hugging Face Hub 的任意模型进行基准测试，并获得可复现的结果。借助该软件包，我们对 100 多个斯堪的纳维亚语或多语言模型进行了基准测试，并在交互式在线排行榜[1],中展示了结果，同时还对这些结果进行了分析。分析表明，斯堪的纳维亚大陆语言（丹麦语、瑞典语和挪威语）之间存在显著的跨语言迁移，而斯堪的纳维亚大陆语言与岛屿斯堪的纳维亚语（冰岛语和法罗语）两组语言之间的跨语言迁移则十分有限。基准测试结果还表明，挪威、瑞典和丹麦对语言技术的投入，促成了性能超过 XLM-RoBERTa 和 mDe-BERTaV3 等超大规模多语言模型的语言模型。我们发布了该软件包[2] 和排行榜[3]的源代码。

1 引言

近年来，单语语言的数量显著增加。模型的数量在斯堪的纳维亚语言中显著增加（Møller-høj，2020；Højmark-Bertelsen，2021；Sarnikowski，2021；Enevoldsen 等，2021；Abdaoui 等，2020；Kummervold 等，2021年；Malmsten 等，2020年；Snæbjarnarson 等，2023），以至于从业者难以为当前任务选择最佳模型，语言研究人员也难以确保他们的研究工作确实在既有成果的基础上有所改进。

除了模型数量不断增加之外，Sahlgren 等人（2021年）还强调，考虑到这些语言和文化的相似性，采用联合斯堪的纳维亚语言模型可能是更适合斯堪的纳维亚国家的策略。事实上，Faarlund（2019年）甚至认为，丹麦语、挪威语和瑞典语非常相似，应当被视为同一种语言。

“斯堪的纳维亚”一词所包含的语言存在争议（oxf，2021年）。根据斯堪的纳维亚大陆语言（丹麦语、瑞典语和挪威语）与斯堪的纳维亚岛屿语言（冰岛语和法罗语）之间的区分（Haugen，1976；Faarlund，2019年），这一划分基于相互可理解性和句法结构，我们在本文中重点关注斯堪的纳维亚大陆语言，同时仍支持斯堪的纳维亚岛屿语言。除了这是一种标准区分之外，我们的选择还基于对这两个群体之间跨语言迁移的实验，这些实验将在第3.3节中介绍。下文中，我们将使用“斯堪的纳维亚”一词来指代这五种语言的集合，并在适用时使用大陆/岛屿的区分。

为推动单语斯堪的纳维亚模型和多语模型的改进，我们推出了 ScandEval，这是一个针对斯堪的纳维亚模型的基准测试，并提供 Python 软件包、命令行界面（CLI）以及配套的在线排行榜。该排行榜列出了语言模型在斯堪的纳维亚大陆

[1]https://scandeval.github.io[2] https://github.com/saattrupdan/ScandEval[3]https://github.com/ScandEval/ scandeval.github.io

语言的数据集上进行基准测试的结果，如第 4 节所述。

近期研究（Khanuja 等人，2021年；Pires 等人，2019年；Lauscher 等人，2020年）表明，当语言足够相似时，多语模型的表现可以优于单语模型；而当语言差异过大时，多语模型的表现则不如单语模型。这表明，与其使用 XLM-RoBERTa（Conneau 等人，2020年）等大规模多语模型，斯堪的纳维亚语言或许能通过构建“本地多语”模型获益。基于此，我们检验以下假设：

- **假设 1：** 大陆斯堪的纳维亚语言之间存在显著的跨语言迁移。

- **假设 2：** 大陆斯堪的纳维亚语言组与岛屿斯堪的纳维亚语组之间不存在显著的跨语言迁移。

据我们所知，这是首个面向任何斯堪的纳维亚语言的基准测试工具，也是首个收录该工具得分的在线排行榜。我们的贡献如下：

1. 我们构建了一个新的斯堪的纳维亚大陆地区语言问答数据集，命名为 ScandiQA。

2. 我们为所有斯堪的纳维亚语言构建了一个新的语言可接受性数据集，命名为 ScaLA。

3. 我们开发了一个 Python 软件包和命令行界面 scandeval，用于在斯堪的纳维亚语言数据集上对语言模型进行可复现的基准测试。

4. 我们统一了基准测试中使用的所有数据集，以便在不同语言和数据集之间进行一致的评估。这些统一后的数据集也可在 Hugging Face Hub[4] 上获取。

5. 我们在基准测试中的斯堪的纳维亚大陆地区语言数据集上，对 Hugging Face Hub[5] 上的所有斯堪的纳维亚语和部分多语言语言模型进行基准测试，并在线排行榜中公布所有得分。

[4]https://huggingface.co/ScandEval[5]https://hf.co

2 相关工作

近年来，已经发布了许多（非英语）自然语言理解基准测试（Wang 等人，2018；Sarlin 等人，2020年；Rybak 等人，2020年；Ham 等人，2020a；Shavrina 等人，2020年；Wilie 等人，2020年；Xiang 等人，2021年；Koto 等人，2020年；Safaya 等人，2022年；Augustyniak 等人；Khashabi 等人，2020年；Ham 等人，2020b；Xu 等人，2020年；Dumitrescu 等人，2021年），我们与这些工作有着相同的目标，即推动各自语言的自然语言处理发展。具体到斯堪的纳维亚语言，SuperLim 基准测试（Adesam 等人，2020年）是一项瑞典语自然语言理解基准测试，包含多个难度较高的任务。不过，SuperLim 基准测试中的大多数数据集仅包含测试集。

XGLUE（Liang 等人，2020年）数据集是另一种多语言自然语言理解基准。该数据集与 ScandEval 的不同之处在于，XGLUE 的所有训练数据均为英语，且大多数测试集都没有斯堪的纳维亚语言版本。

Isbister 和 Sahlgren（2020年）通过将 GLUE 基准（Wang 等人，2018年）中的 STS-B 数据集机器翻译成瑞典语，构建了一个瑞典语相似度基准。除了只涉及单一任务和单一语言之外，由于经过了翻译，该数据集的质量不如黄金标准语料库，作者也指出了这一点。

3 方法

本节详细介绍我们的基准测试方法，包括数据集设置、评估流程以及模型评分。我们还将说明如何开展跨语言迁移实验。

3.1 微调设置

微调时，我们将学习率设为$2 \cdot 10^{-5}$，预热步数设为 100，批量大小设为 32。如果 GPU 显存不足以支持使用此批量大小微调模型，我们就将批量大小减半，并将梯度累积次数加倍，从而保持有效批量大小不变。重复此过程，直到批次能够装入显存。

我们采用线性学习率调度，并在 10,000 个训练步数后设置截距（每个训练步数包含 32 个样本）；此外，我们采用提前停止策略（Plaut 等人，1986），如果验证损失未能继续下

降 90 个训练步数，则停止训练过程。我们使用 AdamW 优化器（Loshchilov 和 Hutter，2018），其第一动量为 $\beta_1 = 0.9$，第二动量为 $\beta_2 = 0.999$；并在所有任务中始终优化交叉熵损失。此外，我们固定随机种子，以确保结果可复现。

微调本身使用 transformers 软件包（Wolf 等人，2020年）。对于命名实体识别任务，我们使用 AutoModelForTokenClassification 类，它将语言模型编码器对每个词元生成的嵌入进行线性投影，得到该词元的实体 logits。对于分类任务，我们使用 AutoModelForSequenceClassification 类，它将语言模型编码器对每篇文档生成的嵌入进行线性投影，得到该文档的类别 logits。最后，对于问答任务，我们使用 AutoModelForQuestionAnswering 类，它将语言模型编码器对每个词元生成的嵌入进行线性投影，得到该词元对应答案起始位置和结束位置的 logits。

3.2 自助法评估

对于每个模型和数据集，我们将以下流程重复 10 次，从而为每种模型与数据集组合生成一个得分：(a) 为本次迭代指定一个唯一的随机种子；(b) 在训练集上微调模型；(c) 在测试集的自助抽样版本（即有放回抽样）上评估模型。评估得分为这些得分的均值 μ，并附上 95% 置信区间 I_{10}，计算方法如下

$$I_N := \mu \pm \frac{1.96}{N-1} \sum_{i=1}^{N} \texttt{score}_i. \qquad (1)$$

通过改变随机种子并使用自助法测试数据集，我们能够捕捉两种噪声：一是模型新增层随机初始化带来的噪声，二是测试集中的噪声。因此，对于每种模型与数据集的组合，我们都能获得更可靠的真实均值置信区间。

为了汇总所有数据集的得分，我们首先计算每个（模型、语言、任务）三元组的特定语言任务得分，即模型在该语言各项任务上的得分平均值。[6] 根据这些特定语言的得分，我们接着计算每个（模型、语言）对的语言得分，即该语言所有任务的特定语言任务得分的平均值。最终的 ScandEval 得分通过对各语言得分取平均值得到，以强调对斯堪的纳维亚模型的训练，而非单语模型的训练。

3.3 跨语言迁移

为检验第 1 节中提出的假设 1 和假设 2，我们引入了一种衡量一组语言“联合跨语言迁移”的方法，所指的是该组内任意两种语言之间跨语言迁移的总体情况。

为此，我们首先引入一组非斯堪的纳维亚语言作为对照组：英语、德语、荷兰语、芬兰语、俄语和阿拉伯语。通过考察对照组语言与斯堪的纳维亚语言的合并集合，我们旨在找出将这些语言划分为两组的最佳方案：ScandEval 基准组和非基准组。划分的“优劣”通过以下方式衡量：在每种基准语言的数据集上对各语言选出的“代表性”模型进行基准测试，并衡量这些基准测试结果的双簇聚类质量。

例如，如果丹麦语和瑞典语构成基准组，而其余语言属于非基准组，我们就会在 ScandEval 中丹麦语和瑞典语的部分上对每种语言的代表性模型进行基准测试，然后根据这些基准测试值计算聚类 {{丹麦语、瑞典语}、{挪威语、冰岛语、法罗语 ...}} 的 F统计量，计算方式为组间方差与组内方差之比。[7] 接着，我们可以将这个 F统计量与另一种聚类方式的 F统计量进行比较，例如，其中基准组由丹麦语和挪威语构成。

在为每种语言挑选代表性模型时，我们在 Hugging Face Hub 上找到了规模大致相同的预训练语言模型，这些模型都仅使用单语数据进行预训练。我们注意到目前不存在法罗语语言模型，因此对于这种语言，我们不纳入任何模型，但仍会纳入法罗语基准测试

[6] 这个均值只有在挪威语中才有实际意义用于命名实体识别任务和语言可接受性任务，因为这两项任务在书面挪威语和新挪威语中均有提供。

[7] 严格来说，基准测试组中的每种语言都会得到一个 F统计量，但我们只使用这些 F统计量的平均值。

当法罗语属于基准测试组时的数据集。完整的模型列表见附录。

因此，我们可以将第一个假设重新表述为：斯堪的纳维亚大陆语言均属于表现最佳的基准组；将第二个假设重新表述为：岛屿斯堪的纳维亚语不属于表现最佳的基准组。

3.4 统一基准测试数据集

由于我们关注比较模型在不同语言上的表现，因此确保基准测试中使用的所有数据集具有相同的格式和相同的规模。

我们的目标是选择一个训练数据规模，在能够区分模型与能够在合理时间内对模型进行基准测试之间取得平衡。我们使用命名实体识别数据集、情感分类数据集和语言可接受性数据集的截断版本，对与第3.3节相同的模型进行了基准测试。基于这些结果，我们通过定性分析发现，使用1,024个训练样本既能区分不同模型，又能在合理时间内完成模型基准测试。图1和图2展示了区分能力与基准测试速度之间的权衡，涵盖 AngryTweets 数据集（Pauli 等，2021年）。其他数据集的其余图表可见附录。

使用较小训练数据集的另一项好处是，这能凸显模型预训练权重的重要性，而非微调过程。此外，我们希望测试数据集尽可能大，以确保对模型进行更稳健的评估，因此根据最小数据集中可用样本的数量，将测试样本数定为 2,048。最后，验证集选定为 256 个样本，以便在训练期间进行合理评估，同时避免耗时过长。所有这些数据集及其划分均可在 Hugging Face Hub 上获取。

4 ScandEval 任务

为了恰当地评估预训练模型的性能，理想情况下，我们需要在许多不同的任务上对其进行评估。遗憾的是，斯堪的纳维亚语言在许多下游任务方面并没有多少公开可用的数据集。

为解决这一问题，我们构建了两个新的斯堪的纳维亚语数据集：ScaLA 和 ScandiQA，分别用于*Linguistic Acceptability*(LA) 和问答(QA)数据集。这些新任务还补充了现有的命名实体识别(NER)和情感分类(SENT)基准测试数据集。除了评估模型在这些任务上的表现外，我们还对每个模型的推理速度进行了基准测试。下文各小节将对此进行更详细的介绍。

4.1 命名实体识别

对于命名实体识别任务，我们采用 CONLL 中使用的四个类别（Tjong Kim Sang 和 De Meulder，2003）：PER、LOC、ORG 和 MISC，分别对应人名、地点、组织和其他实体。

由于这是一个令牌分类任务，且语言模型通常使用不同的令牌化器，因此我们也必须确保对这些令牌化器进行统一处理。我们使用与待基准测试模型关联的预训练令牌化器对文档进行令牌化；为确保评估的一致性，我们将每个词中除第一个令牌之外的所有令牌都替换为空实体 O。例如，如果带有 LOC 标签的词“Københavns”被令牌化为 [“København”和“s”], ，那么我们会分别为这些令牌分配 LOC 和空标签 O。这样可以确保每个文档中的（非空）标签数量保持一致。

在评估指标方面，我们使用微平均 F1 得分，这是命名实体识别中的标准做法。我们还报告无杂项类别分数，即将预测结果和标签中的 MISC 类替换为空标签 O 后计算的微平均 F1 得分。该无杂项类别分数不用于任何汇总得分，仅用于在各个数据集上进行比较。

对于丹麦语，我们使用 DaNE 数据集（Hvin- gelby 等，2020年）。该数据集是丹麦语依存树库（Kromann 和 Lynge，2004年）的命名实体识别标注版本。DaNE 已采用 CoNLL 格式，因此我们不对数据进行预处理。

对于挪威语，我们使用书面挪威语和新挪威语的 NorNE 数据集（Jørgensen 等，2020年）；这些数据集也是挪威语依存句库的命名实体识别标注版本（Øvrelid 和 Hohle，2016年）。除 PER、LOC、ORG 和 MISC 标签外，这些数据集还包含 GPE_LOC、GPE_ORG、PROD、DRV 和 EVT 标签。我们分别将这些标签转换为 LOC、ORG、MISC、MISC 和 MISC。

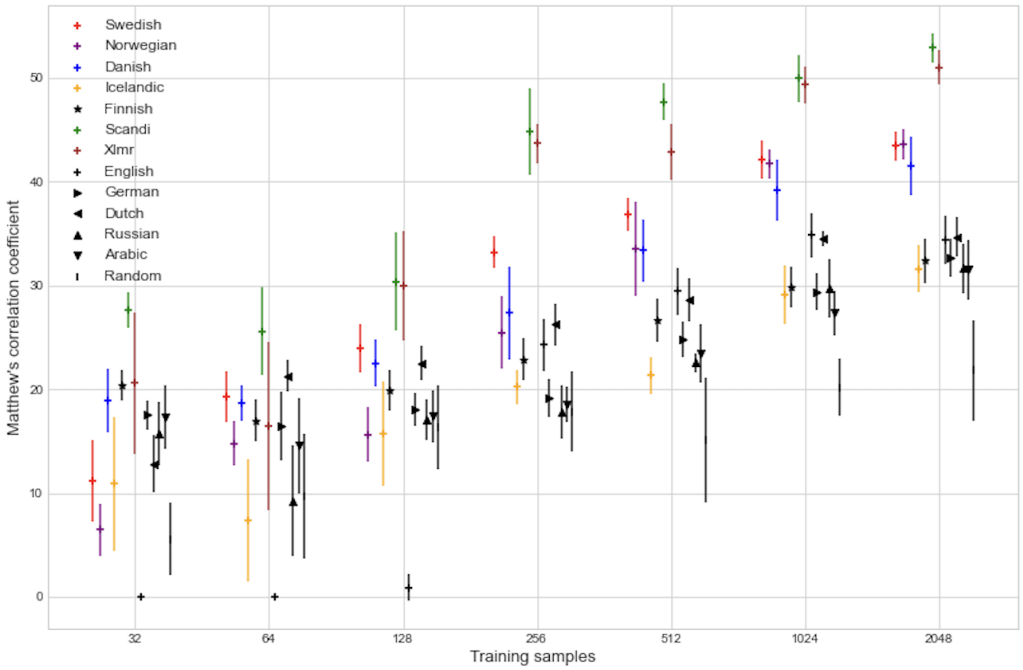


图 1：不同模型在训练样本数量不同的 AngryTweets 数据集上的性能图。

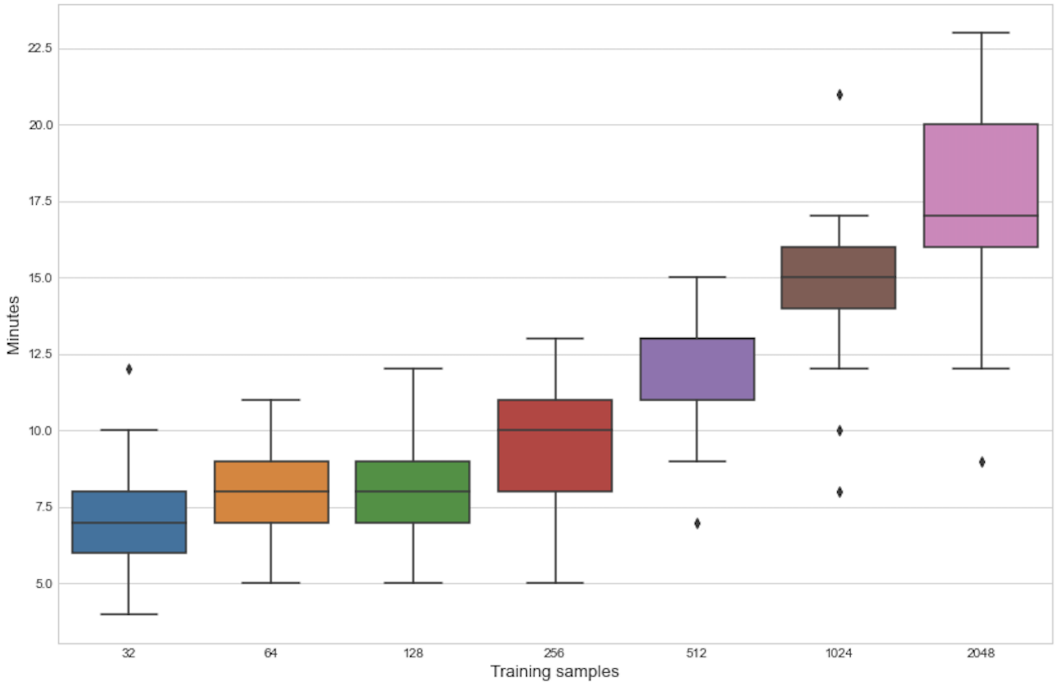


图 2：不同模型在训练样本数量不同的 AngryTweets 数据集上的训练时间箱线图。

最后，瑞典语没有对应依存句库的命名实体识别标注版本，但有 SUC3 数据集，这是对斯德哥尔摩—于默奥语料库进行命名实体识别扩充后的版本（Gustafson-Capková 和 Hartmann，2006年）。该数据集未采用 CoNLL 格式，而是以 XML 格式发布，其中<name> XML 标签包含相应词语的命名实体识别标签

这些标签覆盖的词语[8]。该数据集包含 animal、event、inst、myth、other、person、place、product 和 work 这些命名实体识别标签，分别转换为 MISC、MISC、ORG、MISC、MISC、PER、LOC、MISC 和 MISC。

4.2 情感分类

我们将情感分类任务视为一个三分类任务，类别包括正面、中性和负面。模型评估以马修斯相关系数（Matthews，1975）为主要指标，并报告宏平均 F1 分数作为次要指标。我们选择马修斯相关系数作为主要指标，因为研究表明，它比宏平均 F1 分数更可靠（Chicco 和 Jurman，2020），同时也是 GLUE（Wang 等人，2018）和 SuperGLUE（Sarlin 等人，2020）基准测试采用的标准指标。

对于丹麦语，我们使用情感分类数据集 AngryTweets（Pauli 等人，2021），其中包含众包标注的丹麦语推文。为遵守 Twitter 的使用条款，我们对推文进行了完全匿名化处理：将所有用户提及替换为 @USER，将所有链接替换为[LINK]，并打乱推文顺序。

对于挪威语，我们纳入了情感二分类数据集 NoReC（挪威语评论语料库）（Velldal 等，2018年）；该数据集基于从挪威网站抓取的评论。

最后，对于瑞典语，我们使用 Svensson（2017年）提出的情感二分类数据集，该数据集基于来自瑞典网站 www.reco.se 和 se.trustpilot.com 的评论。类比 NoReC，我们将该数据集命名为瑞典评论语料库（SweReC）。

4.3 语言可接受性

基于将 CoLA（语言可接受性语料库）数据集（Warstadt 等，2019年）纳入 GLUE 基准（Wang 等，2018年），我们为斯堪的纳维亚语言构建了新的语言可接受性数据集。该任务通常被表述为二分类任务，即要求模型预测给定句子在语法上是否正确。

我们将新数据集命名为斯堪的纳维亚语言可接受性（ScaLA），并发布适用于丹麦语、书面挪威语、新挪威语、瑞典语、冰岛语和法罗语。这些数据集各包含 1,024 个训练样本、256 个验证样本和 2,048 个测试样本，符合第 3.4 节的规定。ScaLA 数据集基于通用依存关系数据集的丹麦语、挪威语、瑞典语、冰岛语和法罗语版本（Kromann 和 Lynge，2004年；Øvrelid 和 Hohle，2016年；Nivre 等，2006年；Rögnvaldsson 等，2012年；Jónsdóttir 和 Ingason，2020 年；Arnardóttir 等，2020 年）。

首先，我们假设通用依存数据集中的文档在语法上都是正确的。我们通过人工检查丹麦语部分中的随机样本，验证了这一假设。我们通过删除 一个词或调换两个连续词来创建负例，每个负样本只进行一种这样的“破坏”。

然而，以这种简单方式破坏文档，并不总能得到语法错误的样本。例如，从句子“Den røde bil er stor”（红色的汽车很大）中删除单词“rød”（红色），并不会得到错误的句子“Den bil er stor”（汽车很大）。

为确保生成的句子确实在语法上不正确，我们对可删除或调换的词施加限制。我们采取了较为保守的方法，系统地检查了具有特定词性标签的词被破坏后的情况，并且只允许那些在测试中始终会导致语法错误的破坏方式。由此，我们制定了以下限制：

1. 我们不会删除形容词、副词、标点符号、限定词或数词，因为这样生成的句子在大多数情况下仍然符合语法。

2. 如果名词或专有名词相邻的是另一个名词或专有名词，我们不会将其删除，因为这通常也不会导致句子不合语法。

3. 交换两个相邻词语时，我们要求它们具有不同的词性标签。

4. 我们不会交换标点符号或符号。

5. 如果我们交换第一个词，就能确保交换后的词大小写正确。

[8]这些 <ne> XML 标签也是命名实体识别标签，但它们是由 SpaCy（Honnibal 等，2020年）模型自动生成的，因此并非金标准。

由于这些数据集提供了金标准词性标注令牌，我们能够强制执行这些限制。

4.4 问答

由于我们不知道这些语言是否存在现成的数据集，因此还为斯堪的纳维亚大陆语言构建了新的问答数据集。我们将这些数据集统称为ScandiQA，并为每种斯堪的纳维亚大陆语言分别发布。

这些数据集基于 MKQA 数据集（Longpre 等，2021年），而该数据集又基于自然问题（NQ）数据集（Kwiatkowski 等，2019年）。NQ 数据集包含用户输入 Google 搜索引擎的问题，以及与搜索结果相关联的 HTML 页面。在许多情况下，这些问题都有对应的答案（所谓的短答案），答案会出现在 HTML 中；有时还会标注短答案所在的段落（所谓的长答案）。

MKQA 数据集包含 10,000 个问题及其短答案的人工翻译，涵盖 26 种语言，其中包括丹麦语、挪威语和瑞典语。除了加入这些译文外，MKQA 数据集还修正了原始 NQ 数据集中的许多错误，例如补充原始数据集中缺失的答案，或更正原始数据集中选定的短答案。

MKQA 数据集主要缺少的是上下文段落，我们按如下方式将其添加到数据集中。对于每个 MKQA 样本，我们首先在 NQ 数据集中定位相应的样本。如果该样本有长答案，我们就将其用作初始的（英语）上下文。否则，如果 NQ 数据集和 MKQA 数据集均未登记答案，我们则使用 HTML 中与问题余弦相似度最大的段落，其中我们使用 Sentence Transformer（Reimers 和 Gurevych，2019 年）模型 all-mpnet-base-v2 对文档进行嵌入。[9]

在最后一种情况下，NQ 中该样本没有长答案，但 MKQA 中有答案，我们希望找出 HTML 中包含 MKQA 答案的段落。遗憾的是，MKQA 的答案不会原样出现在 HTML 中（例如，所有日期都统一为 YYYY-MM-DD 格式）。因此，我们首先根据 MKQA 答案构建一个答案候选项列表，其中涵盖了英语中日期和数字的大多数书写方式。然后，我们找出包含任一答案候选项、且与问题的余弦相似度最高的段落；文档的嵌入方式如上所述。

因此，上述过程会得到一个包含答案的英语上下文段落。接下来，我们使用 DeepL 翻译 API[10]将该上下文段落翻译为丹麦语和瑞典语。由于在进行该实验时 DeepL 尚不支持挪威语，我们改用 Google 翻译 API[11] 将上下文段落翻译为挪威语。上下文翻译完成后，我们接着提取与给定斯堪的纳维亚大陆地区语言相关的、已翻译上下文中的所有答案候选项，并将答案改为出现在翻译后上下文中的答案候选项。如果翻译后的上下文中没有出现任何答案候选项，则丢弃该样本。

MKQA 数据集还包含没有答案的样本，我们也将这些样本纳入 ScandiQA 数据集。对于这些样本，我们只需使用上述翻译后的上下文段落。最终的数据集包含 7,810 个丹麦语样本、7,798 个瑞典语样本和 7,813 个挪威语样本。我们也单独发布了该数据集[12]，，并构建了一个 ScandEval 版本，其训练集/开发集/测试集规模与其他 ScandEval 数据集相同。在 ScandEval 版本中（如第 3.4 节所述，包含 1,024/256/2,048 个训练集/验证集/测试集样本），我们只纳入包含答案的样本，否则我们发现，对于此任务而言，1,024 个样本的数据集规模太小。

需要指出的是，由于该数据集是由最初以英语撰写的数据集翻译而来，因此它并不能完美代表大陆斯堪的纳维亚语，因为其中许多问题和答案都涉及美国特有的话题。这可能意味着预训练多语言模型比单语模型更具优势，但我们将这一问题留待未来研究。

[9] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

[10] https://www.deepl.com/pro-api [11] https://cloud.google.com/translate/ [12] 该数据集可在 https://huggingface.co/datasets/alexandrainst/scandi-qa 获取，源代码可在 https://github.com/ alexandrainst/ScandiQA 查看。

4.5 推理速度

除了评估模型的预测表现外，我们还使用 Pyinfer 软件包（Pierse，2020年）对微调模型的推理速度进行了基准测试，并报告平均每秒推理次数。具体做法是，记录将一篇含有 2,600 个字符的文档[13] 输入模型运行一百次的平均推理时间，并将这一过程重复 10 次。我们还按照第3.2节所述计算了置信区间。所有这些计算均使用 AMD Ryzen Threadripper 1920X 12核CPU 完成。

5 基准测试包和命令行界面

为了让每位语言研究人员都能以可复现且一致的方式对其语言模型进行基准测试，我们开发了一个名为 scandeval 的 Python 软件包，可对 Hugging Face Hub 上提供的任何预训练语言模型进行基准测试。

scandeval 软件包既以命令行界面的形式实现，也以 Python 软件包的形式实现，因此既可轻松用作独立的基准测试工具，也可集成到其他 Python 脚本中。该软件包采用了一种非常固执己见的基准测试方式，这意味着可更改的参数很少。这是经过深思熟虑的设计决定，旨在确保对所有模型进行一致的基准测试。该软件包遵循第3.1节中所述的超参数设置。更多信息请参阅 scandeval文档。

6 实验

使用 scandeval 软件包，我们对 Hugging Face Hub 上提供的 100 多个斯堪的纳维亚语言预训练模型进行了基准测试。除了这些模型，我们还纳入了若干多语言模型，以便进行公平比较。最后，为了更好地解读结果，我们还在这些数据集上对随机初始化的 XLM-RoBERTa-base 模型（Conneau 等人，2020年）和 ELECTRA-small 模型（Clark，2019年）进行了基准测试，这有助于更清楚地了解预训练模型在预测中能够利用多少“外部知识”。对所有这些模型进行基准测试大约需要 1000 GPU小时使用 GeForce RTX 2080 Ti GPU，排放量约为 40kgofCO_2 当量[14]。

6.1 基准测试结果

我们已将所有基准测试结果及其对应的置信区间汇总在在线排行榜中。这些得分按照第3节所述的方法计算，各语言表现最佳的前5个模型以及总体表现最佳的前5个模型见表1。

从表1可以看出，NB-BERT-large[15]（Kummervold 等，2021年）是挪威语以及总体表现最佳的模型，DFM-encoder-large-v1[16] 是表现最佳的丹麦语模型，而 KB-BERT-large[17]（Malmsten 等，2020年）在瑞典语上的表现最佳。

得分排名前5的超大规模多语言模型包括 RemBERT（Chung 等，2020年）和 mDeBERTaV3（He 等，2021年）。其余排名前5的模型包括 NB-RoBERTa-base-scandi[18]，DanskBERT（Snæbjarnarson 等，2023）、NB-BERT-base（Kummervold 等，2021年）、Nor-BERT2（Kutuzov 等，2021年）、KB-BERT-base（Malmsten 等，2020年）以及 AI-Nordics-BERT-large[19]。

6.2 跨语言迁移

本实验考察了斯堪的纳维亚模型的跨语言迁移能力，并检验了我们在第 1 节提出的两个假设。实验采用了第 3.3 节所述的方法。对于岛屿斯堪的纳维亚语，纳入的任务包括 ScaLA 数据集的冰岛语和法罗语版本、冰岛语命名实体识别数据集 MIM-GOLD-NER（Ingólfsdóttir 等，2020年），以及命名实体识别数据集 WikiANN（Rahimi 等，2019年）中的法罗语部分。基准测试结果见表 2，所有原始得分见附录。结果证实了我们的两个假设，因为我们发现，具有最大 F统计量的语言组是

[13]该文档为“这是一份虚构文档。”，重复 100 次。

[14]功耗为 250 W/h（Techpowerup.com），在丹麦的碳效率为 0.16 kg/kWh（Ritchie 等，2022年）。[15] https://huggingface.co/NbAiLab/ nb-bert-large[16] https://huggingface.co/chcaa/ dfm-encoder-large-v1[17] https://huggingface.co/KBLab/ megatron-bert-large-swedish-cased-165k[18] https://huggingface.co/NbAiLab/ nb-roberta-base-scandi[19] https://huggingface.co/AI-Nordics/ bert-large-swedish-cased

Rank	总体	丹麦语	挪威语	瑞典语
1	NB-BERT-large	DFM-encoder-large-v1	NB-BERT-large	KB-BERT-large
2	DFM-encoder-large-v1	NB-BERT-large	NB-BERT-base	NB-BERT-large
3	RemBERT	DanskBERT	NB-RoBERTa-base-scandi	KB-BERT-base
4	mDeBERTaV3-base	RemBERT	NorBERT2	AI-Nordics-BERT-large
5	NB-RoBERTa-base-scandi	mDeBERTaV3-base	mDeBERTaV3	RemBERT

表1：在斯堪的纳维亚大陆语言类别中表现最好的五个预训练模型。

斯堪的纳维亚大陆语言组。

基准组	F统计量	基准组	F统计量
丹麦语，挪威语	16.81	丹麦语，瑞典语，冰岛语	4.36
da, sv	15.48	da, sv, fo	10.72
da, is	4.76	da, is, fo	5.48
丹麦语，法罗语	7.29	挪威语，瑞典语，冰岛语	3.11
no, sv	8.14	no, sv, fo	5.57
挪威语，冰岛语	3.73	挪威语，冰岛语，法罗语	3.64
no, fo	2.70	sv, is, fo	4.26
sv, is	4.48	da, no, sv, is	6.97
sv, fo	7.59	da, no, sv, fo	25.40
is, fo	21.84	da, no, is, fo	4.97
丹麦语，挪威语，瑞典语	**33.34**	丹麦语、瑞典语、冰岛语、法罗语	5.21
da, no, is	4.27	no, sv, is, fo	3.38
丹麦语、挪威语、法罗语	11.56	丹麦语、挪威语、瑞典语、冰岛语、法罗语	7.53

表2：展示斯堪的纳维亚语言模型之间跨语言迁移的F统计量。其中，da代表丹麦语，no代表挪威语，sv代表瑞典语，is代表冰岛语，fo代表法罗语。

7 讨论

我们注意到，第6.1节中呈现的基准测试结果表明，挪威和瑞典国家图书馆的工作，以及丹麦丹麦语基础模型项目的工作，都取得了成效；具体而言，它们的模型NB-BERT-large（Kummervold 等，2021年）、KB-BERT-large（Malmsten 等，2020年）和DFM-encoder-large-v1的表现均优于多语言模型。

这似乎表明，在特定语言层面大力投入语言技术是值得的。从同一张表中我们还可以看到，挪威语模型在丹麦语、挪威语和瑞典语中均位列表现最佳的两个模型之列，这表明语言迁移的规模可能相当大，而第6.2节的跨语言迁移实验也支持这一点。这说明，联合大陆斯堪的纳维亚语方案或许能够提升现有单语模型在斯堪的纳维亚大陆地区语言中的表现。

8 结论

本文介绍了一个针对斯堪的纳维亚语言的基准测试框架，并提供了一个 Python 软件包和命令行界面，scandeval 可用于对 Hugging Face Hub 上提供的任何模型进行基准测试。该基准测试包含四项任务：命名实体识别、情感分类、语言可接受性和问答。我们还发布了两个新数据集 ScaLA 和 ScandiQA，分别用于语言可接受性和问答任务。我们已使用基准测试中的斯堪的纳维亚大陆语言数据集对 100 多个模型进行了基准测试，并在在线排行榜中展示了结果。通过分析基准测试结果，我们发现斯堪的纳维亚大陆语言之间存在显著的跨语言迁移，而斯堪的纳维亚大陆语言组与岛屿斯堪的纳维亚语组之间则没有明显的迁移。这就是在线排行榜只纳入斯堪的纳维亚大陆语言，同时在 scandeval 软件包中继续支持岛屿斯堪的纳维亚语的理由。

参考文献

2021年。牛津英语词典。牛津大学出版社。https://www.lexico.com/ definition/scandinavia。

Amine Abdaoui、Camille Pradel 和 Grégoire Sigel。2020年。按需加载：多语言 BERT 的精简版本。载于*SustaiNLP*：简单高效自然语言处理研讨会论文集，第 119–123 页，线上。计算语言学协会。

Yvonne Adesam、Aleksandrs Berdicevskis 和 Felix Morger。2020年。SwedishGLUE——迈向用于评估自然语言理解模型的瑞典语测试集。瑞典语系。

Þórunn Arnardóttir、Hinrik Hafsteinsson、Einar Freyr Sigurðsson、Kristín Bjarnadóttir、Anton Karl In-gason、Hildur Jónsdóttir 和 Steinþór Steingríms-son。2020年。面向 Penn 格式成分树库的通用依存转换流水线。载于第四届通用依存研讨会论文集（*UDW 2020*），第 16–25 页。

Lukasz Augustyniak、Kamil Tagowski、Albert Sawczyn、Denis Janiak、Roman Bartusiak、Adrian Dominik Szymczak、Arkadiusz Janz、Piotr Szyma´nski、Marcin W.atroba、Mikołaj Morzy等。就是这样：设计并编纂 LEPISZCZE，一个全面的波兰语自然语言处理基准。载于第三十六届神经信息处理系统大会数据集与基准赛道。

Davide Chicco 和 Giuseppe Jurman。2020年。二元分类评估中，马修斯相关系数（MCC）相较于 F1 分数和准确率的优势。*BMCgenomics*，21：1–13。

Hyung Won Chung、Thibault Fevry、Henry Tsai、Melvin Johnson 和 Sebastian Ruder。2020年。重新思考预训练语言模型中的嵌入耦合。载于国际学习表征会议。

Kevin Clark、Minh-Thang Luong、Quoc V Le 和 Christopher D Manning。2019年。ELECTRA：将文本编码器预训练为判别器，而非生成器。载于国际学习表征会议。

Alexis Conneau、Kartikay Khandelwal、Naman Goyal、Vishrav Chaudhary、Guillaume Wenzek、Francisco Guzmán、Édouard Grave、Myle Ott、Luke Zettle- moyer 和 Veselin Stoyanov。2020年。大规模无监督跨语言表征学习。载于计算语言学协会第*58*届年会会议论文集，第8440–8451页。

Stefan Daniel Dumitrescu、Petru Rebeja、Beata Lorincz、MihaelaGaman、AndreiAvram、MihaiIlie、Andrei Pruteanu、Adriana Stan、Lorena Rosia、Cristina Iacobescu 等，2021年。LiRo：罗马尼亚语任务的基准测试与排行榜。载于 第三十五届神经信息处理系统大会数据集与基准赛道（第一轮）。

Kenneth Enevoldsen、Lasse Hansen 和 Kristoffer L Nielbo。2021年。DaCy：丹麦语自然语言处理的统一框架。会议论文集 *http://ceur-ws. org* 国际标准连续出版物号，1613:0073。

Jan Terje Faarlund。2019年。斯堪的纳维亚大陆语言的句法。牛津大学出版社。

Sofia Gustafson-Capková 和 Britt Hartmann。2006年。《斯德哥尔摩-于默奥语料库 2.0 版手册》。斯德哥尔摩大学。https://spraakbanken.gu.se/parole/Docs/SUC2.0-manual.pdf。

Jiyeon Ham、Yo Joong Choe、Kyubyong Park、Ilji Choi 和 Hyungjoon Soh。2020a。KorNLI 和 Ko-rSTS：面向韩语自然语言理解的新基准数据集。载于计算语言学协会 *EMNLP 2020* 论文集：研究成果，第 422–430 页，线上。计算语言学协会。

Jiyeon Ham、Yo Joong Choe、Kyubyong Park、Ilji Choi 和 Hyungjoon Soh。2020b。KorNLI 和 KorSTS：面向韩语自然语言理解的新基准数据集。载于计算语言学协会 *EMNLP 2020* 研究成果，第 422–430 页。

Einar Haugen。1976年。斯堪的纳维亚语言：历史导论。马萨诸塞州剑桥：哈佛大学出版社。

Pengcheng He、Jianfeng Gao 和 Weizhu Chen。2021年。DeBERTaV3：使用 ELECTRA 式预训练和梯度解耦嵌入共享来改进 DeBERTa。*arXiv* 预印本 *arXiv:2111.09543*。

Matthew Honnibal、Ines Montani、Sofie Van Lande-glem 和 Adriane Boyd。2020年。spaCy：Python 中的工业级自然语言处理。

Rasmus Hvingelby、Amalie Brogaard Pauli、Maria Bar-rett、Christina Rosted、Lasse Malm Lidegaard 和 Anders Søgaard。2020年。DaNE：丹麦语命名实体资源。载于第*12*届语言资源与评估会议论文集，第4597–4604页，法国马赛。欧洲语言资源协会。

Malte Højmark-Bertelsen. 2021年。Ælæctra——迈向更高效的丹麦语自然语言处理。https://github.com/MalteHB/-l-ctra/。

Svanhvít L Ingólfsdóttir、Ásmundur A Gudjónsson 和 Hrafn Loftsson。2020年。《冰岛语命名实体识别：标注语料库与模型》。载于国际统计语言与语音处理会议，第 46–57 页。Springer。

Tim Isbister 和 Magnus Sahlgren。2020年。《为什么不直接翻译？首个瑞典语语义相似度评估基准测试》。《style id='1'>arXiv 预印本 arXiv:2009.03116。

Hildur Jónsdóttir 和 Anton Karl Ingason。2020年。从原始文本创建符合通用依存规范的冰岛语平行依存树库。载于第十二届语言资源与评测会议论文集，第2924–2931页。

Fredrik Jørgensen、Tobias Aasmoe、Anne-Stine Ruud Husevåg、Lilja Øvrelid 和 Erik Velldal。2020年。NorNE：为挪威语标注命名实体。载于第十二届语言资源与评测会议论文集，第4547–4556页，法国马赛。欧洲语言资源协会。

Simran Khanuja、Melvin Johnson 和 Partha Talukdar。2021年。MergeDistill：使用预训练蒸馏合并语言模型。载于计算语言学协会年会与国际计算自然语言学习联合会议 *ACL-IJCNLP 2021* 论文集（发现），第2874–2887页，线上。计算语言学协会。

Daniel Khashabi、Arman Cohan、Siamak Shakeri、Pedram Hosseini、Pouya Pezeshkpour、Malihe Alikhani、Moin Aminnaseri、Marzieh Bitaab、Faeze Brahman、Sarik Ghazarian、Mozhdeh Gheini、Arman Kabiri、Rabeeh Karimi Mahabadi、Omid Memarrast、Ahmadreza Mosallanezhad、Erfan Noury、Shahab Raji、Mohammad Sadegh Rasooli、Sepideh Sadeghi、Erfan Sadeqi Azer、Niloofar Safi Samghabadi、Mahsa Shafaei、Saber Sheybani、Ali Tazarv 和 Yadollah Yaghoobzadeh。2020年。ParsiNLU：波斯语语言理解挑战集。计算语言学协会学报，9：1147–1162。

Fajri Koto、Afshin Rahimi、Jey Han Lau 和 Timothy Baldwin。2020年。IndoLEM 和 IndoBERT：面向印度尼西亚语自然语言处理的基准数据集和预训练语言模型。载于第*28*届国际计算语言学会议论文集，第757–770页，西班牙巴塞罗那（线上）。国际计算语言学委员会。

Matthias Trautner Kromann 和 Stine Kern Lynge。2004年。《丹麦语依存树库 v. 1.0》。

Per E Kummervold、Javier De la Rosa、Freddy Wetjen 和 Svein Arne Brygfjeld。2021年。将国家数字图书馆付诸实践：挪威语 Transformer 模型的案例。载于第*23*届北欧计算语言学会议（*NoDaLiDa*）会议论文集，第20–29页，冰岛雷克雅未克（线上）。林雪平大学电子出版社，瑞典。

Andrey Kutuzov、Jeremy Barnes、Erik Velldal、Lilja Øvrelid 和 Stephan Oepen。2019? No, 2021年。面向挪威语的大规模上下文化语言建模。载于第*23*届北欧计算语言学会议（*NoDaLiDa*）论文集，第30–40页。

Tom Kwiatkowski、Jennimaria Palomaki、Olivia Redfield、Michael Collins、Ankur Parikh、Chris Alberti、Danielle Epstein、Illia Polosukhin、Jacob Devlin、Kenton Lee 等。2019年。Natural Questions：问答研究基准。计算语言学协会汇刊，7：453–466。

Anne Lauscher、Vinit Ravishankar、Ivan Vuli´c 和 Goran Glavaš。2020年。从零到英雄：论多语言 Transformer 零样本语言迁移的局限性。载于*2020*年自然语言处理实证方法会议（*EMNLP*）会议论文集，第 4483–4499 页，线上。计算语言学协会。

Yaobo Liang、Nan Duan、Yeyun Gong、Ning Wu、Fen-fei Guo、Weizhen Qi、Ming Gong、Linjun Shou、Daxin Jiang、Guihong Cao、Xiaodong Fan、Ruofei Zhang、Rahul Agrawal、Edward Cui、Sining Wei、Taroon Bharti、Ying Qiao、Jiun-Hung Chen、Winnie Wu、Shuguang Liu、Fan Yang、Daniel Campos、Ran-gan Majumder 和 Ming Zhou。2020年。XGLUE：用于跨语言预训练、理解和生成的新基准数据集。载于*2020*年自然语言处理实证方法会议（*EMNLP*）会议论文集，第 6008–6018 页，线上。计算语言学协会。

Shayne Longpre、Yi Lu 和 Joachim Daiber。2021年。MKQA：一个语言多样化的多语言开放领域问答基准测试。计算语言学协会学报，9：1389–1406。

Ilya Loshchilov 和 Frank Hutter。2018年。解耦权重衰减正则化。载于国际学习表征会议。

Martin Malmsten、Love Börjeson 和 Chris Haffenden。2020年。在瑞典国家图书馆玩转文字——打造瑞典语 BERT。*arXiv* 预印本 *arXiv:2007.01658*。

Brian W Matthews。1975年。T4 噬菌体溶菌酶二级结构预测结果与观测结果的比较。《生物化学与生物物理学报》（*BBA*）—蛋白质结构，405(2)：442–451。

Jens Dahl Møllerhøj. 2020年。Nordic BERT。https:// github.com/certainlyio/nordic_bert。

Joakim Nivre、Jens Nilsson 和 Johan Hall。2006年。Talbanken05：一个包含短语结构和依存标注的瑞典语树库。载于第五届国际语言资源与评估会议（*LREC'06*）论文集，意大利热那亚。欧洲语言资源协会（ELRA）。

Lilja Øvrelid 和 Petter Hohle。2016年。《挪威语通用依存关系》。载于第十届国际语言资源与评估会议（*LREC'16*）会议论文集，第1579–1585页，斯洛文尼亚波尔托罗日。欧洲语言资源协会（ELRA）。

Amalie Brogaard Pauli、Maria Barrett、Ophélie Lacroix 和 Rasmus Hvingelby。2021年。《DaNLP：用于丹麦语自然语言处理的开源工具包》。载于第*23*届北欧计算语言学会议（*NoDaL-iDa*）会议论文集，第460–466页，冰岛雷克雅未克（线上）。林雪平大学电子出版社，瑞典。

Charles Pierse。2020年。Pyinfer。https://github.com/cdpierse/pyinfer。

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019年. 多语言 BERT 有多多语言？ In第*57*届计算语言学协会年会论文集, pages 4996–5001, 佛罗伦萨, 意大利. 计算语言学协会.

David C Plaut, Steven J Nowlan, and Geoffrey E Hinton. 1986年. 关于反向传播学习的实验. 技术报告 *CMU-CS-86-126*.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019年. 大规模多语言迁移用于命名实体识别. In第*57*届计算语言学协会年会论文集, pages 151–164, 佛罗伦萨, 意大利. 计算语言学协会.

Nils Reimers 和 Iryna Gurevych。2019年。Sentence-BERT：使用孪生 BERT 网络生成的句子嵌入。在*2019*年自然语言处理实证方法会议与第*9*届自然语言处理国际联合会议（*EMNLP-IJCNLP*）会议论文集中，第 3982–3992 页。

Hannah Ritchie、Max Roser 和 Pablo Rosado。2022年。能源。*OurWorld in Data*。Https://ourworldindata.org/energy。

Eiríkur Rögnvaldsson、Anton KarlIngason、Einar Freyr Sigurðsson 和 Joel Wallenberg。2012年。冰岛语历史解析语料库（IcePaHC）。载于第八届语言资源与评测国际会议（*LREC'12*）会议论文集，第 1977–1984 页。

Piotr Rybak、Robert Mroczkowski、Janusz Tracz 和 Ireneusz Gawlik。2020年。KLEJ：波兰语理解综合基准测试。载于计算语言学协会第*58*届年会会议论文集，第 1191–1201 页，线上。计算语言学协会。

Ali Safaya、Emirhan Kurtulu¸s、Arda Goktogan 和 Deniz Yuret。2022年。Mukayese：土耳其语自然语言处理反击。在计算语言学协会：*ACL 2022*年会议论文集（研究成果），第846–863页。

Magnus Sahlgren、Fredrik Carlsson、Fredrik Olsson 和 Love Börjeson。2021年。反正基本上是同一种语言：支持构建北欧语言模型的理由。载于第*23*届北欧计算语言学会议（*NoDaL-iDa*），第367–372页，冰岛雷克雅未克（线上）。林雪平大学电子出版社，瑞典。

Paul-Edouard Sarlin、Daniel DeTone、Tomasz Malisiewicz 和 Andrew Rabinovich。2020年。Superglue：利用图神经网络学习特征匹配。在*IEEE/CVF*计算机视觉与模式识别会议论文集，第4938–4947页。

PhilipTamimi Sarnikowski。2021年。丹麦语 Transformer。GitHub。https://github.com/ sarnikowski。

Tatiana Shavrina、Alena Fenogenova、Emelyanov Anton、Denis Shevelev、Ekaterina Artemova、Valentin Malykh、Vladislav Mikhailov、Maria Tikhonova、Andrey Chertok 和 Andrey Evlampiev。2020年。RussianSuperGLUE：俄语理解评测基准测试。在*2020*年自然语言处理经验方法会议（*EMNLP*）会议论文集，第4717–4726页，线上。计算语言学协会。

Vésteinn Snæbjarnarson、Annika Simonsen、Goran Glavaš 和 Ivan Vuli´c。2023。通过近缘语言迁移至低资源语言：以法罗语为例。在第*24*届北欧计算语言学会议（*NoDaLiDa*）会议论文集中，法罗群岛托尔斯港。瑞典林雪平大学电子出版社。

Kristoffer Svensson。2017。使用卷积神经网络进行情感分析：对瑞典语评论中的情感进行分类。学士学位论文。

Techpowerup.com。英伟达 GeForce RTX 2080 Ti 规格。https://www.techpowerup.com/gpu-specs/geforce-rtx-2080-ti. c3305。访问日期：2023年4月1日。

Erik F. Tjong Kim Sang 和 Fien De Meulder。2003。CoNLL-2003共享任务简介：语言无关的命名实体识别。在*HLT-NAACL 2003*自然语言学习第七届会议论文集中，第142–147页。

Erik Velldal、Lilja Øvrelid、Eivind Alexander Bergem、Cathrine Stadsnes、Samia Touileb 和 Fredrik Jør- gensen。2018年。NoReC：挪威语评论语料库。载于第十一届国际语言资源与评测会议会议论文集（*LREC2018*），日本宫崎。欧洲语言资源协会（ELRA）。

Alex Wang、Amanpreet Singh、Julian Michael、Felix Hill、Omer Levy 和 Samuel Bowman。2018年。GLUE：用于自然语言理解的多任务基准及分析平台。载于*2018*年 *EMNLP Workshop Black- boxNLP: Analyzing and Interpreting Neural Net- works for NLP* 论文集，第353–355 页，比利时布鲁塞尔。计算语言学协会。

Alex Warstadt、Amanpreet Singh 和 Sa muel Bow- man。2019年。神经网络可接受性判断。计算语言学协会汇刊，7:625–641。

Bryan Wilie、Karissa Vincentio、Genta Indra Winata、Samuel Cahyawijaya、Xiaohong Li、Zhi Yuan Lim、Sidik Soleman、Rahmad Mahendra、Pascale Fung、Syafri Bahar 和 Ayu Purwarianti。2020年。IndoNLU：用于评估印度尼西亚语自然语言理解的基准测试和资源。载于计算语言学协会亚太分会第一届会议暨第十届自然语言处理国际联合会议论文集，第843–857页，中国苏州。计算语言学协会。

Thomas Wolf、Lysandre Debut、Victor Sanh、J ulien Chaumond、Clement Delangue、Anthony Moi、Pier- ric Cistac、Tim Rault、Remi Louf、Morgan Funtow- icz、Joe Davison、Sam Shleifer、Patrick von Platen、Clara Ma、Yacine Jernite、Julien Plu、Canwen Xu、Teven Le Scao、Sylvain Gugger、Mariama Drame、Quentin Lhoest 和 Alexander Rush。2020年。Transformer：先进水平的自然语言处理。载于*2020*年自然语言处理实证方法会议：系统演示论文集，第38–45页，线上。计算语言学协会。

Beilei Xiang、Changbing Yang、Yu Li、Alex Warstadt 和 Katharina Kann。2021年。CLiMP：中文语言模型评估基准测试。载于欧洲计算语言学协会分会第*16*届会议论文集：主会场，第2784–2790页，线上。计算语言学协会。

Liang Xu、Xuanwei Zhang、Lu Li、Hai Hu、Chenjie Cao、Weitang Liu、Junyi Li、Yudong Li、Kai Sun、Yechen Xu、Yiming Cui、Cong Yu、Qianqian Dong、Yi n Tian、Dian Yu、Bo Shi、Jun jie Zeng、Rongzhao Wang、Weijian Xie、Yanting Li、Yina Patterson、Zuoyu Tian、Yiwen Zhang、He Zhou、Shaoweihua Liu、Qipeng Zhao、Cong Yue、Xinrui Zhang、Zhen-Yi Yang、Kyle Richardson 和 Zhenzhong Lan。2020年。CLUE：中文语言理解评测基准。载于国际计算语言学会议。

跨语言迁移实验

语言	Hugging Face 模型 ID	# 参数
丹麦语	vesteinn/DanskBERT	124M
瑞典语	KB/bert-base-swedish-cased	125M
挪威语	patrickvonplaten/norwegian-roberta-base	125M
冰岛语	mideind/IceBERT	124M
英语	roberta-base	125M
德语	deepset/gbert-base	110M
荷兰语	pdelobelle/robbert-v2-dutch-base	117M
芬兰语	TurkuNLP/bert-base-finnish-cased-v1	125M
俄语	DeepPavlov/rubert-base-cased	178M
阿拉伯语	asafaya/bert-base-arabic	111M

表 3：跨语言迁移实验所用模型在 Hugging Face Hub 上的模型 ID。

模型	丹麦语得分	挪威语得分	瑞典语得分	冰岛语得分	法罗语 得分
丹麦语	**63.87 ± 1.26**	53.74 ± 3.73	52.08 ± 2.70	30.39 ± 1.55	45.26 ± 1.35
挪威语	46.30 ± 2.83	**58.78 ± 1.44**	46.90 ± 2.79	28.85 ± 1.45	43.35 ± 2.20
瑞典语	45.81 ± 2.96	47.32 ± 2.66	**69.29 ± 1.40**	28.69 ± 1.61	43.63 ± 1.90
冰岛语	30.20 ± 1.23	28.68 ± 2.91	36.80 ± 2.14	**71.00 ± 1.50**	**48.26 ± 4.76**
芬兰语	32.55 ± 1.47	30.71 ± 2.14	38.94 ± 1.51	16.33 ± 1.89	36.87 ± 1.17
英语	34.11 ± 2.11	30.92 ± 2.69	39.24 ± 1.92	28.39 ± 2.41	40.75 ± 1.59
德语	28.13 ± 2.04	27.58 ± 2.90	37.62 ± 4.18	26.13 ± 1.63	41.02 ± 1.46
荷兰语	31.78 ± 1.62	28.27 ± 2.51	35.06 ± 1.87	26.21 ± 1.79	40.83 ± 1.70
俄语	33.91 ± 1.88	33.55 ± 2.17	39.14 ± 2.33	29.96 ± 1.58	43.17 ± 1.66
阿拉伯语	22.89 ± 1.82	19.98 ± 2.24	25.40 ± 2.69	10.33 ± 2.19	35.33 ± 1.57

表4：跨语言迁移实验中使用的原始基准测试结果。

B 训练数据规模实验

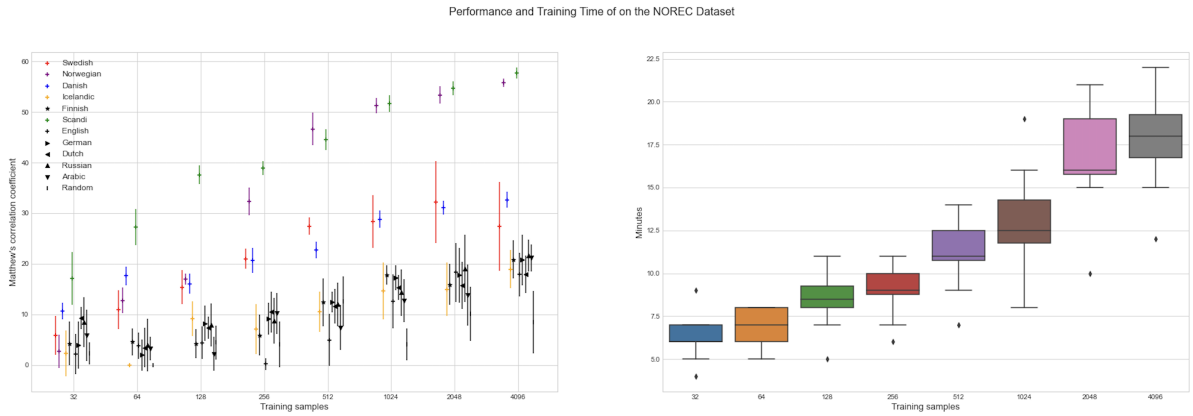


图 3：NoReC 数据集的训练数据规模实验结果。

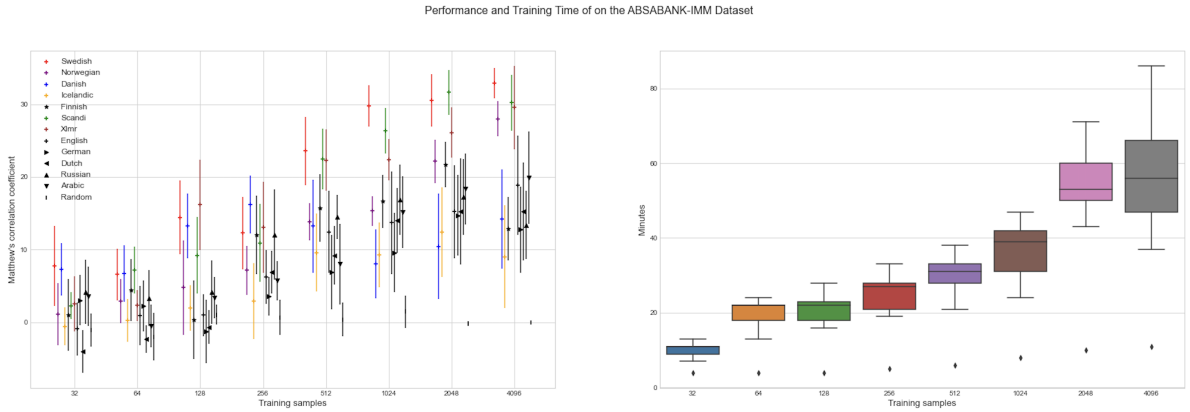


图 4：Absabank-Imm 数据集的训练数据规模实验结果（Adesam 等人，2020年）。

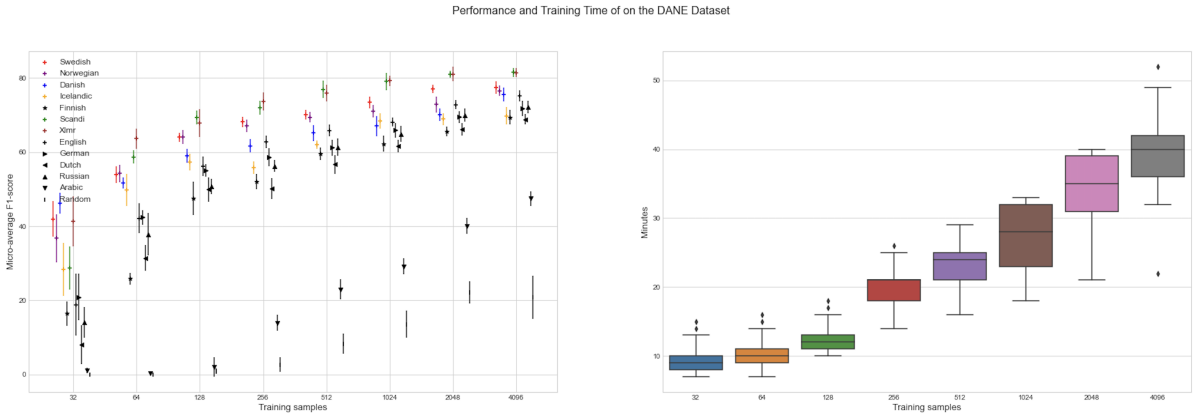


图 5：DaNE 数据集的训练数据规模实验结果。

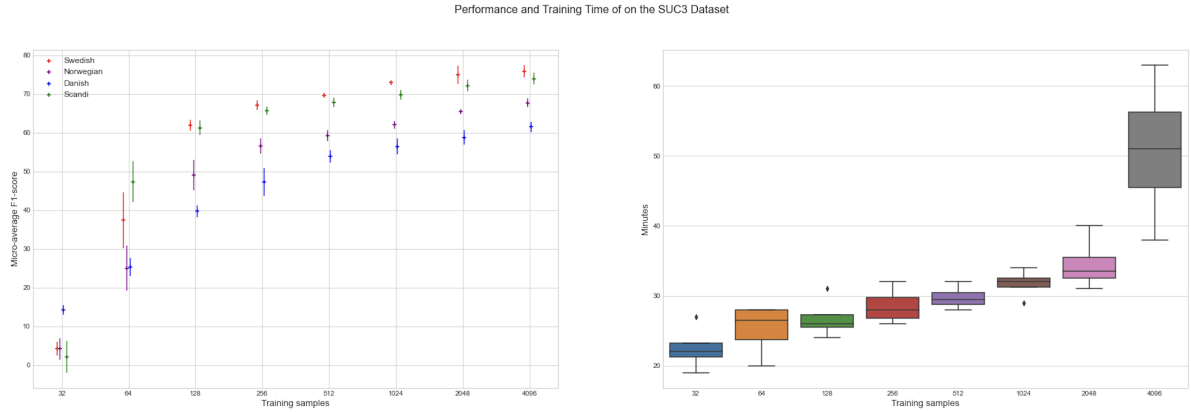


图 6：SUC3 数据集的训练数据规模实验结果。

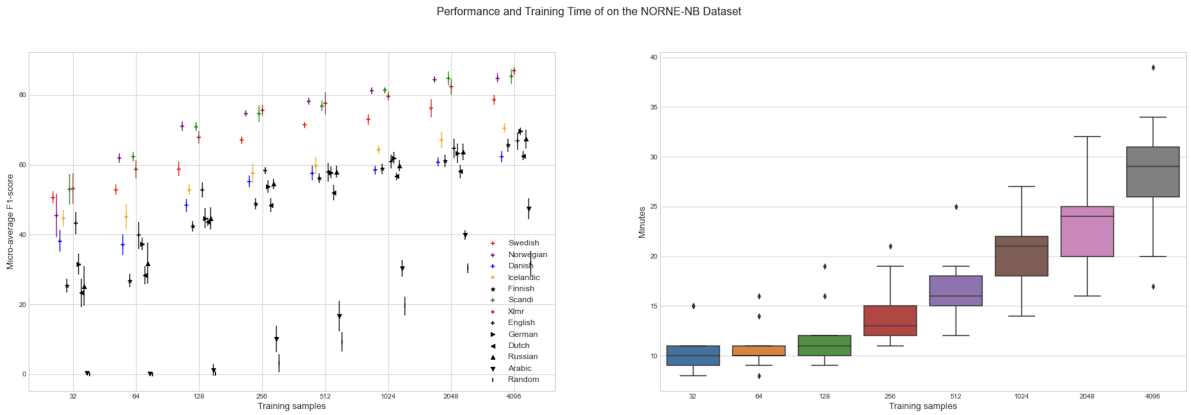


图 7：NorNE-NB 数据集的训练数据规模实验结果。

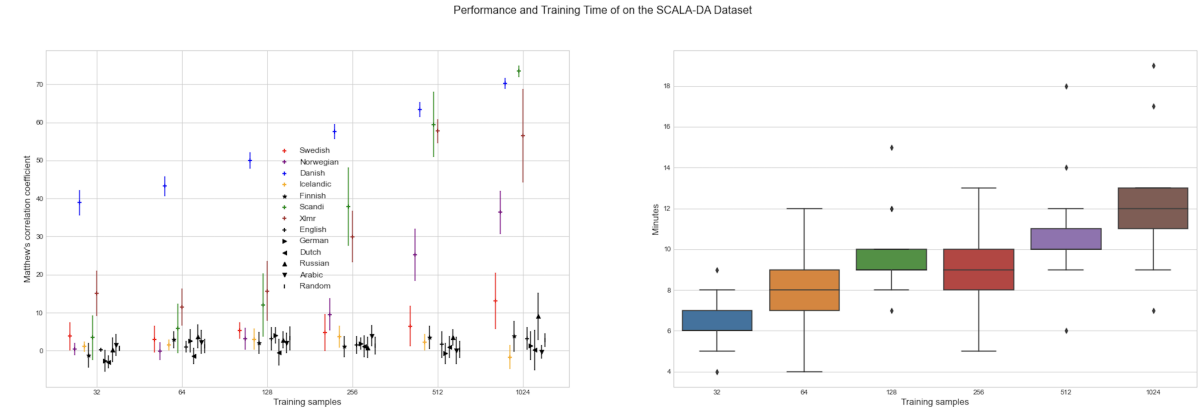


图 8：ScaLA-DA 数据集的训练数据规模实验结果。



图9：ScaLA-SV 数据集的训练数据规模实验结果。

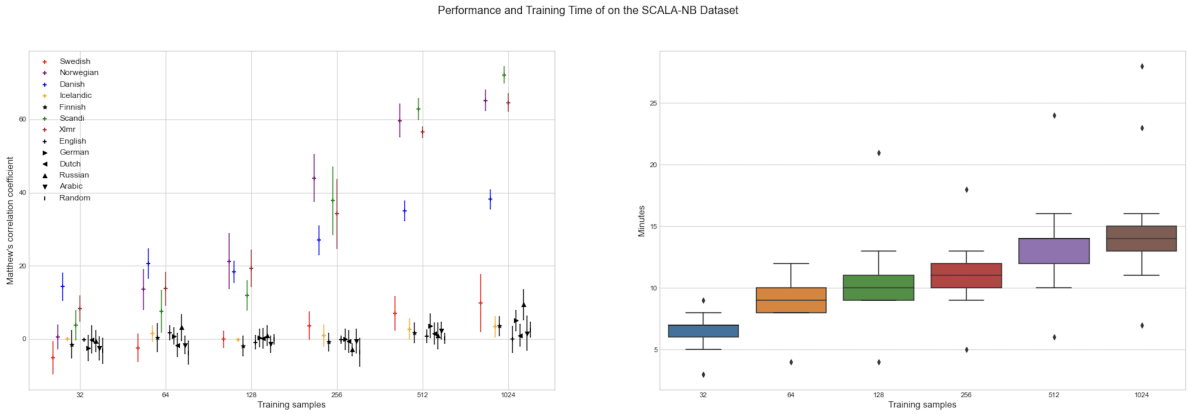


图10：ScaLA-NB 数据集的训练数据规模实验结果。

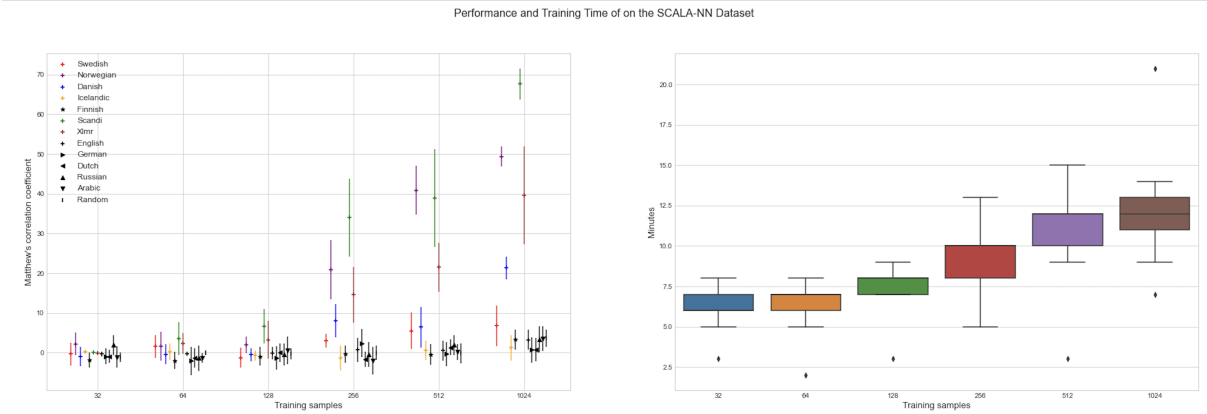


图11：ScaLA-NN 数据集的训练数据规模实验结果。