# 带有真实标签的语言变体识别

**Marcos Zampieri[1], Kai North[1],Tommi Jauhiainen[2],Mariano Felice[3]Neha Kumari[4],Nishant Nair[4],Yash Bangera[4]**

[1]美国乔治·梅森大学， [2]芬兰赫尔辛基大学[3]英国剑桥大学， [4]美国罗切斯特理工学院

mazgla@rit.edu

arXiv:2303.01490v1 [cs.CL] 2 Mar 2023

## 摘要

语言识别是许多信息检索和自然语言处理应用中的重要第一步。然而，大多数公开可用的语言识别数据集都是基于这样一种假设编制的：每个实例的金标准标签由文本的检索来源决定。研究表明，这一假设存在问题，尤其是在非常相似的语言（例如克罗地亚语和塞尔维亚语）以及国家语言变体（例如巴西葡萄牙语和欧洲葡萄牙语）的情况下，因为文本可能不含有能够区分特定语言或变体的标记。为克服这一重要局限，本文介绍了 DSL True Labels（DSL-TL），这是首个由人工标注、用于语言变体识别的多语言数据集。DSL-TL 共包含 12,900 个实例，涵盖葡萄牙语（分为欧洲葡萄牙语和巴西葡萄牙语）、西班牙语（分为阿根廷西班牙语和卡斯蒂利亚西班牙语）以及英语（分为美式英语和英式英语）。我们训练了多个模型来区分这些语言变体，并详细展示了实验结果。本文提出的数据和模型为开发稳健且更公平的语言变体识别系统提供了可靠的基准。我们向研究界免费开放 DSL-TL。

## 1 引言

语言识别是指自动识别给定文本或文档所使用语言的任务（Jauhiainen 等人，2019c）。这项任务是许多信息检索和自然语言处理应用中不可或缺的预处理步骤。语言识别通常被建模为监督式文本分类任务，典型的语言识别系统通常遵循以下四个主要步骤（Lui，2014）：

(a) 表示：选择一种文本表示方式（例如，字符、词或两者的组合）；

（b）语言建模：从文本中为每种语言构建一个模型；

（c）分类：定义一个函数，以最佳方式表示文本与各语言模型之间的相似度；

（d）预测：计算得分最高的模型，以确定给定文本所使用的语言。

21世纪初，语言识别普遍被认为是一个已解决的问题（McNamee，2005年），因为字符 n 元语言模型在标准的当代文本（如报纸文本）中区分彼此差异较大的语言（如阿拉伯语、英语、芬兰语和日语）时能够达到完美表现。过去十年中，随着信息检索应用中出现了更具挑战性的场景，人们重新燃起了对这一任务的兴趣。这些场景包括识别用户生成内容（如微博）中的超短非标准文本所使用的语言（Tromp 和 Pechenizkiy，2011年），以及识别网络查询所使用的语言（Anand，2014年；Ceylan 和 Kim，2009年）。语言混合和语码转换等语言现象，也给最先进的语言识别系统带来了其他挑战；在这些现象中，两种或更多种语言会混杂在文本或社交媒体帖子中（Solorio 等人，2014年；Molina 等人，2016年）。

区分非常相似的语言、方言以及同一种语言的国家变体，是语言识别中另一个重要且具有挑战性的场景，已有多项研究对此进行了探讨（Tiedemann 和 Ljubeši´c，2012年；Lui 和 Cook，2013年；Bouamor 等，2019年）。在这一场景中，系统需要对一组彼此密切相关的语言（如保加利亚语和马其顿语）、方言（如

阿拉伯语的不同方言），或同一种语言的国家变体（如巴西葡萄牙语和欧洲葡萄牙语）进行细致建模，才能准确地区分它们。这一挑战一直是相似语言、变体和方言自然语言处理（VarDial）系列研讨会的主要议题（G˘a-man 等，2020；Chakravarthi等人，2021；Aepli 等，2022），其相关基准竞赛自2014年以来每年举办。VarDial竞赛为学界提供了多个包含多种语言和方言的数据集，有助于建立重要的语言识别基准。

如第 2 节所述，为 VarDial 及类似竞赛收集的数据集的主要局限在于，每个实例的金标准标签并非通过人工标注获得。最广泛使用的数据集 DSL 语料库集合（DSLCC）（Tan 等人，2014年）就是一个例子。该语料库包含从多个新闻网站收集的新闻文本，并将网站所属领域作为语言变体的替代指标。例如，从托管于特定国家域名（如 .br 和 .pt）下的新闻网站收集的所有内容，分别被标记为巴西葡萄牙语和欧洲葡萄牙语。尽管这是一个直接明了的假设，能够生成大量准确的金标准标签，但在文章被不同国家转载的情况下，这一假设已被证明存在问题；对于在世界各地广泛使用的语言，尤其是英语，这个问题尤为突出（Zampieri等人，2014年）。此外，多项研究（Ács 等人，2015；Goutte 等人，2016年）使用 DSLCC 评估了母语者识别语言变体的表现，并得出结论：许多实例中并未包含任何能够帮助人们区分不同变体的标记。

为了解决这一局限，本文介绍了 DSL True Labels（DSL-TL），这是首个经过人工标注的语言变体识别数据集。据我们所知，目前还没有带有真实标签、用于语言变体识别或一般语言识别的人工标注数据集，而我们的数据集填补了这一空白。我们收集了 DSLCC 和其他新闻语料库中的实例，并通过众包平台为每个实例收集多个人工判断。最后，我们在这个新数据集上训练并评估了多个机器学习模型。

本文的贡献如下：

1. 对语言变体识别以及一般语言识别提出了一种新颖的问题形式化。2. 发布 DSL-TL，这是首个由人工标注的语言识别数据集。[1] 3. 在这一新数据集上对多个语言识别模型进行了评估。

本文其余部分的组织如下。第2 节讨论语言变体识别领域的先前研究，包括 VarDial 竞赛和现有数据集。第3 节详细介绍 DSL-TL 数据集的构建步骤，从数据收集到标注。第4节介绍实验中使用的语言识别模型，第5 节则展示这些模型在新的 DSL-TL 数据集上的结果。最后，第6节总结我们的研究，并探讨未来工作的方向。

## 2 相关工作

正如近期的一项综述所述（Jauhiainen 等人，2019c），多项语言识别研究报告称，在各种场景下均取得了近乎完美的表现。Ljubeši´c 和 Kranj˘ci´c（2014年）训练了多种传统机器学习分类器，以区分用四种相关语言书写的社交媒体帖子（推文）：波斯尼亚语、克罗地亚语、黑山语和塞尔维亚语。其最佳模型——高斯朴素贝叶斯（GNB）分类器——准确率达到 97.1%。Martadinata等人（2016年）使用马尔可夫模型识别摘录内容分别取自印度尼西亚语、爪哇语、巽他语或米南加保语的维基百科文章。该模型的准确率达到 95.75%。Li等人（2018年）在包含 97 种语言的多个数据集上训练了卷积神经网络（卷积神经网络）。在多个数据集上进行训练时，该模型的准确率始终超过 95%。

不过，正如 VarDial 共享任务报告中所述，区分同一语言不同变体的语言变体识别系统，其表现差异更大（G˘aman 等人，2020；Chakravarthi等人，2021；Aepli 等，2022）。自 2014年 起，VarDial 研讨会已举办了多项语言变体识别共享任务，接下来将对此进行介绍。

[1] https://github.com/LanguageTechnologyLab/ DSL-TL

### 2.1 VarDial共享任务

VarDial-2014 的相似语言辨识（DSL）共享任务（Zampieri等人，2014年）中，八支团队开发了用于区分若干语言组中的相似语言及语言变体的系统。表现最佳的模型采用了一种与朴素贝叶斯分类器相似的概率模型，并结合了多个支持向量机。研究报告称，该模型区分欧洲葡萄牙语和巴西葡萄牙语的准确率为 91%，区分卡斯蒂利亚语和阿根廷西班牙语的准确率为 95.6%（Goutte 等人，2014年）。同一模型区分英式英语和美式英语的准确率为 52.2%（Zampieri等人，2014年）。

DSL 在 VarDial-2015 和 2016年继续举办（Zampieri等人，2015；Malmasi等人，2016）。这项共享任务的两届赛事都在原始数据集的基础上进行了扩展。2015年版新增了若干种语言，并移除了命名实体，以确定它们对性能的影响（Zampieri等人，2015）。2016年版纳入了法语的不同变体，并以阿拉伯语方言识别为题，向18支队伍发起挑战（Malmasi等人，2016）。表现最好的系统（Malmasi和Zampieri，2016；Ionescu和Popescu，2016；Eldesouki等人，2016；Adouane等人，2016）在区分埃及、海湾、黎凡特、现代标准阿拉伯语和马格里布方言时，准确率介于49.7%和51.2%之间。表现最好的两个系统分别是集成模型，以及一个使用字符级和词级n元语法训练的单一SVM（Malmasi和Zampieri，2016；Eldesouki等人，2016）。

自2017年起，VarDial继续举办识别其他语言变体的共享任务（Zampieri等人，2017年，2018年，2019年；Gˇaman 等人，2020；Chakravarthi等人，2021；Aepli 等，2022）。这些共享任务的表现与2014至2016年的情况一致，支持向量机以及基于字符级和词级n元语法训练的模型通常优于其他方法。2017年，在德语语言变体识别中，基于字符级n元语法训练的集成模型或单个支持向量机取得了最高成绩，F1 分数为0.662（Malmasi and Zampieri，2017年）；2018年，在荷兰语和佛兰芒语识别中，F1 分数为0.660（Çöltekinand Rama，2017年）；2020年，在罗马尼亚语识别中，F1 分数为0.787（Çöltekin，2020）。基于字符级n元语法训练的朴素贝叶斯模型在2019年中文识别中也报告了最高的F1分数，为0.908（Jauhiainen等人，2019a），2021年罗马尼亚语识别的得分为0.777（Jauhiainen 等人，2021），2022年意大利语识别的得分为0.9（Jauhiainen 等人，2022a）。因此，语言识别远未成为已解决的问题，其表现因方言群体和语言变体而有很大差异。数据集质量以及语言变体之间的相似程度，是造成这种表现差异的原因。

### 2.2 可用数据集

VarDial共享任务以及其他类似共享任务中使用的数据集（Zubiaga 等，2016年；Rangel 等，2017年）包含数千个句子，涵盖不同语言或方言群体，主要采样自地方报纸和社交媒体。例如，葡萄牙语、西班牙语和英语（Zampieri 等，2014年、2015）、阿拉伯语（Mal-masi 等，2016年）、中文（Zampieri 等，2019年）、罗马尼亚语（Gˇaman 等，2020）和意大利语（Aepli 等，2022）。然而，正如引言中所讨论的，这些数据集包含的实例，其真实标签是根据文本的发布地点确定的（例如英国、美国等）。因此，这些数据集中的每个句子都被归为某种类别，例如美式英语或英式英语，而根据数据集中包含的金标准标签，只有其中一个标签被视为正确。

以这种方式表述自动语言识别问题的弊端在于，许多句子不一定只属于一种语言变体（Goutte等人，2016年）。英语变体和其他语言的变体也存在这种情况。例如，2014年的 DSL 数据集（Zampieri等人，2014年）包含了一些真实标签错误的实例。含有英式英语典型特征的文章可能刊登在美国报纸上，反之亦然，从而导致数据标注错误。 Goutte等人（2016年）进一步表明，人工标注者在 DSL 2014 和 2015 年数据集（Zampieri等人，2014年，2015年）上的表现无法达到具有竞争力的水平。单个人工标注者的准确率平均仅略高于 50%，且表现因语言而异。因此，依赖摘录来源或单个标注者来确定二元真实标签存在问题。我们通过 DSL-TL 弥补这一局限，收集了首个每个实例都包含多个标注的人工标注语言识别数据集。

| 语言 | 变体 A | 变体 B | 两者或皆非 | 总计 |
|---|---|---|---|---|
| 葡萄牙语 | 1,317（欧洲葡萄牙语） | 3,023（巴西葡萄牙语） | 613（pt） | 4,953 |
| 西班牙语 | 2,131（es-ES） | 1,211 (阿根廷西班牙语) | 1,605 (es) | 4,947 |
| 英语 | 1,081 (英式英语) | 1,540 (en-US) | 379（英语） | 3,000 |
| **总计** | | | | **12,900** |

表 1：DSL-TL 的类别划分及实例总数。

## 3 数据集：DSL-TL

### 3.1 理由与动机

上述基准竞赛所用数据集的主要局限在于，每个实例（一句话或一个段落）仅包含一个真实标签，该标签依据文本发表所在的国家（例如英国、美国等）来指定。因此，数据集中的每个句子要么被归为美式英语，要么被归为英式英语，而在评估语言识别系统时，只有一个标签被视为正确。此类任务表述的问题在于，正如先前研究所示（Goutte等人，2016年），许多句子实际上无法识别，因为它们可能属于多种语言变体。例如，并非所有刊登在英国报纸上的句子都包含英式英语独有的特征，例如拼写惯例（如 analyze、neighbor）或词汇选择（如 trousers、rubbish），使英语母语者能够据此识别句子属于英式英语。其他英语变体以及其他语言的变体也存在同样的情况。Goutte 等人（2016 年）指出，如果每篇文本只能指定一个标签，母语者在识别语言变体方面的表现会很差。

由于人类表现在人类执行此任务时往往低于随机水平，因此，期望自动系统在区分同一语言的国家变体时能够达到 100% 的表现是不公平的。为应对这一重要限制，我们在语言变体识别中引入真实标签的使用。真实标签旨在通过使用亚马逊土耳其机器人（AMT）为每个数据点收集并汇总多项人工判断，来捕捉每个句子中变体特有特征的存在或缺失。向标注员展示了数据集中的句子，并要求他们为给定句子赋予以下三种标签之一。

- 当实例中至少存在变体 X 的一项特征时，标注为变体 X；
- 当实例中至少出现一个变体 Y 的特征时，标记为变体 Y；
- 当实例中未出现变体 X 和变体 Y 的特征，或两者出现的特征数量相同时，标记为两者兼有/两者皆无。

标注员每条标注获得 3 至 5 美分的报酬。关于数据收集和标注的更多细节将在下文介绍。

### 3.2 数据

DSL-TL 包含 12,900 个实例，分属多种语言变体，如表 1所示。这些实例的长度从 1 到 3 个句子不等，内容摘自报纸文章。英语文章来自 Zellers 等人(2019年) 提供的新闻文章合集（下文简称 True News）；葡萄牙语和西班牙语文章则来自 DSLCC（Tan 等人，2014年）。这两个数据集都包含从各国多家报纸收集的数据。我们从原始数据集中随机选取实例，并确保各种语言变体的数量均等：葡萄牙语和西班牙语变体各选取 2,500 个，英语变体各选取 1,500 个。DSL-TL 中最终的 12,900 个实例被随机划分为训练集、开发集和测试集，比例为 70%、20% 和 10%，如表 2所示。最后，表 3列出了 DSL-TL 中的实例示例。

| 变体 | 训练集 | Dev | Test | 总计 |
|---|---|---|---|---|
| 葡萄牙语 | 3,467 | 991 | 495 | 4,953 |
| 西班牙语 | 3,467 | 985 | 495 | 4,947 |
| 英语 | 2097 | 603 | 300 | 3,000 |
| 总计 | | | | 12,900 |

表 2：DSL-TL 的训练集、开发集和测试集划分分别占实例总数的 70/20/10%。

| | | 金标准标签 | | |
|---|---|---|---|---|
| 语言 | 句子 | Old | DSL-TL | 标记 |
| 葡萄牙语 | 从**克拉科兰迪亚**到海滨的大型豪宅.....<br>Neoenergia集团旗下的Celpe公司申请的**调整** 预计将产生以下影响<br>平均8 ,67%.<br>毕尔巴鄂竞技的**球队**与1976-77赛季参加决赛的那支球队截然不同<br>并且输掉了决赛。 | 巴西葡萄牙语<br>巴西葡萄牙语<br>欧洲葡萄牙语 | 巴西葡萄牙语<br>巴西葡萄牙语<br>欧洲葡萄牙语 | 克拉科兰迪亚<br>调整<br>团队 |
| 西班牙语 | 我把钥匙交给了一位**向导**，必须很快回来，因为四月<br>情况.....<br>即便如此，健康饮食每天仍吸引着更多**拥护者** ，这要归功于.....绿色餐<br>厅。<br>这些违规行为是由**西班牙工人社会党** 以及，尤其是<br>左翼的立法者提出的..... | 阿根廷西班牙语<br>es-ES<br>阿根廷西班牙语 | 阿根廷西班牙语<br>es-ES<br>es-ES | 向导<br>支持者<br>PSOE |
| 英语 | 一位 Rose Hill 的殡葬服务负责人，以精心安排完美的**个性化**<br>**送别仪式**<br>十个象征性剪影在**郡**各处展出，作为.....的一部分<br>活动。<br>在没有任何准备的情况下进入比赛的最后一天，似乎非常不像圣徒队的作风，甚至有些不自然，<br>赛季。 | 英式英语<br>英式英语<br>英式英语 | 英式英语<br>两者都/两者都不<br>英式英语 | 个性化<br>郡<br>类似圣徒队的 |

表 3：DSL-TL 中英语、葡萄牙语和西班牙语的实例示例。“旧标签”表示原始数据集标签，以及 DSL-TL 语料库中的新标签。标注员识别出的语言标记以粗体显示。这里只展示了这些实例的部分内容。

### 3.3 标注

标注员通过 AMT 众包招募，分别位于使用这些语言变体的六个国家，即阿根廷、巴西、葡萄牙、西班牙、英国和美国。我们要求标注员根据实例所使用的语言变体，标注其母语或非母语变体。标注员将实例标记为欧洲葡萄牙语（pt-PT）或巴西葡萄牙语（pt-BR）、卡斯蒂利亚语（es-ES）或阿根廷西班牙语（es-AR），以及英式英语（en-GB）或美式英语（en-US）。标签分布如表 1所示。

我们请标注员为每个实例标注他们认为最具代表性的变体标签。他们有三个选项可选：（1）语言变体 A；（2）语言变体 B；或（3）两者或皆非。最初，我们为数据集中的 12,900 个实例各收集了三份标注。如果三位标注员给出的标签相同，或两位标注员的标签与原始金标准标签一致，我们就认为该金标准标签正确；西班牙语和葡萄牙语的原始金标准标签来自 DSLCC，英语的则来自 True News。由此，6,426 个实例获得了三位标注员的标注。对于其余 6,474 个实例，我们又收集了两份人工标注，目标是至少有三位标注员对标签达成一致，或有两位标注员的标签与原始数据集的金标准标签一致。

最后，还要求标注员识别影响其判断的语言标记或专名实体。

在总计 12,900 个实例中，有 3,386 个实例包含语言标记。英语、西班牙语和葡萄牙语的标记数量分别为 270、2,378 和 738。

## 4 模型

我们在 DSL-TL 语料库上训练了经典机器学习模型以及基于 Transformer 的模型，具体内容将在以下章节中介绍。正如 Medvedeva 等人（2017年）和 Jauhiainen 等人（2019c）所讨论的，深度学习模型在语言识别方面尚未表现出明显优于传统机器学习模型的效果，因此我们借此机会测试来自不同机器学习范式的方法。

这些模型在两个赛道中进行了评估。

- **赛道**一包含九个标签——六种语言变体，以及每种语言对应的 both、none 类别。

- **赛道**二包含六个标签——仅包括六种语言变体

### 4.1 朴素贝叶斯

我们介绍了使用朴素贝叶斯系统最新版本进行的实验，该系统此前曾被Jauhiainen 等人(2022b) 以及Jauhiainen 等人(2022a) 使用。对于每个语言对，我们将共同实例作为第三种“语言”，在赛道一中按照通常的分类设置进行实验。

| | 多语言 BERT | XLM-R | XLM-R-LD |
|---|---|---|---|
| type | BERT-base | RoBERTa-base | RoBERTa-base |
| 语料库<br>size | 维基百科<br>33亿（102 种语言） | CC 数据<br>2.5 TB（100 种语言） | LID<br>7万（20 种语言） |
| #层数 | 12 | 12 | 12 |
| #注意力头数 | 12 | 16 | 16 |
| #层大小 | 768 | 768 | 768 |
| #参数 | 110M | 250M | 250M |

表 4：多语言 BERT、XLM-R 和 XLM-R-LD 模型的比较。Lang 是 languages（语言）的缩写。CC 数据指 CommonCrawl 数据。

在使用优化后的参数，并将开发数据作为额外训练数据加入后，我们首次在测试数据上运行，赛道一的宏平均 F1 分数为 0.505。由于字符三元组在开发数据上的表现最佳，因此仅使用了字符三元组。表 5 展示了赛道一中各语言变体的统计数据。宏平均 F1 分数显然受共同实例较低 F1 分数的影响。对于赛道二，我们修改了系统，使其在优化过程中进行评估时忽略共同实例。训练集和开发集中的共同实例均被加入每种语言中的两个变体。这次，最佳字符 n 元语法范围为二至五。该系统取得了 0.794 的宏平均 F1 分数，也优于使用预训练语言模型的深度学习系统。

对于赛道二，我们还通过加入 DSLCC v1.0 语料库中的训练数据，尝试了开放变体。加入 DSL 共享任务第一届的英语实例后，开发集上的结果变差；而加入葡萄牙语和西班牙语实例则提升了结果。这可能是因为英语变体取得的结果已经远高于 DSL 2014 上最佳系统的结果。开放式朴素贝叶斯运行的宏平均 F 分数为 0.803。

### 4.2 自适应朴素贝叶斯

除了传统的朴素贝叶斯识别器外，我们还采用了自适应语言模型（Jauhiainen 等人，2019b），方式与Jauhiainen 等人（2022a）在 ITDI 共享任务的获胜系统中采用的方式类似（Aepli 等，2022）。对于自适应版本的朴素贝叶斯分类器，我们使用与非自适应版本相同的惩罚修正项和字符 n 元语法范围。利用开发数据，我们优化了自适应过程中使用的分割数量以及学习轮数。分割数量表示每次识别运行后，置信度最高的测试数据中要加入的部分的大小。我们将分割数量设为 512，并进行四轮自适应。表5 展示了赛道一的结果。宏平均 F1 分数为 0.503，略高于未采用自适应的相同系统所取得的 0.501。在赛道二中也观察到类似的性能提升，自适应朴素贝叶斯取得了 0.799。

### 4.3 深度学习模型

我们还实验了几种预训练大语言模型（LLMs）。这些大语言模型均为多语言模型，包括多语言 BERT[2]（mBERT）（Devlin 等，2019年）、XLM-RoBERTa[3]（XLM-R）（Liu 等，2019年）以及 XLM-R 语言检测[4] （XLM-R-LD）。XLM-R-LD 是在包含20种语言、9万个实例的语言识别数据集[5] （LID）上微调的 XLM-R 模型。这些实例取自多种来源，包括亚马逊评论以及 2012年至2017年的 SemEval 任务。这三个模型均在训练集和开发集上进行训练，两个数据集之间没有重叠（表2）。英语、西班牙语和葡萄牙语的训练集分别包含2097、3467和3467个实例。开发集中，英语包含599个实例，西班牙语包含989个实例，葡萄牙语包含991个实例。模型以 2e-5的学习率训练5个轮次。我们的模型汇总见表4。

## 5 结果

在本节中，我们展示赛道一和赛道二中所有模型获得的结果。表5 展示了赛道一中各模型在

[2]https://huggingface.co/bert-base-multilingual-cased[3]https://huggingface.co/xlm-roberta-base[4]https://huggingface.co/papluca/xlm-roberta-base-language-detection[5]https://huggingface.co/datasets/papluca/language-identification

精确率、召回率和 F1 分数方面的结果，以及所有语言的宏平均值。

| 变体 | 模型 | 召回率 | 精确率 | F1 |
|---|---|---|---|---|
| 英式英语 | NB | 0.754 | 0.705 | 0.729 |
| | ANB | 0.772 | 0.721 | 0.746 |
| | 多语言 BERT | 0.760 | 0.807 | 0.783 |
| | XLM-R | 0.750 | 0.842 | 0.793 |
| | XLM-R-LD | 0.771 | 0.798 | 0.784 |
| en-US | NB | 0.750 | 0.848 | 0.796 |
| | ANB | 0.731 | 0.851 | 0.786 |
| | 多语言 BERT | 0.867 | 0.795 | 0.829 |
| | XLM-R | 0.829 | 0.776 | 0.801 |
| | XLM-R-LD | 0.797 | 0.782 | 0.790 |
| en | NB | 0.267 | 0.190 | 0.222 |
| | ANB | 0.233 | 0.146 | 0.179 |
| | 多语言 BERT | 0.278 | 0.333 | 0.303 |
| | XLM-R | 0.231 | 0.200 | 0.214 |
| | XLM-R-LD | 0.233 | 0.233 | 0.233 |
| 阿根廷西班牙语 | NB | 0.481 | 0.427 | 0.452 |
| | ANB | 0.579 | 0.458 | 0.512 |
| | 多语言 BERT | 0.551 | 0.489 | 0.518 |
| | XLM-R | 0.551 | 0.489 | 0.518 |
| | XLM-R-LD | 0.511 | 0.519 | 0.515 |
| es-ES | NB | 0.636 | 0.679 | 0.657 |
| | ANB | 0.612 | 0.716 | 0.660 |
| | 多语言 BERT | 0.651 | 0.670 | 0.660 |
| | XLM-R | 0.689 | 0.752 | 0.719 |
| | XLM-R-LD | 0.684 | 0.694 | 0.689 |
| es | NB | 0.327 | 0.336 | 0.331 |
| | ANB | 0.34 | 0.351 | 0.345 |
| | 多语言 BERT | 0.442 | 0.468 | 0.455 |
| | XLM-R | 0.454 | 0.442 | 0.448 |
| | XLM-R-LD | 0.444 | 0.429 | 0.436 |
| 巴西葡萄牙语 | NB | 0.662 | 0.762 | 0.708 |
| | ANB | 0.609 | 0.795 | 0.689 |
| | 多语言 BERT | 0.718 | 0.799 | 0.756 |
| | XLM-R | 0.753 | 0.786 | 0.769 |
| | XLM-R-LD | 0.739 | 0.796 | 0.767 |
| 欧洲葡萄牙语 | NB | 0.533 | 0.442 | 0.483 |
| | ANB | 0.555 | 0.442 | 0.492 |
| | 多语言 BERT | 0.459 | 0.496 | 0.477 |
| | XLM-R | 0.492 | 0.657 | 0.562 |
| | XLM-R-LD | 0.488 | 0.613 | 0.544 |
| pt | NB | 0.136 | 0.118 | 0.126 |
| | ANB | 0.153 | 0.100 | 0.121 |
| | 多语言 BERT | 0.214 | 0.051 | 0.082 |
| | XLM-R | 0.000 | 0.000 | 0.000 |
| | XLM-R-LD | 0.000 | 0.000 | 0.000 |
| 宏平均 | NB | 0.505 | 0.501 | 0.501 |
| | ANB | 0.509 | 0.509 | 0.503 |
| | 多语言 BERT | **0.549** | **0.545** | **0.540** |
| | XLM-R | 0.528 | 0.549 | 0.536 |
| | XLM-R-LD | 0.519 | 0.541 | 0.529 |

表 5：赛道一中，使用朴素贝叶斯、自适应朴素贝叶斯、多语言 BERT、XLM-R 和 XLM-R-LD 在测试集上对各语言变体的召回率、精确率和 F1 分数。平均值采用宏平均报告。最佳平均结果以粗体显示。

在赛道一中，我们展示了多语言 BERT 取得了最佳结果，平均 F1 分数为 0.540，其次是 XLM-R，平均 F1 分数同样为 0.540。就各个语言的表现而言，我们观察到所有模型在两种英语变体上的结果都最好，尤其是在 en- US 上，XLM-R 模型取得了高达 0.829 的 F1 分数。考虑到英语数据集是三种语言中最小的数据集，这多少有些令人惊讶。

最后，对于所有语言，与各语言变体所获得的分数相比，所有模型在“两者皆是或皆非”类别（en、es 和 pt）上的结果都非常低。这表明，由于缺乏任何特定特征，该类别极难建模。此前发布的语言识别数据集尚未经过人工标注；其中并不包含这样的类别。因此，“两者皆是或皆非”上的结果需要进一步研究。

为了进一步理解我们表现最佳的模型——多语言 BERT——的预测情况，在图1中，我们绘制了赛道一设置下该模型的混淆矩阵。混淆矩阵显示，混淆不会发生在每种语言的三个标签之外，这证明了该模型在区分不同语言方面的高性能。我们观察到，葡萄牙语“两者皆是或皆非”类别的预测表现不同于其他任何语言的类别，且没有任何正确预测。

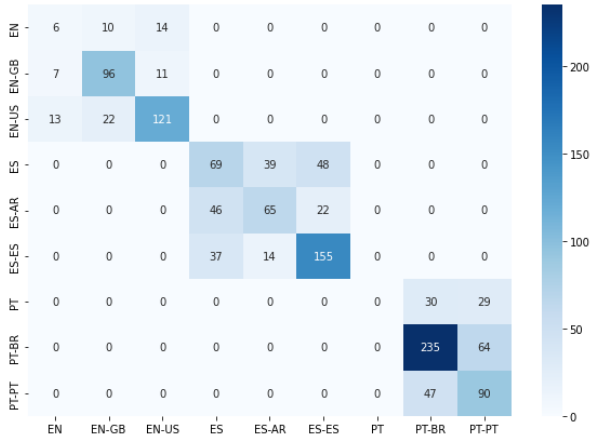


图 1：展示多语言 BERT 在赛道一中类别预测的混淆矩阵。

在表 6中，我们展示了赛道二中所有模型的结果。除了赛道一使用的五个模型外，我们还纳入了一个朴素贝叶斯（开放式挪威语书面语）变体，该模型使用从 DSLCC 检索到的额外原始数据进行训练。不出所料，额外训练数据的使用提升了该系统的性能，使其以 0.803 的平均得分在所有模型中名列第一

F1 分数。尽管如此，我们观察到，在赛道二中，三种朴素贝叶斯变体的表现优于深度学习系统，这印证了先前研究的发现（Medvedeva 等人，2017年；Jauhiainen 等人，2019c）。语言识别本质上是一项模式匹配任务，而非语义理解任务。我们认为，与更复杂的基于文本嵌入的表示相比，这往往更有利于相对简单的字符 n 元语法模型。

| 变体 | 模型 | 召回率 | 精确率 | F1 |
|---|---|---|---|---|
| 英式英语 | NB | 0.921 | 0.761 | 0.833 |
| | 开放式挪威语书面语 | 0.877 | 0.758 | 0.813 |
| | ANB | 0.93 | 0.752 | 0.831 |
| | 多语言 BERT | 0.828 | 0.842 | 0.835 |
| | XLM-R | 0.795 | 0.921 | 0.854 |
| | XLM-R-LD | 0.802 | 0.851 | 0.826 |
| en-US | NB | 0.808 | 0.933 | 0.866 |
| | 开放式挪威语书面语 | 0.821 | 0.901 | 0.859 |
| | ANB | 0.795 | 0.939 | 0.861 |
| | 多语言 BERT | 0.889 | 0.872 | 0.880 |
| | XLM-R | 0.935 | 0.827 | 0.878 |
| | XLM-R-LD | 0.886 | 0.846 | 0.866 |
| 阿根廷西班牙语 | NB | 0.857 | 0.726 | 0.786 |
| | 开放式挪威语书面语 | 0.789 | 0.789 | 0.789 |
| | ANB | 0.887 | 0.724 | 0.797 |
| | 多语言 BERT | 0.772 | 0.662 | 0.713 |
| | XLM-R | 0.750 | 0.654 | 0.699 |
| | XLM-R-LD | 0.765 | 0.684 | 0.722 |
| es-ES | NB | 0.791 | 0.896 | 0.840 |
| | 开放式挪威语书面语 | 0.864 | 0.864 | 0.864 |
| | ANB | 0.782 | 0.915 | 0.843 |
| | 多语言 BERT | 0.800 | 0.874 | 0.835 |
| | XLM-R | 0.794 | 0.859 | 0.825 |
| | XLM-R-LD | 0.809 | 0.864 | 0.836 |
| 巴西葡萄牙语 | NB | 0.716 | 0.873 | 0.787 |
| | 开放式挪威语书面语 | 0.696 | 0.924 | 0.794 |
| | ANB | 0.702 | 0.897 | 0.788 |
| | 多语言 BERT | 0.766 | 0.779 | 0.773 |
| | XLM-R | 0.823 | 0.809 | 0.816 |
| | XLM-R-LD | 0.810 | 0.796 | 0.803 |
| 欧洲葡萄牙语 | NB | 0.774 | 0.564 | 0.652 |
| | 开放式挪威语书面语 | 0.876 | 0.58 | 0.698 |
| | ANB | 0.825 | 0.568 | 0.673 |
| | 多语言 BERT | 0.504 | 0.489 | 0.496 |
| | XLM-R | 0.599 | 0.620 | 0.609 |
| | XLM-R-LD | 0.570 | 0.591 | 0.581 |
| 宏平均 | NB | 0.811 | 0.792 | 0.794 |
| | 开放式挪威语书面语 | **0.820** | **0.803** | **0.803** |
| | ANB | 0.820 | 0.799 | 0.799 |
| | 多语言 BERT | 0.760 | 0.753 | 0.755 |
| | XLM-R | 0.783 | 0.782 | 0.780 |
| | XLM-R-LD | 0.774 | 0.772 | 0.772 |

表6：在赛道二测试集上，使用朴素贝叶斯、自适应朴素贝叶斯、多语言 BERT、XLM-R 和 XLM-R-LD 对各语言变体进行评估所得的召回率、精确率和 F1 分数。平均值报告为宏平均。最佳平均结果以粗体显示。

赛道二的平均结果和各类别的单独结果都显著高于赛道一，这再次证明了对“两者皆是或皆非”类别进行建模的挑战，以及该类别对模型整体性能的影响。不过，在这种任务设定下，赛道二为 DSL-TL 提供了最重要的基线结果，因为这些结果是在仅区分语言变体时获得的。我们认为，“两者皆是或皆非”对现实世界中的语言识别系统重要性要低得多。

## 6 结论与未来研究

本文介绍了 DSL-TL，这是首个用于语言变体识别以及更广泛的语言识别、由人工标注的数据集。该数据集包含以三种语言撰写的报纸文本，并标注了六种语言变体和一个“两者皆是或皆非”类别。我们评估了多种模型在该数据集上的表现，包括经典机器学习方法（朴素贝叶斯）的不同变体以及多种深度学习系统（多语言 BERT、XLM-R 和 XLM-R-LD）。从性能来看，我们观察到朴素贝叶斯系统的表现与深度学习模型不相上下，这证实了先前研究的发现（Medvedeva 等人，2017年；Jauhiainen 等人，2019c）。在某些情境下，朴素贝叶斯的表现甚至超过了深度学习模型。

这是看待该问题的一种全新方式。我们的研究结果表明，在“两者或皆非”类别的处理和计算建模方面仍有改进空间。尽管与语言变体标签相比，这一类别对现实应用的重要性较低，但该类别的低分数凸显了在这一新颖的语言识别场景中对其进行建模的挑战。

这个新数据集为语言识别研究开辟了多条途径。它使研究群体能够对语言识别系统进行更加公平的评估，从而减少潜在偏差。我们预计，DSL-TL 中提出的真实标签策略将成为语言变体识别领域的新标准，有助于提升信息检索和自然语言处理应用的性能。这些应用在处理语言变异时面临困难，例如虚拟助手（如 Alexa、Siri）、机器翻译系统、文本与多媒体检索系统等。

未来，我们希望扩大这个数据集的规模，以进一步研究数据集大小对性能的影响。我们还将

对极为相似的语言群组进行同样的标注，例如波斯尼亚语、克罗地亚语和塞尔维亚语。DSL-TL 是 2023年 VarDial研讨会同期举行的一项同名竞赛的官方数据集。[6] 本文所呈现的结果将作为该竞赛的基线结果。

## 致谢

我们感谢 DSLCC 的创建者提供数据。我们还感谢 VarDial DSL-TL共享任务的参与者提供反馈。本研究部分得到了 RIT 授予的 GWBC种子基金、芬兰科学院（资助决定编号341798）以及芬兰研究影响基金会通过其与 Lingsoft 合作开展的 Tandem产学合作资助的支持。

## 参考文献

Judit Ács、László Grad-Gyenge 和 Thiago Bruno Rodrigues de Rezende Oliveira。2015。用于相似语言辨析的双层分类器。载于*LT4VarDial*论文集。

Wafia Adouane、Nasredine Semmar 和 Richard Johansson。2016。ASIREM 参与 2016年相似语言辨析共享任务。载于*VarDial*论文集。

Noëmi Aepli、Antonios Anastasopoulos、Adrian-Gabriel Chifu、William Domingues、Fahim Faisal、Mihaela Gaman、Radu Tudor Ionescu 和 Yves Scherrer。2022。2022年VarDial评测活动的研究发现。载于*VarDial*论文集。

Supriya Anand。2014年。印度语言查询转写形式的语言识别。载于*FIRE*会议论文集。

Houda Bouamor、Sabit Hassan 和 Nizar Habash。2019年。MADAR 阿拉伯语细粒度方言识别共享任务。载于*WANLP*会议论文集。

Çağrı Çöltekin 和 Taraka Rama。2017年。图宾根团队在 VarDial 2017 共享任务中的表现：语言识别和跨语言句法分析实验。载于*VarDial*会议论文集。

Hakan Ceylan 和 Yookyung Kim。2009年。搜索引擎查询的语言识别。载于*ACL*会议论文集。

BharathiRaja Chakravarthi、Mihaela Gˇaman、Radu Tu- dor Ionescu、Heidi Jauhiainen、Tommi Jauhiainen、Krister Lindén、Nikola Ljubeši´c、Niko Partanen、Ruba Priyadharahsini、Christoph Purschke、Eswari Rajagopal、Yves Scherrer 和 Marcos Zampieri。2021。VarDial 评测活动 2021 年结果。载于*VarDial* 论文集。Çağrı Çöltekin。2020。领域迁移下的方言识别：区分罗马尼亚语和摩尔多瓦语的实验。载于*VarDial* 论文集。Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2019。BERT：用于语言理解的深度双向 Transformer 预训练。载于*NAACL*。Mohamed Eldesouki、Fahim Dalvi、Hassan Sajjad 和 Kareem Darwish。2016。QCRI @ DSL 2016：利用文本特征识别阿拉伯语口语方言。载于*VarDial* 论文集。Cyril Goutte、Serge Léger 和 Marine Carpuat。2014。NRC 系统：区分相似语言。载于*VarDial* 论文集。Cyril Goutte、Serge Léger、Shervin Malmasi 和 Marcos Zampieri。2016。区分相似语言：评测与探索。载于*LREC* 论文集。Mihaela Gˇaman、Dirk Hovy、Radu Tudor Ionescu、Heidi Jauhiainen、Tommi Jauhiainen、Krister Lindén、Nikola Ljubeši´c、Niko Partanen、Christoph Purschke、Yves Scherrer 和 Marcos Zampieri。2020。VarDial 评测活动 2020 年报告。载于*VarDial* 论文集。Radu Tudor Ionescu 和 Marius Popescu。2016。UnibucKernel：一种基于多种字符串核的阿拉伯语方言识别方法。载于*VarDial* 论文集。Tommi Jauhiainen、Heidi Jauhiainen 和 Krister Lindén。2021。基于朴素贝叶斯的罗马尼亚语方言识别实验。载于*VarDial* 论文集。Tommi Jauhiainen、Heidi Jauhiainen 和 Krister Lindén。2022a。使用自适应朴素贝叶斯进行意大利语语言与方言识别及法语区域变体检测。载于*VarDial* 论文集。Tommi Jauhiainen、Heidi Jauhiainen 和 Krister Lindén。2022b。优化用于阿拉伯语方言识别的朴素贝叶斯。载于*WANLP* 论文集。Tommi Jauhiainen、Krister Lindén 和 Heidi Jauhiainen。2019a。使用自适应语言模型区分普通话与瑞士德语变体。载于*VarDial* 论文集。Tommi Jauhiainen、Krister Lindén 和 Heidi Jauhiainen。2019b。用于文本语言与方言识别的语言模型自适应。自然语言工程，25(5):561–583。

[6] https://sites.google.com/view/vardial-2023/ 共享任务

Tommi Jauhiainen、Marco Lui、Marcos Zampieri、Timothy Baldwin 和 Krister Lindén。2019c。文本自动语言识别：综述。《人工智能研究杂志》，65:675–782。

Yitong Li、Timothy Baldwin 和 Trevor Cohn。2018年。领域中有什么？使用对抗训练学习领域鲁棒文本表示。载于*NAACL* 论文集。

Yinhan Liu、Myle Ott、Naman Goyal、Jingfei Du 等。2019年。RoBERTa：一种稳健优化的 BERT 预训练方法。*arXiv* 预印本 *arXiv:1907.11692*。

Nikola Ljubeši´c 和 Denis Kranj˘ci´c。2014年。在 Twitter 用户中区分极为相似的语言。载于*IS-LT* 论文集。

Marco Lui。2014年。通用语言识别。博士论文，墨尔本大学。

Marco Lui 和 Paul Cook。2013年。根据国家方言对英语文档进行分类。载于澳大利亚语言技术协会论文集。

Shervin Malmasi 和 Marcos Zampieri。2016年。语音转录文本中的阿拉伯语方言识别。载于*VarDial*论文集。Shervin Malmasi 和 Marcos Zampieri。2017年。访谈转录文本中的德语方言识别。载于*VarDial*论文集。Shervin Malmasi、Marcos Zampieri、Nikola Ljubeši´c、Preslav Nakov、Ahmed Ali 和 Jörg Tiedemann。2016年。相似语言区分与阿拉伯语方言识别：第三届DSL共享任务报告。载于*VarDial*论文集。Puji Martadinata、Bayu Distiawan Trisedya、Hisar Maruli Manurung 和 Mirna Adriani。2016年。利用维基百科数据构建印度尼西亚地方语言检测工具。载于*WLSI*论文集。Paul McNamee。2005年。语言识别：一个已解决的问题，适合本科教学。《大学计算科学期刊》，20(3):94–101。Maria Medvedeva、Martin Kroon 和 Barbara Plank。2017年。稀疏传统模型何以胜过稠密神经网络：相似语言辨识这一有趣案例。载于*VarDial*论文集。Giovanni Molina、Nicolas Rey-Villamizar、Thamar Solorio、Fahad AlGhamdi、Mahmoud Ghoneim、Abdelati Hawwari 和 Mona Diab。2016年。代码切换数据语言识别首届共享任务概述。载于*CodeSwitch*论文集。

Francisco Rangel、Paolo Rosso、Martin Potthast 和 Benno Stein。2017 年。PAN 2017 年第五届作者画像任务概述：Twitter上的性别与语言变体识别。*CLEF*论文集。Thamar Solorio、Elizabeth Blair、Suraj Maharjan、Steven Bethard、Mona Diab、Mahmoud Ghoneim、Abdelati Hawwari、Fahad AlGhamdi、Julia Hirschberg、Alison Chang 和 Pascale Fung。2014年。代码切换数据语言识别首届共享任务概述。载于*CodeSwitch*论文集。Liling Tan、Marcos Zampieri、Nikola Ljubešic 和 Jörg Tiedemann。2014年。合并可比数据源以区分相似语言：DSL语料库合集。载于*BUCC*论文集。Jörg Tiedemann 和 Nikola Ljubeši´c。2012年。高效区分密切相关语言。载于*COLING*论文集。Erik Tromp 和 Mykola Pechenizkiy。2011年。基于图的短文本N元语法语言识别。载于*Benelearn*论文集。Marcos Zampieri、Shervin Malmasi、Nikola Ljubeši´c、Preslav Nakov、Ahmed Ali、Jörg Tiedemann、Yves Scherrer 和 Noëmi Aepli。2017年。VarDial 2017评测活动的研究发现。载于*VarDial*论文集。Marcos Zampieri、Shervin Malmasi、Preslav Nakov、Ahmed Ali、Suwon Shuon、James Glass、Yves Scherrer、Tanja Samardži´c、Nikola Ljubeši´c、Jörg Tiedemann、Chris van der Lee、Stefan Grondelaers、Nelleke Oostdijk、Antal van den Bosch、Ritesh Kumar、Bornini Lahiri 和 Mayank Jain。2018年。语言识别与形态句法标注：第二届VarDial评测活动。载于*VarDial*论文集。Marcos Zampieri、Shervin Malmasi、Yves Scherrer、Tanja Samardži´c、Francis Tyers、Miikka Silfverberg、Natalia Klyueva、Tung-Le Pan、Chu-Ren Huang、Radu Tudor Ionescu、Andrei Butnaru 和 Tommi Jauhiainen。2019年。第三届VarDial评测活动报告。载于*VarDial*论文集。Marcos Zampieri、Liling Tan、Nikola Ljubeši´c 和 Jörg Tiedemann。2014年。DSL 2014年共享任务报告。载于*VarDial*论文集。Marcos Zampieri、Liling Tan、Nikola Ljubeši´c、Jörg Tiedemann 和 Preslav Nakov。2015年。DSL共享任务2015概述。载于*LT4VarDial*论文集。Rowan Zellers、Ari Holtzman、Hannah Rashkin、Yonatan Bisk、Ali Farhadi、Franziska Roesner 和 Yejin Choi。2019年。防范神经网络生成的假新闻。载于神经信息处理系统大会论文集。Arkaitz Zubiaga、Inaki San Vicente、Pablo Gamallo、José Ramom Pichel、Inaki Alegria、Nora Aranberri、Aitzol Ezeiza 和 Víctor Fresno。2016年。Tweetlid：一种

推文语言识别基准。语言资源与评估，50(4)：729–766。