PALI：波斯-阿拉伯文字的语言识别基准测试

SinaAhmadi Milind Agarwal Antonios Anastasopoulos乔治梅森
大学计算机科学系{sahmad46,magarwa,antonis}@gmu.edu

摘要

波斯-阿拉伯文字是一系列在全球各地的不同语言社群中广泛采用和使用的文字系统。使用这类文字识别不同语言，对语言技术至关重要，但在低资源环境下颇具挑战。因此，本文探讨了使用波斯-阿拉伯文字识别语言所面临的挑战，尤其是在采用“非常规”书写方式的双语社区中。为应对这一问题，我们采用一组监督式技术，将句子分类到相应的语言类别。在此基础上，我们还提出了一种层次模型，专门针对分类器更容易混淆的语言簇。实验结果表明，我们提出的解决方案效果显著。[1]

1 引言

从历史上看，阿拉伯征服所带来的领土扩张在世界范围内引发了诸多持久变化，尤其从民族语言学角度来看，当时的当地语言面临着生存危机（Wasserstein，2003年）。阿拉伯语作为行政语言——一种行政语言——使许多语言受到影响，并在书写或词汇等方面以多种方式发生变化。几个世纪以来，波斯语通过增添字素扩展了阿拉伯文字，例如 <پ>（<p>，U+067E）和 <گ>（<g>，U+06AF）以适应该语言的音系。因此，古典阿拉伯文字的主要扩展变体之一便是波斯-阿拉伯文字；直到今天，许多其他语言也逐渐采用了这种文字系统，主要分布于西亚、中亚和南亚（Khansir 和 Moza- fari，2014年）。使用波斯-阿拉伯文字系统的一些语言包括乌尔都语、库尔德语、普什图语、阿塞拜疆土耳其语、信德语和维吾尔语，以及历史上曾使用这种文字的许多其他语言，例如奥斯曼土耳其语。除此之外，还有一些文字系统直接从阿拉伯文字采用而来，并未受到波斯语改造的影响，例如用于一些非洲语言（如斯瓦希里语和沃洛夫语）的阿贾米文字、南亚使用的佩贡文字和爪夷文字，以及历史上用于一些欧洲语言的阿尔哈米亚多文字。

语言识别是指在文档、句子和子句等不同层级上检测文本所用语言的任务。鉴于这项任务在自然语言处理（NLP）中的重要性，例如在机器翻译和信息检索中的应用，人们已对其进行了广泛研究，并发现它有助于情感分析和机器翻译等多种应用（Jauhiainen 等人，2019年）。不过，这项任务对不同的设置和语言而言难度并不相同。研究表明，对于较短文本或语言学上及书写形式上关系密切的语言，语言识别更具挑战性，例如法尔西语与达里语，或库尔德语的各种变体（Malmasi 等人，2015; Zampieri 等人，2020年）。

此外，双语社区中一些资源较少的语言，由于缺乏针对其母语的行政或教育支持，或技术工具有限，在书写方面面临各种挑战。这导致文本内容的书写方式不规范，即不遵循该语言的传统文字系统或正字法，而是依赖行政上的“优势语言”所使用的文字系统或正字法。例如，克什米尔语或库尔德语有时分别使用乌尔都语或波斯语的文字书写，而不是采用其既定的波斯-阿拉伯文字正字法。这进一步增加了识别这些语言的难度：文字系统相似容易造成混淆，数据稀缺也妨碍了数据驱动方法的应用。因此，准确识别使用波斯-阿拉伯文字的语言，至今仍是一项挑战，尤其是对于代表性不足的语言而言。

[1]数据和模型可在此处获取 https://github.com/sinaahmadi/PersoArabicLID

语言	ISO 639-3	WP	文字系统类型	变音符号	ZWNJ	主要
阿塞拜疆土耳其语	azb	azb	辅音音素文字	✓	✓	波斯语
吉拉基语	glk	glk	辅音音素文字	✓	✓	波斯语
马赞德兰语	mzn	mzn	辅音音素文字	✓	✓	波斯语
普什图语	pus	ps	辅音音素文字	✓	✗	波斯语
戈拉尼语	hac	-	字母表	✗	✗	波斯语、阿拉伯语、索拉尼语
北库尔德语（库尔曼吉方言）	kmr	-	字母表	✗	✗	波斯语、阿拉伯语
中库尔德语（索拉尼语）	ckb	ckb	字母表	✗	✗	波斯语、阿拉伯语
南库尔德语	sdh	-	字母表	✗	✗	波斯语、阿拉伯语
俾路支语	bal	-	辅音音素文字	✓	✗	波斯语、乌尔都语
布拉灰语	brh	-	辅音音素文字	✓	✗	Urdu
克什米尔语	kas	ks	字母表	✓	✗	Urdu
信德语	snd	sd	辅音音素文字	✓	✗	Urdu
萨莱基语	skr	skr	辅音音素文字	✓	✗	Urdu
托尔瓦利语	trw	-	辅音音素文字	✓	✗	Urdu
旁遮普语	pnb	pnb	辅音音素文字	✓	✗	Urdu
波斯语	fas	fa	辅音音素文字	✓	✓	-
阿拉伯语	arb	ar	辅音音素文字	✓	✗	-
Urdu	urd	ur	辅音音素文字	✓	✓	-
维吾尔语	uig	ug	字母表	✗	✗	-

表 1：本文研究的所选语言的波斯-阿拉伯文字。第 2 列和第 3 列分别显示这些语言在 ISO 639-3 中以及其专属维基百科（WP）（如有）的代码。“附加符号”和“零宽非连接符（ZWNJ）”列分别指示附加符号（哈拉卡特）和零宽非连接符作为单个字符的使用情况。

因此，我们选取了若干使用波斯-阿拉伯文字的语言，并在表1中进行了汇总。其中，大多数语言面临的挑战不仅与数据稀缺有关，也与非规范书写有关。因此，我们将这些语言的语言识别任务分为两种设置：a）文本按照该语言的文字系统或正字法书写，即规范书写；b）文本因使用行政优势语言的文字系统或正字法而存在一定程度的异常。鉴于波斯-阿拉伯文字主要用于巴基斯坦、伊朗、阿富汗和伊拉克的本土语言，我们还纳入了乌尔都语、波斯语和阿拉伯语，因为它们主要作为行政优势语言使用。此外，纳入更多样化的语言集合，有助于揭示哪些语言更容易相互混淆。尽管我们也纳入了维吾尔语，但需要注意的是，它主要通行于双语社区，即中国，而当地的非规范书写并非波斯-阿拉伯文字; 因此，我们仅考虑维吾尔语的规范书写。

贡献 本文探讨了使用波斯-阿拉伯文字或其变体书写的语言的语言识别问题。我们介绍了如何收集数据并利用文字映射生成合成含噪句子（§2）。我们实现了几种分类技术，并提出一种层次模型方法，以解决不同语言簇之间的混淆。对于噪声环境，所提出的方法优于其他技术，其宏平均 F_1 值介于 0.88 至 0.95 之间（§3）。

2 方法论

鉴于所选语言大多资源匮乏，收集数据，更重要的是，识别以规范和非规范方式书写的文本，是一项艰巨的任务。为解决这一问题，我们重点从网络上的各种来源收集数据，尤其是维基百科。[2] 随后，我们提出一种生成合成数据的方法，以期反映并建模非规范书写中出现的各种噪声。为此，我们采用一种简单的技术，将某种语言的文字系统中的字符映射到另一种语言（即优势语言）的文字系统中。最后，我们讨论了为这一任务构建基准测试所做的工作，并提出一种层次模型，用于消除相似语言之间的混淆。

[2] https://www.wikidata.org

2.1 数据收集

如表1 所示，除戈拉尼语、北库尔德语和南库尔德语、俾路支语、布拉灰语及托尔瓦利语外，所有语言都有使用其波斯-阿拉伯文字的专属维基百科页面。因此，我们使用现有语言对应的维基百科转储文件作为语料库。[3] 另一方面，对于北库尔德语和南库尔德语、俾路支语及布拉灰语，我们通过爬取表A.2中列出的地方新闻网站来收集数据。此外，我们使用 Uddin 和 Uddin (2019年) 的托尔瓦利语语料库、Ahmadi(2020年) 的戈拉尼语语料库、Esmaili 等人(2013) 的中库尔德语语料库，以及 Tehseen 等人(2022年) 的旁遮普语语料库。对于波斯语、阿拉伯语和乌尔都语，我们使用 Tatoeba 数据集。[4]

数据收集完成后，我们首先将各种格式转换为原始文本，然后进行文本预处理，使用正则表达式删除与格式样式相关的特殊字符，并移除电子邮件、电话号码和网站网址等信息。我们还将数字转换为拉丁字母形式的数字，因为波斯－阿拉伯文字文本中通常混用多种数字，即波斯语数字 <۰۱۲۳۴۵۶۷۸۹>、阿拉伯语数字 <٠١٢٣٤٥٦٧٨٩> numerals 以及拉丁字母形式的数字。这样做是为了确保后续用于语言识别任务的句子中包含多种数字。由于部分所选语言使用两种文字系统，例如旁遮普语使用古木基文和沙穆基文书写，克什米尔语使用天城文和波斯－阿拉伯文字书写，我们还使用了若干正则表达式，从语料库中移除存在文字系统切换和语码转换的句子以及引用的句子。鉴于检测此类切换的复杂性，我们指出，清理后的语料库中可能仍存在文字系统切换和语码转换词语。

我们通过统一字符的 Unicode 编码来完成文本预处理。Unicode 编码不一致通常是由于使用了代码绑定不同的键盘所致；此前，一些语言的预处理中也已纳入了这一处理步骤（Ahmadi, 2019; Doctor et al., 2022年）。例如，在库尔德语中，可能会使用 <ے> (U+06D2) 和 <ي> (U+064A)，而非 <ی>(U+06CC)；也可能会使用 <ك>(U+0643)，而非 <ک> (U+06A9)。如表 1 所示，考虑到零宽非连接符字符（ZWNJ，U+200C）的使用情况，我们还会在预处理步骤中考虑将其移除。[5] 最后，我们使用正则表达式在句子级别对语料库进行分词——级别。

表A.3 列出了收集的语料库中出现频率最高的 10 个三元组，其中包含许多可用于指示语言的词缀和连词。

2.2 文字映射

假设噪声文本使用优势语言的文字系统或正字法书写，我们便将给定语言的波斯-阿拉伯文字映射为优势语言的文字系统，例如，将克什米尔语文字映射为乌尔都语文字，或将中库尔德语文字映射为波斯语和阿拉伯语文字。为此，我们依据字符的视觉相似性和 Unicode 编码进行如下映射：

- 如果两种语言的文字系统中存在两个字素，例如在信德语和乌尔都语中，<ھ>(U+06BE)，或在萨莱基语和乌尔都语中，<ٹ> (U+0679)，则无论它们在两种语言中的发音如何，我们都会将它们对应起来。

- 若主导文字系统中没有完全相同的字素，则将视觉上最相似的字符映射到源字符。例如，在乌尔都语中，与布拉灰语的 <ڵ>(U+06B7) 最相似的字符是 <ل>(U+0644)。同样，吉拉基语中的 <ۆ> (U+06CB)会映射到波斯语中与之相似的 <و>(U+0648)。通过这种方式，一个字符可以映射到源语言中的多个其他字符。

- 有些映射遵循正字法规则，尤其适用于那些会因所在词语中的位置不同而发生变化的字符。例如，库尔德语中的元音前面会出现一个起首的哈姆扎，即 <ـئ>(U+0626)，如 <ـئۆ> /oː/和<ـئێ> /eː/。我们也纳入了这类规则。

- 由于数据收集过程中数字已统一（§2.1），我们也将拉丁数字随机映射到波斯语和阿拉伯语的数字。

根据优势语言，针对每种源语言和优势语言，人工创建文字映射。需要注意的是，除了不带附加符号的字符外，如果附加符号（包括哈拉卡特）属于字素的一部分，也会纳入带附加符号的字符，例如戈拉尼语和信德语中的 <ڈ>(U+068E)，但不包括 <ا́>。可分离的哈拉卡特，如法塔哈、卡斯拉和达马，不纳入文字映射。表A.1根据所选语言与阿拉伯语、波斯语和乌尔都语这三种主要的波斯-阿拉伯文字使用语言之间的关系，列出了这些语言中使用的字符集。

[3] 截至 2023 年 1 月 20 日的转储文件。

[4] https://tatoeba.org [5] 我们从网络上的各种来源获取信息关于所选语言中的常见书写习惯，尤其是 https://scriptsource.org。

2.3 合成数据生成

利用文字映射，我们以“干净数据集”中的句子（即所收集语料库中的句子）为基础生成合成句子，模拟非规范书写方式。具体做法是根据我们的映射，随机将干净句子中的字符替换为目标文字系统中的替代字符。为了评估噪声对语言识别的影响，我们合成了不同噪声水平的数据，范围从20%到100%；具体噪声水平根据可能替换的数量设定。表2展示了一个北库尔德语干净句子，以及不同噪声水平下对应的合成噪声句子。

因此，数据集分类如下：

1. 干净数据集：包含语料库中的原始句子，未注入任何噪声。相当于数据中的噪声比例为 0%。其中包括所有选定的语言，以及乌尔都语、波斯语、阿拉伯语和维吾尔语。

2. 噪声数据集：包含带有不同程度噪声字符的句子，噪声比例从 20% 起，每次递增 20%，直至 100%。无论是否使用，可移除变音符号都会在噪声水平为 100% 时被移除；这也适用于严格使用附加符号的克什米尔语。我们将所有噪声水平的数据合并为一个单独的数据集，称为全部。由于波斯语、乌尔都语、阿拉伯语和维吾尔语不存在非规范书写方式，因此不将它们纳入噪声数据。

3. 合并：将干净数据集与全部数据集合并后的结果。

干净数据集和噪声数据集每种语言均包含 10,000 个句子，但布拉灰语、托尔瓦利语和俾路支语除外，其语料库中分别只有 549、1371 和 1649 个句子。因此，我们将这些语言各 500 个句子纳入测试集，并以四倍系数对剩余句子进行上采样，即将剩余句子各复制四次，然后将其用作训练集。同样，对于克什米尔语和戈拉尼语，其语料库中分别有 6340 和 8742 个句子；我们将 2000 个句子加入测试集，并对剩余句子进行上采样，使训练集达到 8000 个句子。

为避免不存在噪声的主导语言（即乌尔都语、波斯语、阿拉伯语）以及维吾尔语之间的数据失衡，

噪声 %	句子
干净数据集	在比利时举办的第二届库尔德摄影师展览
	在比利时举办的第二届库尔德摄影师展览
20	在比利时举办的第二届库尔德摄影师展览
40	شكراً لكم على مشاهدة هذا العرض التقديمي.
60	无法辨识的文本
80	比利时第二届库尔德摄影师展览
100	比利时第二届库尔德摄影师展览

表2：北库尔德语（库尔曼吉方言）中的一个句子，以及基于不同噪声等级生成的合成噪声样本。

我们从各自的干净语料库中额外加入了10,000个样本。因此，MERGED数据集中的每种语言均包含20,000个干净数据集和含噪句子。

2.4 基准测试

我们将语言识别视为一个概率分类问题，其中预测每个句子属于某个特定类别（即某种语言）的概率。按照前几节所述，我们将各数据集中的句子按80/20的比例划分为训练集和测试集。两个集合均来自同一数据。

作为基线系统，我们使用 fastText 的预训练语言识别模型–`lid.176`该模型使用来自维基百科、Tatoeba 和 SETimes 的数据进行训练，涵盖176种语言，包括所有所选语言，但不包括俾路支语、布拉灰语、吉拉基语、戈拉尼语、北库尔德语（波斯-阿拉伯文字系统）、南库尔德语和托尔瓦利语。此外，我们还使用 fastText 训练了一个模型，其词向量大小为64，字符 n-gram 的最小和最大长度为2至6，学习率为1.0，训练25轮，并采用分层 softmax 损失。

除与 fastText 相关的基线模型和我们自己的模型外，我们还报告了精确率、召回率和 F_1 分数，以便将谷歌的 CLD3（Salcianu 等人，2020年）、`Franc`[6] 以及 `Langid.py` （Lui 和 Baldwin，2012）等最先进的方法作为基准进行比较。我们还提供另外两个从头训练的基线模型，它们使用字符 n二元至四元语法特征：多项式朴素贝叶斯模型（MNB——类别先验不均匀且通过学习得到，不进行拉普拉斯平滑），以及多层感知机（MLP），其最大迭代次数为 500，包含一个大小为 500 的隐藏层，批次大小为 1000。

[6] https://github.com/wooorm/franc/

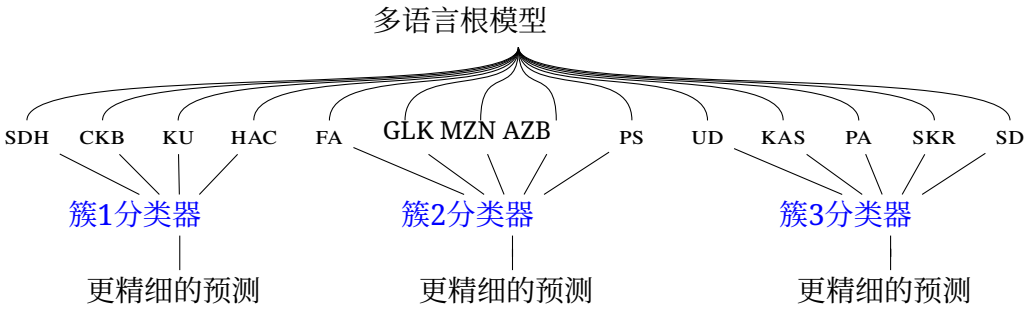

图1：我们层次模型的架构。如果根模型预测为南库尔德语（SDH）、戈拉尼语（HAC）、北库尔德语（KMR）或中库尔德语（CKB），则该样本会被送入一个较小的专家分类器，该分类器经过训练，可区分这四种关系密切的语言。簇2和簇3中的语言也采用同样的处理方式。如果根模型预测为未聚类语言，即没有可用的分支，则层次模型会给出与根模型相同的标签。

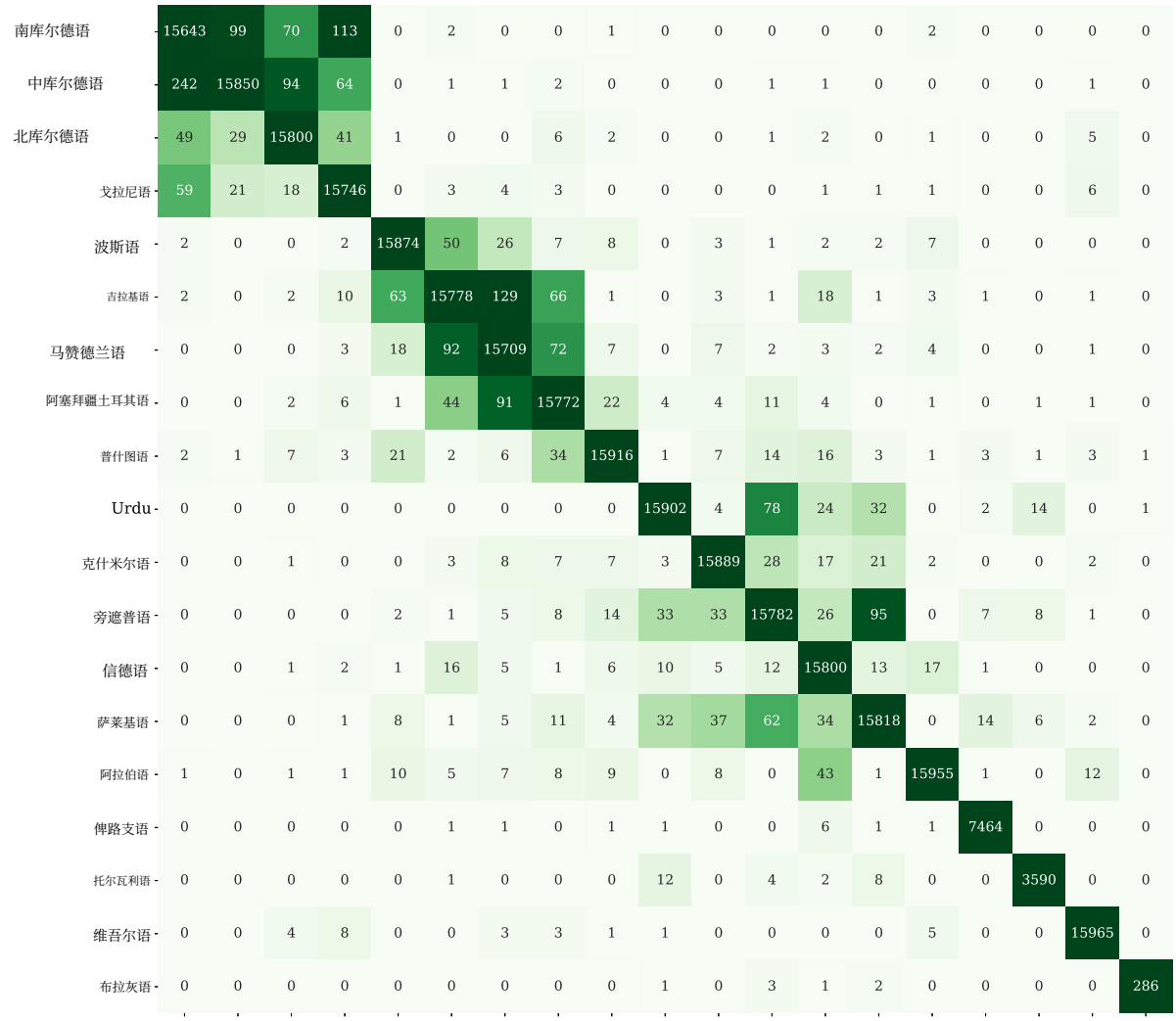

图2：多语言根模型在训练数据集上的混淆矩阵。行标签表示我们定制的 fastText 模型的预测，列表示真实标签（训练数据集），每个单元格中的计数表示模型对某个（预测，真实标签）组合所作出的预测次数。根据该混淆矩阵，我们识别出三个高度混淆的语言簇，如第2.5节所述。

2.5 层次建模

层次建模（图1）旨在通过训练专门用于区分少量语言的专家分类器，解决模型对高度相关语言的混淆问题。为此，我们检查表现最佳的模型（在训练数据上）的混淆矩阵，并找出模型在预测时容易混淆的语言簇。上一节所述的定制 fastText 模型作为根分类器。我们根据其混淆矩阵识别出以下三个簇（图2）：

1. 簇 1 包含南库尔德语、中部库尔德语、北库尔德语和戈拉尼语 2. 簇2 包含波斯语、吉拉基语、马赞德兰语、阿塞拜疆土耳其语和普什图语 3. 簇 3 包含乌尔都语、克什米尔语、旁遮普语、信德语和萨莱基语

层次树中的每个子单元都是一个 fastText 模型，使用相关簇中各语言的数据从头训练，并采用与根模型相同的参数设置。

3 结果

在表3中，我们报告了所有数据集、6 个最先进基线和定制训练基线、我们的根 fastText 模型（Root）以及层次混淆消解模型（Hier）的精确率、召回率和 F_1分数。与预训练的 fastText 基线、Google的CLD3 相比，我们的根 fastText 模型表现出显著优势，`langid.py`、`Franc`多项式朴素贝叶斯和多层感知机。

3.1 最先进模型与简单基线

在我们的测试集上，面对全部 19 种语言和各种噪声设置，三种最先进的模型（CLD3、`langid.py`、`Franc`）的 F_1 分数均未超过 0.15。事实上，在混合噪声设置（40% - 全部）下，它们的 F_1 分数往往极低（$0 \leq F_1 < 0.1$）。尽管这些模型支持乌尔都语、波斯语、阿拉伯语和信德语，且 `Franc` 还支持中库尔德语，结果仍是如此。这表明，尽管最先进的预训练模型声称覆盖数百种语言，其语言识别质量依然很差，进一步说明语言识别问题远未得到解决。与这三种模型相比，MNB 和多层感知机模型在所有噪声级别（20% 噪声除外）下表现更好，甚至在 8 种噪声设置中的 7 种下胜过 fastText 的大型预训练模型 `lid.176` 在 8 种噪声设置中的 7 种下，使其成为比 `lid.176` 模型更强的基线。

3.2 使用 fastText 进行层次建模

说到我们的两个模型，即定制的 fastText 模型（Root）和层次混淆消解模型（Hier），显然，与任何基线相比，这两个模型的表现都明显更好，优势巨大。由于层次模型是在 MERGED数据集上训练的，该数据集包含噪声句子和干净句子，并且比干净数据集（0% 噪声）设置多四个类别，因此 Root 模型在干净数据集设置下表现更好是自然的。然而，对于任何现实的噪声水平（从20%到MERGED），层次模型的表现都优于 Root 模型。

为了检验这些细微的改进，我们采用单尾 Z 检验，在 0.01 的显著性水平下比较根模型和层次模型，并报告各噪声水平下的统计显著性结果。我们采用 Z 检验，是因为样本数量大于 30，且样本方差可作为总体方差的可靠估计。零假设是根模型与层次模型之间不存在显著差异（μ_0： $f_{root} = f_{hier}$），备择假设则认为层次模型的性能显著且严格高于根模型（μ_1： $f_{root} < f_{hier}$）。我们为根模型的 F_1 分数 f_{root}计算单尾 99% 置信区间。根据单尾 Z 检验，如果层次模型的 F_1 分数 f_{hier}严格高于该区间的上界，我们就可以拒绝零假设，并得出 F_1分数之间的差异具有统计显著性的结论。表4报告了我们的假设检验结果。我们发现，在所有噪声设置下，我们的层次化混淆消解方法所带来的优势在 99% 置信度下均具有统计显著性。因此，我们证实，利用混淆信息的层次化方法无需重新训练整个模型，即可用于提升模型在噪声数据上的性能；该方法在测试集上也表现良好，带来了具有统计显著性的改进。

3.3 特定语言性能

在表5中，我们报告了两个表现最佳的系统在不同噪声级别下的语言级别分数：我们定制的 fastText 模型和我们的混淆消解层次模型。在所有语言和噪声级别下，层次模型仅在

噪声	指标	Hier	根	fastText	CLD3	langid.py		Franc	MNB	MLP
0%	精确率	0.72	**0.91**	0.16	0.03	0		0.02	0.43	0.47
	召回率	0.70	**0.89**	0.07	0.05	0		0.02	0.14	0.16
	F_1 分数	0.72	**0.90**	0.10	0.04	0		0.02	0.21	0.24
20%	精确率	**0.92**	0.92	0.30	0.08	0.13		0.13	0.08	0.03
	召回率	**0.89**	0.89	0.32	0.18	0.18		0.18	0.05	0.05
	F_1 分数	**0.91**	0.90	0.31	0.11	0.15		0.15	0.06	0.04
40%	精确率	**0.91**	0.90	0.17	0.04	0		0.01	0.51	0.49
	召回率	**0.88**	0.88	0.07	0.05	0		0	0.09	0.11
	F_1 分数	**0.90**	0.89	0.10	0.05	0		0	0.16	0.19
60%	精确率	**0.91**	0.90	0.17	0.04	0		0	0.45	0.54
	召回率	**0.88**	0.87	0.07	0.05	0		0	0.12	0.09
	F_1 分数	**0.89**	0.88	0.09	0.04	0		0	0.20	0.15
80%	精确率	**0.90**	0.90	0.16	0.03	0		0	0.25	0.33
	召回率	**0.88**	0.87	0.06	0.05	0		0	0.12	0.15
	F_1 分数	**0.89**	0.88	0.08	0.04	0		0	0.16	0.21
100%	精确率	**0.90**	0.90	0.15	0.03	0		0	0.44	0.44
	召回率	**0.88**	0.87	0.06	0.05	0		0	0.08	0.11
	F_1 分数	**0.89**	0.88	0.08	0.03	0		0	0.13	0.17
ALL	精确率	**0.90**	0.89	0.15	0.03	0		0	0.28	0.51
	召回率	**0.87**	0.86	0.06	0.05	0		0	0.16	0.10
	F_1 分数	**0.88**	0.88	0.08	0.04	0		0	0.20	0.17
合并	精确率	**0.95**	0.95	0.28	0.06	0.11		0.11	0.15	0.15
	召回率	**0.94**	0.94	0.27	0.16	0.16		0.16	0.08	0.07
	F_1 分数	**0.95**	0.94	0.27	0.09	0.13		0.13	0.10	0.10

表 3：不同噪声设置下所有语言识别模型的精确率、召回率和 F_1 分数对比。我们的层次分类（Hier）模型和根模型在所有噪声级别下均表现最佳，位列前两名。fastText、多项式朴素贝叶斯（MNB）和多层感知机（MLP）在不同噪声级别下排名第三。报告所有方法的精确率、召回率和 F_1 分数，以提供基准。对于精确到小数点后两位相同的数值，加粗条目表示性能严格更优。

噪声	测试样本	Δ	显著
0	33500	-0.188	✗
20	25500	0.005	✓
40	25500	0.006	✓
60	25500	0.007	✓
80	25500	0.007	✓
100	25500	0.007	✓
ALL	27806	0.007	✓
合并	69304	0.002	✓

表 4：与根模型相比，我们的层次建模方法在所有噪声水平下的 F_1 分数提升（正值 Δ）均具有统计显著性，显著性水平为 $=0.01$，即置信度为 99%。

128 种设置中的 5 种表现较差。其余设置下，它的表现要么与根模型持平，要么更好。粗体条目表明，层次模型在三个已识别的簇中，在噪声环境（20%-全部）下带来的提升最大。正如预期，对于不属于任何高度混淆簇的语言，即 AR、BAL、TRW、UG 和BRH，层次模型与根模型会产生相同的预测，因此在各个噪声级别下得分相同。在表A.4中，我们还根据预训练 fastText 模型的预测，提供了不同噪声级别下的一些语言识别示例，并将其与我们的模型进行比较。

	0%		20%		40%		60%		80%		100%		ALL		合并	
	M1	M2	M1	M2	M1	M2	M1	M2	M1	M2	M1	M2	M1	M2	M1	M2
簇 1																
SDH	0.95	**0.96**	0.95	**0.96**	0.94	**0.95**	0.93	**0.94**	0.93	**0.94**	0.94	0.94	0.94	0.94	0.95	**0.96**
CKB	0.95	0.95	0.94	0.94	0.92	**0.94**	0.91	**0.93**	0.91	**0.93**	0.91	**0.92**	0.92	**0.93**	0.95	0.95
KU	0.95	0.95	0.93	**0.94**	0.93	0.93	0.92	**0.93**	**0.93**	0.92	0.92	0.92	0.92	**0.93**	0.95	0.95
HAC	0.94	0.94	0.94	0.94	0.93	0.93	0.92	0.92	0.92	0.92	0.92	0.92	0.91	**0.92**	0.94	0.94
簇2																
FA	0.97	**0.98**	-	-	-	-	-	-	-	-	-	-	-	-	0.97	**0.98**
GLK	0.92	**0.94**	0.88	**0.89**	0.88	**0.89**	0.88	**0.9**	0.88	**0.89**	0.88	**0.89**	0.92	0.92	0.92	**0.94**
MZN	0.92	0.92	0.85	**0.86**	0.85	**0.86**	0.85	**0.87**	0.85	**0.86**	0.85	**0.87**	0.92	**0.93**	0.92	0.92
AZB	0.91	0.91	0.86	**0.87**	0.85	**0.86**	0.86	**0.87**	0.86	**0.87**	0.85	**0.86**	0.9	**0.91**	0.91	0.91
PS	0.96	0.96	0.94	**0.95**	0.94	**0.95**	0.94	**0.95**	0.94	**0.95**	0.94	0.94	0.96	0.96	0.96	0.96
簇 3																
UD	0.96	**0.97**	-	-	-	-	-	-	-	-	-	-	-	-	0.96	**0.97**
KAS	0.94	**0.95**	0.9	**0.91**	0.9	**0.91**	0.9	**0.91**	0.9	**0.91**	0.87	**0.88**	**0.91**	0.9	0.94	**0.95**
PA	0.91	0.91	**0.87**	0.86	0.86	0.86	0.86	0.86	0.85	**0.86**	0.85	0.85	0.87	0.87	0.91	0.91
SD	0.93	**0.94**	0.89	**0.91**	0.88	**0.89**	0.87	**0.89**	0.87	**0.89**	0.87	**0.89**	0.91	0.91	0.93	**0.94**
SKR	**0.92**	0.91	0.85	0.85	0.84	**0.85**	0.84	**0.85**	0.85	0.85	0.84	**0.85**	0.86	**0.88**	**0.92**	0.91
AR	0.98	0.98	-	-	-	-	-	-	-	-	-	-	-	-	0.98	0.98
BAL	0.98	0.98	0.94	0.94	0.94	0.94	0.94	0.94	0.95	0.95	0.95	0.95	0.97	0.97	0.98	0.98
TRW	0.95	0.95	0.87	0.87	0.89	0.89	0.88	0.88	0.88	0.88	0.87	0.87	0.91	0.91	0.95	0.95
UG	0.99	0.99	-	-	-	-	-	-	-	-	-	-	-	-	0.99	0.99
BRH	0.84	0.84	0.7	0.7	0.67	0.67	0.68	0.68	0.68	0.68	0.65	0.65	0.63	0.63	0.84	0.84

表5：我们的层次模型（M1）和根模型（M2）在语言级别上的 F_1 分数。与根模型相比，我们的层次模型在三个簇（从上往下的前 3 个部分）中的语言上，在不同噪声水平下的 F_1 分数均有所提升。虚线单元格表示该语言只有传统文字，因此不属于合成数据设置。

4 相关工作

建模方法 语言识别通常被建模为多类别文本分类任务，借助简单的字节、字符或词级 n-gram 特征，在不同语言和语言变体以及数据有限的场景中，已取得最先进的性能（Jauhiainen 等，2017年）。模型或分类器的选择高度依赖于数据来源、领域以及每种语言的数据量。诸如支持向量机（Ciobanu 等，2018; Malmasi 和 Dras，2015）和多项式朴素贝叶斯（King 等，2014;Mathur 等，2017年）这样的简单线性分类器，在数据和计算资源有限的情况下，能在不同领域提供强有力的基线。如果有大量数据可用，则可以使用聚合分类器（Baimukan 等，2022年）和神经模型，但它们在区分相似的语言变体和方言时仍面临困难，而且容易过拟合（Medvedeva 等，2017; Criscuolo 和 Aluísio，2017; Eldesouki 等，2016）。

在我们的论文中，我们提出了一种层次化的语言识别方法，用于识别噪声环境中常见的易混淆语言对，并通过小型分类单元纠正此类误判。这种建模方法可用于扩大语言覆盖范围，并提升现有预训练模型的性能，而无需重新训练计算开销巨大的大型模型。在我们的研究中，我们发现该方法在噪声数据环境下带来了统计显著的改进。

相似语言与变体 语言识别是一个研究充分的问题，有时甚至被认为 已经解决; 但实际上，世界上大多数语言都不受当前系统支持。这种代表性不足影响大规模数据挖掘工作，并进一步加剧低资源语言的数据短缺。提升语言识别系统语言覆盖范围的一个关键瓶颈，是区分相似语言、语言变体和方言的能力。正如本文所述，当一个语言社群采用优势语言的不常见文字系统时，这一问题会更具挑战性。近期，已有研究致力于区分北欧语言（Haas 和 Derczynski，2021年），

阿拉伯语方言（Nayel 等人， 2021; Abdul-Mageed 等人， 2020; Salameh 等人，2018）以及意大利语和法语的区域语言变体（Jauhiainen 等人， 2022;Camposampiero 等人，2022年）。Haas 和 Derczynski （2021年）例如尝试了多种建模和特征提取方法，以尽可能准确地区分六种北欧语言：丹麦语、瑞典语、挪威语（新挪威语）、挪威语（书面挪威语）、法罗语和冰岛语。他们发现，从 fastText 中提取的 Skip-gram 嵌入信息丰富，能够区分密切相关的语言。值得注意的是，尽管论文提出的方法提升了所选语言的识别效果，但这六种所选北欧语言都有大量训练数据（50K+ 个句子），且现成工具已经支持这些语言，例如 `langid.py`。这与我们的工作不同：我们的工作将先前未受支持的语言和变体纳入语言识别系统并进行评估。

在区分相似语言和方言时，朴素贝叶斯和逻辑回归等较浅层的线性分类器往往优于多层感知机或卷积神经网络等神经模型（Chakravarthi 等， 2021;Aepli 等， 2022; Ceolin，2021年）。这一点得到了以下事实的印证：非神经的经典机器学习方法在 VarDial 2021 和 2022年的大多数共享任务中胜出，这些任务涵盖了类型多样的语言，包括达罗毗荼语系语言、罗马尼亚语方言、意大利语和法语的地域变体（Jauhiainen 等， 2022; Camposampiero 等，2022年）以及乌拉尔语系语言（Chakravarthi 等，2021年）。由于相似语言或变体的数据有限，神经建模方法的表现有时也可能不及非神经基线，如乌拉尔语系语言识别和意大利语方言识别共享任务中所报告的那样（Chakravarthi 等， 2021; Aepliet 等，2022年）。

5 结论

我们聚焦于双语社区中使用的语言。在这些社区中，人们常常使用非规范的主导波斯-阿拉伯文字系统，而非规范且更适用的波斯-阿拉伯文字变体书写系统。我们探讨这一情境在数据收集和语言识别方面的独特挑战，以及最先进的系统面对这类非规范书写系统中的数据时出现的性能问题。这一点可从噪声数据集与干净数据集/混合设置之间相差20个百分点的性能差异中看出，体现在F_1 分数上。我们提出的层次化方法优于定制训练的fastText系统、简单的MNB和MLP，以及谷歌的CLD3、 `Franc` 和 `langid.py`等最先进的语言识别系统。在分析根多语模型的混淆矩阵后，我们发现，采用层次化模型可带来统计显著的改进。

6 局限性

部分所选语言使用多种文字系统，例如旁遮普语或库尔德语。这会影响自动预处理的采集数据的质量。因此，我们认为数据集中包含少量但确实存在的语码转换文本。此外，由于我们专注于波斯-阿拉伯文字，因此未纳入这些语言使用其他文字系统书写的文本。尽管一种语言可能受到多种优势语言的影响，且合成数据的生成考虑了各种文字映射，但各个优势语言的影响仍有待分析。为此，应针对每种优势语言定义更细粒度的分类任务。

此外，还可以将波斯语的变体（如达里语和法尔西语），以及所选语言的次方言纳入此任务。同样，我们的层次化方法也可应用于其他文字系统，尤其是被多种语言采用的文字系统，例如西里尔字母和拉丁字母。最后，还可以基于我们采集的数据实现并微调其他技术。

通常而言，预期所提出的模型在更多数据上训练时会表现得更好。我们也认为，层次模型相较于根模型的改进受到训练集规模的限制。如果能获得更多真实的噪声数据，我们在所有噪声设置以及干净数据设置下的性能都有可能提升。

致谢

本研究得到了美国国家科学基金会 DEL/DLI 项目资助（资助编号 BCS-2109578）以及美国国家人文基金会资助（资助编号 PR-276810-21）。作者还感谢匿名审稿人，以及乔治梅森大学（GMU）研究计算办公室；所有计算实验均在该办公室完成。

参考文献

Muhammad Abdul-Mageed、Chiyu Zhang、Houda Bouamor 和 Nizar Habash。2020年。NADI 2020：首届细分阿拉伯语方言识别共享任务。载于第五届阿拉伯语自然语言处理研讨会论文集，第 97–110 页。

Noëmi Aepli、Antonios Anastasopoulos、Adrian-Gabriel Chifu、William Domingues、Fahim Faisal、Mihaela Gaman、Radu Tudor Ionescu 和 Yves Scherrer。2022年。VarDial评测活动 2022 的研究结果。载于第九届相似语言、变体和方言自然语言处理研讨会论文集，第 1–13 页，韩国大韩民国庆州。计算语言学协会。

Sina Ahmadi。2019年。一种基于规则的库尔德语文本转写系统。*ACM* 亚洲及低资源语言信息处理汇刊（*TALLIP*），18(2)：1–8。

Sina Ahmadi。2020年。构建扎扎-戈拉尼语族语料库。载于第七届相似语言、变体和方言自然语言处理研讨会论文集，*VarDial@COLING 2020*，西班牙巴塞罗那（线上），*2020*年*12*月*13*日，第 70–78 页。国际计算语言学委员会（ICCL）。

Nurpeiis Baimukan、Houda Bouamor 和 Nizar Habash。2022年。用于阿拉伯语方言识别的方言数据层次聚合。载于第十三届语言资源与评测会议论文集，第4586–4596页。

Giacomo Camposampiero、Quynh Anh Nguyen 和 Francesco Di Stefano。2022年。用于意大利语言和方言识别的逻辑回归奇案。载于第九届相似语言、变体与方言自然语言处理研讨会论文集，第86–98页，韩国庆州。计算语言学协会。

Andrea Ceolin。2021年。比较卷积神经网络和浅层模型在语言识别方面的性能。载于第八届相似语言、变体与方言自然语言处理研讨会论文集，第102–112页，乌克兰基辅。计算语言学协会。

Bharathi RajaChakravarthi、Gaman Mihaela、Radu Tu- dor Ionescu、Heidi Jauhiainen、Tommi Jauhiainen、Krister Lindén、Nikola Ljubešić、Niko Partanen、Ruba Priyadharshini、Christoph Purschke、Eswari Rajagopal、Yves Scherrer 和 Marcos Zampieri。2021年。VarDial 2021评测活动的研究成果。载于《第八届相似语言、变体与方言自然语言处理研讨会论文集》，第 1–11 页，乌克兰基辅。计算语言学协会。

Alina Maria Ciobanu、Marcos Zampieri、Shervin Malmasi、Santanu Pal 和 Liviu P. Dinu。2018 年。使用 SVM 集成区分印度-雅利安语。收录于第五届研讨会论文集，主题为相似语言、变体和方言自然语言处理（*VarDial2018*），第 178–184 页，美国新墨西哥州圣菲。计算语言学协会。

Marcelo Criscuolo 和 Sandra Maria Aluísio。2017年。运用词级卷积神经网络区分相似语言。载于《第四届相似语言、变体和方言自然语言处理研讨会（*VarDial*）论文集》，第 124–130 页，西班牙瓦伦西亚。计算语言学协会。

Raiomond Doctor、Alexander Gutkin、Cibu Johny、Brian Roark 和 Richard Sproat。2022年。《波斯-阿拉伯文字的字形规范化》。*arXiv* 预印本 *arXiv:2210.12273*。

Mohamed Eldesouki、Fahim Dalvi、Hassan Sajjad 和 Kareem Darwish。2016年。QCRI @ DSL 2016：使用文本特征识别口语阿拉伯语方言。收录于第三届相似语言、变体与方言自然语言处理研讨会（*VarDial3*），第221–226页，大阪，日本。COLING 2016 组织委员会。

Kyumars Sheykh Esmaili、Donya Eliassi、Shahin Salavati、Purya Aliabadi、Asrin Mohammadi、Somayeh Yosefi 和 Shownem Hakimi。2013年。为索拉尼库尔德语构建测试集。收录于*2013*年 *ACS* 国际计算机系统与应用会议（*AICCSA*），第1–7页。IEEE。

René Haas 和 Leon Derczynski。2021年。区分相似北欧语言。收录于第八届相似语言、变体与方言自然语言处理研讨会论文集，第 67–75 页，乌克兰基辅。计算语言学协会。

Tommi Jauhiainen、Heidi Jauhiainen 和 Krister Lindén。2022年。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国区域变体检测。收录于第九届相似语言、变体与方言自然语言处理研讨会论文集，第 119–129 页，韩国庆州。计算语言学协会。

Tommi Jauhiainen、Krister Lindén 和 Heidi Jauhiainen。2017年。使用 285 种语言评估语言识别方法。收录于第 *21* 届北欧计算语言学会议论文集，第 183–191 页，瑞典哥德堡。计算语言学协会。

Tommi Jauhiainen、Marco Lui、Marcos Zampieri、Ti m- othy Baldwin 和 Krister Lindén。2019年。文本中的自动语言识别：综述。《人工智能研究杂志》，65：675–782。

Ali Akbar Khansir 和 Nasrin Mozafari。2014年。波斯语对印度语言的影响。《语言研究理论与实践》，4(11)：2360。

Ben King、Dragomir Radev 和 Steven Abney。2014年。使用相似语言组进行句子语言识别的实验。载于《

首届将自然语言处理工具应用于相似语言、变体和方言研讨会论文集，第146–154页，爱尔兰都柏林。计算语言学协会和都柏林城市大学。

MarcoLui 和 Timothy Baldwin。2012 年。langid.py：开箱即用的语言识别工具。载于 *ACL 2012* 系统演示论文集，第 25–30 页，韩国济州岛。计算语言学协会。

Shervin Malmasi 和 Mark Dras。2015 年。使用分类器集成进行语言识别。载于语言技术在密切相关语言、变体和方言中的联合研讨会论文集，第 35–43 页，保加利亚希萨尔。计算语言学协会。

Shervin Malmasi、Mark Dras 等。2015 年。针对波斯语和达里语文本的自动语言识别。载于*PACLING* 论文集，第 59–64 页。

Priyank Mathur、Arkajyoti Misra 和 Emrah Budur。2017 年。LIDE：从文本文档中进行语言识别。*CoRR*，abs/1701.03682。

Maria Medvedeva、Martin Kroon 和 Barbara Plank。2017年。稀疏传统模型优于稠密神经网络：区分相似语言时的一个奇特案例。载于第四届相似语言、变体和方言自然语言处理研讨会论文集（*VarDial*），第156–163页，西班牙瓦伦西亚。计算语言学协会。

Hamada Nayel、Ahmed Hassan、Mahmoud Sobhi 和 Ahmed El-Sawy。2021年。基于机器学习的阿拉伯语方言识别方法。载于第六届阿拉伯语自然语言处理研讨会论文集，第287–290页，乌克兰基辅（线上）。计算语言学协会。

Mohammad Salameh、Houda Bouamor 和 Nizar Habash。2018年。阿拉伯语方言细粒度识别。载于 第*27*届国际计算语言学会议（*COLING 2018*），第1332–1344页。计算语言学协会（ACL）。

Alex Salcianu、Andy Golding、Anton Bakalov、Chris Alberti、Daniel Andor、David Weiss、Emily Pitler、Greg Coppola、Jason Riesa、Kuzman Ganchev、Michael Ringgaard、Nan Hua、Ryan McDonald、Slav Petrov、Stefan Istrate 和 Terry Koo。2020年。紧凑型语言检测器 v3（CLD3）。

Amina Tehseen、Toqeer Ehsan、Hannan Bin Liaqat、Amjad Ali 和 Ala Al-Fuqaha。2022年。使用上下文化词表示对沙穆基文进行神经网络词性标注。沙特国王大学学报：计算机与信息科学。

Naeem Uddin 和 Jalal Uddin。2019年。迈向托尔瓦利语机器翻译：低资源语言形态句法挑战分析。收录于第二届低资源语言机器翻译技术研讨会，第 6–10 页，都柏林，爱尔兰。欧洲机器翻译协会。

David J Wasserstein。2003年。为什么阿拉伯语取得了成功，而希腊语却未能成功？穆罕默德之后近东地区的语言变迁。*Scripta ClassicaIsraelica*，22：257–272。

Marcos Zampieri、Preslav Nakov 和 Yves Scherrer。2020年。相似语言、变体和方言的自然语言处理：一项综述。*NaturalLanguageEngineering*，26(6)：595–612。

A 所选语言

字素	库尔德语	戈拉尼语	维吾尔语	阿拉伯语	阿塞拜疆语	吉拉基语	马赞德兰语	波斯语	乌尔都语	克什米尔语	旁遮普语	萨莱基语	托尔瓦利语	普什图语	信德语	布拉灰语	俾路支语
Arabic	ئ ا ب پ ت ج چ ح خ د ر ز ژ س ش ع غ ف ڤ ق ک گ ل م ن ه و ۆ ی ێ ە	ئ ا ب پ ت ج چ ح خ د ر ز ژ س ش ع غ ف ڤ ق ک گ ل م ن ه و ۆ ی ێ ە ھ ۵ ڵ ۊ ۋ ۑ ۉ ۇ ۈ ۅ ۍ ې ے	ئ ا ب پ ت ج چ ح خ د ر ز ژ س ش ع غ ف ڤ ق ک گ ل م ن ه و ۆ ی ێ ە ھ ۵ ڵ ۊ ۋ ۑ ۉ ۇ ۈ ۅ ۍ ې ے	ئ ا ب پ ت ج چ ح خ د ر ز ژ س ش ع غ ف ڤ ق ک گ ل م ن ه و ۆ ی ێ ە ھ ۵ ڵ ۊ ۋ ۑ ۉ ۇ ۈ ۅ ۍ ې ے	ئ ا ب پ ت ج چ ح خ د ر ز ژ س ش ع غ ف ڤ ق ک گ ل م ن ه و ۆ ی ێ ە ھ ۵ ڵ ۊ ۋ ۑ ۉ ۇ ۈ ۅ ۍ ې ے	ئ ا ب پ ت ج چ ح خ د ر ز ژ س ش ع غ ف ڤ ق ک گ ل م ن ه و ۆ ی ێ ە ھ ۵ ڵ ۊ ۋ ۑ ۉ ۇ ۈ ۅ ۍ ې ے	ئ ا ب پ ت ج چ ح خ د ر ز ژ س ش ع غ ف ڤ ق ک گ ل م ن ه و ۆ ی ێ ە ھ ۵ ڵ ۊ ۋ ۑ ۉ ۇ ۈ ۅ ۍ ې ے	ئ ا ب پ ت ج چ ح خ د ر ز ژ س ش ع غ ف ڤ ق ک گ ل م ن ه و ۆ ی ێ ە ھ ۵ ڵ ۊ ۋ ۑ ۉ ۇ ۈ ۅ ۍ ې ے	ئ ا ب پ ت ج چ ح خ د ر ز ژ س ش ع غ ف ڤ ق ک گ ل م ن ه و ۆ ی ێ ە ھ ۵ ڵ ۊ ۋ ۑ ۉ ۇ ۈ ۅ ۍ ې ے	ئ ا ب پ ت ج چ ح خ د ر ز ژ س ش ع غ ف ڤ ق ک گ ل م ن ه و ۆ ی ێ ە ھ ۵ ڵ ۊ ۋ ۑ ۉ ۇ ۈ ۅ ۍ ې ے	ئ ا ب پ ت ج چ ح خ د ر ز ژ س ش ع غ ف ڤ ق ک گ ل م ن ه و ۆ ی ێ ە ھ ۵ ڵ ۊ ۋ ۑ ۉ ۇ ۈ ۅ ۍ ې ے	ئ ا ب پ ت ج چ ح خ د ر ز ژ س ش ع غ ف ڤ ق ک گ ل م ن ه و ۆ ی ێ ە ھ ۵ ڵ ۊ ۋ ۑ ۉ ۇ ۈ ۅ ۍ ې ے	ئ ا ب پ ت ج چ ح خ د ر ز ژ س ش ع غ ف ڤ ق ک گ ل م ن ه و ۆ ی ێ ە ھ ۵ ڵ ۊ ۋ ۑ ۉ ۇ ۈ ۅ ۍ ې ے
Persian	ئ ا ب پ ت ج چ ح خ د ر ز ژ س ش ع غ ف ڤ ق ک گ ل م ن ه و ۆ ی ێ ە	ئ ا ب پ ت ج چ ح خ د ر ز ژ س ش ع غ ف ڤ ق ک گ ل م ن ه و ۆ ی ێ ە	ئ ا ب پ ت ج چ ح خ د ر ز ژ س ش ع غ ف ڤ ق ک گ ل م ن ه و ۆ ی ێ ە			ئ ا ب پ ت ج چ ح خ د ر ز ژ س ش ع غ ف ڤ ق ک گ ل م ن ه و ۆ ی ێ ە	ئ ا ب پ ت ج چ ح خ د ر ز ژ س ش ع غ ف ڤ ق ک گ ل م ن ه و ۆ ی ێ ە	ئ ا ب پ ت ج چ ح خ د ر ز ژ س ش ع غ ف ڤ ق ک گ ل م ن ه و ۆ ی ێ ە	ئ ا ب پ ت ج چ ح خ د ر ز ژ س ش ع غ ف ڤ ق ک گ ل م ن ه و ۆ ی ێ ە	ئ ا ب پ ت ج چ ح خ د ر ز ژ س ش ع غ ف ڤ ق ک گ ل م ن ه و ۆ ی ێ ە	ئ ا ب پ ت ج چ ح خ د ر ز ژ س ش ع غ ف ڤ ق ک گ ل م ن ه و ۆ ی ێ ە	ئ ا ب پ ت ج چ ح خ د ر ز ژ س ش ع غ ف ڤ ق ک گ ل م ن ه و ۆ ی ێ ە	ئ ا ب پ ت ج چ ح خ د ر ز ژ س ش ع غ ف ڤ ق ک گ ل م ن ه و ۆ ی ێ ە	ئ ا ب پ ت ج چ ح خ د ر ز ژ س ش ع غ ف ڤ ق ک گ ل م ن ه و ۆ ی ێ ە	ئ ا ب پ ت ج چ ح خ د ر ز ژ س ش ع غ ف ڤ ق ک گ ل م ن ه و ۆ ی ێ ە	ئ ا ب پ ت ج چ ح خ د ر ز ژ س ش ع غ ف ڤ ق ک گ ل م ن ه و ۆ ی ێ ە
Urdu	ئ ا ب پ ت ج چ ح خ د ر ز ژ س ش ع غ ف ڤ ق ک گ ل م ن ه و ۆ ی ێ ە	ئ ا ب پ ت ج چ ح خ د ر ز ژ س ش ع غ ف ڤ ق ک گ ل م ن ه و ۆ ی ێ ە	ئ ا ب پ ت ج چ ح خ د ر ز ژ س ش ع غ ف ڤ ق ک گ ل م ن ه و ۆ ی ێ ە			ئ ا ب پ ت ج چ ح خ د ر ز ژ س ش ع غ ف ڤ ق ک گ ل م ن ه و ۆ ی ێ ە	ئ ا ب پ ت ج چ ح خ د ر ز ژ س ش ع غ ف ڤ ق ک گ ل م ن ه و ۆ ی ێ ە	ئ ا ب پ ت ج چ ح خ د ر ز ژ س ش ع غ ف ڤ ق ک گ ل م ن ه و ۆ ی ێ ە	ئ ا ب پ ت ج چ ح خ د ر ز ژ س ش ع غ ف ڤ ق ک گ ل م ن ه و ۆ ی ێ ە	ئ ا ب پ ت ج چ ح خ د ر ز ژ س ش ع غ ف ڤ ق ک گ ل م ن ه و ۆ ی ێ ە	ئ ا ب پ ت ج چ ح خ د ر ز ژ س ش ع غ ف ڤ ق ک گ ل م ن ه و ۆ ی ێ ە	ئ ا ب پ ت ج چ ح خ د ر ز ژ س ش ع غ ف ڤ ق ک گ ل م ن ه و ۆ ی ێ ە	ئ ا ب پ ت ج چ ح خ د ر ز ژ س ش ع غ ف ڤ ق ک گ ل م ن ه و ۆ ی ێ ە	ئ ا ب پ ت ج چ ح خ د ر ز ژ س ش ع غ ف ڤ ق ک گ ل م ن ه و ۆ ی ێ ە
Language-specific	ڵ ۆ ێ	ھ ۵ ڵ ۊ ۋ ۑ ۉ ۇ ۈ ۅ ۍ ې ے	ھ ۵ ڵ ۊ ۋ ۑ ۉ ۇ ۈ ۅ ۍ ې ے	ھ ۵ ڵ ۊ ۋ ۑ ۉ ۇ ۈ ۅ ۍ ې ے		ھ ۵ ڵ ۊ ۋ ۑ ۉ ۇ ۈ ۅ ۍ ې ے	ھ ۵ ڵ ۊ ۋ ۑ ۉ ۇ ۈ ۅ ۍ ې ے	ھ ۵ ڵ ۊ ۋ ۑ ۉ ۇ ۈ ۅ ۍ ې ے	ھ ۵ ڵ ۊ ۋ ۑ ۉ ۇ ۈ ۅ ۍ ې ے	ھ ۵ ڵ ۊ ۋ ۑ ۉ ۇ ۈ ۅ ۍ ې ے	ھ ۵ ڵ ۊ ۋ ۑ ۉ ۇ ۈ ۅ ۍ ې ے	ھ ۵ ڵ ۊ ۋ ۑ ۉ ۇ ۈ ۅ ۍ ې ے	ھ ۵ ڵ ۊ ۋ ۑ ۉ ۇ ۈ ۅ ۍ ې ے	ھ ۵ ڵ ۊ ۋ ۑ ۉ ۇ ۈ ۅ ۍ ې ے	ھ ۵ ڵ ۊ ۋ ۑ ۉ ۇ ۈ ۅ ۍ ې ے	ھ ۵ ڵ ۊ ۋ ۑ ۉ ۇ ۈ ۅ ۍ ې ے
总数/唯一数	35/7	40/12	34/7	35/0	41/4	40/3	38/1	37/0	43/0	50/6	46/3	48/5	49/8	52/13	58/21	44/1	32/3

表 A.1：所选语言所使用的波斯-阿拉伯文字，并对阿拉伯语、波斯语和乌尔都文字进行了比较概述。请注意，语言特有字符是指某种语言独有、且未用于阿拉伯语、波斯语或乌尔都语的字符。这一点也在最后一行中有所展示。

语言	网站
俾路支语	https://sunnionline.us/balochi/
布拉灰语	https://talarbrahui.com
北库尔德语	https://badinan.org
南库尔德语	https://shafaq.com/ku

表 A.2：用于爬取所收集数据的地方新闻网站。

KMR	CKB	SDH	HAC	UIG	ARB	AZB	GLK	MZN	FAS	URD	KAS	PNB	SKR	TRW	PUS	SND	BRH	BAL
␣و␣	␣نی	␣له	␣و␣	␣ئا	␣ال	␣ای	␣ان	␣ون	␣ست	␣یں	␣چھ	给␣	他们␣	␣سی	␣د␣	␣جي	␣نا	␣نت
␣ێن	␣له	␣له	␣به	␣نك	␣أن	␣یر	ستا	␣نه	␣می	␣ہے	␣اَ	␣اں	␣دے	␣سی	␣په	␣جي	␣نا	␣اں
␣ان	انی	␣ئه	␣نه	␣ىك	␣لی	یند	ای␣	␣و␣	است	␣می	␣مَن	␣دے	␣یں	␣می	␣په	␣ان	␣ان	␣ان
␣هه	␣ان	␣وه	␣ان	ىنی	␣من	␣ین	تان	␣تا	␣را	میں	␣چھِ	␣وں	␣وں	␣می	␣او	␣مِ␣	␣اس	انت
␣ئه	␣به	␣ان	␣جه	␣نی	الم	␣ده	␣ته	␣وه	␣را	ہیں	␣ہِ	␣تے	␣تے	␣مھ	␣او	␣ءِ␣	␣ٹی	␣نی
کو␣	␣ه‌ی	␣و␣	␣ه‌ی	ىلی	␣في	␣او	␣␣	␣ه‌ا	␣اس	␣一个	␣ه‌	␣دی	␣دی	␣کی	␣د␣ه	␣جو	␣اٹ	␣که
␣یا	␣ئه	ه‌یل	␣که	␣ان	␣في	␣ان	␣ه‌ا	␣ره	␣به	␣ہے	␣کھ	␣دی	␣دے	␣ھی	␣کې	␣ج␣ن	␣او	␣ان
␣ل␣	ه‌وه	␣یل	␣جه	␣ئی	␣سا	␣ای	ایس	␣ای	␣ند	␣ه‌ا	␣مَ	␣وچ	␣دا	␣م␣ی	␣له	␣آه	␣آ␣	انی
␣نا	␣و␣	کرد	انی	ئ␣ی	␣مي	␣و␣	你可否␣		␣به	␣ہو	اَک	␣نے	␣تے	␣تھ	␣دې	␣هن	␣ا␣ن	␣که
␣دا	ئ␣ی	␣یگ	␣ئا	ىری	␣من	␣دا	␣سه	␣ته	␣می_	␣نے	␣کَھ	␣وچ	␣دی	␣بھ	␣کې	␣جو	␣دا	␣ێں

表 A.3：所选语言的收集的语料库中最常见的 10 个三元组。␣ 和 _ 分别表示空格和零宽非连接符。在这些三元组中，可以看到许多词缀和连词，例如北库尔德语（KMR）中的␣و␣（“和”），以及戈拉尼语（HAC）中的␣که（“那”）。

语言噪声 %		预测 (@1)		句子
		lid.176	我们的	
旁遮普语	0	Urdu	阿塞拜疆语	并将世俗主义和社会主义以民主的美好形式呈现出来。
萨莱基语	0	旁遮普语	萨莱基语	语言与文学研究大多是定性的。
信德语	0	信德语	信德语	许多次，一位女性伴侣，与她有柏拉图式关系
俾路支语	0	Urdu	俾路支语	拉瓦ˆ格ˇ战争坦ˊ阿·加图布·拉赫姆·耶特，哈克·阿哈尔·亚尼亚
阿塞拜疆语	0	波斯语	阿塞拜疆语	北库尔德语与南库尔德语
吉拉基语	0	波斯语	吉拉基语	无法确定（文本无法辨读）
波斯语	0	波斯语	波斯语	·老年是练习它的时光 ·青年是获取智慧的时光
维吾尔语	0	维吾尔语	维吾尔语	无法识别的文本
南库尔德语	0	索拉尼语	南库尔德语	在冠状病毒大流行期间，人们一直在等待司法程序的恢复
克什米尔语	20	Urdu	克什米尔语	。苏达ˊ·拉ˈ妮 ̣是ˆ一ˊ位ˊ印 ̣度电影女演员ˊ，ˈ她 ̣曾ˊ在 ̣电影中出演。
克什米尔语	100	Urdu	克什米尔语	。苏达·拉妮是一位印度电影女演员。
索拉尼语	20	波斯语	索拉尼语	今年毕业生比例高于去年
索拉尼语	100	阿拉伯语	索拉尼语	ريژەي دەرجانی ئەمصال لە صالي پشطر زياطره

表 A.4：所选语言中的一些示例，以及 fastText 的预训练模型（lid.176）的预测结果，并与我们使用 fastText 在所收集数据上训练的模型之一的预测结果进行比较。