文章编号：1003-0077（2017）00-0000-00

利用领域外数据对口语风格短文本的相近语种识别研究

何峻青[1,2] 黄娴[3] 赵学敏[1] 张克亮[3]

（1.中科院声学研究所 语言声学与内容理解实验室，北京市 100190；2.中国科学院大学，北京市 100190；信息工程大学洛阳校区，河南省洛阳市 471003）

摘要：该文以维吾尔语和哈萨克语这一组相近语言为例，在哈语语料受限的情况下，使用领域外语料增补原始语料，经同化后提高了在口语风格短文本上进行语种识别的精确度。该文分析了维、哈两种语言的词形学特点，提出了多条特征，构建了一个最大熵分类器，在测试集上识别维语和哈语口语风格短文本的精确度达到 95.7%，而 CNN 分类器的精确度仅为 69.1%。实验结果证明本系统对其他语种口语风格短文本的语种识别亦具有适用性。

关键词：领域外数据；口语风格短文本；字符的 n 元特征

中图分类号：TP391 **文献标识码：**A

A Study on Discrimination Between Similar Languages on Short Conversational Texts with Out-of-domain Data

Junqing He[1,2], Xian Huang[3], Xuemin Zhao[1] and Keliang Zhang[3]

(1.Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190, China ; 2. University of Chinese Academy of Sciences, Beijing 100190,China ;3. Information Engineering University, Luoyang Division 471003,China)

Abstract: This paper aims at identifying limited-resourced similar languages such as Uyghur and Kazakh on short conversational texts. Since collected Uyghur and Kazakh data are severely imbalanced, we leveraged a compensation strategy and an assimilation method, found appropriate out-of-domain data to build a balanced Uyghur and Kazakh corpus. Then we constructed a maximum entropy classifier based on morphologic features to discriminate between the two languages and investigated the contribution of each feature. Experimental results suggest that the MaxEnt classifier effectively discriminates between Uyghur and Kazakh on the test set with the accuracy being 95.7%. In contrast, the CNN classifier performs much poorer than MaxEnt classifier with accuracy being only 69.1%. Though much less preprocessing is applied, the classifier outperforms the champion of the VarDial'2016 DSL shared task on test sets B1 and B2 by 0.6% and 1.2%.

Key words: out-of-domain data; short conversational texts; morphological features

0 引言

语种识别(language identification，LID) 是自然语言处理的一个重要分支，旨在识别一个文本内容所属的语种。自 Cavnar&Trenkle 首先提出基于 n 元特征的文本分类方法[1]以来，语种识别研究得到了快速发展，在大量训练数据和格式规范的文本上取得了高精确度[2]和高覆盖率[3]的成绩，语种识别也被认为是一项基本已经解决的任务。然而识别基于少量的数据、多语种混合输入、语码转换（在两种或两种以上语言间转换）、相近语言（语言变体、方言）、非常短的文本（如推特的推文）仍然是该领域的瓶颈[4-5]。在本研究中，我们遇到了训练数据严重不平衡、相近语言以及文本非常短这三个问题，在训练数据受限的情况下识别维吾尔语和哈萨克语的口语风格短文本。

维吾尔语（以下简称维语） 和哈萨克语（以下简称哈语）是典型的相近语言，都属于阿尔泰语系突厥语族，都是粘连语，在中东和中国西北部广泛使用。[6]认为维语和哈语在句子层面的相似程度超过 80%，在词层面的相似程度则达到 90%以上。区分这两种语言的困难在于：（1）两种语

基金项目：国家自然科学基金（批准号：11590771, 11590770-4, 11722437, 61650202, U1536117, 61671442, 11674352, 11504406, 61601453）；国家重点研发计划(2016YFB0801203, 2016YFC0800503, 2017YFB1002803)，新疆维吾尔自治区重大科技专项(2016A03007-1)。

言都用阿拉伯字母按照从右至左的顺序书写；（2）共享字母多达 26 个，另外还有两个字母看上去一模一样；（3）词汇和句法有很多重叠之处，仅靠查询字典来区分两种语言难度极大；（4）都包含大量前后缀，导致词干提取和识别困难。

本文定义的“口语风格短文本”包括手机短信、微信等聊天工具的聊天记录以及推特、脸书、微博等社交平台上的发言。对这类文本进行语种识别存在很大难度，原因主要如下：（1）每条文本长度太短，大多数句子的长度仅为 3-9 个词；（2）文本中存在大量的拼写和语法错误，大大增加了词干提取和错误更正的代价；（3）广泛使用了缩略语和俚语表达，普通字典中并未收入这些内容；（4）收集口语风格短文本费时费力，经常存在语料不足的问题；（5）人们为了输入方便，很多情况下未遵守语言使用规范，造成了语料中的字符远超过标准字符总数。本研究收集的维、哈口语风格短文本语料中包含了超过 100 种字符，进一步增加了区分维语和哈语口语风格短文本的难度。

本研究旨在构建出一个相近语种识别系统，即使在训练数据受限的情况下也能够识别口语风格短文本所属的相近语种（语言变体、方言）。文章共分为六部分，第一部分为背景介绍；第二部分简要总结了相关研究；第三部分介绍了维、哈口语风格短文本语料库的构建；第四部分详细介绍了分类特征的设计、相近语种识别系统的构建、以及评测标准的拟定；第五部分通过一系列实验检测了数据增补策略的有效性、各个特征在相近语种识别过程中的贡献、传统机器学习和深度学习分类器的性能比较、以及本系统对其他相近语种（语言变体、方言） 的识别效果；第六部分为结论。

1 相关研究

最早进行相近语种识别的研究有[7]。该文首先提出了识别相近语言的重要性和难度，并提出了利用一个半监督的模型来识别印度尼西亚语和马来语。此后该领域受到越来越多学者的关注，研究范围包括多种南斯拉夫语言[8-9]、汉语变体[10]、葡萄牙语变体[11]、西班牙语变体[12]、英语变体[13]、以及阿拉伯语方言[14]等。2014 年至 2017 年 Marcos Zampieri 等人在 COLING(2014)、RANLP(2015)、COLING(2016)、EACL(2017)下组织了“运用自然语言处理工具识别相近语言、语言变体和方言”工作坊系列(Workshop Series on Applying NLP tools to Similar Languages, Varieties and Dialects, VarDial), 允许参赛者使用相同的数据来比较不同的相近语种识别方法的效果。每一届工作坊的共享任务提供若干组相近语言（语言变体、方言）语料，每种语言（变体或方言）有 18000 个句子作为训练集，2000 个句子作为开发集，此外还有 1000 条句子作为测试集。四年来，VarDial 工作坊提供训练和测试的语种（变体、方言）种类不断增加，共享任务亦越来越多样化。关于这几届 VarDial 工作坊共享任务的语料、参赛系统采用的方法以及评测结果可见[5, 15-18]。综合来看，字符的 n 元特征为最有效的特征，效果最佳的分类模型包括支持向量机(SVM)、逻辑回归(logistic regression)，然而深度学习方法取得的效果并不理想[17:5, 18:11]。

对于短文本的语种识别，[19-22]采用了通过额外语义(additional semantics)来扩充短文本表征(short text representation)的方法，额外语义来自数据采集或者一个更大规模的知识源。[23]介绍了设在 SEPLN2014 下的推特文本语种识别任务的情况。

虽然 n 元模型在已有大量文本数据的情况下能取得非常好的效果，但是当某个领域的数据很少的时候，则面临严重的数据稀疏问题。传统处理数据稀疏问题的方法包括构建与领域不那么相关的模型或者构建使用专门领域技术的模型，但结果并没有明显改进[24:221]。另一个处理方法则为使用大量别的任务或领域的数据，即领域外数据(out-of-domain data)，来改进领域内的语言模型。[24]、[25]分别使用领域外数据来训练语音识别语言模型和统计机器翻译语言模型，均取得了较好的效果。[26]讨论了处理不均衡数据(imbalanced data)的多种方法。

区分维语和哈语的研究有[27]。该研究以特有字符为特征区分维语、哈语和柯尔克孜语。该方法在 70 个词以上的文本中达到了 97.7%的精确度，然而对于少于 10 个词的文本对哈语的识别率降到了 65.31%，原因在于哈萨克语的特殊字符比其他两个语种要少得多。针对短文本，有必要提取特有字符以外更多的有效特征来区分维语和哈语。

在本研究中，我们试图探讨以下四个问题：

(1) 区分相近语种时，如何解决有的语种资源受限的问题？
(2) 本文提出的特征是否有效？各个特征对系统的贡献如何？
(3) 传统机器学习分类器和深度学习分类器对维、哈语这一组相近语言的口语风格短文本的识别性能孰优孰劣？
(4) 本研究构建的相近语种识别系统是否能够有效识别其他相近语种（方言、变体）？

2 维、哈语口语风格短文本语料库构建

2.1 语料收集

随着社交网络的普及和手机等聊天工具的推广，人们越来越多地使用即时信息来交流，对口语风格短文本的自然语言处理具有重要意义。我们从新疆收集了匿名来源手机用户共计 48460 条手机短信作为训练集，将同样来源一天内收集到的 973 条手机短信作为测试集。经过维语和哈语语言专家的辨别和标注，确定训练集中包含了 48432 条维语短文本和 148 条哈语短文本，测试集中包含了 687 条维语短文本和 286 条哈语短文本。训练集中维语和哈语文本数量的比例达到了 327:1，数量严重失衡，在训练相近语种识别系统前有必要平衡两个语种语料的数量。

2.2 哈语口语风格短文本增补和同化

平衡维、哈两种语言的语料规模可以通过删减维语语料或者增加哈语语料的办法来达成。考虑到将维语语料删减到 148 条，数据过少会严重影响训练的效果，我们决定增补哈语语料。我们没有继续收集更多的匿名短信，原因在于此来源语料的获取具有相当难度，而且语言专家需要浏览超过 300 条文本才能筛选到 1 条哈语文本，若以此方法获取四万余条哈语文本将耗费巨大的人力物力。

前面提到[24]、[25]使用领域外数据来训练语音识别语言模型和统计机器翻译语言模型取得了较好的效果，我们决定使用领域外哈语短文本来补充哈语语料。为了获取哈语的口语风格短文本，我们选择爬取哈语论坛[1]上的文本，没有选择爬取哈语新闻网页或者推特推文的原因为：（1）新闻网页上的内容为正式的书面语，文本较长，与口语风格短文本在词和字符层面的重合率较小；（2）虽然推特上的推文完全符合口语风格短文本的特点，但中国人极少使用推特，同时，即便是哈萨克人使用推特发布的推文也可能使用了其他语言；（3）该哈语论坛中的内容经哈语专家鉴定内容基本全部为哈语，内容以对话风格为主，符合我们选取语料的标准。

基于以上理由，我们从该论坛爬取了 70909 个网页。爬取下来网页中的文本有长有短，分属于文学、经济、娱乐等主题。为使爬取的数据最大可能接近训练语料，我们进一步清洗爬取的内容，从中选取不超过 14 个词的短文本，获得了 339,609 条符合要求的哈语文本。在此基础上我们随机选取了 48,000 条文本来匹配维语训练文本的规模。

通过对哈语语料的增补和同化，我们最终构建了一个包含 49119 条维语和 48286 条哈语口语风格短文本的语料库，基本达到了数量平衡、风格一致的要求。语料库分为训练集（维语文本 48432 条、哈语文本 48000 条） 和测试集（维语文本 687 条、哈语文本 286 条）。图 1 为增补哈语语料前后训练集中两个语种口语风格短文本的数量对比情况。图 2 为经增补后的维、哈口语风格短文本语料库的构成。

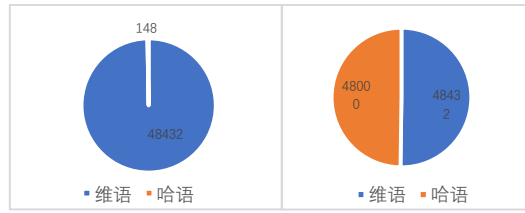


图 1 语料增补前后训练集中维、哈语短文本数量对比

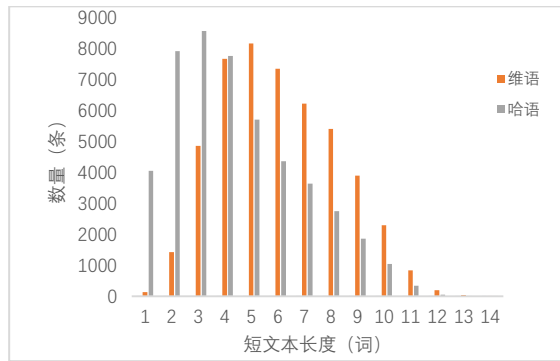


图 2 经增补后的维哈口语风格短文本语料库构成

3 相近语种识别系统构建

3.1 特征提取

由于本研究使用语料长度很短，使用词汇的 n 元特征会造成特征过于稀疏的问题。通过重点分析维语和哈语的词形学（morphology）特点，我们设计了以下特征：

(1) 特有字符。虽然两种语言共享多达 26 个字母，但仍有少量字符不同。一旦在文本中找到了属于某个语种的特有字符，就可判定该文本属于对应的那个语种。

(2) 字符的 n 元特征。虽然维语和哈语有很多共有字符，但是各自字符排序和组合有一定特点，这些特点可以有效帮助区分语种。从已有相关研究可以看出，该特征为也是相近语种识别系统中最常用的特征。

(3) 前缀和后缀。维语和哈语都有许多词缀，但是在许多情况下，两种语言使用的词缀不同。例如，两种语言中表达相同意义的单词往往以不同的字符开头，在维语中以“ya”开头的单词在哈语中通常以“ja”开头，“o”在哈语中可以作为单词首字母而在维语中则不行。维语中的“-lar”和哈语中的“-dar”表达同样意思，但是拼写不同。需要注意的一点是，口语短文本中存在大量的拼写错误，可

[1] http://bbs.senkazakh.com

能会导致该特征难以提取。因此我们将每个词的前 n 个和后 n 个字符作为特征，n 的范围为 1-3。

(4) 词的一元特征。词的一元特征即该单词出现的频率。如果一个文本中包含了某个语种的高频词，那么该文本就更可能属于该高频词对应的语种。

(5) 文本长度的 bin 值。 按照文本的长度将其划分为不同的 bin 值类型。

3.2 分类器

本文认为相近语种识别任务实际上为将相近语种的文本进行分类，因此本研究中的相近语种识别系统即用来区分相近语种的分类器。

目前常见的分类器有传统的机器学习分类器和新兴的神经网络分类器。最大熵（the maximum entropy, MaxEnt）是机器学习算法中的最佳模型之一。最大熵分类器以最大熵原理作为模型学习的准则，即在所有满足约束条件集合的模型中，选取熵最大的模型。对输入的特征$x \in X$，使用特征函数$f(x,y)$模拟特征和对应类别标签$y \in Y$的关系，并求解模型 P(Y|X)。在满足特征函数关于经验分布$\tilde{p}(X,Y)$的期望$E_{\tilde{p}}$与关于模型P(Y|X)的期望E_p相等的约束下，选取条件熵H(P)最大的模型。具体公式如下：

$$C \equiv \{p \in P | E_p(f) = E_{\tilde{p}}(f)\} \quad (1)$$

$$E_{\tilde{p}}(f) = \sum_{x,y} \tilde{P}(x,y) f(x,y) \quad (2)$$

$$E_p(f) = \sum_{x,y} \tilde{P}(x) P(y|x) f(x,y) \quad (3)$$

$$\tilde{P}(x,y) = \frac{\#(x,y)}{N}, \tilde{P}(x) = \frac{\#(x)}{N} \quad (4)$$

$$H(P) = -\sum_{x,y} \tilde{P}(x) P(y|x) log P(y|x) \quad (5)$$

上述公式中 N 为总样本数，#为频数。$f(x,y)$通常为二值函数，当(x,y)同时出现时为 1，否则为 0。预测的时候，分类器计算每个样本的分值，选择分值最高的类别作为标签。由于最大熵分类器将特征依赖度考虑在内，该过程更近似于人类决策的过程，我们使用了斯坦福分类器工具包[2]构建了一个基于最大熵的分类器作为相近语种识别系统。

随着卷积神经网络（convolutional neural networks, CNN）成功地应用于图像识别[28]和文本分类[29]任务，CNN 成为目前最流行的深度学习分类器。我们基于字符矢量(character embeddings)构建了一个 CNN 分类器来测试该分类器识别维、哈语口语风格短文本的表现情况。

对于输入的每个句子，将每个字符表示为固定长度的字符矢量，则该句子表示为一个矩阵S。对每个句子矩阵 S 过一个卷积神经网络的过程如下：对于高度为 z 的一个卷积核 Wm，用它以 1 为滑动步长，在整个矩阵中从上至下滑动，每一步计算重合部分的两个矩阵的点积及经过激活的值x_i，最后得到一个长度为 N-z+1 的向量 X，N 为句子所包含的字符的数目。然后使用最大池化，取其中最大值得到一个元素c_m。使用多个不同高度的卷积核进行卷积，卷积核的宽度都为词向量长度，将结果拼接得到一个特征向量 s。之后将特征向量 s 经过一个全连接层，再使用 Softmax 归一化，预测该文本分别属于维语和哈语的概率。公式如下：

$$x_i = ReLU(W_m \cdot S_{i:i+z-1}) + b_m \quad (6)$$

$$X = [x_1, x_2, \dots, x_{N-z+1}] \quad (7)$$

$$c_m = \max(X) \quad (8)$$

$$s = [c_1, c_2, \dots, c_k] \quad (9)$$

$$y' = \text{Softmax}(Us + b) \quad (10)$$

上述公式中，·为点积操作，[...]表示元素拼接，ReLU 表示规整线性单元(Rectified Linear Unit)，k 为卷积核的总数，m 为第 m 个卷积，b_m为对应卷积核的偏置。U 为全连接层的参数矩阵，b 为偏置，均为可训练参数。

3.3 评价标准

本文采用准确率（Precision，P）、召回率（Recall，R）和精确度（Accuracy，A）来评价系统的性能，公式如下所示。准确率衡量系统正确判断样本类别的能力，召回率描述系统检索正确样本的能力，精确度表示所有类别的正确样本能被正确分类的比例，是整体的评价指标。公式中 TP 指正确预测为某种语言的样本数，FP 指预测为该语言实际上不是该语言的样本数，TN 指正确预测为不是该语言的样本数，FN 指预测不是该语言但实际是该语言的样本数。

$$Precision = \frac{TP}{TP+FP} \quad (1)$$

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (3)$$

4 实验和结果

在完成维、哈语口语风格短文本语料库和相近语种识别系统后，我们进行了四组实验，来寻求 1 中所提出问题的答案。

4.1 增补的哈语语料的可用性实验

针对问题（1），我们使用最大熵分类器来测试哈语语料增补前后相近语种识别的效果。使用原有语料训练的分类器对测试语料的识别结果和使用经增补后的语料的识别结果如表 1 所示：

训	维语	哈萨克语	A(%)

[2] https://nlp.stanford.edu/software/classifier.html

练集	P(%)	R(%)	P(%)	R(%)	
原始语料	89.0	99.3	97.6	70.6	90.9
最终语料	98.5	95.1	89.0	96.5	95.5

表 1 增补后哈萨克语料的可用性实验结果

从实验结果来看，利用未经增补的训练集训练分类器后，维语的召回率高达 99.3%，哈语的召回率则仅有 70.6%。哈语训练语料经过增补后，两种语言的召回率接近了（分别为 95.1%和 96.5%），精确度从 90.9%上升到了 95.5%，证明增补策略有效，同时显示了需要进行相近语种识别的语种训练语料规模均衡的重要性。

4.2 选取特征的重要性实验

针对问题（2），为了考察 3.1 中提出的每个特征的重要性，我们分别测试了（1）所有特征、（2）所有特征减去特殊字符、（3）所有特征减去字符的 n 元特征（n=1,2,3,4）、（4）所有特征减去前后缀、（5）所有特征减去词的一元特征、以及（6）所有特征减去 bin 值的分类结果。本实验使用了最大熵分类器在增补后的维哈训练语料上训练，实验结果如表 2 所示：

特征	维吾尔语		哈萨克语		
	P(%)	R(%)	P(%)	R(%)	A(%)
所有特征	98.5	95.2	89.3	96.5	95.6
-特殊字符	98.3	98.5	87.9	96.2	95.0
-字符 n 元特征	97.8	90.5	80.7	95.1	91.9
-前后缀	98.5	94.6	88.2	96.5	95.2
-词的一元特征	**98.5**	**95.3**	**89.6**	**96.5**	**95.7**
-bin 值	98.5	95.0	89.0	96.5	95.5

表 2 特征的重要性实验结果

从表 2 可以看出，在减去每个特征（词的一元特征除外）后系统的性能都有不同程度的下降，尤其是移除了字符的 n 元特征后，系统精确度下降最多，说明这些特征都对本任务起到作用，字符的 n 元特征对本任务贡献最大。相反，当系统移除了词的一元特征后，精确度还有少许提升，意味着词层面的特征非但没有起到帮助作用，反而降低了对维、哈语文本的语种识别效果。在后续的实验中，我们默认选取除去词的一元特征以外的所有特征。

4.3 传统机器学习分类器与深度学习分类器表现比较实验

针对问题（3），我们分别构建了最大熵分类器和 CNN 分类器来识别维语和哈语口语风格短文本。最大熵分类器使用了词的一元特征以外的所有特征。CNN 分类器使用了 50 维的字符矢量(character embedding)，并进行均匀分布的随机初始化，取值范围为(-0.5, 0.5)，卷积核的宽度分别设为[1,2,3,4]，数目分别为[50, 200,300,500]。卷积层后用了一个随机丢弃(dropout)层和最大池化(max-pooling)层，丢弃概率(dropout rate)设为 0.5。表 3 列出了两个分类器的表现：

分类器	维吾尔语		哈萨克语		
	P(%)	R(%)	P(%)	R(%)	A(%)
最大熵分类器	98.5	95.3	89.6	96.5	95.7
CNN 分类器	30.7	10.1	71.4	93.6	69.1

表 3 分类器有效性实验结果

从表 3 可以看出，在识别维语和哈语的口语风格短文本这一任务中，最大熵分类器精确度明显高于 CNN 分类器。在 VarDial'2016 DSL 共享任务中，参赛队伍 mitsls、Uppsala 分别使用了基于字符层面的 CNN 和词层面的 CNN，结果精确度和 F1 值均低于大多数同时参赛的传统机器学习分类器（如基于 SVM、逻辑回归的分类器） 的识别效果[18: 7-8]。

神经网络分类器在识别多语种文本时取得的高精确度[2]与在处理相近语种时的低精确度形成了鲜明对比，原因值得探求。通过错误分析，我们认为 CNN 分类器结果难以令人满意的原因有两点：1）CNN 分类器用太多的卷积核作为参数，对训练语料的规模要求高，4 万条左右的训练文本难以使 CNN 分类器学到足够的特征规律；2）CNN 分类器的鲁棒性较差，在处理维语文本时，由于入库的维语语料中包含有大量拼写错误，含有拼写错误的字符被当作集外词（Out of Set Vocabulary），无对应的字符向量，导致卷积核无法识别出特征字符序列，所以 CNN 分类器对维语的识别效果较差。相比而言，哈语语料主要来自网络论坛，误拼错误要少得多，所以 CNN 分类器识别哈语的精确率和召回率比维语要高得多。

4.4 本系统对其他相近语种（语言变体、方言）的适用性测试

针对问题（4），即测试本系统识别其他相近语种（语言变体、方言）的口语风格短文本的性能，

我们使用了最大熵分类器来识别 VarDial'2016DSL 共享任务子任务一下的两个领域外口语风格短文本测试集 B1、B2。

VarDial'2016 DSL 共享任务子任务一提供了 12 种语言（语言变体）的新闻短文本作为训练语料，每种语料提供 18000 个句子作为训练集，2000 个句子作为开发集。测试集分为一个领域内测试集(A)，两个领域外测试集(B1, B2)。B1（波斯尼亚语、克罗地亚语和塞尔维亚语）和 B2（巴西葡萄牙语和欧洲葡萄牙语）两个测试集各包含 100 个推特用户的推文，平均每个用户 98.88 和 50.47 条推文。选取识别 B1、B2 两个测试集来测试本系统性能的原因在于，这两个测试集中的文本同样属于口语风格短文本，可以较好地考察本系统识别其他相近语种（语言变体、方言）口语风格短文本的适用性。

我们对测试语料做了简单的预处理，清除了其中的链接、@符号以及标签。然后使用最大熵分类器，选取了除词的一元特征以外的所有特征，字符的 n 元特征中的 n 设置为 1 到 7，对 B1 和 B2 进行了语种分类。为了与当时的参赛系统进行比较，本次实验使用了 VarDial'2016DSL 共享任务中的评价指标：精确度(A)和 F1 值。F1 值的计算公式如下：

$$F1 = \frac{2PR}{P+R} \quad (14)$$

其中，P 为准确率，R 为召回率。

本系统和当时参加 VarDial'2016DSL 共享任务子任务 1 前五名对 B1、B2 的分类结果如表 4 和表 5 所示：

参赛系统	A	F1	方法及主要特征
本系统	0.926	0.926	最大熵，字符 n 元特征（n=1-7）等复合特征
GW-LT3	0.920	0.919	逻辑回归,字符 n 元特征/单词 n 元特征
nrc	0.914	0.913	两阶段支持向量机，字符的 n 元特征（n=1-6）
UniBueNLP	0.898	0.897	逻辑回归，单词的 n 元特征（n=1-2）
UPV_UA	0.888	0.886	字符串核（string kernel）及核区分分析（ kernel discriminant analysis）
tubasfs	0.862	0.860	支持向量机，字符的 n 元特征（n=1-7）

表 4: 对 B1 测试集相近语种识别排名前五的系统和本系统的表现

参赛系统	A	F1	方法及主要特征
本系统	0.926	0.926	最大熵，字符 n 元特征（n=1-7）等特征
GW-LT3	0.920	0.919	逻辑回归,字符 n 元特征/单词 n 元特征
nrc	0.914	0.913	两阶段支持向量机,字符 n 元特征（n=1-6）
UniBueNLP	0.898	0.897	逻辑回归,单词 n 元特征（n=1-2）
UPV_UA	0.888	0.886	字符串核（string kernel）及核区分分析（ kernel discriminant analysis）
tubasfs	0.862	0.860	支持向量机，字符 n 元特征（n=1-7）

表 5: 对 B2 测试集相近语种识别排名前五的系统和本系统的表现

从表 4 和表 5 可以看出，GW_LT3 在当时的评测中排名第一，该系统使用了字符的 n 元特征（n=2-6）和单词的 n 元特征（n=1-3），用词频对那些特征进行加权，并作了复杂的预处理。对比之下，本系统作的预处理少得多，对 B1、B2 进行相近语种识别的精确度分别比该系统高 0.6%和 1.2%。由此，本系统不仅能够在口语风格短文本上有效区分维语和哈语，对于其他语种的口语风格短文本也能作很好的区分。

对比 nrc 和 tubasfs 系统，这两个系统都使用了支持向量机分类器，特征也都使用了字符的 n 元特征，n 分别为 1-6 和 1-7, 本系统处理 B1、B2 的精确度均优于这两个系统，显示出在该任务中使用复合特征的最大熵分类器分类效果要优于使用字符的 n 元特征的支持向量机分类器。

此外，本系统在处理维、哈语料时，字符的 n 元特征（n=1-4）就取得了 95.7%的精确度，而在处理本任务中的 B1 和 B2 测试集时，n 用到了 1 到 7，精确度才分别达到 92.6%和 89.0%。其中一个原因在于 VarDial'2016DSL 共享任务子任务一提供的训练语料效果不如我们自建的维、哈语口语风格短文本训练语料。因此，识别口语风格短文本所属语种时，网络论坛的语料比新闻语料更适合作训练语料。

5 结论

本研究构建了一个维语和哈语口语风格短文本语料库，在此基础上训练出了一个最大熵分类器，对维语、哈语的口语风格短文本进行语种识别。为了解决语料严重不平衡的问题，我们使用了语料增补和同化的方法，从在线论坛爬取长度相近的、领域外口语风格文本来增补训练语料。实验结果证明增补和同化方法有效，并且在区分口语风格短文本时，论坛上爬取的文本比新闻文本更适合作训练语料。

本文设计了一个最大熵分类器对口语风格短文本进行相近语言语种识别。从实验结果看，字符层面的形态特征有效而单个词的频率反而导致更糟糕的结果。此外，本系统不仅能够有效区分维、哈语口语风格短文本，针对 VarDial'2016DSL 共享任务子任务 1 中三种南斯拉夫语言和葡萄牙语的两个变体的口语风格短文本的语种识别也取得了非常好的效果。

而对于区分维语和哈语这一组相近语言来说，CNN 分类器并未取得理想的效果，这与文献[18, 17]提出的现象一致。我们就此做了一定的错误分析，在未来的工作中我们会继续探求 CNN 分类器在处理相近语言（语言变体、方言）效果不尽人意的原因，并尝试提出改进方法。

参考文献

[1] Cavnar W B, Trenkle J M. N-Gram-Based Text Categorization[C]// Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval. Las Vegas, 1994:161-175

[2] Simões A, Almeida J J, Byers S D. Language identification: a neural network approach[C]//Proceedings of the 3rd Symposium on Languages, Applications and Technologies, SLATE'14. Dagstuhl, 2014:252-265.

[3] Brown R. Non-linear Mapping for Improved Identification of 1300+ Languages[C]//Proceedings of Conference on Empirical Methods in Natural Language Processing. 2014: 627–632.

[4] Ljubešić N, Kranjcic D. Discriminating Between Very Similar Languages Among Twitter Users[C]//Proceedings of 9th Language Technology Conference Information Society-IS 2014. 2014:90-94.

[5] Zampieri M, Tan L, Ljubešić N, Tiedemann J, Nakov P. Overview of the DSL Shared Task 2015[C]//Proceedings of LT4VarDial Workshop. 2015.

[6] 王玲, 达瓦·伊德木草, 吾守尔·斯拉木. 维哈柯及蒙语多文种语言相似性考查研究[J]. 中文信息学报, 2013, 27(6):180-187.

[7] Ranaivo-Malancon B. Automatic identification of close languages–case study: Malay and Indonesian[C]//Proceedings of ECTI Transactions on Computer and Information Technology. 2006: 126-134.

[8] Ljubešić N, Mikelic N, Boras D. Language identification: How to distinguish similar languages?[C]//Proceedings of the 29th International Conference on Information Technology Interfaces. 2007.

[9] Tiedemann J, Ljubešić N. Efficient discrimination between closely related languages[C]//Proceedings of COLING 2012. 2012:2619-2634.

[10] Huang C R, Lee, L H. Contrastive approach towards text source classification based on top-bag-of-word similarity[C]//Proceedings of PACLIC 2008. 2008: 404-410.

[11] Zampieri M, Gebre B G. Automatic identification of language varieties: The case of Portuguese[C]// Proceedings of KONVENS2012. Vienna, 2012:233–237.

[12] Zampieri M, Gebre B G, and Diwersy S. N-gram language models and POS distribution for the identification of Spanish varieties[C]//Proceedings of TALN2013. Sable d'Olonne, 2013: 580–587.

[13] Lui M, Cook P. Classifying English documents by national dialect[C]//Proceedings of Australasian Language Technology Workshop. 2013:5–15.

[14] Zaidan O F, Callison-Burch C. Arabic dialect identification[J]. Computational Linguistics. 2013.

[15] Tan L, Zampieri M, Ljubešić N, and Tiedemann J. Merging comparable data sources for the discrimination of similar languages: The DSL corpus collection[C]// Proceedings of The Workshop on Building and Using Comparable Corpora (BUCC). Reykjavik, 2014.

[16] Zampieri M, Tan L, Ljubešic N, and Tiedemann J. A Report on the DSL Shared Task 2014[C]//Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects. Dublin,2014: 58–67.

[17] Zampieri M, Malmasi S, Ljubešić N, et al. Findings of the VarDial Evaluation Campaign 2017[C]//Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects. Valencia,2017:1–15.

[18] Malmasi S, Zampieri M, Ljubešić N, et al. Discriminating Between Similar Languages and Arabic Dialect Identification: A Report on the Third DSL Shared Task[C]// Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects. Osaka,2016: 1–14.

[19] Phan X H, Nguyen L M, Horiguchi S. Learning to Classify Short and Sparse Text & Web with Hidden Topics from Large-scale Data Collections[C]//Proceedings of WWW 2008. Beijing, 2008:91–100.

[20] Řehůřek R, Kolkus M. Language identification on the web:Extending the dictionary method[C]//Proceedings of International Conference on Intelligent Text Processing and Computational Linguistics. Heidelberg, 2009 357–368.

[21] Tromp E, Pechenizkiy M. Graph-based n-gram language identification on short texts[C]//Proceedings of the 20th Machine Learning conference of Belgium and The Netherlands. 2011:27–34.

[22] Dai Z, Sun A, Liu X Y. Crest: cluster-based representation enrichment for short text classification[C]//Proceedings of PAKDD 2013.2013:256–267.

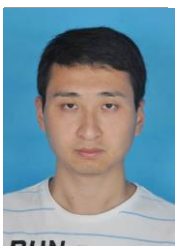
[23] Zubiaga A, Vicente I S, Gamallo P, et al. Overview of TweetLID: Tweet Language Identification at SEPLN 2014[C]//Proceedings of the Tweet Language Identification Workshop, TweetLID2014. Girona, 2014: 1–11.

[24] Iyer R, Ostendorf M, Gish H. Using out-of-domain data

to improve in-domain language models[J]. IEEE Signal Processing Letters, 1997,*4*(8):221-223.
[25] Haddow B, Koehn P. Analysing the effect of out-of-domain data on SMT systems[C]//Proceedings of The Workshop on Statistical Machine Translation. Association for Computational Linguistics, 2012:422-432.
[26] He H, Garcia E A. Learning from imbalanced data[J]. IEEE Transactions on Knowledge & Data Engineering. 2009,21(9):1263-1284.
[27] 买买提依明·哈斯木, 吾守尔·斯拉木, 维尼拉·木沙江等. 基于统计专用字符的维、哈、柯文文种识别研究[J]. 中文信息学报, 2015, 29(2):111-117.
[28] Krizhevsky A, Sutskever I, Hinton, G E. ImageNet classification with deep convolutional neural networks[C]//Proceedings of International Conference on Neural Information Processing Systems. Curran Associates Inc,Vol.25,2012:1097-1105.
[29] Kim Y. Convolutional Neural Networks for Sentence Classification[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014:1746–1751.

何峻青（1994—），在读博士，主要研究领域为自然语言处理。
E-mail：hejunqing@hccl.ioa.ac.cn

黄 娴（1984—），硕士，在读博士，讲师，主要研究领域为自然语言处理。
E-mail：a77huang852yi@sina.com



赵学敏（1984—），博士，副研究员，主要研究领域为自然语言处理。
E-mail：zhaoxuemin@hccl.ioa.ac.cn