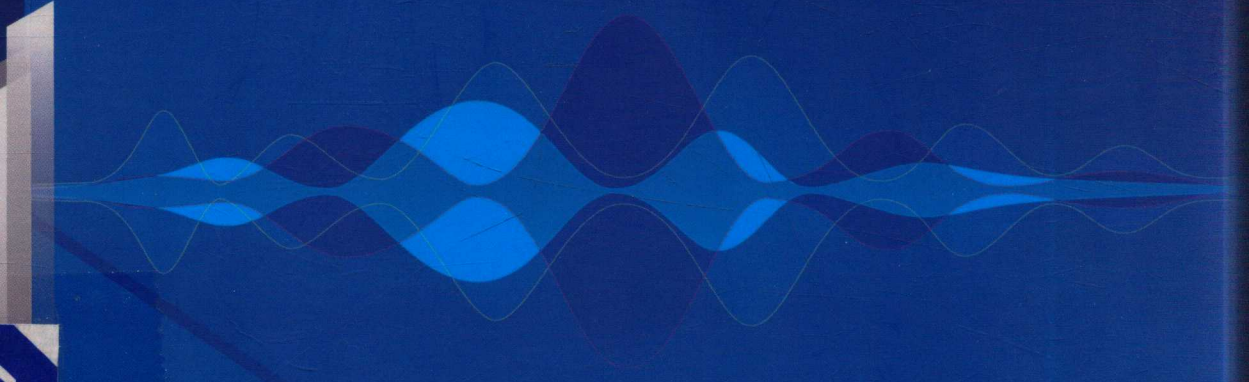
语言变体语料库构建及计算模型

徐 凡 著



科学出版社

语言变体语料库构建及计算模型

徐 凡 著

科学出版社

北 京

内 容 简 介

近年来，语言变体研究受到了计算语言学界的广泛关注。本书从人工智能技术和语言变体相结合的视角，结合自然语言处理、机器学习、神经网络、语音识别、语料库语言学等相关技术，以作者的一系列研究成果为内容主线，系统介绍语言变体(相似语言)研究的语料库构建及计算模型。全书共9章，探索了相似语言及变体语料库的构建规范、多模态(语音和文本)语料库的标注、语言变体文字和语音处理的计算模型。本书对相似语言及语言变体中的关键技术进行较为深入的研究，提出相关问题的一些解决方法，并设计相应的算法和实验。实验表明，本书提出的这些方法有助于提高相似语言的分析性能，同时减少对大规模语料库的依赖性，为今后的相似语言变体分析研究奠定了重要基础，为同类研究提供了参考。

本书可作为从事自然语言处理、计算语言学、数据挖掘研究的科研、管理等相关人员的参考书，也可供高等院校语言学、智能科学与技术、管理科学与工程等教育类、信息类和管理类相关研究生及本科生使用。

图书在版编目（CIP）数据

语言变体语料库构建及计算模型 / 徐凡著. —北京：科学出版社，2019.6
ISBN 978-7-03-059955-1

Ⅰ. ①语… Ⅱ. ①徐… Ⅲ. ①语料库-翻译学-研究 Ⅳ. ①H059

中国版本图书馆 CIP 数据核字（2019）第 279319 号

责任编辑：阚 瑞 / 责任校对：赵桂芬
责任印制：吴兆东 / 封面设计：迷底书装

科学出版社 出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com

北京虎彩文化传播有限公司 印刷

科学出版社发行 各地新华书店经销

*

2019 年 6 月第 一 版 开本：720×1000 1/16
2019 年 6 月第一次印刷 印张：10 1/2
字数：200 000

定价：68.00 元

(如有印装质量问题，我社负责调换)

前　言

随着普通话的逐步推进和日益普及，国内许多省市地区的家乡话(语言变体、方言或相似语言)有逐步被普通话所同化的趋势。但是，方言作为中华民族的一种优秀非物质文化遗产不应该随之而消失。它不仅是一种语言现象，更是一种社会文化现象。同时，方言还具有珍贵的史学及语言学研究价值。当一种语言消失后，与之对应的整个文明也会消失。如今，弱势方言正面临着强势语言、全球化、互联网等的冲击，正处于逐渐消失之中。因此，方言的保护和研究具有重要的现实意义。

当前，人工智能和大数据技术正以前所未有的速度渗透至各行各业，影响到人们生活的方方面面。正如微软亚洲研究院副院长周明提到，“语言智能是人工智能皇冠上的明珠”，充分说明了语言处理的重要性。随着自然语言处理(natural language processing，NLP)技术与语料库语言学研究方面的逐步深入，同时伴随高性能计算机的出现，大规模语言变体研究成为可能。

本书作者从 2010 年开展相关的研究，主要开展过隐式篇章关系识别、篇章论元识别、篇章连贯性建模、文本相似度建模、词对齐抽取、赣方言智能处理等方面的研究，同时在英汉篇章语料库语言学方面也有一定的积累。本书研究得到国家自然科学基金面上项目《赣方言篇章平行语料库构建及计算模型研究》(项目编号：61772246)、国家自然科学基金青年科学基金项目《汉语篇章连贯性分析计算模型研究》(项目编号：61402208)、江西省社会科学研究规划项目《江西方言平行语料库构建及计算模型》(项目编号：17YY05)、江西省高校人文社会科学研究项目《赣方言篇章平行语料库构建及共时与历时研究》(项目编号：YY17211)资助，取得了一系列阶段性研究成果。

感谢参与以上这些项目的课题组成员所作出的贡献，他们分别是博士研究生颜为之，硕士研究生徐雄飞、杜舒静、罗健、孙玉梅和汪苏琪。由于作者能力有限，书中难免存在疏漏和不足，殷切希望读者批评指正。

徐　凡

2018 年 7 月

目　录

第 1 章　相似语言及变体研究概述

本章首先介绍相似语言及变体的相关概念，然后从计算语言学和自然语言处理视角，针对国内外该领域的研究现状进行综述，具体包含语料库资源建设、计算模型和评测指标三个方面的内容。

1.1　相似语言及变体相关概念

英国语言学家 Hudson 把语言变体定义为“社会分布相似的一套语言项目”[1]，意指由具备相同社会特征的人在相同的社会环境中所普遍使用的某种语言表现形式。这套语言项目可以是整个语音、语法和词汇系统，也可以仅仅是某个特定的词语、特定的语法成分或规则[2]。

语言变体是一个内涵很宽泛的概念，大至一种语言的各种方言，小至一种方言中某一项语音、词汇或句法特征，只要有一定社会分布的范围，就是一种语言变体。语言变体受到复杂的社会因素制约，社会语言学对语言变体的研究一般认为，讲话人的社会阶层(class)和讲话风格(style)是语言变体的重要基础，而讲话人的性别对语言变体也产生重要影响。根据使用者来划分的变体称为方言，根据语言使用来划分的变体称为语体或语域。语言变体是社会语言学研究的重要课题。为清晰起见，下面进一步阐述与之相关的几个概念。

(1) 地域方言：简称“方言”，是指同一语言在不同地域的分支。地域方言之间最显著的区别一般在语音方面，但语法和词汇也有不少区别。地域方言的差别是很大的，许多地域都存在着无法相互通话的地域方言。事实上，地域方言与独立语言之间是没有明确界限的，有许多地域变体(如汉语的地域变体)，究竟是方言还是独立语言都存在争议。

(2) 社会方言：是指在同一地区居住的因年龄、性别、职业、宗教、受教育程度等社会因素的不同而产生的小社团语言差异。社会方言最常见的显著差异体现在词汇方面，有时也有发音和音位的差异。一般来说，社会方言的差别较小，可以自由通话。例如，行话、术语、黑话等，它是按照使用的人群来划分的。

(3) 标准语：出于教育、社会事务等目的而设立的标准化的方言。例如，普通话，它是以北京语音为标准音，以北方话(官话)为基础方言，以典范的现代白话文著作为语法规范的现代标准汉语。

需要说明的是，以上这些概念主要是从社会语言学角度出发，并构成其主要研究内容。与之不同，本书主要从计算语言学和自然语言处理视角分析语言变体，旨在通过计算机等信息化手段构建大规模的语言变体语料库，并提出相应的计算模型，开发对应的计算机处理平台，从定性层面拓展至定量层面。

相似语言及变体分析(language varieties and dialects analysis, LVDA)①旨在研究语言的演化现象，其已经成为自然语言处理(NLP)[3]和中文信息处理(Chinese information processing，CIP)[4]的一个重要任务，尤其在语音识别和社会媒体相关信息分析领域具有重要作用。相似语言及变体分析成为自然语言处理相关系统的首个关键步骤，因为人们在处理特定语言的文本和语音前需要知道此文本和语音所属的语言类型。

由于相似语言及变体分析的应用范围非常广泛，受到了国际计算语言学界和产业界的高度重视，因此其相关语料库的构建及计算模型研究逐步受到广泛关注。各大高校和科研院所都从不同角度从事相关方面的评测及研究工作。近年来，相似语言及变体研讨组在国际计算语言学界的顶级和重要会议上(COLING、EACL、RANLP)举办了多种形式的相似语言国际测评会议，如 LT4VarDial(Language Technology for Closely Related Languages, Varieties and Dialects)国际评测。

到目前为止，LT4VarDial 已经成功举办了五次，分别是：VarDial @ COLING 2014、LT4VarDial@RANLP 2015、VarDial@COLING 2016、VarDial@EACL 2017 和 VarDial@COLING 2018，并且收录了很多高质量的研究论文。而且近年来关于相似语言及变体分析的研究仍有很多高质量的研究成果出现。鉴于此，综述这方面的工作有重要意义。接下来，本章的其他部分对相似语言及变体的综合研究成果进行整体上的介绍，以便使读者对该领域有更清晰和完整的认识。

1.2　相似语言及变体研究综述

本节分别从语料库资源建设、传统及当前基于深度学习的计算模型、评测指标等方面阐述相似语言及变体的国内外研究现状。

1.2.1　语料库资源建设

本节主要介绍国内外相似语言及变体语料库资源建设情况，同时对现有语料的优缺点进行分析。

① 本书不具体区分语言、相似语言、语言变体和方言几个概念。

1. 国外语言变体语料库

在国际上，国外变体语言的研究刚刚起步，主要起源于 VarDial @ COLING 2014 国际评测，其提供了 6 组 13 种相似语言的文本语料库，分别是：波斯尼亚语、克罗地亚语、塞尔维亚语；印度尼西亚语、马来西亚语；捷克语、斯洛伐克语；巴西葡萄牙语、欧洲葡萄牙语；西班牙半岛语、阿根廷西班牙语；美式英语、英式英语。随后，LT4VarDial@RANLP 2015 提供了两种数据集：DSL(discriminating between similar languages)2.0 和 DSL 2.1。其中，DSL 2.0 包括保加利亚语、马其顿语、塞尔维亚语、克罗地亚语、波斯尼亚语、捷克语、斯洛伐克语、西班牙半岛语、阿根廷西班牙语、巴西葡萄牙语、欧洲葡萄牙语、马来西亚语、印度尼西亚语等；DSL 2.1 包括 DSL 2.0 的语言以及墨西哥西班牙语和中国澳门葡萄牙语。

VarDial@COLING 2016 和 VarDial@EACL 2017 提供了六组语言识别，分别是：波斯尼亚语、克罗地亚语、塞尔维亚语；马来西亚语和印度尼西亚语；葡萄牙语：巴西和葡萄牙；西班牙语：阿根廷、墨西哥、西班牙；法语：法国、加拿大；阿拉伯方言，同时还设置了新闻语料和社会媒体语料的方言识别两种任务。

2. 国内语言变体语料库

一般而言，汉语变体语言(方言)通常包括七大方言[5,6]，分别是：北方方言、吴方言、湘方言、赣方言、客家方言、闽方言和粤方言。目前，针对汉语方言的语料库研究主要还处于语音语料库的建设阶段。其中，北京海天瑞声科技有限公司标注的 King-ASR-384-4①和数据堂标注的赣方言普通话语音识别语料库②是两个代表性的赣方言语音语料库。为清晰起见，表 1-1 列举了这两个语料库的一些参数，主要包括语料的时间、人数、年龄分布、采集方式、片区等。

表 1-1　代表性赣方言语音语料库

语料	时间	人数	年龄分布	采集方式	片区
King-ASR-384-4	520.7 小时	516	18～30 岁；31～45 岁；46～60 岁	手机	未知
赣方言普通话语音识别语料库	未知	50	16～30 岁(45%)；31～45 岁(45%)；46～55 岁(10%)	未知	未知

从表 1-1 的数据可以得知，北京海天瑞声科技有限公司标注的 King-ASR-384-4 是国内较全的代表性的汉语方言语音语料库，其规模一般都在 500 小时以

① http://kingline.speechocean.com/exchange.php?id=16491&act=view

② http://blog.csdn.net/shujutang/article/details/21621401

上，而且语音录制人数也在 500 人以上，同时年龄分布也较为合理(18～30 岁、31～45 岁、46～60 岁)。除此之外，一些研究院所也标注了小规模的汉语方言语料库，例如，“863”四大方言(上海、广州、重庆和厦门普通话)的语音语料库[7]、兰州方言语料库[8]、福建方言语料库[9]、胶东半岛方言语料库[10]、湖南安仁方言语料库[11]、湖南双峰方言语料库[12]、湖南永州方言语料库[13]、湖南客家方言语料库[14]、武汉方言语料库[15,16]、天津方言语料库[17]等。但是，总体而言这些语料库规模一般都不大，另外这些语料库也都不可公开获取，不利于研究者间共享数据，从而限制了相关计算模型的研究。

1.2.2　计算模型

针对变体语言处理的计算模型，主要包括语音处理和文本处理两个方面。限于篇幅，本节主要介绍国内外变体语言(方言)文本处理的研究现状，而语音处理的内容放置在第 8 章和第 9 章加以介绍。

1. 国外语言变体计算模型

国外相似语言识别的计算模型主要集中在句子级别的语言识别 VarDial @ COLING 2014、LT4VarDial@RANLP 2015、VarDial@COLING 2016、VarDial@EACL 2017 和 VarDial@COLING 2018 和国际评测中。其中在 VarDial@EACL 2017 中，相比于汉语相似语言(方言)，外语的语言变体之间差异一般较小，双方基本能互相沟通，类似于普通话和中国香港、中国台湾、中国澳门、马来西亚、新加坡等的“汉语”之间的差别。下面主要从传统模型和基于深度学习的模型两大类型加以介绍。

1) 传统模型

文献[18]研究了印度语言的文本识别；文献[19]利用基于字符的 n-gram 特征来识别马来西亚语和印度尼西亚语文本；文献[20]针对中国大陆、台湾和新加坡三种语言(或语言变体/方言)的篇章所隶属的方言类型进行识别。他们针对篇章中 top-bag-of-word 进行统计，并根据这些单词在相应篇章中的共现情况进行方言识别；文献[21]利用概率图模型方法表明了 uni-gram 特征与字符 n-gram 特征有利于巴西语的文本识别；文献[22]和文献[23]表明了 uni-gram 结合 Naïve Bayes 分类器有利于南斯拉夫语文本的识别；文献[24]验证了 uni-gram 在句子级别的印欧语言识别中的重要性；文献[25]和文献[26]针对英语变体语言的识别，研究表明了 bag-of-words 特征更优于句法或字符序列特征。此外，文献[27]研究了西班牙语的识别，文献[28]～[31]采用传统机器学习模型研究了阿拉伯语识别中一些特征抽取工作。在最近的 DSL 2017 工作中，代表性的方法有基于两阶段的方法：第一阶段采用 n-gram，利用支持向量机(support vector machine, SVM)分类器对语

言所属的组别进行分类；第二阶段仍采用 SVM 分类器对组内的语言进行分类，基于 CNN(convolutional neural networks)的方法、基于语言模型困惑度的方法、基于 LSTM(long short-term memory)的方法。文献[32]～ [35]分别描述了 DSL shared task 2014, 2015, 2016 和 2017 评测任务的相关计算模型，并介绍了相关的组织和评测工作。文献[35]的评测包括四个任务：相似语言的识别、阿拉伯方言识别、德国方言识别和跨语言依存分析。其中，第一个任务目的是识别来自新闻体裁的短文本所属的相似语言类型；第二个任务目的是识别阿拉伯语音转换后文本所属语言类型，包括 Egyptian, Gulf, Levantine, North African 和 modern standard Arabic (MSA)这五种类型[34]；第三个任务目的是识别 Basel, Bern, Lucerne 和 Zurich 四个地区的德语方言类型；第四个任务目的是通过已知某种语言中标注好的依存分析结果来分析其他相似语言文本的依存分析。

2) 基于深度学习的模型

文献[36]进行了 DSL 2016 语言识别工作，同时采用线性 SVM 分类器和基于深度学习的语言识别两种模型。模型虽然简单，但取得了相对最好的语言识别性能，在部分测试集中取得了排名第一的成绩。虽然文献[35]提到基于神经网络的方法在这种小数据集的 DSL 任务中的性能并不太好，需要大量训练数据去调整模型参数。但是由于此模型简单高效(略低于传统方法)，后继可以将此深度模型加以扩充，为此先着重分析此深度学习模型(图 1-1)。

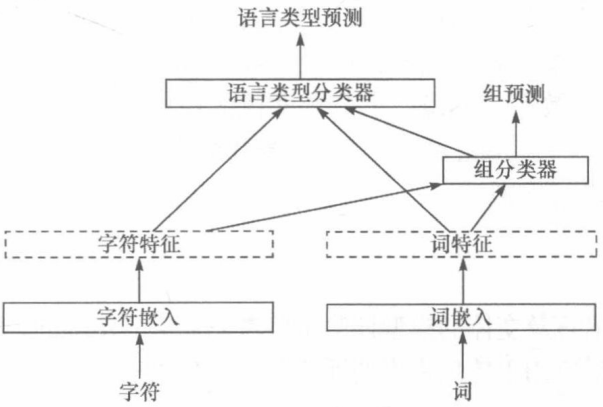


图 1-1　基于深度学习的相似语言识别框架

该模型针对 VarDial@COLING 2016 的国际评测任务，抽取了字符级别和单词级别的分布式表示(向量)，然后将这两种向量进行拼接(concatenation)，再根据 softmax 分类先对这些语言所在的组进行识别，同时将识别出的组信息再次融合至前面的向量，进行更细粒度的语言分类。

2. 国内语言变体计算模型

在国内，清华大学研发的非特定人汉语数码串连续语音识别系统可以识别普通话和四川话两种语言。中国科学院自动化研究所于 2002 年发布的面向不同计算平台和应用的“天语”中文语音系列产品——PattekASR，开启了国产中文语音识别系统的篇章，结束了 IBM 推出的可以识别上海话、广东话和四川话的 ViaVoice 语音系统的市场垄断地位。国内科大讯飞股份有限公司首次将深度神经网络(deep neural network，DNN)技术应用于语音云平台，研发出一个方言语音识别引擎，用于方言口音适配。其中，讯飞输入法已率先支持粤语、四川话、河南话、东北话、天津话、湖南话、山东话、武汉话、合肥话九种方言，具有一定的开创意义。

此外，文献[37]针对中国、新加坡、马来西亚的方言句子进行识别，研究结果表明基于字符的 bi-gram 要显著优于欧洲语言的方言识别中较好的基于字符或单词的 uni-gram 特征，同时提出了基于互信息和词语对齐的特征提取方法，并将这些特征有效地与 bi-gram 特征加以融合，在新闻和 Wikipedia 两种语料下分别验证了这些特征的有效性。文献[38]提出了一种基于联合多样性密度的汉语方言辨识方法，该方法首先将方言预分类为多个小类，然后将各小类方言进行多示例包生成，并通过期望最大多样性密度算法进行多示例学习，最后利用平均最近距离算法进行模式分类，并在闽方言、粤方言和吴方言上的实验结果表明该模型具有一定的辨识度。文献[39]标注了首个篇章级别的江西省方言语料库，共采集了江西省 19 个县市构成的方言点，针对 310 篇文章进行录音，合计 131.5 小时的新闻、故事、散文、小说、书信等文体的篇章级别江西方言语料库。该文抽取出方言文本中的 n-gram 特征，训练 SVM 分类器，并针对句子级别的江西方言所属的方言点进行识别。

1.2.3　评测指标

目前，相似语言及变体的模型评测主要考虑算法的正确度和 F1 值两个性能指标，这些指标广泛应用于信息检索的评测[40]。正确度表示为

$$\text{Accuracy} = \frac{\text{TruePositive} + \text{TrueNegative}}{\text{All}} \tag{1-1}$$

其中，TruePositive 代表本来是正样例，同时分类成正样例的个数；TrueNegative 代表本来是负样例，同时分类成负样例的个数；All 代表样例总个数。

F1 值由准确率(precision)和召回率(recall)共同体现，可表示为

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{1-2}$$

其中，

$$\text{Precision} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalsePositive}} \tag{1-3}$$

$$\text{Recall} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalseNegative}} \tag{1-4}$$

式中，FalsePositive 代表本来是负样例，但被分类成正样例的个数(通常叫误报)；FalseNegative 代表本来是正样例，但被分类成负样例的个数(通常叫漏报)。

1.3 本 章 小 结

本章对主流的相似语言及变体分析技术工作进行了分类、对比和综述，分别从语料库构建、传统及当前深度学习计算模型、评测指标等方面对相似语言及变体的研究现状进行了综述。由此可见，当前国际评测对汉语的变体语言语料库仍然很缺乏。本书的研究一方面可以丰富国际评测中的中文语料库；另一方面也可以促进汉语的变体语言计算模型发展。

参 考 文 献

[1] Hudson R A. Sociolinguistics [M]. Beijing: Foreign Language Teaching and Research Press, 2000.
[2] 陈松岑. 社会语言学导论[M]. 北京: 北京大学出版社, 1985.
[3] 宗成庆. 统计自然语言处理[M]. 2 版. 北京: 清华大学出版社, 2013.
[4] 朱巧明, 李培峰, 吴娴, 等. 中文信息处理技术教程[M]. 北京: 清华大学出版社, 2005.
[5] 田慧刚. 汉语方言区划分概说[J]. 学术界, 1993,3: 90-96.
[6] 李新魁. 汉语各方言的关系和特点[J]. 学术研究, 1991, 2: 87-93.
[7] 李爱军, 王天庆, 殷治纲. 863 语音识别语音语料库 RASC863——四大方言普通话语音库[C]//第 6 届全国现代语音学学术会议论文集, 2003: 274-277.
[8] 杨鸿武, 梁青青, 郭威彤, 等. 一个面向言语工程的兰州方言语料库[J]. 西北师范大学学报(自然科学版), 2009, 6: 54-59.
[9] 王勇卫. 试论福建省方言有声语言资源数据库的建设[J]. 价值工程, 2014, 21: 243-244.
[10] 张绍麒, 姜岚, 张文峰, 等. 胶东半岛方言电子语音语料库的研制与方言电子语音词典编纂[C]//辞书与数字化研讨会论文集, 2014: 163-167.
[11] 何铃玉. 湖南安仁方言语气词研究及其语料库建设[D]. 长沙: 湖南师范大学, 2015.
[12] 李斌. 用 ELAN 自建汉语方言多媒体语料库及其应用研究[D]. 长沙: 湖南师范大学, 2013.
[13] 唐亚纯. 基于 HMM 模型的永州方言数字语音识别系统的研究[D]. 长沙: 湖南大学, 2012.
[14] 陈立中. 湖南客家方言音韵研究[D]. 长沙: 湖南师范大学, 2002.
[15] 罗斌. 武汉方言语音识别系统研究[D]. 武汉: 武汉工程大学, 2015.
[16] 周杨. 计算机汉语方言辨识的理论与方法探讨——以黄孝片方言为例[D]. 武汉: 华中科技大学, 2008.

[17] 胡琼. 基于隐马尔科夫模型的天津方言语音合成[D]. 上海: 上海交通大学, 2011.

[18] Murthy K N, Kuma G B. Language identification from small text samples [J]. Journal of Quantitative Linguistics, 2006, 13(1): 57-80.

[19] Malancon B R. Automatic identification of close languages-case study: Maly and Indonesian [J]. ECTI Transactions on Computer and Information Technology, 2006, 2: 126-134.

[20] Huang C R, Lee L H. Contrastive approach towards text source classification based on top-bag-of-word similarity [C]// Proceedings of the 22nd Pacific Asia Conference on Language, Information and Computation, PACLIC, 2008: 404-410.

[21] Zampieri M, Gebre B G. Automatic identification of language varieties: The case of Portuguese [C]// Proceedings of KONVENS, 2012: 233-237.

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

[23] Ljubesic N, Kranjcic D. Discriminating between closely related languages on twitter [J]. Informatica, 2015, 39(1): 1-8.

[24] Purver M. A simple baseline for discriminating similar languages [C]// Proceedings of the LT4VarDial, 2015: 1-5.

[25] Grefenstette G. Comparing two language identification schemes [C]// Proceedings of Analisi Statistica Dei Dati Testuali, 1995: 263-268.

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

[27] Maier W, Rodriguez C G. Language variety identification in Spanish tweets [C]// Proceedings of the LT4closelang Workshop, 2014: 25-35.

[28] Elfardy H, Diab M T. Sentence level dialect identification in Arabic [C]// Proceedings of ACL, 2013: 456-461.

[29] Zaidan O F, Burch C C. Arabic dialect identification [J]. Computational Linguistics, 2014, 40(1): 171-202.

[30] Salloum W, Elfardy H, Salloum L A, et al. Sentence level dialect identification for machine translation system selection [C]// Proceedings of ACL, 2014: 772-778.

[31] Tillmann C, Onaizan Y A, Mansour S. Improved sentence-level Arabic dialect classification [C]// Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, 2014: 110-119.

[32] Zampieri M, Tan L, Ljubesic N, et al. A report on the DSL shared task 2014 [C]// Proceedings of the VarDial Workshop, 2014: 58-67.

[33] Zampieri M, Tan L, Ljubesic N, et al. Overview of the DSL shared task 2015[C]// Proceedings of the LT4VarDial, 2015: 1-9.

[34] Malmasi S, Zampieri M, Ljubešić N, et al. Discriminating between similar languages and Arabic dialect identification: Areport on the third DSL shared task 2016 [C]// Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects, 2016: 1-14.

[35] Zampieri M, Malmasi S, Ljubesic N, et al. Findings of the VarDial evaluation compaign[C]// Proceedings of the EACL VarDial Workshop, 2017: 1-15.

[36] Coltekin C, Rama T. Discriminating similar languages: Experiments with linear SVMs and neural networks[C]// Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects, 2016: 15-24.

[37] Xu F, Wang M W, Li M X. Sentence-level dialects identification in the greater China region [J]. International Journal on Natural Language Computing, 2016, 5(6): 9-20.

[38] 顾明亮, 张世形, 张浩, 等. 基于联合多样性密度的汉语方言辨识[J]. 计算机工程与应用, 2016, 52(10): 161-166.

[39] Xu F, Wang M W, Li M X. Building parallel monolingual Gan Chinese dialects corpus[C]// Proceedings of the 11th Edition of the Language Resources and Evaluation Conference, 2018: 244-249.

[40] Manning C D, Schütze H, Raghavan P. Introduction to Information Retrieval[M]. Cambridge: Cambridge University Press, 2008.