DOI:10.19651/j.cnki.emt.2106694

基于双字典类标签语言模型的电力调度语音识别*

赵 晴[1] 李庭瑞[1] 罗 睿[1] 李 锐[1] 韩天宇[2] 韩东升[2]
(北京中电飞华通信有限公司 北京 100070; 2.华北电力大学 电子与通信工程系 保定 071000)

摘 要: 语言模型的效果关系到电力调度语音识别系统的识别准确性。为了提高电力调度语音识别的精度,提出一种基于双字典(通用字典和电力调度领域词字典)的类标签语言模型,该模型以 n-gram 语言模型为基础加以改进并添加类标签信息,进而提升电力调度语音识别的准确性;同时提出了一种基于双字典的分词、词性标注的联合系统,用于语料的分词、类标签标注任务,从而提高基于双字典的类标签语言模型对电力调度语言的适应性。最后,在采集到的电力调度指令集上对所提语言模型和常用统计语言模型进行了对比实验。此外还通过实验对联合系统和其他分词、词性标注系统进行了对比。仿真结果表明联合系统的分词、词性标注效率更高。考虑了语义信息、字典、分词和词性标注系统等综合因素的类标签语言模型在电力调度语言识别的准确率更高,词错误率只有 4.14%。

关键词: 语言模型;电力调度;分词;词性标注
中图分类号: TP391 **文献标识码**: A **国家标准学科分类代码**: 520.6010

Power dispatching speech recognition based on double dictionary class label language model

Zhao Qing[1] Li Tingrui[1] Luo Rui[1] Li Rui[1] Han Tianyu[2] Han Dongsheng[2]
(1. Beijing Fibrlink Communications Co., Ltd., Beijing 100070, China; 2. Department of Electronic and Communication Engineering, North China Electric Power University, Baoding 071000, China)

Abstract: The accuracy of power dispatching speech recognition system is related to the effect of language model. In order to improve the accuracy of power dispatching speech recognition, proposes a class label language model based on double dictionaries (general dictionary and power dispatching domain word dictionary). The model improves the n-gram language model and adds class label information, so as to improve the accuracy of power dispatching speech recognition. At the same time, a joint method of word segmentation and part of speech tagging based on double dictionaries is proposed. The system is used for word segmentation and class label labeling of corpus, and then improves the adaptability of class label language model based on double dictionary to power dispatching language. Finally, the comparison experiments between the proposed language model and the common statistical language models are carried out on the collected command set of power dispatching. In addition, the joint system and other word segmentation and part of speech tagging systems are compared by experiments. The simulation results show that the efficiency of word segmentation and part of speech tagging is higher in the joint system. Considering the comprehensive factors of semantic information, dictionaries, word segmentation and part of speech tagging system, the error rate of the proposed model in power dispatching language recognition is only 4.14%.

Keywords: language model;power dispatching;word segmentation;part of speech tagging

0 引 言

随着电网的规模和结构形态的快速发展,中国电网已经发展成为跨区互联大电网,随之带来的是物理电网和信息系统高度融合的问题,电网对信息化、智能化的要求与日俱增。电网信息化程度的不断提升,电子调度系统对“统一调度,各级管理”的需求,导致电力调度中心处理的信息量不断增加。虽电力调度中心已经建设了完备的调度交换通信网络,能够满足日常的调度通信需求,但调度过程中各类问题也不断产生。问题的根源之一在于调度指令不具备

收稿日期:2021-05-14
* 基金项目:国家自然科学基金(61771195)项目资助

可视化的功能，需要依靠调度人员人为的记录、操作，使得调度指令需反复确认，沟通效率不高。电力调度语音识别技术有助于调度人员对调度指令的理解，是电力调度系统中的关键技术之一。调度指令中含有丰富的电力调度领域词汇，例如：电力调度简写词汇、异常地点、操作设备、检修信息以及特殊符号等[1]，因此如何准确识别出调度指令中的领域词汇和通用词汇，是电力调度语音识别系统的主要工作。

语音识别系统由声学模型、发音字典、语言模型和解码器等核心模块组成。声学模型、发音字典从声学以及发音角度对输入单元进行建模，语言模型从语义角度对输入单元进行建模；解码器是在由声学模型、发音字典以及语言模型构成的加权有限状态转换器（WFST）中寻找最优句子[2-3]。

电力调度语音识别系统的研究集中在声学模型、语言模型等方面。文献[4]和[5]分别从汉语声调建模单元、小规模词汇量场景两个方面研究了基于高斯混合-隐马尔可夫模型（GMM-HMM）的电力调度声学模型，没有考虑到语言模型对语音识别系统的影响。文献[6]在小规模调度指令场景下，设计了简单的语法规则，完成了电力调度语音交互系统的设计，但该研究设计的语义信息不全。文献[7]研究了基于双向编码器的 Transformer 模型（BERT）的语言模型在电力调度语音识别中的应用，在 BERT 模型的输入特征向量的基础上，删除了片段特征、添加了关键字特征和命名实体特征。从上述文献可看出语言模型的研究大多只考虑电力调度指令的语义信息，没有考虑字典、分词和词性标注系统等综合因素对语言模型的影响。

为此，本文针对电力调度系统中的语义信息、语言模型的字典、分词和词性标注系统等综合因素，在统计语言模型 n-gram 的基础上，提出了基于双字典（通用字典、电力调度领域词字典）的类标签语言模型。为解决类标签语言模型训练中遇到的通用词汇和电力调度领域词词汇多切分歧义的问题，提出一种基于双字典的分词、词性标注的联合系统来处理类标签语言模型的训练语料，以提升基于双字典的类标签语言模型对电力调度语言的适应性。最后对联合系统和基于双字典的类标签语言模型进行了实验测试，实验结果表明联合系统在分词、词性标注实验，语言模型在困惑度和语音识别上的实验中有着较为明显的优势。

1　基于双字典的类标签语言模型

1.1　基于双字典的类标签语言模型的特点

n-gram 模型是应用广泛的一种统计语言模型。在 n-gram 中，每一个词的出现只依赖于该词的前 $n-1$ 个词，降低了整个语言模型的复杂度，但 n 的取值并非越大越好[8]。n-gram 语言模型可以分为词级别的语言模型和类级别的语言模型。相对与词级别的语言模型，类级别的语言模型在词性区分明显的电力调度语言中有着更大优势，但该语言模型在增加电力调度语音场景的适应性的同时会带来致命的缺陷即语言模型的困惑度会急速下降[9]。由于电力调度语言和通用语言的差异性，为了解决类级别的 n-gram 语言模型困惑度的降低，提出一种以双字典（通用字典和电力调度领域词字典）为基础的类标签语言模型，对 n-gram 语言模型加以改进，用以提高语言模型在电力调度环境中的有效性。

相比类级别的语言模型和词级别的语言模型，基于双字典的类标签语言模型有着如下优点。

1）遇到训练语料中不存在的词序列组合，类标签语言模型能够通过类别间概率和词与类别间的概率间接计算出该词序列组合概率，从而有效避免数据稀疏情况。

2）以双字典为基础来分词、训练模型的方法，能够有效降低字典词向量相似度，且能实时添加新词汇信息至电力调度领域词字典，增强语音识别系统在多专业领域中的应用。

3）类标签语言模型能有效回溯词的类别信息以及领域字典权重信息，提高语言模型的有效性以及泛化性。

1.2　基于双字典的类标签语言模型原理

在一元类标签语言模型中，词与词间的组合概率如下所示：

$$P(w_{t+1}\mid w_t)=P(C_{t+1}\mid C_t)P(w_t\mid C_t)P(w_{t+1}\mid C_{t-1}) \tag{1}$$

其中，$P(C_{t+1}|C_t)$表示类别之间的概率；$P(w_t|C_t)$表示词与类别之中的概率。接着，将展开基于双字典的类标签语言模型及其训练的原理。

首先定义 w_i，c_j 表示的通用字典和电力调度领域词字典中的词和词类别，并且用 N_w 表示两个字典中词的数量，N_c 表示两个字典中类别的数量，其中通用字典和领域词字典的类别种类不重复，且词与词类别的映射是多对多的关系。式(2)描述了词 w_i 与词类别 c_j 的关系。

$$c_j=C(w_i),j\in\{0,1,\cdots,N_c-1\} \tag{2}$$

然后可以得到由词序列 $(w_0,w_1,\cdots,w_{t-1})$ 所对应的一个类标签语言模型的类别序列，公式表达式如下所示：

$$g_i=G(w_0,w_1,\cdots,w_{t-1})=\{C(w_0),C(w_1),\cdots,C(w_{t-1})\},i\in\{0,1,\cdots,N_g-1\} \tag{3}$$

其中，N_g 表示类标签语言模型中类别序列的数量，由于一个词能够属于多个类别，$(w_0,w_1,\cdots,w_{t-1})$ 可以映射类标签语言模型的多个类别序列。假设词的概率 $P(w_i)$ 由词类别所决定，则：

$$P(w_t\mid w_0,w_1,\cdots,w_{t-1})=\sum_{\forall c\,c\in C(w_t)}P(w_t\mid c)\cdot P(c\mid w_0,w_1,\cdots,w_{t-1}) \tag{4}$$

$$P(c\mid w_0,w_1,\cdots,w_{t-1})=\sum_{\forall g\,g\in G(w_0,w_1,\cdots,w_{t-1})}P(c\mid g)\cdot P(g\mid w_0,w_1,\cdots,w_{t-1}) \tag{5}$$

式(4)表示确定词序列 $(w_0,w_1,\cdots,w_{t-1})$ 的下一个词为 w_t 的概率，式(5)表示确定词序列 $(w_0,w_1,\cdots,w_{t-1})$ 的

下一个词类别的概率。由于词可以同时属于通用字典和电力调度领域词字典,两个字典同时存在的词和词类别的映射关系为一对多的关系。定义 $k(w_i)$ 为词与电力调度领域词字典的权重,$k(w_i)=-1$ 时表示词属于通用字典且通用字典中词性信息适合电力调度语音场景,$k(w_i)=1$ 表示词属于电力调度领域词字典且电力调度领域词字典中词性信息适合电力调度语音场景,则:

$$C(w_i)=\begin{cases}C_1(w_i), & k(w_i)=-1\\ C_2(w_i), & k(w_i)\neq -1\end{cases} \tag{6}$$

其中,$C(w_i)$ 表示词 w_i 在通用字典中的类别;$C_2(w_i)$ 表示词 w_i 在电力调度领域词字典中的类别。加入了词与电力调度领域词字典的权重,词与词类别的映射变为了多对一的关系,且词的概率 $P(w_i)$ 完全由词类别决定,则式(4)和(5)可以化简为:

$$P(w_t\mid w_0,w_1,\cdots,w_{t-1})=P(w_t\mid C(w_t))P(C(w_t)\mid w_0,w_1,\cdots,w_{t-1}) \tag{7}$$

$$P(c\mid w_0,w_1,\cdots,w_{t-1})=P(C(w_t)\mid G(w_0,w_1,\cdots,w_{t-1})) \tag{8}$$

以三元类标签语言模型为例,采用最大对数似然估计的方法,求解式(7)和(8)中的分。首先,看一下三元类标签语言模型的词与类别之间的关系:

$$P(w_t\mid w_{t-2},w_{t-1})=P(w_t\mid C(w_t))\cdot P(C(w_t)\mid w_{t-2},w_{t-1}) \tag{9}$$

其中,$P(C(w_t)\mid w_{t-2},w_{t-1})$ 可以由二元类标签语言模型通过式(8)计算得来,接着对式(9)采用最大似然估计方法,得到训练语料的对数似然:

$$LL_{\text{sum}}^{g_i}=\sum_{k=0}^{N_{ww}}N(w_{t-2},w_{t-1})\cdot\log N(w_{t-2},w_{t-1})+\sum_{n=0}^{N_{ccc}}N(C(w_{t-2}),C(w_{t-1}),C(w_t))\cdot\log P(C(w_i)\mid C(w_{t-2}),C(w_{t-1})) \tag{10}$$

其中,N_{ww} 表示二元类标签语言模型的词序列的数量;N_{ccc} 表示三元类标签语言模型中词类别的数量;$N(w_{t-2},w_{t-1})$ 表示训练语料中出现连续词序列 (w_{t-2},w_{t-1}) 的数量。式(10)中的第 1 项不受词和类别之间的影响,对于特定的训练语料是固定的,第 2 项中 $P(C(w_t)\mid C(w_{t-2}),C(w_{t-1}))$ 可以用如下公式表示:

$$P(C(w_t)\mid C(w_{t-2}),C(w_{t-1}))=\frac{N(C(w_{t-2}),C(w_{t-1}),C(w_t))}{N(C(w_{t-2}),C(w_{t-1}))} \tag{11}$$

其中,$N(C(w_{t-2}),C(w_{t-1}),C(w_t))$ 表示训练语料中出现连续类别序列 $C(w_{t-2}),C(w_{t-1})$ 的数量;$N(C(w_{t-2}),C(w_{t-1}))$ 表示训练语料中出现连续类别序列 $(C(w_{t-2}),C(w_{t-1}),C(w_t)$ 的数量。

1.3　基于双字典的类标签语言模型训练过程

基于双字典的类标签语言模型的训练过程如图 1 所示。首先将训练语料经过基于双字典的分词、词性标注的联合系统得到分词后带类标签的语料。其中标记的主要信息包括词的领域词字典权重以及词的类别信息,最后使用由语言模型训练工具训练类标签语言模型。训练过程中使用了 Katz 回退平滑技术处理数据稀疏导致的估计不准问题[10]。

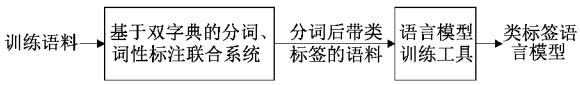

图 1　类标签语言模型的训练过程

2　基于双字典的分词、词性标注联合系统

训练语言模型前需要对未分词的语料进行分词,一般语言模型的分词与字典的词条保持一致,否则字典中不被语言模型包含的词条将成为无效符号[11]。为了处理训练基于双字典的类标签语言模型的训练语料,提出一种基于双字典的分词、词性标注联合系统,在提高电力调度语言场景下的分词和词性标注的精度的同时,还能减少分词产生的切分歧义。基于双字典的分词、词性标注联合系统的结构如图 2 所示。

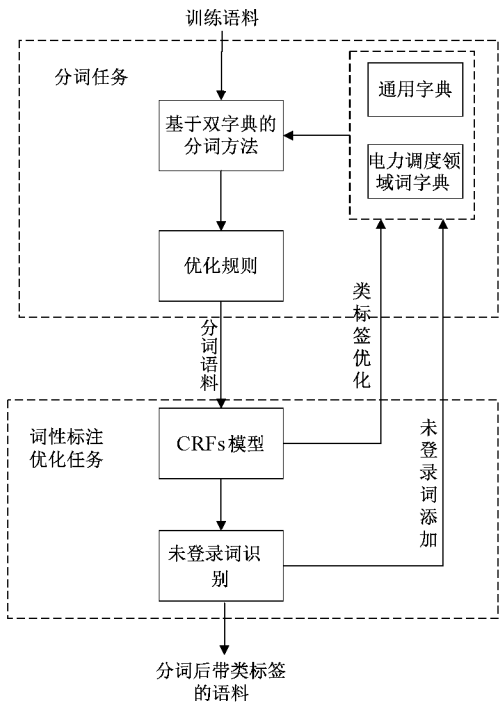

图 2　基于双字典的分词、词性标注联合系统的结构

2.1　分词任务

该联合系统的分词任务采用双字典机制(通用字典和电力调度领域词字典)和规则优化的方法。另对双字典中的词进行分类,分类标准如表 1 所示。

由于算法简单,基于字典的分词系统具有分词速度快的天然优势,然而字典机制采用的匹配算法中存在较多切

表 1　字典词汇分类标准

字典种类	词类别	词类别标签
通用字典	名词	N
	动词	V
	介词	P
	形容词	A
	数量词	Q
电力调度领域词字典	设备名词	EN
	地点	PN
	专业词汇	SN
	特殊符号	SS
	人名	NN

分歧义问题[12]。例如:“检修改冷备用”中的“冷”和“备用”在通用字典中可以单独构词,在电力调度领域词字典中也可以构词为“冷备用”。为更好地消除切分歧义,在训练联合系统时,引入如下优化规则。

当一个词的字粒度 $N_{w_i} \leqslant 2$,类别 $C_{w_i}=N$ 即词类别为名词,则为 $w'=w_t+w_{t-1}+w_{t+2}$,其中 w_{t+1},w_{t+2} 为 w_t 后的两个词,w'为 w_t,w_{t+1},w_{t+2} 拼接而成的新词,若电力调度领域词字典中存在 w',则 w'的领域字典权重 $k_{w'}=1$,w_t,$w_{t\ 1}$,$w_{t\ 2}$ 的领域字典权重可由式(12)求得。

$$k_{w_i}=\frac{N_{w_i}}{Location_{w_i}\times N_{w'}} \tag{12}$$

其中,$Location_{w_i}$ 是词 w_i 在新词 w'中的位置。若电力调度领域词字典中不存在 w',则 w'的领域字典权重 $k_{w'}=-1$。

2.2　词性标注优化任务

该联合系统的词性标注任务主要用于训练语料的类标签标记和字典优化。在分词任务中,会对分词后的语料进行粗略的词性标注,该词性标注的适应能力不高,识别未登录词的能力弱,为了进一步提高词性标注的效率,增加词性标注优化任务。人工智能技术在计算机视觉、目标检测、图像处理等领域都取得了很好的成绩[13-15]。自然语言处理领域中,条件随机场(CRFs)模型和 BERT 模型等人工智能技术表现卓越。本文的词性标注优化任务就由 CRFs 模型和未登录词识别组成。

CRFs 模型是一种结合了最大熵模型和隐马尔可夫模型特点的无向图模型。在词性标注优化任务中,选择词性信息作为 CRFs 模型的特征,“上下文”窗口的大小为 5,解码过程采用向前-向后算法,以迭代的方法训练 CRFs 模型[16]。CRFs 模型除可提高词性标注效率的作用外,还能将字典中词的错误类标签信息修改为优化后的类标签信息。

未登录词识别是本联合系统针对电力调度领域词字典中不存在的领域词汇,例如调度人员的更替和新型的调度设备等新涌现的词汇的任务,可根据规则识别未登录词和类别并添加到字典中。考虑电力调度语音多特有名词的特点,设置如下规则,用以标注未登录词的词性。

1)检测到“路”、“街”、“站”等地点特征明显的单一词汇,回溯前面的名词词汇,直至词汇标注为动词或形容词,将动词或形容词到单一词汇之间的名词整合起来,并标注为地点(PN)。

2)检测到两个动词之间的连续词性标注为名词(N)的数量>3 且不含地点特征明显的词汇时将动词间的名词整合起来,并标注为设备名词(EN)。

3　实验结果与分析

3.1　词性标注优化任务

为了训练语料能够反映电力调度领域中文本多样性,实验时将训练语料中的标点符号全部用空格取代,并将采集到的样本数据以 7∶3 的比例分成训练集和测试集。其中语料数据统计如表 2 所示。

表 2　语料数据统计

数据集	词汇种类	词汇种类占比/%
训练集	通用字典词汇	20
	电力调度领域词字典词汇	78
	未登录词汇	2
测试集	通用字典词汇	20
	电力调度领域词字典词汇	78
	未登录词汇	2

本文将实验分为 3 部分,分别是联合系统、基于双字典的类标签语言模型的实验和该语言模型在电力调度语音识别中的实验,其中训练联合系统、类标签语言模型的工具分别采用 FudanNLP 自然语言处理工具和 SRILM 语言模型训练工具。分词、词性标注联合系统实验采用的指标为准确率(Precision)、召回率(Recall)和 F 值(F-Value),其中 F 值是准确率和召回率两者的一个调和平均值;类标签语言模型实验采用的指标是困惑度(Perplexity)。

3.2　实验结果分析

首先为了验证本文联合系统在电力调度语言场景下的分词、词性标注的有效性,加入分词、词性标注的管道系统和基于领域词字典的分词、词性标注的联合系统实验,与本文联合系统形成对照。各分词、词性标注模型实验结果如表 3 所示。

从表 3 可知,联合系统的分词、词性标注精度优于另外两个模型,说明基于双字典的分词、词性标注联合系统能够有效地处理分词过程出现的切分歧义中,尤其提高处理电力调度领域词汇中经常出现的多义组合型切分歧义的效率。可以看出联合系统将分词、词性标注相结合能够有效提高整体效率,提升模型在电力调度领域的适应性,未登录词汇的分词和词性标注精度可以通过加大训练语料规模,

表 3　各分词、词性标注模型实验结果

分词、词性标注系统	词汇种类	分词精度			词性标注精度		
		P/%	R/%	F/%	P/%	R/%	F/%
分词、词性标注管道系统	通用字典词汇	85.50	85.94	85.72	82.24	82.47	82.35
	电力调度领域词字典词汇	85.21	85.63	85.42	82.08	82.36	82.22
	未登录词汇	76.42	76.73	76.57	75.06	75.32	75.19
基于领域词字典的分词、词性标注联合系统	通用字典词汇	86.37	86.66	86.51	83.36	83.72	83.54
	电力调度领域词字典词汇	88.27	88.53	88.40	83.15	83.43	83.29
	未登录词汇	78.02	78.47	78.24	76.19	76.63	76.41
基于双字典的分词、词性标注联合系统	通用字典词汇	90.21	90.72	90.46	86.42	86.93	86.67
	电力调度领域词字典词汇	91.32	91.65	91.48	87.26	87.73	87.49
	未登录词汇	81.04	81.46	81.25	78.21	78.46	78.33

优化未登录词汇识别的规则等方式加以提升。

在分词、词性标注实验后,接着进行实验比较在各分词、词性标注模型对不同语言模型的影响,其中类标签语言模型是三元模型。各语言模型的困惑度,如表 4 所示。

表 4　各种语言模型的困惑度

分词、词性标注系统	语言模型	困惑度
分词、词性标注管道系统	Bi-gram	2 563
	Tri-gram	1 974
	基于双字典的类标签语言模型	1 708
基于领域词字典的分词、词性标注联合系统	Bi-gram	2 336
	Tri-gram	1 792
	基于领域词字典的类标签语言模型	1 644
基于双字典的分词、词性标注联合系统	Bi-gram	2 218
	Tri-gram	1 558
	基于双字典的类标签语言模型	1 376

从表 4 可知,语言模型的困惑度不仅与语言模型的种类有关、还跟语言模型训练前分词所选分词、词性标注模型有关,训练语料采用基于双字典的分词、词性标注联合系统进行处理后训练出来的基于双字典的类标签语言模型的困惑度最低,语言模型效果最好。

3.3　电力调度语音识别效果

上文对联合系统和类标签语言模型的性能进行了实验测试,接下来将进行基于双字典的类标签语言模型在电力调度语音识别系统中应用实验,来测试该语言模型在实际电力调度语音识别中的性能。

本次实验中声学模型采用基于 Kaldi 上的 nnet3 神经网络框架训练的声学模型,将训练好的声学模型、语言模型以及相关字典构成 WFST 解码器后,用于测试电力调度语音识别。实验使用词错误率(WER)和句错误率(SER)作为评价指标。

$$\mathrm{WER}=100\times\frac{Sub+Del+Ins}{Num}\% \tag{13}$$

式(13)为词错误率的公式,表示识别出的词序列中需替换、删除、插入的词的数量与标准词序列总数的比值。其中,Sub 表示需要替换词的个数,Del 表示需要删除词的数量,Ins 表示需要插入词的数量,Num 表示词序列中词的总数。

$$\mathrm{SER}=100\times\frac{Sen_{error}}{Sen_{total}}\% \tag{14}$$

式(14)为句错误率的公式,其中 Sen_{error} 表示错误句子的数量,Sen_{total} 表示为总的句子的数量。

电力调度语音识别测试结果如表 5 所示。从表 5 中可以看出,采用了联合系统处理的语料,然后训练出的基于双字典的类标签语言模型在电力调度语音识别任务中有着较低的词错误率和句错误率,能够显著地提高电力调度语音识别的准确率,分词、词性标注模型能够影响语言模型,进而影响电力调度语音识别的性能。此外,基于双字典的类标签语言模型相比 Bi-gram、Tri-gram 在电力调度语音识别任务中词错误率和句错误率更低,这与语言模型困惑度测试中的结果相对应,进一步说明了本文提出的采用联合系统训练出来的类标签语言模型能够提高语音识别在电力调度领域中的应用。

表 5　电力调度语音识别测试结果

分词、词性标注系统	语言模型	WER/%	SER/%
分词、词性标注管道系统	Bi-gram	12.16	23.42
	Tri-gram	10.73	19.97
	基于双字典的类标签语言模型	8.48	15.05
基于双字典的分词、词性标注联合系统	Bi-gram	9.43	19.32
	Tri-gram	7.37	15.56
	基于双字典的类标签语言模型	4.14	7.27

4　结　　论

本文针对电力调度语言与通用语言在语义上的差异性，提出一种以 n-gram 语言模型为基础加入改进的基于双字典的类标签语言模型，从而提高语音识别在电力调度场景的性能。同时提出一种基于双字典的分词、词性标注联合系统用于语言模型所需语料的分词、词性标注等预处理任务，进而提高语言模型对电力调度语音的适应性。实验结果表明，本文提出的方法能够提高语言模型的性能，且有效地提升了电力调度语音识别的准确率。此外，本文提出基于双字典的分词、词性标注的联合系统在电力调度领域词汇的分词、词性标注任务也有着较为出色的表现。

参考文献

[1] 王适乾. 电力调度控制系统中语义解析技术研究[D]. 济南：山东大学，2018.

[2] 于镭，林冉腾. 基于香橙派的智能语音识别系统的设计[J]. 电子测量技术，2019，42(19)：36-40.

[3] MENDIS C，DROCPPO J，MALEKI S，et al. Parallelizing WFST speech decoders [C]. IEEE International Conference on Acoustics，Speech and Signal Processing，2016：5325-5329.

[4] 易雪蓉. 电力系统下语音识别的研究与应用[D]. 武汉：武汉工程大学，2018.

[5] 鄢发齐，王春明，窦建中，等. 基于隐马尔可夫模型的电力调度语音识别研究[J]. 武汉大学学报(工学版)，2018，51(10)：920-923.

[6] 杨柳青. 语音人机交互及其在智能调度中的应用[D]. 济南：山东大学，2013.

[7] 陈蕾，郑伟彦，余慧华，等. 基于 BERT 的电网调度语音识别语言模型研究[J/OL]. 电网技术，2021：1-8.

[8] MAIPRADIT R，HATA H，MATSUMOTO K. Sentiment classification using N-gram IDF and automated machine learning [J]. IEEE Software，2019，36(5)：65-70.

[9] KANG R，ZHANG H，HAO W，et al. Learning Chinese word embeddings with words and subcharacter N-Grams[J]. IEEE Access，2019，47(6)：87-92.

[10] ARISOY E，CHEN S F，RAMABHADRAN B，et al. Converting neural network language models into back-off language models for efficient decoding in automatic speech recognition[J]. IEEE/ACM Transactions on Audio，Speech，and Language Processing，2014，22(1)：184-192.

[11] 唐琳，郭崇慧，陈静锋. 中文分词技术研究综述[J]. 数据分析与知识发现，2020，4(Z1)：1-17.

[12] 梁喜涛，顾磊. 中文分词与词性标注研究[J]. 微机发展，2015(2)：175-180.

[13] 王斌. 基于人工智能技术在高校验实室建设中应用与研究[J]. 国外电子测量技术，2020，39(7)：33-37.

[14] 王晓雷，李栋豪，郑晓婉，等. 基于 RBF 神经网络的跌倒检测算法研究[J]. 电子测量与仪器学报，2019，33(11)：185-191.

[15] 蓝金辉，王迪，申小盼. 卷积神经网络在视觉图像检测的研究进展[J]. 仪器仪表学报，2020，41(4)：167-182.

[16] ZHANG M，YU N，FU G. A simple and effective neural model for joint word segmentation and POS tagging [J]. IEEE/ACM Transactions on Audio，Speech，and Language Processing，2018，26(9)：1528-1538.

作者简介

赵晴，工程师，主要从事能源互联网、电力信息通信技术研究与管理工作。

E-mail：zhaoqing@sgitg.sgcc.com.cn

李庭瑞，工程师，主要从事电力通信技术研究与管理工作。

E-mail：litingrui@sgitg.sgcc.com.cn

罗睿，工程师，主要从事电力通信技术研究工作。

E-mail：luorui@sgitg.sgcc.com.cn

李锐，工程师，主要从事电力通信技术研究工作。

E-mail：lirui1@sgitg.sgcc.com.cn

韩天宇，工学硕士，主要研究方向为语音识别、自然语言处理。

E-mail：932167541@qq.com

韩东升，工学博士，副教授，博士生导师，主要研究方向为无线通信网络与新技术、能源互联网信息通信技术。

E-mail：handongsheng@ncepu.edu.cn