融合形态结构与语法关系的藏语语言模型

Tibetan Language Model Integrating Morphological Structure and Grammatical Relations

一级学科：计算机科学与技术

学科专业：计算机应用技术

作者姓名：更太加

指导教师：党建武 教授

答辩日期	2020 年 07 月 29 日		
答辩委员会	姓名	职称	工作单位
主席	李爱军	研究员	中国社会科学院语言研究所
委员	张为	教授	天津大学
	熊德意	教授	天津大学
	魏建国	教授	天津大学
	陈彧	教授	天津理工大学

天津大学智能与计算学部

二〇二〇年七月

独创性声明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作和取得的研究成果，除了文中特别加以标注和致谢之处外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得 天津大f 或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名: 签字日期: 2020 年 8 月 6 日

学位论文版权使用授权书

本学位论文作者完全了解 天津大f 有关保留、使用学位论文的规定。特授权 天津大f 可以将学位论文的全部或部分内容编入有关数据库进行检索，并采用影印、缩印或扫描等复制手段保存、汇编以供查阅和借阅。同意学校向国家有关部门或机构送交论文的复印件和磁盘。

(保密的学位论文在解密后适用本授权说明)

学位论文作者签名: 导师签名:

签字日期: 2020 年 8 月 6 日 签字日期: 2020 年 8 月 6 日

摘　要

语言是现实生活中最主要的信息交流方式。语言模型是语言研究中的一项基础工作，能够提供有效的词表征以及词序列的概率化表示，可以应用于语音识别、机器翻译、手写体识别和句法分析等相关研究。目前，语言模型在英语、汉语和日语等语料相对充足语言领域已经取得了比较理想的效果。而针对藏语的相关研究仍处在初级阶段，由于藏语语料资源的匮乏和研究人员的稀少，严重制约了藏语语言模型的研究发展。在此背景下，本文从藏语自身的语言特点着手：一方面构建了藏语语料库，以验证本文研究结果的合理性；另一方面是从藏语形态结构出发，解决在有限的语料中获取更加有效的信息来补充资源缺乏的问题。

藏语作为资源匮乏的语言之一，目前没有公开的、标准的音频和文本数据资源。根据藏语拉萨方言的特点和藏语文本的特殊性，本文考虑了音素平衡以及文本域问题，构建了藏语的音频和文本语料库。基于藏语句子中一些虚词接续错误和低频词问题，本文重点关注了藏语中后缀对虚词的影响，以及形态动词对低频词的影响。

在上述基础上，首先，本文提出了藏语静态形态结构关系的语言模型。与其他语言不同，藏语中特有的静态形态结构关系（即后缀对虚词接续关系）会严重影响藏语句子的语义理解。具体地，除了字本身的信息之外，字的后缀信息能够使其更加准确接续正确的虚词。因此，本文将静态形态结构融入到字的信息中，以纠正句子中一些语法错误，从而使句子语义能够准确表达。其次，本文提出了藏语动态形态结构关系的语言模型。我们发现在语料中有一些动态形态结构关系（即藏语中的形态屈折变化词），这类词比较特殊且重要，对句子的语义会产生重要影响，尤其是在语音识别中的同音字，预测错误的可能性较大。由于词类中候选词越多，其对应的候选词权重越低，被选中的概率就越低。为此，我们对藏语中的形态动词进行加权，使其不但能够被分配到更高的词类中，而且能更加准确地表示句子语义。最后，本文提出了融合静态和动态形态结构的藏语语言模型。经统计发现，静态形态结构关系可以纠正句子中语法错误的问题，而动态形态结构可以使句子中形态动词的权重发生变化，这两种结构具有互补的关系，能够进

一步增强对藏语句子语义的理解。我们有效融合了静态和动态形态结构，不仅考虑到后缀对虚词的影响，而且对形态动词进行了加权以增强句子的语义理解，比仅考虑单个特性的模型在性能上有所提升。

综上所述，通过构建藏语语料库以及对其分析，我们发现语法和低频词问题。进而将后缀对虚词的语法问题以及形态动词对低频词的影响应用于藏语语言模型的研究中，可以有效提升藏语句子的识别和理解能力。除了语音识别，本文的工作还可被应用到手写体识别、机器翻译和句法分析等藏语自然语言处理的不同任务上，希望该工作能为藏语信息处理研究做出一点绵薄之力。

关键词： 藏语语言模型，静态形态结构，藏语语法，动态形态结构，自动语音识别

ABSTRACT

Language is the most important way of information exchange in real life. Language model is a basic work in language research. It can provide effective word representation and probabilistic representation of word sequences. It can be applied to related researches such as speech recognition, machine translation, handwriting recognition, and syntax analysis. As the core component of the natural language processing system, the language model can provide word representation and overall and has achieved relatively ideal results in the relatively redundant language field of training corpus. The study of Tibetan language models is still in its infancy. Considering the lack of Tibetan corpus resources and the scarcity of researchers, the existing work is basically applied to English, Chinese and Japanese research methods. In this context, starting from the model structure of the deep neural network, a series of systematic and in-depth studies are carried out. On the one hand, it is to verify the effectiveness of the model we build, on the other hand, from the Tibetan morphological structure, to solve the problem of obtaining more effective information in a limited corpus to supplement the lack of resources.

Tibetan language is a kind of low resource language, there are currently no open standard audio and text data resources. Based on the characteristics of the Tibetan Lhasa dialect and the particularity of the Tibetan text, we considered the phoneme balance and text domain issues, resulting in a Tibetan audio and text corpus. Based on the continuation errors and insertion of some functional words in Tibetan sentences, we focus on the influence of suffixes on functional words in Tibetan and the influence of morphological verbs on additional words.

On the basis of the above, first of all, we propose a language model of the static morphological structure of Tibetan. We have found that, unlike other languages, the unique static morphological and structural relationship in Tibetan (that is, the suffix-to-functional continuation relationship) will seriously affect the semantic understanding of Tibetan sen-

tences. Specifically, in addition to the information of the character itself, we also integrate the suffix information of the character, so that the character can be more accurately connected to the correct functional word. Therefore, considering the static morphological structure will correct some grammatical errors in the sentence, the sentence semantics can be accurately expressed. Secondly, a language model of the dynamic morphological structure of Tibetan is proposed. We found that there are some dynamic morphological and structural relationships in the corpus (i.e, morphological inflections in Tibetan). In Tibetan, morphological inflectional change words are special and very important. Certain words will have an important influence on the semantics of sentences, especially homophones in speech recognition. This may be the reason for prediction errors. The more expected word pairs in this category, the lower the probability of being replaced. Transformation, we transform the morphological verbs in Tibetan, the transformation can not only be assigned to a higher part of speech, predict the probability estimate, and Semantics can be more accurate. Finally, a Tibetan language model combining static and dynamic morphological structures is proposed. Our statistical corpus found that the static morphological structure relationship can refer to the problem of grammatical errors in the sentence, and the dynamic morphological structure can change the weight of the morphological verb in the sentence. This effectively integrates the static and dynamic morphological structure. Influence, and the morphological verbs have been increased, and the performance has been improved than considering the characteristics.

To summarize, through the establishment and analysis of the Tibetan corpus, we have discovered some characteristics that have an important influence on the Tibetan language. The study of the effect of long suffixes on functional words and morphological verbs on sentences mapping Tibetan language model can effectively improve the recognition and understanding of Tibetan language. In addition to speech recognition, our work can be applied to the field of Tibetan natural language processing such as handwriting recognition, machine translation and syntactic analysis. We hope that through this work we will make contribution to the research of Tibetan information processing in the future research.

KEY WORDS： Tibetan language model, Static morphological structure, Tibetan grammar, Dynamic morphological structure, Automatic speech recognition

目　录

第1章 绪论

语言是人类最重要的交际工具，是人与人进行思想交流的媒介，它会影响人类社会的发展。一直以来，人们渴望着用语言与机器沟通，自从 1952 年美国 AT&T 贝尔实验室成功研制了世界上第一个能识别十个英文数字发音的实验系统以后，人和机器用语言交流成为可能。1960年英国的Denes等人成功研究出了第一个计算机语音识别系统，开启了人类对语音识别的研究。在对于信息化飞速发展的时代，“语音识别”离我们的生活越来越近，简单来说就是让机器识别和表达语言的语音内容，让机器具有语言的听觉功能。现在的一些科技发达国家已经开始应用语音识别产品，比如：吩咐机器人做一些家务，问一些问题，这些都是基于语音识别技术实现的。随着人工智能的不断进步，将来有一天我们可以用语音和机器进行日常交流，甚至可能会指责洗衣机衣服洗得不干净。

1.1 研究背景和意义

随着计算机技术和信息技术的发展，尤其是深度学习的出现，语音识别、图形图像、机器翻译和视频处理等领域取得了成功应用。深度学习方法首先在语音识别领域取得了较大的突破。语音识别产品应用于各种商业系统和应用软件，如Google语音搜索、Bing语音搜索、Siri语音助手、百度语音助手和科大讯飞语音助手等。作为人机交互工具，语音正在改变人们的生活。

当前，最新的语音识别系统需要大量的资源。世界上约有7100多种不同的语言[1]，由于很多语言使用的人数较少或缺少书写的文字，正在逐渐消失[2]，我们将这些语言称为资源稀缺（Low-Resource）语言。由于资源稀缺，这类语言利用现有的处理语言的技术时面临着很多问题和挑战[3]。资源稀缺语言语音识别的研究已成为语音识别领域中一个非常活跃的研究方向，这种研究不仅可以促进语音识别的进一步发展，还可以利用这些技术来保护这些语言，有利于其文化的保护、传播和传承。由此，资源稀缺语言的研究具有重要的价值和意义。近年来，深度学习技术不断改进语音识别效果的性能，使得稀缺资源的语言语音识别效果也得

到了很大的提升。在语音识别研究中，除了声学模型，语言模型也很重要，所以我们在研究中有必要对语言模型进行研究。

语言模型可分为传统的文法型语言模型和基于统计的语言模型。考虑文法的语言模型是人工编制的语言学文法，文法规则来源于语言学家掌握的语言学知识和领域知识，但这种语言模型不能处理大规模真实文本。为满足处理大规模的文本这一需求，基于统计的语言模型应运而生。统计语言模型是将一些词序列，根据算法确定哪个词序列的可能性更大，或给定若干个词，预测下一个最可能出现的词语。其目的是建立一个能够描述给定词序列在句子中出现的概率分布，在语音识别、机器翻译、信息检索、文本生成等相关研究中得到广泛应用[4–6]。

统计语言模型是计算一个词序列可以构成一个句子的概率的模型。例如，图1-1在语音识别过程中，计算机通过使用声学模型和查找发音词典，可以将语音信号转化为文字，但转化结果不是唯一的[7]。假设语音识别系统的输入音频为“Recognize Speech”，计算机或者可以识别出正确的结果“Recognize Speech”，或者会识别出错误的结果“Wreck a nice beach”。对待两个均符合词法的句子，计算机很难通过判断得出正确的结果。此时语言模型就可以通过计算两个句子的概率，帮助计算机筛选出更符合语法语音的结果，所以语言模型是很多自然语言处理相关任务中必不可少的一部分。

对于语音识别，语言模型是提供字或词间的上下文信息和语义信息，由于语音信号的复杂性，不同音的发音间存在叠接现象，有些单音节词若没有上下文信息让人分辨也很困难，语言模型可以提高声学模型的区分度，语言模型可以是语言中一些规则或语法结构，也可以是表现字词的上下文间的统计模型。

近年来，随着信息技术的飞速发展，国内一些少数民族语言的语音识别也得到了相应的发展。如蒙语、藏语、维吾尔语和哈萨克语等少数民族语言，这种进步也会相应影响少数民族地区的经济和文化。藏语作为少数民族语言，我们有必要对其进行研究，要学习技术成熟语言的基础上，应用藏语自身的一些语言特征

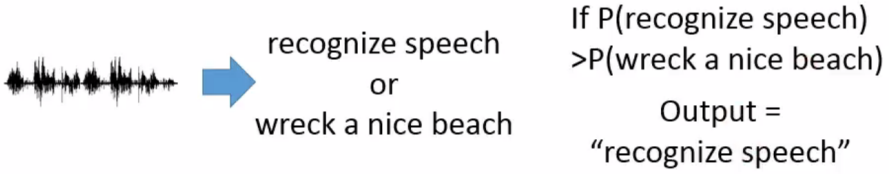

图 1-1　　语言模型在ASR中的应用。

来辅助藏语语音识别。

1.2 研究的现状

统计语言模型在很多研究领域都发挥着重要的作用[8–11]，语言模型的目的是建立一个能够描述给定词序列在语言中可能出现组合的概率分布，常见的方法有N-gram模型方法、决策树方法、最大熵模型、最大熵马尔科夫模型和条件随机域等等，这些方法的缺点是无法解决长距离信息、维数问题和序列问题等。传统的 N-gram 语言模型由于其容易理解、模型结构简单、训练速度快等优点在相关研究人员中盛行好几十年。但是 N-gram 模型可能面临严重的数据稀疏问题，相继各种平滑技术[12–15]被提出来解决数据稀疏问题。

自Hinton 等人提出有效地训练深度神经网络算法开始[16,17]，深度学习技术逐渐流行并在多个领域取得显著成果。Bengio提出NNLM，将几个历史词拼在一起作为输入，将当前词放在输出层作为目标。为了解决词典的高维数问题，NNLM利用了映射层，对输入进行降维。NNLM属于连续型模型，自带平滑，对相同的词历史有一定的聚类功能，一定程度上增加了模型的鲁棒性[18]。

随着人工智能的发展，循环神经网络语言模型（RNNLM）在很多语音、自然语言处理相关领域表现出了很好的性能，从而超过了传统的 N 元语法模型（N-gram），成为主流的语言模型建模方法。循环神经网络（Recurrent Neural Network, RNN）[19,20]序列数据处理上表现出了很好的性能。Mikolov等人在NNLM上的基础上，提出了RNNLM，因RNN在序列建模上有很大的优势。RNNLM将词历史抽象成一个state，降低了输入维数。此外，为了解决输出维数过大的问题，将输出层的词进行聚类，通过因式分解，降低了计算复杂度，将其应用在语音识别[21–23]领域。RNNLM虽然解决了传统N-gram 模型中存在的数据稀疏和维数灾难问题，但缺乏对长距离信息的描述能力。随后由Hochreater和Schmidhuber在1997年提出Long Short Term Memory(LSTM)[24]和Gated Recurrent Units(GRU)[25]引入弥补了RNNLM长距离信息的问题[26]。

循环神经网络与传统的 N-gram 语言模型相比，可以从一定程度上解决数据稀疏问题，但是需要大量的训练数据作为支撑。为了减小对大数据的依赖，一部分研究者希望通过增加循环神经网络的输入特征，来丰富词向量表达，进而提高语言模型的性能并帮助循环神经网络学习到更多有用的上下文相关信息[27,28]。对

于语言模型的研究大多数是基于词，因为词是能够最小独立运用的语言单位。目前也有一些研究者将词进行分解，利用词的形态特征进行分析，应用比词小的粒度作为单元，理论上，输入粒度越小，就需要越少的训练数据，所以选用比词更小粒度的子词（Subword）作为输入单元，获取到词的形态结构，可以一定程度上解决数据匮乏的问题。子词[29–31]包括词素、字符等比词更小粒度的单位。在文献[32]中，在循环神经网络的输入层和输出层使用了词的形态结构，但是这种方法的一个缺点是需要使用额外的工具来划分词素，且词素划分的准确性将会影响实验结果。

这里的形态结构包含静态形态结构和动态形态结构，静态形态结构是字符内在关系或组合方式形成的固定关系，而动态形态结构是根据上下文关系进行变化的结构关系。在[5]中通过将传统的词嵌入（ Word Embedding）级降到字符级，避免了大规模的嵌入（Embedding）计算和低频词的问题，通过Highway Network技术构建更深的网络，得到了不错的结果。该模型由两个部分组成，char-level 作为输入，输入给CNN，通过CNN和Highway Network的输出，输入给 RNNLM，但最终预测仍然是词。这里获取的是静态结构中字符的结构信息，通过以多种语言语料作为测试进行实验，可以获得语义和语法信息。

随着技术的发展和改进，一些研究者提出词和字结构信息相结合的语言模型。提出了一种混合字(Character)级别和词(Word)级别的语言模型，通过一种gate机制来选择字级别结构信息用来表示一个词向量，还是直接用词级别来表示一个词向量。字(Character)级别结构模型的优势在于解决低频词的问题，很多已有的模型都是用字结构来作为基本单元。实验结果表明gate机制的混合语言模型有效地利用了字母级别输入对罕见词和未登录词的刻画，并在多个英文语料上性能超过了词级别的语言模型[33]。

对于藏语语言模型相关研究较少，藏语属于低资源（Low Resource）语言，缺乏公开的数据集进行研究[34]。先前的研究应用传统N-gram方法以及一些平滑方法（线性插值、Kneser-Ney 平滑和加法平滑）来进行处理[35]。藏语中字（Character）指藏语中分隔符切分的一个单元，类似于汉语中的一个汉字，因为藏语中部件（Radical）是指30个字母，词是指具有语义的词（需要分词），我们这里的字是介于他们之间的一个粒度，即一个音节。藏语中以音节点区分[36]。我们之前的研究也利用形态结构来探讨了如何解决藏语循环神经网络语言模型训练过程中存在的数据匮乏问题。一，针对藏语语料不足的问题，提出了三种解决方案，从字结构上获取更多的信息[78]；二，提出自适应的方法，解决语料不足的问题[38]；三，应

用了LSTM方法获取字的更长的历史信息[43]。

以上研究中我们利用的是静态的形态结构关系，且未考虑藏语中的词法和语法关系对句子的影响。为了解决这一问题，我们将从融合静态形态结构和动态形态结构的角度出发，尝试应用藏语中后缀对虚词接续关系及语义关系，以及词法理论来获取到更多的有用信息。

1.3 研究的内容和创新点

本研究主要是研究藏语资源匮乏的情况下，如何利用藏语中静态形态结构和动态形态结构，修正藏语语言模型中出现的语法和词法错误，从而提高句子语义的准确表示。在藏语自然语言处理相关研究问题中，藏语语言模型发挥了重要作用，比如自动语音识别，手写体字符识别和机器翻译等。例如，在微信中，输入标准的汉语语音就可以解码出准确的文字，而藏语中目前为止还未实现，类似应用在实际生活中较多，如语音搜素引擎、智能机器人等。所以，藏语语言模型能够有效帮助机器来理解用户，正确识别出用户的表达，可以提供更好的用户体验和沟通，因此藏语语言模型研究很重要，具有很高的研究价值和应用价值。

藏语语言模型的研究起步较晚，基本都是借鉴英文和中文的方法，且基于N-gram方法[31,39]。我们之前的研究首先应用了深度学习的方法，在模型中应用RNN方法获取了基于字符的一些静态形态结构信息，取得较好的效果[37,40,41]。由于语言模型在具体任务上获取的信息不同，比如机器翻译任务上我们需要语言规则来获取目标语言信息，在语音识别中我们需要预测下一个词的信息。因此，我们有必要构建一个数据库来验证我们的方法。

主要研究工作创新点包括以下几点：

构建语料。由于藏语没有公开的数据集，需要构建符合我们研究需要的音频和文本数据库。音频语料根据拉萨方言元音长短对立和后缀影响的特点，我们考虑了音素平衡。对于文本数据集，我们从互联网上爬取藏文文本数据，并对文本进行去噪处理，参照[5,21]，整理出标准的藏语语料库。我们的语料库域以新闻为主，所以数据域为新闻的测试集相对于其他内域的测试集效果比较显著。为了验证，我们提出了组合基字（Combination Tibetan Radical, CTR）的卷积神经网络（CNN）方法，这种方法验证了我们的文本数据集具有藏语特色，且符合我们研究的需求。

利用静态形态结构关系，引入后缀信息对藏语虚词的影响，用深度学习方法构建后缀接续虚词的藏文语言模型。针对现有语言模型研究方法，我们发现，对于预测藏语句子时，虚词判断错误会导致句子的语义发生变化，从而使句子表达错误信息。为确保纠正句子能够准确表达语义是语言模型的基本任务，本文将藏语后缀对虚词的接续关系这一信息加入语言模型。我们提出了基于显式后缀判定虚词的方法，解决了显式后缀在句子中非自由虚词的判定错误的问题。我们研究发现藏语中还有一些字符存在隐式后缀，通常是省略不写的，而这些隐式的后缀会影响自由虚词的接续。由此，我们提出了基于隐式后缀判定虚词的方法，不但可以解决显式后缀对非自由虚词的影响，还可以解决隐式后缀对自由虚词的影响，使句子的语义更加准确。

利用动态的形态结构关系，引入藏语动词的形态屈折变化对句子的影响，提升这些词在语言模型任务中的权重。这里的动态形态结构指的是藏语中的形态动词，它的形态屈折变化比较特殊，且对句子的语义会产生影响，尤其是出现同音字时，预测错误的可能性就较大。所以，有必要对这些词进行加权，突出这类词在句子中的重要性。我们首先观察到，与其它语言不同，藏语包含许多形态动词，这些动词很少出现在句子中，但是，在字符的准确预测中起关键作用。这种属性通常被现有方法所忽略，并使得传统的训练策略在构建准确的藏语语言模型方面不太有效。因此，我们通过字频重新计算形态动词来调整判别权重，提出了形态特征的藏语语言模型。

融合静态和动态形态结构对藏语句子的影响，提出了考虑语法和形态动词相结合的藏语语言模型。根据静态和动态的形态结构关系，研究了藏语语法和形态动词对语言模型的影响，我们在研究中对后缀和虚词的语法关系进行了分析，考虑了藏语语法对模型的影响。针对低频词问题，我们分析发现，藏语中形态动词基本都属于低频词或罕见词，因此，我们对形态动词进行了加权。这种方法可以验证语言特点融入后对句子语义的影响其效果，所以融入语言自身的特征可以提高性能。

1.4 章节关系和安排

语言模型最早是应用在语音识别任务上，当然它仍然在现代语音识别系统中发挥着核心的作用，并且也被广泛应用到其他的自然语言处理任务之中[42]。最原

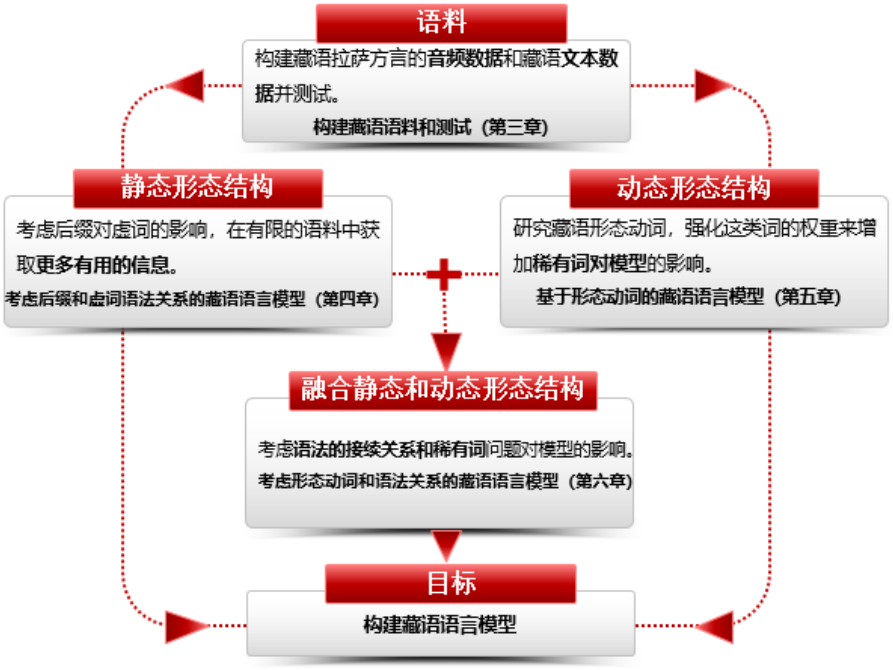

图 1-2　　章节关系图与研究内容

始的语言模型是采用参数评估技术来实现的，这个技术在很多的自然语言处理任务中都被使用。对于藏语语言模型的研究，现有的工作基本都是英语和汉语语言模型研究的方法直接应用[35]。对于藏语语言模型的研究，也有基于形态结构来研究，缓解了藏语数据的稀疏问题[27,38,43]。从语言自身的角度而言，任何语言都有它自身的特点。基于这点出发，本文着重藏语语法特点和形态结构关系对句子的影响，强化语言的特点来构建有效的模型。具体而言，本文针对藏语预料匮乏问题，对现有方法基础上补充了藏语特有的一些信息。图 1-2从可用信息角度体现了不同章节的主要研究内容及章节关系。

本文的章节安排如下：

第一章　首先对语言模型的研究背景和意义进行介绍，然后，对现有的藏语语言模型方法进行了系统总结和分析。最后，介绍了本文的主要研究内容和创新点。

第二章　首先具体介绍了本文所研究的语言模型问题以及藏语语言模型的现状，然后进一步介绍藏语语言模型存在的问题。最后，介绍了本文所用的评价指标。

第三章　介绍了本文第一个创新点，即藏语语料库的构建以及建模测试，我

们提出了基于DNN方法的藏语不同音素集声学模型研究方法，以及藏语文本数据库的构建，并通过我们提出的基于形态结构的组合基字相关藏语语言模型在数据集上进行了验证，证明了我们构建的语料库符合我们研究的需求。

第四章　对本文第二个创新点，利用静态形态结构获取信息，即考虑后缀与虚词语法关系的藏语语言模型，进行了详细介绍。相对于直接借用现有英语和汉语的研究方法，我们的方法可以获取更多的语法信息。

第五章　详细介绍了本文第三个创新点，即利用动态形态结构进行加权，利用藏语形态动词对稀有词进行加权。与传统的方法相比我们的方法取得较好的效果，而且可以极大提升句子语义的理解。

第六章　介绍了本文第四个创新点，融合静态形态结构和动态形态结构获取信息，即考虑形态特征和语法关系信息。将我们提出的考虑语法和基于形态动词的方法进行融合，进一步补充和强化了语法信息和稀有词信息，有效解决了数据稀疏前提下获取更多的有用信息，在我们数据集上也取得了最好的结果。

第七章　对全文的主要内容、创新性和研究结果进行了总结，并对各个创新点在其他相关问题的应用前景进行了展望。

第2章 语言模型概述

语言模型在研究自然语言处理的各个任务中都占有很重要的地位，为了更好地理解本文后续的主要研究内容和创新点，本章首先对语言模型进行介绍。然后，分析藏语语言模型中存在的问题。最后，介绍数据集的评价标准。

2.1 语言模型的简介

语言模型（language model，LM）是人工智能（Artificial Intelligence，AI）中一项基础性工作，在自然语言处理（Natural Language Processing ，NLP）、语音识别（Speech Recognition）、机器翻译（Machine Translation， MT）和手写体识别（Hand Writing Recognition）等领域广泛应用[5,6,8,44]。语言模型是对一段文本的概率进行估计即针对文本X，计算$P(X)$的概率，所以语言模型处理的是序列数据的预测问题。

要判断一段文字是不是准确的一句话，可以通过这些词的概率分布来判断其存在的可能性。语言模型中的词是有顺序的$W =< w_1 w_2 \cdots w_n >$，通过计算该词序列的概率$P(W)$来判断这句话是不是一句合理的自然语言，关键是看这些词的排列顺序是不是正确的。其中，

$$P(W) = P(w_1 w_2 \cdots w_N) = \prod_{i=1}^{N} P(w_i | w_1^{i-1}), \tag{2-1}$$

$w_1^{i-1} =< w_1 w_2 \cdots w_{i-1} >$表示词 w_i的先前词的序列，$P(w_i|w_{i-1})$表示在给定的先前词的历史信息w_1^{i-1}条件下预测到词w_i的概率，与此同时需要满足以下条件，

$$\begin{cases} P(w_i|w_{i-1}) > 0 \\ \sum\limits_{w} P(w_i|w_1^{i-1}) = 1 \end{cases} \tag{2-2}$$

传统的语言模型基本是基于 N 元语法模型（N-gram），这种方法简单有效，所以在很多自然语言处理任务中广泛使用。然而，N-gram 语言模型存在一定的局限性，即数据稀疏和不能够获取长距离信息问题。随着人工智能的发展，神

经网络的出现缓解了数据稀疏问题和长距离信息获取问题，尤其是循环神经网络（Recurrent Neural Network, RNN）、Long Short Term Memory(LSTM)[24]和Gated Recurrent Units(GRU)[25]等在处理序列数据的时候表现出了良好的性能，在从一定程度上解决了 N-gram语言模型的数据稀疏问题。

2.1.1 N-gram语言模型

N-gram 是最为普遍的统计语言模型。简单的N元语言模型通过词汇的共同出现频率(Word Co-occurence Frequencies)来进行估计概率，在很多自然语言处理研究领域都广泛应用。它的基本思想是将文本里面的内容进行大小为 N 的滑动窗口操作，形成长度为 N 的子序列，对所有子序列的出现频度进行统计。N-gram是基于一个假设：第n个词出现与前n-1个词相关，而与其他任何词不相关,这也是隐马尔可夫模型的基本假设之一。整个句子出现的概率就等于各个词出现的概率乘积，各个词的概率可以通过语料统计计算得到，通常N-gram取自文本或语料库。

$$P(w_i|w_1^{i-1}) = P_{NG}(w_i|w_{i-N+1}^{i-1}), \tag{2-3}$$

满足上述条件的语言模型为 N 元语法模型（N-gram），N=1时称为unigram，即为：

$$P(w_1, w_2, \cdots, w_m) = \prod_{i=1}^{m} P(w_i), \tag{2-4}$$

N=2称为bigram，即为：

$$P(w_1, w_2, \cdots, w_m) = \prod_{i=1}^{m} P(w_i|w_{i-1}), \tag{2-5}$$

N=3称为trigram，即为：

$$P(w_1, w_2, \cdots, w_m) = \prod_{i=1}^{m} P(w_i|w_{i-2}w_{i-1}), \tag{2-6}$$

假设下一个词的出现依赖它前面的一个词，即 bigram，假设下一个词的出现依赖它前面的两个词，即 trigram，以此类推。值得注意的是：这里的实际含义是句首词概率，实际应用中应加入起始符$< s >$，如对于 bigram 而言，同理应在句尾加入结束符$< /s >$，加入起始符和结束符的意义在于对句中任意长度的部分序列均进行建模。利用极大似然估计（Maximum Likelihood Estimation, MLE）计算句子的生成概率。

为了计算概率$P(w_i|w_{i-1})$，通过统计文本频率代表概率统计，公式如下，

$$P(w_i|w_{i-1}) = \frac{c(w_{i-1}w_i)}{\prod_{w_i} w_{i-1}w_i}, \tag{2-7}$$

N-gram方法中N越大，模型困惑度（Perplexity, PPL）越小，模型效果越好。从直观上是说依赖的词越多，我们获得的信息量越多，对未来的预测就越准确。然而，容易使某些词在模型中的出现概率显著降低，从而出现稀疏问题。

N-gram最大的问题就是稀疏问题（Sparsity）。例如，在bigram中，若词库中有20k个词，那么两两组合（ C_{20k}^{20} ）就有近2亿个组合。其中的很多组合在语料库中都没有出现，根据极大似然估计得到的组合概率将会是0，从而整个句子的概率就会为0。最后的结果是，我们的模型只能计算零星的几个句子的概率，而大部分的句子算得的概率是0，这显然是不合理的。因此，我们要进行数据平滑（data Smoothing），基本思想是“劫富济贫”，即提高低概率（如零概率），降低高概率，尽量使概率分布趋于均匀。它的本质是重新分配整个概率空间，使已经出现过的N-gram的概率降低，补充给未曾出现过的N-gram。

加法平滑方法（Additive smoothing） 。也叫拉普拉斯平滑，这种方法在上个世纪前半叶由G.J.Lidstone, W.E.Johnson和H.Jeffreys等人提出和改进。即强制让所有的N-gram至少出现一次，只需要在分子和分母上分别做加法即可。这个方法的弊端是，大部分N-gram都是没有出现过，很容易为它们分配过多的概率空间。

$$P(w_n|w_{n-1}) = \frac{C(w_{n-1}w_n) + 1}{C(w_{n-1}) + |V|}, \tag{2-8}$$

加法平滑算法虽然简单，但是应用起来效果很差。在2-8基础上做了一点小改动，原本是加1，现在加上一个小于1的常数K。

$$P(w_n|w_{n-1}) = \frac{C(w_{n-1}w_n) + k}{C(w_{n-1}) + k|V|}, \tag{2-9}$$

缺点是这个常数仍然需要人工确定，对于不同的语料库K 可能不同。

古德-图灵（Good-Turing）估计法。Good-Turing估计法是很多平滑技术的核心。这种方法在1953年由古德（I.J.Good）引用图灵（Turing）的方法而提出来的，古德图灵平滑是一种非常巧妙的平滑方式，直观上讲，它通过高频事件优化低频事件的概率表示，并最终通过层层迭代获取到未登录词的概率。基本思想是：用观察计数较高的N元语法数重新估计概率量的大小，并把它指派给那些具有零计数或者较低计数的N元语法。

$$c^* = (c+1)\frac{N_{c+1}}{N_c}, \tag{2-10}$$

c^*是Good-Turing平滑计数，c是某个N元语法出现的频数，同样N_{c+1}是所有发

生次数为c + 1的元组个数，一般来说，发生次数为c的元组个数多于发生次数为c + 1的元组个数，为了消除混淆发生次数和元组个数，我们假设，元组就是字典的key，元组个数是字典的value，元组个数是key的个数，不太严谨，但是好理解。N_c是出现次数为c的N-gram词组的个数，是频数的频数，如下所示，

$$N_c = \sum_{b:c(b)=c} 1, \tag{2-11}$$

Katz平滑法。 Katz平滑方法通过加入高阶模型与低阶模型的结合，扩展了Good-Turing估计方法。是根据低阶的语法模型分配由于减值而节省下来的剩余概率给未见事件，这比将剩余概率平均分配给未见事件合理。其基本思想是：对于任何一个出现了 c 次的 N 元语法，都假设它出现了c^*次

$$c^* = (c+1)\frac{n_{c+1}}{n_c}, \tag{2-12}$$

其中，训练语料中出现次数为 c 次的 N 元语法的数目用n_c表示。实现高阶模型与低阶模型的结合Good-Turing方法无法实现，而 Katz 平滑算法就是 Good-Turing 估计方法进行了扩展，通过加入高阶模型和低阶模型的结合。

Katz 平滑算法会对 N 元语法出现的次数进行判断，假设 N 元语法出现次数小于或者等于 k，则以一定的回退率d_r 进行回退；如果出现的次数大于 k，则不进行回退，即 $d_r = 1$ 。Katz 平滑算法的公式是，

$$\begin{cases} P_{Katz}(w_i|w_{i-N+1}^{i-1}) = d_r P(w_i|w_{i-N+1}^{i-1}) & if(c(w_{i-N+1}^{i-1})) > 0 \\ P_{Katz}(w_i|w_{i-N+1}^{i-1}) = \alpha(w_{i-N+1}^{I-1}) P_{Katz}(w_i|w_{i-N+1}^{i-1}) & if(c(w_{i-N+1}^{i-1})) = 0 \end{cases} \tag{2-13}$$

以上所述，当使用Good-Turing估计时一般需要平滑n_c，比如，对于那些值非常小的n_c。然而，在Katz平滑方法中这种处理并不需要，因为只有当计数cdk时才使用Good-Turing估计，而对于这些c值来说，n_c一般是比较合理的。

Katz平滑方法属于后备（back-off）平滑方法。这种方法的中心思想是，当某一事件在样本中出现的频率大于k时，运用最大似然估计经过减值来估计其概率。当某一事件的频率小于k时，使用低阶的语法模型作为代替高阶语法模型的后备，而这种代替必须受归一化因子α的作用。

Kneser-Ney平滑方法。R.Kneser和H.Ney提出了一种扩展的绝对减值算法，用一种新的方式建立与高阶分布相结合的低阶分布。在前面的算法中，通常用平滑后的低阶最大似然分布作为低阶分布。然而，只有当高阶分布中具有极少的或没有计数时，低阶分布在组合模型中才是一个重要的因素。因此，在这种情况下，

应最优化这些参数，以得到较好的性能。Knerser-Ney 平滑算法的公式如下，

$$P_{KN}(w_i|w_{i-N+1}^{i-1}) = \frac{c(w_{i-N+1}^{i}) - D}{\sum_{w_j} c(w_{j-N+1}^{j})} + \gamma(w_{i-N+1}^{i-1})P_{KN}(w_i|w_{i-N+2}^{i-1}), \tag{2-14}$$

其中，

$$P_{KN}(w_i|w_{i-N+2}^{i-1}) = \frac{N_{l+}(\bullet, w_{i-N+2}^{i})}{N_{l+}(\bullet, w_{i-N+2}^{i-1}), \bullet)}, \tag{2-15}$$

Knerser-Ney 平滑算法相对其他平滑方法取得了较好的结果，所以，在 N 元语法模型中被广泛应用。

综上所述，N-gram 方法有很多优点：

第一，可以直接处理自然语言，对参数空间进行了优化，具有很强的解释性。

第二，它包含前 n-1 个词的全部信息，不会产生丢失和遗忘。

第三，它还具有计算逻辑简单的优点。

但是，N-gram 方法也存在本质上的缺陷：

第一，N-gram 方法无法获取长距离信息，当 n 过大时会出现数据稀疏问题，训练速度慢，生成的模型大，会影响任务性能，所以，在实际应用中我们常用 bigram 或 trigram。

第二，N-gram 基于频次进行统计，没有足够的泛化能力。

随着深度学习的出现，神经网络语言模型逐渐取代传统的统计自然语言模型成为主流[45]。

2.1.2 语言模型自适应方法

在自然语言处理系统中，语言模型的研究也提升了一些任务的性能。语言模型的理论基础已比较完善，在实际应用中经常会遇到一些不好处理的问题。其中，模型对跨领域的脆弱性（brittleness across domains）和独立性假设的无效性（false independence assumption）是两个最明显的问题。我们知道训练语言模型时所采用的语料往往来自多种不同的领域和主题，这些综合性语料难以反映不同领域之间使用规律上的差异，而语言模型对于训练文本的类型、主题和风格等都十分敏感；另外N元语言模型的独立性假设前提是一个文本中的当前词出现的概率只与它前面相邻的n-1个词相关，但这种假设在很多情况下是明显不成立的。因此，为了提高语言模型对语料的领域、主题、类型等因素的适应性，提出了自适应语言模型

（adaptive language model）的概念。对于语言模型自适应方法的研究也提出了一些方法，如基于缓存的语言模型（cache-based LM）[10]、基于混合方法的语言模型（mixture-based LM）[44,46]和基于最大熵的语言模型[47,48]等。

基于缓存的语言模型自适应方法研究针对的是在文本中刚刚出现过的一些词，在后边句子中再次出现的可能性较大，比标准的n元语法模型预测的概率要大。对于这类研究，黄非等提出了利用特定领域中少量自适应语料，在原词表中通过分离通用领域词汇和特定领域词汇，并自动检测词典外领域关键词实现词典自适应，然后结合基于缓存的方法实现语言模型的自适应方法[49]。曲卫民等（2003）基于记忆的自适应语言模型对 n 元语言模型改进 ,通过采用TF-IDF公式代替原有的简单频率统计法，从而在一定程度上消除了高频词的影响。建立基于记忆的扩展二元模型，考虑到了不同的词之间的相互影响，并采用权重过滤法以节省模型计算量，实现了对基于缓存记忆的语言模型自适应方法的改进[50]。张俊林等（2005）也对基于记忆的语言模型进行了扩展，利用汉语语义类词典，将词汇语义上相近或者相关的词汇也引入缓存，在一定程度上提高了原有模型的性能[51]。

基于混合方法的自适应语言模型针对的问题是语料，由于大规模训练语料本身是异源的（Heterogenous），来自不同领域的语料无论在主题（Topic）方面，还是在风格（style）方面，或者同时在这两方面都有一定的差异，而测试语料一般是同源的（Homogeneous），因此，为了获得最佳性能，语言模型必须适应各种不同类型的语料对其性能的影响。

基于最大熵的语言模型，基于缓存的语言模型（Cache-based LM）和基于混合方法的语言模型（Mixture-based LM）自适应方法采用的思路都是分别建立各个子模型，将子模型的输出组合起来。最大熵模型是通过结合不同信息源的信息构建一个语言模型，一个直接将输入层线性变换到输出层的矩阵。每个信息源提供一组关于模型参数的约束条件，在所有满足约束的模型中，选择熵最大的模型。这个模型的动机是因为训练数据集很大时，隐藏层也不得不随之增加，所以采用联合训练最大熵模型进行训练，可以将隐藏层的规模控制得很小。

综上所述，语言模型的自适应方法是改进和提高语言模型性能的重要手段之一。由于语言模型广泛地应用于自然语言处理的各个方面，而其性能表现与语料本身的状况（领域、主题、风格等）以及选用的统计基元等密切相关，因此，其自适应方法也要针对具体问题和应用目的（语音识别、机器翻译、信息检索、语义消歧等）综合考虑。

2.1.3 循环神经网络语言模型

统计语言模型在学习词序列的联合概率时，存在的问题是计算量和存储参数巨大，我们把这个称之为维度灾难，数据稀疏在高维条件下会导致语言模型存在很多为零的条件概率。N-gram 方法的思想是考虑对词序较近的词依赖更大，降低上下文长度，减少了参数数量，但丢失了长距离信息，若 n 元取较大值，则会出现稀疏问题；另外统计语言模型中的变量都是离散型，每个变量的轻微改变都会对联合概率产生很大影响，且基于离散变量的概率分布很难通过汉明距离捕获句子之间的相似性。

2003年 Bengio 等人提出，神经网络语言模型(Neural Network Language Model, NNLM)，其思想是提出词向量的概念，代替 N-gram 方法使用离散变量(高维)，采用具有一定维度的实数向量来进行单词的分布式表示，缓解了维度爆炸问题，同时通过词向量可获取词之间的相似性[23]。NNLM 将联合概率通过拆分为两步来计算：将词汇表中的每个词对应一个分布式向量表示，对句子中的词向量通过函数得到联合概率，在大语料上通过神经网络来学习词向量和联合概率函数的参数。

N-gram 语言模型的问题在于捕捉句子中长期依赖的能力非常有限，NNLM 对输入数据要求固定长度(一般取5-10)，直观上看就是使用神经网络编码的 N-gram 模型，也无法解决长期依赖的问题。为了克服n元模型的缺陷，2010年 Mikolov 提出 RNNLM，因为语言模型任务是一个序列预测问题，RNN 是天然用来解决序列问题的模型，RNNLM 的历史信息是句子前边所有的词，使其可以捕获更长的历史信息。RNN解决的就是获取上下文信息问题，神经网络比n元模型更优秀，主要就是对于历史记忆的的获取。n元模型的历史是精确的 $n-1$个单词的序列，但RNN将这部分记忆放在隐藏层中，并且把稀疏的RNN记忆隐藏层 h射影到低维的连续空间，并且使 h和 $\widehat{h}$能够聚类，于是并不需要精确地匹配历史记忆 h ；另外由于RNN的矩阵和偏置向量共享参数，所以RNNLM只需要从训练集中训练较少的数据，即可使模型具有较高的鲁棒性。

RNNLM的体系结构如图2-1所示。循环神经网络一般有三个层，分别是输入层，隐含层和输出层。输入层由向量w_t和向量s_{t-1}组成，向量w_t表示t时刻输入层的输入，也就是w_t的词向量，w_t的维度为词典的大小V_{word}。s_t表示t时刻隐含层的激活值，它保存了从开始到t时刻的所以现行信息。每个词的概率通过y_t得到，y_t的每一维表示在给定历史词序列$< w_1 w_2 \cdots w_t >$的条件下，$t+1$时刻是词典中每一个词

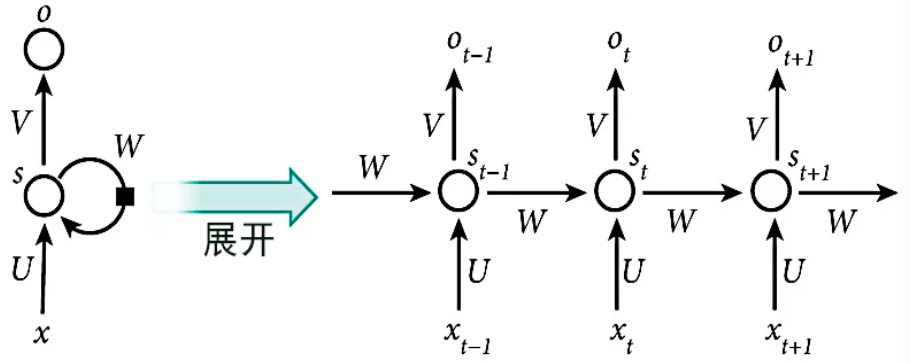

图 2-1　　标准循环神经网络

的概率。

$$s(t) = f(\mathbf{U} \cdot w(t) + \mathbf{W} \cdot s(t-1)), \tag{2-16}$$

$$s_j(t) = f(\sum_i w_i(t) \cdot \upsilon_{ji} + \sum_l s_l(t-1) \cdot w_{jl}), \tag{2-17}$$

$$y(t) = g(\mathbf{V} \cdot s(t)), \tag{2-18}$$

$$y_k(t) = g(\sum_j s_j(t) \cdot \upsilon_{kj}, \tag{2-19}$$

其中对输出层采用softmax函数进行激活，激活函数 sigmoid 函数的优点是非线性激活，经过激活后其输出范围为[0,1]，数据在前向传播的过程中不会太发散，所以曾经被广泛使用。以确保输出的概率满足归一化与大于零的条件：

$$f(x) = sigmoid(x) = \frac{1}{1+e^{-x}}, g(x_k) = softmax(x_k) = \frac{e^{x_k}}{\sum_i e^{x_i}}, \tag{2-20}$$

模型训练采用随机梯度算法(SGD)。最初用比较小的随机数给U 、V 、W三个矩阵赋值，Mikolov在实验中使用平均数0、方差0.1的正态分布进行赋值。每训练一个单词，将U、V、W三个矩阵的参数更新一次。

前馈和循环体系结构之间的主要区别在于历史记录的表示方式，而对于前馈NNLM，历史记录仍只是前几个词，对于循环模型，可以在训练过程中从数据中学习历史记录的有效表示形式。RNN 网络打破了上下文窗口的限制，使用隐藏层的状态概括历史全部语境信息，对比 NNLM 可以捕获更长的依赖，在实验中取得了更好的效果[23]。RNN的隐藏层表示所有以前的历史，而不仅仅是$n-1$个以前的词，因此该模型在理论上可以表示较长的上下文模式。语言也是一种序列信息，循环神经网络获取的是序列信息，所以应用于语言模型中也取得了很好的性能。

对于RNN的训练和对传统的NN训练一样。同样使用BP误差反向传播算法，

不过有一点区别：如果将RNNs进行网络展开，那么参数W，U，V是共享的，而传统神经网络却不是。循环神经网络的前向传播（Forward Propagation）主要是以时间推移的，所以前向传播按照时间顺序向前计算一次即可。对于反向传播（Back Propagation）过程则需要对当前时刻所累积的所有残差进行传递，所以在循环神经网络中所使用的反向传播算法是 BPTT（Back Propagation Through Time）算法。在 BPTT算法中，容易产生梯度消失和梯度爆炸的问题。对于梯度爆炸问题，通常设置一个阈值来限制梯度的增长。虽然循环神经网络理论上可以保存所有的历史信息，但是由于梯度消失问题的存在，循环神经网络只可以对有限步的历史信息进行建模。对于梯度消失问题，循环神经网络结构的改进方法，即长短时记忆模型（Long Short-Term Memory, LSTM）[24]从一定程度上解决了循环神经网络的梯度消失问题，从而可以对更长的先行信息进行建模。

LSTM最早由 Hochreiter & Schmidhuber 在1997年提出，设计初衷是希望能够解决神经网络中的长期依赖问题。LSTM记忆单元具有遗忘门（Forget Gate）、输入门（Input Gate）和输出门（Output Gate），LSTM记忆单元拥有长短时记忆机制[24,52]。其中，遗忘门负责决定保留多少上一时刻的单元状态到当前时刻的单元状态；输入门负责决定保留多少当前时刻的输入到当前时刻的单元状态；输出门负责决定当前时刻的单元状态有多少输出。

$$f_t = f(W_f[h_{t-1}, x_t] + b_f), \tag{2-21}$$

$$i_t = f(W_i[h_{t-1}, x_t] + b_i), \tag{2-22}$$

$$\widetilde{C}_t = tanh(W_c[h_{t-1}, x_t] + b_c), \tag{2-23}$$

$$C_t = f_t C_{t-1} + i_t \widetilde{C}_t, \tag{2-24}$$

$$O_t = W_o[h_{t-1}, x_t + b_o], \tag{2-25}$$

$$h_t = O_t tanh(C_t), \tag{2-26}$$

由于门结构的引入，使得长短时记忆模型从一定程度上缓解了梯度消失的问题，从而获取长距离信息进行建模。第一个Sigmoid门决定了需要从当前状态中舍弃哪些信息，这个门称为“遗忘门”，例如根据当前的状态出现了一个新的主语，那么之前的句子中的主语就应该被丢弃掉。第二个Sigmoid+T函数组成的门决定了哪些信息需要被添加到状态中，这里分为两部分，一个是Sigmoid层决定了将要更新哪些值，这一部分跟第一层完全一样，而T层会创建一个新的信息用来添加到

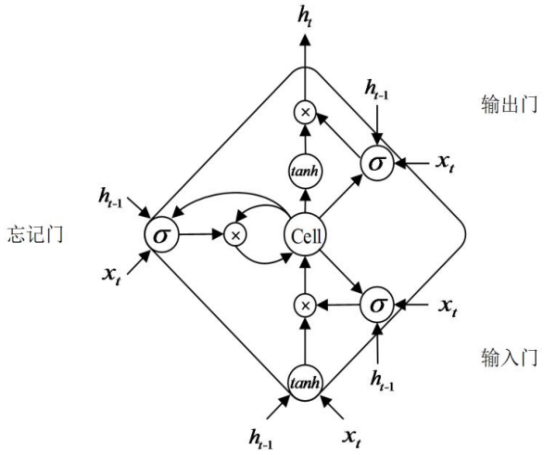

图 2-2　　长短时记忆模型的结构图

状态中。如替换掉之前句子的主语的状态。前边两个门的工作主要是用来更新贯穿线的状态的，第三个门的作用是根据贯穿线上的信息以及当前的输入信息计算模块的输出，更新的依然是哪些信息需要丢弃，哪些信息需要被添加。

我们知道梯度消失会导致神经网络中前面层的网络权重无法得到更新，也就停止了学习。而梯度爆炸会使得学习不稳定，参数变化太大导致无法获取最优参数。在循环神经网络（RNN）中，梯度爆炸会导致网络不稳定，使得网络无法从训练数据中得到很好的学习，最好的结果是网络不能在长输入数据序列上学习。LSTM把原本RNN的单元改造成一个叫做CEC的部件，这个部件保证了误差将以常数的形式在网络中流动，并在此基础上添加输入门和输出门使得模型变成非线性的，并可以调整不同时序的输出对模型后续动作的影响。

2.2　藏语语言模型以及相关研究

形态结构是组成主题的基础属性及其等级结构，对于语言而言是组成词或字的构件或字符。对具有丰富词素结构的语言而言，词汇量本来就很大，这种词汇量的训练会导致数据稀疏，并且难以训练高阶语言模型。对于资源匮乏的语言而言，这个问题尤其具有挑战性。由此对于语言模型的研究，如自动语音识别、手写体识别和机器翻译之类的应用提出了挑战，而且在这些应用中，语言模型对性能产生重大影响。在语音识别中，可以利用语言中丰富的形态和形态结构来缓解数据的稀疏性，如基于词法的子词等[53]。本研究主要对拉萨方言进行研究。

2.2.1 资源丰富的语言

对于形态结构丰富的语言，例如：英语、阿拉伯语、德语和波兰语等语言，除了以词作为处理单元，还有一种方法是借助形态结构进行语言建模。通常，给定文本语料库中可能的字符或部件的数量小于完整单词的数量，这会导致更高的词汇覆盖率。而且，可以适当地组合字符或部件以产生实现较低未登录词（OOV）。此外，字符或部件的平均频率大于完整词的平均频率。

词和字分别是英汉两种语言中的建模单元。汉语中的字有点类似于词或比词小，是汉语的构词单位，相当于英语中的词素（Morpheme）。在形态结构上，汉语中的字和英语中的词大都可以分成更小的音义单位。这些单位都相当于英语中的 Morpheme。Morpheme 可译成词素或字素，相当于汉语的偏旁部首。英语中的构词和汉语中的构字部件都不变，但衍生能力极强，掌握这些基本的构件，对于提高语言学习的效力是很有意义的。

在语言模型（LM）的构建中，神经体系结构是突出的。但是，单词级别的预测通常与子单词级别的信息（字符和字符序列）无关，并且在由有限单词集组成的封闭词汇上运行。对于形态丰富的语言，通过subword-aware获取词信息，并在50多种类型多样的语言上进行了大量实验，这些语言中有各种各样的形态结构，并验证了对于形态丰富的语言而言，会提高困惑度[54]。也有研究提出字符序列和语素序列形态结构进行组合，取得了不错的效果，并且在芬兰、土耳其和俄语上进行了验证[55]。

基于词素的语言模型。基于子词的语言模型是基于词素的语言建模，其中概率估计是针对词素序列进行的，而不是全词序列进行的。通常，通过基于监督或无监督方法进行形态结构分解，从全词生成语素。例如[30]是处理在资源匮乏的情况下关键字的OOV问题。基于词素的子词建模方法在土耳其语资源少的关键字搜索任务中可以有效地恢复OOV关键字，其中混合词和词素解码方法的性能优于传统的基于子词的搜索方法。此外，尽管资源匮乏，形态学的无监督学习也几乎与针对语言设计的基于规则的系统一样有效，分阶段的关键字搜索策略得益于两种形态学分析方法。针对阿拉伯语言的研究，[27]探索了功能丰富的DNN-LM的使用，其中网络的输入是单词和词素及其特征的混合体。[29]结合了语素水平和功能丰富的建模的优点，比较了基于词素序列及其基于特征执行阿拉伯语LVCSR估计的基于流，基于类和因式语言模型的性能。

英语构词的基本单位是词素，无论是词根还是词缀，每个单位都是音义结

合体。词根可分为成词词根和非成词词根，英语中分别叫自由词素（Free Morphem e）和粘附词素（Bound Morpheme)。一个自由词根构成的词，如 sun（太阳）、moon（月亮）、grass （草）、flower （花）等。粘着词根必须有附加成分才能构成词。如 prediction（预测），由词根-dict-和前缀pre-及后缀-tion构成；aquatic（水族的）由词根aqua-和后缀 -ic构成，其中 -dict- 和 aqua- 都是粘着词根。词缀都是不能成词的，像 drinkable中的后缀 -able和 underdevelopment中的前缀 under-分别与单词 able 和 under 是偶合现象，不能看作是成词词素。这些词素都不能再分，否则就是无意义的字母。

基于音素的语言模型。基于子词的语言模型的另一种类型的语言模型，其中主要识别单元是由一个或多个代表语音单位的书面字母组成的音节。音节也可以被识别为语音构建块。通过执行称为音节化的过程，从全字生成音节。在大多数语言中，可以通过应用语言和语音规则来实现音节化。基于音节的语言模型已用于波兰语、德语和英语等语言。对于音素作为建模单元，构建一个标准的音素识别器是关键，所以在以往的研究中音素识别器的好坏会直接影响识别结果。

前面引用的大多数工作要么基于较小的词汇量，要么缺乏适当的优化，例如选择最合适的单元类型，优化整体词汇量，不同单元数和OOV率。由于形态结构的不同和各个语言本身的差异性，构建一个标准、通用和实用的识别器是具有挑战性的。因此，我们有必要对各种语言本身的形态结构进行分析和研究，构建具有语言特点的特征来辅助缓解资源稀疏问题。

2.2.2 藏语以及藏语语言模型

藏语属于汉藏语系藏缅语族藏语支，主要分布在青藏高原地区，包括西藏自治区、青海、甘肃、四川、云南等省份。藏语作为一种少数民族语言，目前中国境内约700万人（2016年）使用藏语。此外，一些邻近的国家印度、巴基斯坦、尼泊尔等也有藏语使用者。在国内，藏语大体可分三大方言，即：卫藏方言（西藏自治区前藏和后藏），康方言（西藏昌都专区、四川甘孜藏族自治州、云南迪庆藏族自治州和青海玉树藏族自治州）以及安多方言（甘肃、青海各藏族自治州、自治县等）。

藏文已有近 1400 年的历史，用藏文记载的经典文献、古籍著述和译作浩如烟海[56]。“藏文”一词写作“bod-yig”（意为“藏族的文字”。藏文作为藏族人民的书面交际工具，历史之悠久在国内仅次于汉文。它是一种拼音文字，属辅音字母

文字型，分辅音字母、元音符号 2 个部分。其中有常用的 30 个辅音字母，4 个元音符号。如图2-3是藏文和国际上通用的藏文拉丁转写。表中元音有5个，因为藏语中一般“a”只是作为元音是存在的，在实际书写中一般略去。

藏语音节也包括声母和韵母，汉语普通话有 39个韵母，按结构可以分为单韵母、复韵母、鼻韵母。汉语的单韵母和藏文的元音应该是相同的，单韵母有 a ,o, e, i ,u ,ü 六个，藏文元音有 a ,o ,e ,i ,u 五个。汉语的复韵母和鼻韵母的功能近似藏文的 10 个后加字和 5 个前加字，10 个后加字尽管从 30 个辅音字母演化而来，但其发音与基本辅音有差异，在基本辅音中，他们的发音为 ga/nga/da /na/ba/ma/a/ra/la/sa/，但在作为后加字或韵母时其发音变为：ag/ang/ad/an/ab/am/a/ar/al/ei/。除了 10 个后加字外，藏文中还有 5个前加字 ga/da/ba/ma/va 和 2 个次后加字，2 个次后加字 da 和 sa 主要起时态和辨义作用，语音上不发生变化，藏文经历第三次文字厘定后，da 几乎省略不被显示出来。 5 个前加字也可称为前韵母，后加字和次后加字可以称为后韵母，这是按其所处位置而言，实际上都应该是韵母范畴，比如前加字 ba 在音节中的发音，bkav 这个音节的安多方言读音同汉语的瓜（gua），其中前加字 ba 的发音和汉语拼音的u 韵母相同。

我们知道汉语和藏语同属汉藏语系，但又处在不同的语支。汉语属于汉藏语系的汉语支，是一种十分有活力的语言，具有灵活的构词法和丰富的词汇，能够反映纷繁的社会现象和表达细腻的思想情感。汉字由象形发展为表意，且动词没有时态变化，被认为是汉藏语系中的一种特殊情况。相对于汉语，藏语有黏着和屈折的特点，藏文是属于拼音文字，语法体系完整而结构严密，其中虚词作为表达语法意义的重要手段。

现代藏语有以下特点：一、藏语语音发音中的变化。随着语言的发展变化，藏语的辅音声母逐渐简化，目前，只有很少的地方还保留着前置辅音构成复辅音。还有就是元音的长度不一，使得元音和声调形成一种互补关系。二、藏语的声调

གསལ་བྱེད་ (字母)	ཀ(ka)	ཁ(kha)	ག(ga)	ང(nga)	ཅ(ca)	ཆ(cha)	ཇ(ja)
	ཉ(nya)	ཏ(ta)	ཐ(tha)	ད(da)	ན(na)	པ(pa)	ཕ(pha)
	བ(ba)	མ(ma)	ཙ(tsa)	ཚ(tsha)	ཛ(dza)	ཝ(wa)	ཞ(zha)
	ཟ(za)	འ(v)	ཡ(ya)	ར(ra)	ལ(la)	ཤ(sha)	ས(sa)
	ཧ(ha)	ཨ(a)					
དབྱངས(声调)	ཨ(a)	ཨི(i)	ཨུ(u)	ཨེ(e)	ཨོ(o)		

图 2-3　　藏文字母及声调

系统有很强的完整性和稳定性，而且有不断增强的趋势。我们知道，在藏语语音识别的任务中，声调信息起到了很好的作用，并且能够提高识别率。这就说明了声调信息的重要性。三、有丰富的虚词结构，这类词在句子表达中起到重要语义作用。藏文是一门语法比较严谨的语言，在应用藏语时都有明确的语法规范。藏文虚词在句子中能够表达句子的语义，错误的虚词接续会导致语义的变化。四、动词具有屈折变化。这类词在句子中一般充当谓语使用，且对句子的语义变化具有影响。五、藏语有敬语和非敬语的区别。

藏语语音机理属性。传统的藏语语音分析研究与现代语音的研究没有太大区别，基本是通过音高、音强、音色来分析语音的性质。但是，在语音的发音机理上藏语具有独特之处，传统语音分析方法认为发音需要的条件是发位（gnas）、发音机理（byed pa）、气（rlung）、意识（rnam rtog）。

发音位是产生声源的基础，藏语的发音位包括胸腔（khog）、喉（mgrin pa）、颚（rkan）、牙（so）、唇（mchu）、舌（lce）、肪腔（spyi bo）、鼻（sna）。气息是发音的动力，通过人的意识来发出语音，是人脑中呈现出的信息传递的过程。

对应藏语字母，可以看出ka，kha，ga，nga，v，sha，等是发音位为喉，ca，cha，ja，nya，zha，ya，等发音位为颚，ta，tha，da，na，tsa，tsha，dza，za，sa，la等发音位为齿，pa，pha，ba，ma，wa等发音位为唇，ra发音位为舌和颅腔，ha，a发音位为胸腔，nga，nya，na，ma发音位为鼻。

四个元音字母（i）读音为（ai），是由喉部发音，并从颅腔发出回响的较紧音；（u）即（au），是由喉和唇发音的，发音时双唇撮合，但不接触，然后降低下唇发出的较紧音；（e，）即（ae）是由喉部发音，发音部位向上抬起，混合有颅腔音的松音；（o）也是由喉咙和双唇发音，发音时下唇上撮发松音。

藏语音节结构。音节是听觉可以区分的语音基本单位，一般有一个或几个音素按一定规律组合而成。藏语文字是拼音文字，藏语的基本单位是音节，音节与音节之间用音节点分隔，一般由 30 个辅音字母和 4 个元音字母组成，根据藏语本身的特点和规律按照字母组合排列而成。

藏语文字书写顺序为从左到右，从上到下，藏文音节有叠加的现象，在叠加书写的结构中以一个辅音字母为中心位置，将字母分为“前加字”，“上加字”，“基字”“下加字”，“元音”“后加字”，“再后加字”，如图2-2所示。

现代藏语文字一般一个音节最少由两个字母构成，即一个辅音字母和一个元音字母，最多用七个字母组成。通常，我们发现书写中只有一个辅音字母，看不见元音，那不是没有元音，而是在新厘定的藏文中，为了方便书写，省略了隐含

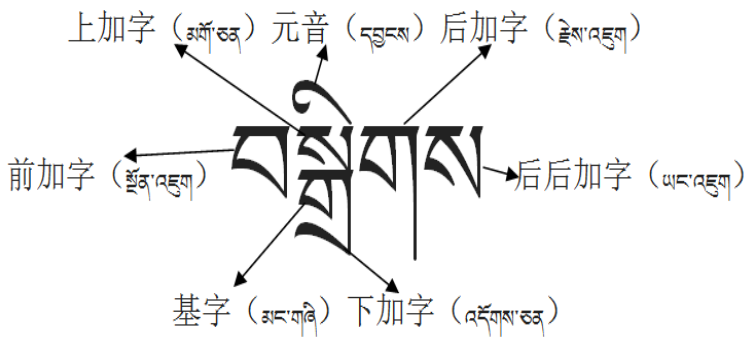

图 2-4 藏文字结构

元音a。藏语中的 a 既是辅音也是元音，同时又是隐式后加字，藏语字母 a 既充当声母又充当韵母，和汉语的零辅音很相似，是属于辅音。

与英语相比，藏语的构词法有些不同。英语使用空格来划分单词，而藏语则需要进行分词。直至目前，还没有现成成熟开源的藏语分词标准，而且语料资源非常有限，所以，在我们的研究中以常用字为单位。构成藏文字一般都1-7个字符（这里不包含梵文），如图2-4藏语中常见的最长字。

对于汉语一个字即是一个字丁，同时也是一个音节，从音节的角度看，GB2312 所包含的就是汉语常用的 6763 个音节。而对于藏语，三十个辅音字母中每一个同元音的结合是一个字，字母和元音以及上加字、下加字相结合而成的 445 个藏文字丁。对于现代藏文字数统计，高定国先生统计字数为19380个，才旦夏茸先生统计字数为17532个，而江狄老师在1998年从100百万字的现代藏语文本中统计出5581个[39,74,96,97,100]。我们可以从江狄老师的研究中发现，现代藏文实际应用的音节大概在5000-6000左右[53]。

2.2.3 藏语语言模型的研究与存在的问题

语音识别就是把语音转换成文字序列。具体来说，是输入一段语音信号，要找一个文字序列（由词或字组成），使得它与语音信号的匹配程度最高。在实际应用中这个匹配程度，一般是用概率表示的。用X表示语音信号，W表示文字序列，则要求解的是下面这个问题：

$$W^* = arg\max_W P(W|X), \tag{2-27}$$

一般认为，语音是由文字产生的，我们利用贝叶斯公式，可以实现条件和结

论的反转：

$$W^* = arg\max_W \frac{P(X|W)P(W)}{P(X)} = arg\max_W P(X|W)P(W), \quad (2\text{-}28)$$

公式2-28是语音识别的核心公式，我们要找W文字序列，就需要使得$P(W)$和$P(X|W)$都大。$P(W)$表示一个文字序列本身的概率，也就是一串词或字本身多接近目标词。$P(X|W)$表示给定文字后语音信号的概率，也就是这句话有多大概率发成这串词音。

藏语语言模型的研究起步较晚，基本都是借鉴英文、日语和中文等资源丰富语言研究的方法，且都用N-gram方法[30,57]。我们之前的研究首先应用了深度学习的方法，在模型中应用RNN（LSTM）方法获取了基于字符的一些结构信息，而且取得了较好效果[39–41]。但是这些方法未考虑词法和语法的影响，现有的研究工作存在的主要问题有：

(1) 由于缺乏公开的标准藏语数据资源，我们有必要构建一个数据库。因为我们构建的语言模型是对针对语音识别任务，要验证模型的有效性必须在语料库上进行实验。声学方面，我们发现现有的藏语拉萨方言的研究音素集存在一些不足，需要进行改进。在语言模型方面，因为没有开源的数据库，需要我们构建一个全新的数据库。

(2) 针对现有语言模型研究方法，发现藏语句子中虚词接续错误，这种错误会导致句子语义发生变化。所以，为了能够准确表达句子的语义，我们有必要利用静态形态结构把藏语后缀对虚词的接续关系这一信息加入语言模型。

(3) 藏语动态形态结构比较特殊，动词的形态屈折变化对语义产生的影响，因为这类词的形态屈折变化对句子时态发生语义上的变化，尤其是在语音识别中的同音字，预测错误的可能性就较大。因此，有必要对这些词进行加权，突出这类词在句子中的重要性。

2.3 评价标准

语言模型的评价标准一般是困惑度（Perplexity, PPL），是衡量模型与真实分布之间的差异。衡量一个语言模型的好坏，最好的方法就是将其应用到具体的问题当中，如语音识别、手写体识别、机器翻译等任务中，然后使用这些任务的评价指标来评估语言模型的性能。我们的工作以语音识别为主，所以我们应用困惑度和语音识别的词错误率（Word Error Rate, WER）进行评价。

2.3.1 困惑度

迷惑度/困惑度/混乱度（preplexity, PPL），其基本思想是给测试集的句子赋予较高概率值，当语言模型训练完之后，测试集中的句子都是正常的句子，那么训练好的模型就是在测试集上的词的概率越高越好，困惑度是对于词序列中的词而定义的：

$$\begin{aligned} PP(S) &= P(w_1 w_2 \cdots w_N)^{-\frac{1}{N}} \\ &= \sqrt[N]{\frac{1}{p(w_1 w_2 \cdots w_N)}} \\ &= \sqrt[N]{\prod_{i=1}^{N} \frac{1}{p(w_i|w_1 w_2 \cdots w_{i-1})}} \end{aligned} \tag{2-29}$$

S代表句子，N是句子长度，$p(w_i)$是第i个词的概率。第一个词的概率就是$p(w_1|w_0)$，而w_0是开始，表示句子的起始，是个占位符。这个式子可以这样理解，PPL越小，$p(w_i)$则越大，我们期望的句子出现的概率就越高。也就说，句子概率越大，语言模型越好，迷惑度越小。可如果我们想一想语言的实质，实际上可以把语言看作一种词汇的序列，而不同的词汇之间千差万别，所以也大致上可以认为它们是离散的符号，这样想来，自然语言实际上就是离散的符号之间的一种特殊的、受语法规则限制的序列。

$$PPL = exp(-\frac{1}{N}\sum_{i=1}^{N} \ln P(W_i|w_1^{i-1})), \tag{2-30}$$

一般情况下，PPL 的值越小说明模型性能越好，越接近我们语料的分布。所以训练语言模型的任务就是寻找 PPL 最小的模型，使得模型更接近于我们真实的语料分布。

2.3.2 语音识别的词错误率

语音识别系统中主要由声学模型和语言模型两个模型组成。声学模型是计算语音到音节的概率，语言模型是计算音节到字的概率。所以衡量语言模型的好坏可以将语言模型应用到语义识别中，通过语义识别的准确率来评价语言模型的性能。

$$PPL = exp(-\frac{1}{N}\sum_{i=1}^{N} \ln P(W_i|w_1^{i-1})), \tag{2-31}$$

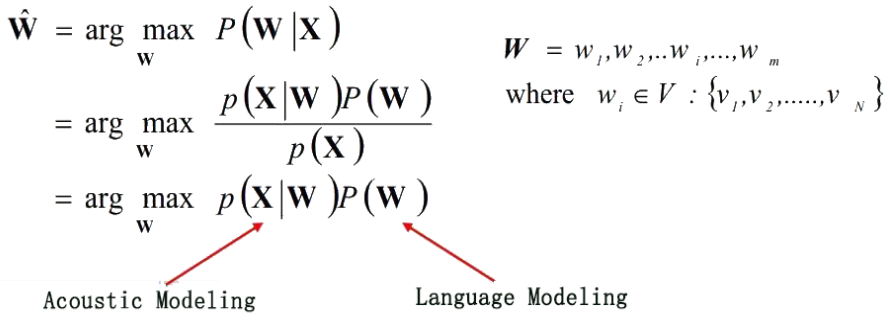

图 2-5　　识别过程要计算声学概率$P(X|W)$ 和语音概率$P(W)$

图2-5 可以看出，计算这两项的值就是语言模型和声学模型的各自任务，而我们研究的是语言模型$P(W)$。为了使识别出来的词序列和标准的词序列之间保持一致，需要进行替换，删除，或者插入某些词，这些插入，替换，删除的词的总个数，除以标准的词序列中词的个数的百分比，即为WER，其计算公式如下所示：

$$WER = \frac{S + D + I}{N} = \frac{Substitutions + Deletions + Insertions}{N}, \tag{2-32}$$

$$WER = 100 \cdot \frac{S + D + I}{N}\% \qquad Accurcy = 100 - WER\%, \tag{2-33}$$

其中， Substitution表示替换，Deletion表示删除，Insertion表示插入，N表示单词数目，错误率越低说明语言模型的性能越好。英文最小单元是词，语音识别应该用”词错误率”（Word Error Rate, WER），中文最小单元是字符，语音识别应该用“字符错误率”（Character Error Rate, CER）。

2.4　本章小结

本章首先简单阐述了语言模型以及主要方法，其中，N-gram作为传统主流方法，加上很多平滑处理方法解决了数据稀疏的问题，在很多任务中得道应用。自适应方法也对不同领域或主题的任务解决了数据本身存在的问题。然而，随着神经网络的崛起，语言模型也随之出现了CNN、RNN、LSTM和GRU等方法，一定程度上解决了数据稀疏和长距离信息问题。但是，因数据量以及语言的差异性，还有很多未解决的问题。其后，本章介绍了形态结构以及相关研究，对于形态丰富语言而言，可以从形态结构上获取更多有用的信息。阐明了英语、德语等语言形态结构上的丰富性以及汉语形态结构上的特点。然后根据藏语的基本属性和特点，

从语音的机理属性介绍了藏语发音机理上独特之处和发音位的构成情况。最后，提出了藏语语言模型研究中存在的问题，第一，因为藏语是低资源语言，能获取到的数据量有限，现有研究方法的基本都是基于传统的方法；第二，藏语语法比较严谨，虚词的接续关系会影响句子的语义；第三，藏语的形态动词对句子的时态及语义也有影响。然后，我们定义了本文实验的评价标准：困惑度和字错误率，由困惑度来评价语言模型的优劣，字错误率来评价藏语拉萨方言的识别率。

第3章　藏语语料的构建和测试

3.1　拉萨方言语音数据库构建和测试

语音识别技术发展了几十年，从简单的模板匹配到现在的噪音下的连续语音识别。构建标准数据库是语音识别理论和技术发展的前提，比较著名的有美国Defense Advanced Research Projects Agency（DARPA）支持的The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus（TIMIT）[58]、Wall Street Journal（WSJ）[59]、Switchboard[60]等数据集。这些数据集包括了音频、音转字和发音字典等，目的是给研究者提供语音识别相关的资源。这些数据集的构建给研究者提供了实验数据和评价标准，通过在同一数据集上应用提出不同方法进行实验和验证，促进了语音识别技术的发展。

藏语作为国内少数民族语言，是低资源（Low Resource）语言[34]，目前尚无公开的标准数据库，而且，在网上数据相对有限，加上研究人员相对较少，获得大规模数据比较困难。我们针对藏语音素平衡构建了拉萨方言数据库，他可以推动了藏语语音识别的发展，对藏语自然语言处理具有重要意义。

语料的选取需要根据语言学的研究成果，并结合藏语在声、韵、调上的特点来选取，选取的语料既能研究拉萨方言的实验语音学，也能应用于拉萨话的韵律建模、语音合成和语音转换。它包括语料的选择、语音录制、标注文本和语音数据管理，我们必须设计一个科学完整的语音语料库。在之前的研究中由于藏语语音结构复杂，有复辅音、辅音韵尾、复合元音等语音现象，清浊音出现规律不统一，还有就是藏语的单音节较多，所以没有公开的数据库。

我们知道设计语料时需要用尽量少的语料覆盖自然语言中的现象。[61]在语料选择时考虑了声调的组合、音段的音联现象、清浊搭配、语句的持续时间等现象。原始文本语料选自2007 年的《西藏日报》文本，以100篇2007年《西藏日报》的新闻稿2 000个句子作为录音的文本。未说明录音人员的的性别以及录音人员，不能体现藏语拉萨方言的语音特点。语料库中共有608条句子，收集了藏语拉萨方言发音标准的4个男生，3个女生的日常口语语音。其中前538句用作模型预训练

初始化模型网络参数，并使用其中带标注的180句语音来回调修正模型，后90句做测试[62]。藏语语音库为解放军外国语学院自建的藏语拉萨话语音语料库，共计49.6h语音数据。所选语料来源于藏语教学过程中使用的录音广播、阅读教材、新闻摘要及口语练习材料[63]。

从以上已有的数据中可以看出，有些没有体现男女生发音者的区别，文本句子太少或只有新闻域，而有些发音者太少，无法获取有效的语音特征。我们构建的语料考虑了拉萨方言的音素平衡，而且录音对象基本都是拉萨本地出生的在校大学生，在一定程度了保障了拉萨方言的质量。

3.1.1 拉萨方言语音数据库构建

音频语料库旨在涵盖所有音节的各种组合，尤其是藏语中两个音节的所有可能性。同时考虑声音片段和节奏，包括多种藏语句子选择中的语调和发声组合。因此，目前对其研究还处在初级阶段。本论文中使用的拉萨方言音频数据库是天津大学天津市认知计算与应用重点实验室和中国社会科学院民族学与人类学研究所合作录制，音频为朗读语音，每个人是读取3126句的拉萨口语文本。朗读文本是由民族所藏语语言学专家设计，考虑了音素平衡，覆盖了拉萨方言日常交流使用的单音节、双音节和三音节词。

该语音语料库的发音人是23名（13男，10女）拉萨方言为母语的大学生。录音之前所有发音人都经过相关培训，以确保录音的质量。语音信号以16kHz采样，16位采样精度。语料库共有38700多句，其中训练集使用36090句，测试使用2664句，训练集和测试集之间没有重叠，表3-1是我们音频语料库的基本信息。

表 3-1　藏语音频数据的基本信息

说话人		语音信号		文本	
男生	女生	采样率	量化精度	训练集	测试集
13	10	16K HZ	16-bit	36,090	2,644

后期的数据校对工作对语料库至关重要，尤其对于一个低资源语言来说，数据质量会影响识别。虽然我们录音之前进行了培训，但是在录制中难免会出现一些错误。我们在检查中发现有些录制的音频和提示语不匹配的，念错句子中的字或词；因操作问题导致的录音不完整；句子录音的前后没有流出足够的静音；因录音环境出现的噪音等问题，需要在后期进行校对，对不合格的音频进行重录或

移除。

3.1.2 藏语音频语料库在不同音素集声学模型上的测试

近年来，深度神经网络（Deep Neural Networks, DNN）声学模型在大词汇量连续语音识别任务中的表现优于传统的高斯混合隐马尔可夫模型（GMM-HMM）。在尝试开发实用的藏语拉萨话语音识别（Automatic Speech Recognition, ASR）系统时，为了获得更好的语音识别准确度，我们使用了基于多种不同音素集的DNN方法研究藏语声学建模的性能。这些音素集是根据藏族拉萨方言的语言学和语音学知识定义的。实验是通过音素使用双语法例语言模型，对20人记录的藏语语料库进行的。音素错误率（Phone Error Rate, PER）结果表明，使用区分韵尾的辅音没有长音设置的声学模型表现最佳，其准确度比基本音素设置高10.43%。此外，我们的结果证实了这样的事实，即对于拉萨藏语声学模型，范例DNN-HMM优于传统的GMM-HMM。

3.1.2.1 拉萨方言音素集

当前，藏语语音分析和识别技术已引起人们的广泛关注[64–66]。藏语字符由词根和其他子部分（例如后缀，前缀等）组成，以表示各种语法类别和语音变化（例如，数字，时态），从而产生大量词汇。后缀随着元音的不同而变化，如果元音不同，即使跟随相同的后缀，发音也会不同。对于音素，不同的音素单元，会影响语音识别性能。

深度学习已成为声学建模中的主要方法[67]。现在人们基本都使用DNN而不是GMM来计算发射概率(Emission Probabilities)并获得令人印象深刻的结果[68–70]。现已证明，具有多隐藏层的神经网络对于声源信号的可变性更鲁棒[71,72]。在DNN-HMM范例中，DNN使用堆叠式受限堆叠受限玻尔兹曼机（Restricted Boltzmann Machine, RBM）进行初始化，并使用交叉熵作为目标函数进行训练。尽管DNN已在ASR域中得到广泛使用，但是在用DNN方法进行藏语连续语音识别方面所做的工作很少。

由此，我们开始设计实用的藏语拉萨方言的ASR系统。但是，我们找不到在藏语声学建模中为适当音素集提供良好选择的参考。为了提高语音识别的准确性，我们采用研究基于不同音素集的DNN方法（基于藏语拉萨方言的语言学和语音学知识）来检测藏语声学模型的性能[73]。采用音素错误率（PER）以评估识别结果

并为我们选择音素提供证据。

3.1.2.2 语言单位

声母，是使用在韵母前面的辅音，跟韵母一起构成一个完整的音节，一般由辅音充当。韵母至少要有一个元音，也可以有几个元音，或元音之后再加辅音。一般韵母由韵头（介音）、韵腹（主要元音）、韵尾三部分组成。藏语的音律体系比较复杂，拉萨方言的口语与书面语有一定的差异，很多方言之间相互都听不懂，但是书面语是相通的。

藏语有30个字母和4个元音组成，在藏语中任何一个单个的字都不表示意义，表示意义的每个单字后面都有一个“a”，现在藏文语法规则为了书写的方便，将“a”进行简化，但是每个懂得藏语言文字的人都明白隐含的“a”。我们在读写的时候也要注意，每个没有元音的字都是有隐含的元音“a”，所以在拉丁转写时或IPA转写，都是要定义一个元音“a”，这样就在藏语传统的4个元音加一个“a”元音即5个元音[74]。

藏语的“音节=声母+声调+韵母”，是组成音节的三要素，“基字+元音”是构成藏语语音的基础。国内的藏语三大方言一般都有a，i，u，e，o五个单元音韵母，拉萨方言元音长短对立实际上，藏语拉萨方言有29个声母，如图3-1所示。表中藏文有五个元音（如“ a”，“ i”，“ u”，“ e”，“ o”）和十个后缀（如“ ga”，“ nga”，“ da”，“ na”，“ ba”，“ ma”，“v”，“ ra”，“ la”和“sa”）。

拉萨方言元音长短对立，有区分词义和语法意义的功能。

1) 元音a在声母和韵尾的变化的影响下有靠前和靠后的变化。
2) 韵尾k，p只闭塞而不而不发生爆破。
3) 韵尾的?在双音节词的前一音节时都会脱落。在语流中往往也会消失。

在29个声母中，需要特别注意的是：

1) R做声母，是舌尖后的半元音或浊擦音。做韵尾是舌尖后的颤音。
2) ɬ，ʂ，h做声母，只出现在高调的音节。
3) ɬ做声母时双音节的第二个音节变读作l。
4) ʂ是一个舌卷清擦音，做声母时双音节的第二个音节变读作r。

3.1.2.3 不同音素集

音素是根据语音的自然属性划分出来的，是构成音节的最小单位或最小的语音片段，是从音质的角度划分出来的最小的线性的语音单位。从声学性质来看，

Lhasa Tibetan initials						
p	c	ts	tɕ	ŋ̊	s	h
p^h	c^h	ts^h	$tɕ^h$	ŋ	ɬ	ɹ
t	k	tʂ	m	l	ʂ	w
t^h	k^h	$tʂ^h$	n	f	ɕ	j
ç						

Lhasa Tibetan finals							
i	y	o:	io	an	oŋ	op	u?
e	ø	u:	im	on	uŋ	up	ir
a	ɛ	y:	em	yn	ip	i?	er
ə	i:	ø:	am	iŋ	ep	e?	ar
o	e:	ɛ:	om	eŋ	ap	a?	or
u	a:	iu	um	aŋ	əp	o?	ur

图 3-1　　拉萨方言的声韵母

音素是从音质角度划分出来的最小语音单位。音素一般用国际音标（IPA）标记音位是一定的语言或方言系统中，能区别不同语言符号的最小语音单位，是根据语音的社会性质划分出来的。音素是根据语音的自然属性划分出来的，在语音学与音韵学中，音素一词所指的是说话时所发出的声音。看两个语音形式是不是同一个音素，只需看发音动作、声学特征是否相同；而看两个语音形式是不是同一个音位，主要是看它们的社会属性，看它们能否区分不同语言符号。音位和音素是集体和个体的关系，一个音位往往包括几个不同的音素。属于同一个音位的不同语音形式，就是这个音位的音位变体。对于音质音位来说，属于同一个音位的不同音质形式（音素），就是音质音位的音位变体。

藏语音素集是根据藏语拉萨方言本身的特点，对音素进行分类，以读音相同的音素来归类的，再加上元音和后加字的组合形式归纳出藏语拉萨方言的语言特点，归结出六个不同的音素集。这些音素集是根据不同的韵律有：区分韵尾的辅音、区分韵尾的辅音没有长音、无鼻化元音、无长音、元音和韵尾合成一个音素。根据藏语实际发音以国际音标来标注的，除了单个音素组成的外，还有一些元音辅音组合而成的音素，元音加后加字，藏语的后加字是特定的，藏语文法中描述“ga，nga，da，na，ba，ma，v ，ra，la，sa”这10 个辅音字母可以做后加字。这10个后加字中“ga，da，ba，ma，v”可以做前加字，“da，sa”是再后加字。在拉萨方言中后加字影响韵母，并决定着声调的升降，再后加字基本和后加字读音

相同。

基本音素（Basic phonemes, BPH）。此音素集通常用于拉萨藏语语音识别，其中包含五十个音素，如图3-2(a)所示。在该表中我们同时给出了与藏语对应的国际音标（International Phonetic Alphabet, IPA）。

Tibetan	IPA	Tibetan	IPA	Tibetan	IPA
ཀ་ཁ་ག་ང་	c	ཝ	w	ི+ར/ལ/ཤ	i:
ཅ་ཆ	c^h	ཟ་འ་ཡ	ts^h	ི+ད/ས	i?
ཏ་ཐ་ད་ན་	h	ར་ཤ་ཟ་ཧ	ŋ	ོ	o
ཨ	j	ཀ་ཁ་ག་ང་ཙ་ཚ་ཛ་ཝ་ཞ་ཟ་འ་ཡ་ར་ལ་ཤ་ས་ཧ་ཨ	ɳ	ོ+ར	o:
ཀ་ཁ་ག་ཨ་ཤ	k	ཨ་ཤ	ɛ	ཱ	u
ཁ་ཤ་ཀ	k^h	ཏ་	ʂ	ཱ+ར/ཨ	u:
ཨ་ལ་ར་ས་ཤ་ཀ་ཨ	l	ཙ་ཛ་ཝ་ཞ	ʦ	ཱ+ལ	y:
མ་ཨ་ཝ་ཤ	m	ཚ་ཛ	$ʦ^h$	ཱ+ད/ས	y?
ན་ཨ་ལ	n	ཅ་ཆ་ཇ་ཉ་ཏ་ཐ་ད	ʨ	ེ+ ན	ø
པ་ཕ་ཐ་ཤ་ཧ་ཨ	p	ཆ་ཇ་ཉ་ཏ	$ʨ^h$	ེ+ ལ	ø:
ཕ་བ་ལ	p^h	ཨ	ʔ	ེ+ད/ས	ø?
ར	r	ཨ+ར	a:	ི+ན	ĩ
ཝ་ས་ཤ་ཟ	s	ཨ	a	ཨ+ན	ɛ
ཏ་ཐ་ཝ་ཟ་ཨ་ཡ་ཤ	t	ི	e	ཨ+ ད/ས	ɛ?
ཐ་ད་ཤ	t^h	ི+ར/ལ	e:	ཨ+ད/ས	ɛ̃
ཀ་ག་ང་ཅ་ཇ་ཉ་ཏ་ད་ན་པ་བ་མ་ཙ་ཛ་ཝ་ཞ་ཟ་འ་ཡ་ར་ལ་ཤ་ས་ཧ་ཨ	ʦ	ི+ད	e?	ེ+ན	ẽ
		ི	i	ཱ+ན	ỹ

(a) 藏语中的基本音素集

Tibetan	IPA	Tibetan	IPA	Tibetan	IPA
ཀ་ཁ་ག་ང་ཤ	c	ང་ཙ་ཛ་ཝ	ŋ	ོ+ར	o:
ཅ་ཆ	c^h	ཀ་ཁ་ག་ང་ཤ་ཧ	ɳ	ཱ	u
ཏ་ཐ་ད་ན་	h	ཀ་ཏ	ɕ	ཱ+ཨorཨ	u:
ཡ	j	ཀ་	ʂ	ཱ+ལ	y:
ཀ་ཏ་ཀ་ཏ་ཐ	k	ཙ་ཚ་ཛ་ཝ	ʦ	ཱ+ཨor+ད	y?
ཁ་ཐ་ཏ	k^h	ཚ་ཛ	$ʦ^h$	ེ+ན	ø
ཨ་ལ་ར་ས་ཤ་ཀ	l	ཅ་ཇ་ཉ་ཤ་ཧ	ʨ	ེ+ལ	ø:
མ་ཨ་ཝ་ཤ	m	ཆ་ཇ་ཉ་ཏ	$ʨ^h$	ེ+ ད+ར+ལ	ø?
ན་ཨ་ལ	n	ད	ʾ	ི+ན	ĩ
པ་ཕ་ཐ་ཤ་ཧ་ཨ	p	ཨ+ར	a:	ཨ+ན	ɛ
ཕ་བ་ལ	p^h	ཨ	a	ཨ+ལ	ɛ:
ར	r	ི	e	ཨ+ ད+ར+ལ	ɛ?
ས་ཝ་ཤ་ཟ	s	ི+ར+ཨ+ལ	e:	ི+ན	ɛ̃
ཏ་ཐ་ཝ་ཟ་ཨ་ཡ་ཤ	t	ི+ད+ར+ལ	e?	ཱ+ན	ỹ
ཐ་ད་ཤ	t^h	ི	i	+ཁ	k kz
ཀ་ག་ང་ཅ་ཇ་ཉ་ཏ་ད་ན་པ་བ་མ་ཙ་ཛ་ཝ་ཞ་ཟ་འ་ཡ་ར་ལ་ཤ་ས་ཧ་ཨ	ʦ	ི+ར+ཨ+ལ	i:	+མ	mmz
ཁ་ག་ཟ	$ʦ^h$	ི+ད+ར+ལ	i?	+ད	n nz
ཝ	w	ོ	o	+བ	p pz

(b) 区分韵尾的辅音的音素集

图 3-2　藏语中基本音素集和区分韵尾的辅音的音素集

区别性辅音韵尾（Distinguishing consonants suffixes, CTS）。为了减少具有相同形式不同发音的后缀的识别错误，我们尝试在元音后区分四个后缀。然后定义了比BPH集合更多的四个音素，它们是“ kz”，“ mz”，“ nz”和“ pz”，如图3-2(b)。

无长音（Changing long vowel, LV）随着藏语拉萨方言的发展，大多数长音（如/ o：/）被更改为发音短音（如/ o /）。然后，我们根据藏语基本音素删除/ y：/以外的所有长音素，如图3-3(a)。

区分韵尾的辅音没有长音（Consonants' suffix and long vowel, CTL）。在这里，我们区分了辅音的后缀，并同时删除了/ y：/以外的长元音。图3-3(b)列出了这种情况的所有音素。

无鼻化元音（Change Nasal vowel, NV）。我们尝试将鼻元音简化为未强调的元音，因为在藏语拉萨方言中，元音的发音就像是轻声。这意味着将七个鼻元音(“e?”，“i ?”，“y ?”，“ø?”，“ĩ”，“ɛ?”，“ɛ̃”)融合到其相应的短元音中。如图3-4中，音素集变为43个音素。

Tibetan	IPA	Tibetan	IPA	Tibetan	IPA
ཀ་ཁ་ག་ང་	c	ང་ཅ་ཆ་ཇ	ŋ	ོ+ར།	o:
ཅ་ཆ	cʰ	ཀ་ཁ་ག་ང་ཅ་ཆ་ཇ་	ɳ	ཨ	u
ཏ་ཐ་ད་ན	h	ཁ་ག	ɕ	ཨ+ར་ཀ།	u:
ཨ	j	ཏ་	ʂ	ཨ+ལ།	y:
ཀ་ཁ་ག་ང་ཅ	k	ཙ་ཚ་ཛ་ཝ	ʦ	ཨ+ས་or་།	yˀ
ཁ་ག་ཀ	kʰ	ཚ་ཛ་	ʦʰ	ོ+ན།	ø
ཆ་ལ་ཤ་ས་ཧ་ཀ་ཅ	l	ཅ་ཏ་ཐ་ད་ན་པ་ཕ	ʨ	ོ+ལ།	ø:
ཨ་མ་ཡ་ཝ	m	ཆ་ཇ་ཉ་ཏ་	ʨʰ	ོ+ད་or་	øˀ
ན་ཅ་ཝ	n	ཨ	ˀ	ི+ན།	ĩ
བ་ཏ་ཐ་ཚ་ཛ་ཝ་ཟ	p	ཨ+ར།	a:	ཨ+ན།	ɛ
ཁ་ཅ་ཏ	pʰ	ཨ	a	ཨ+ལ།	ɛ:
ར	r	ེ	e	ཨ+ད་or་	ɛˀ
ཤ་ས་ཧ་ཀ	s	ེ+ར་or་	e:	ེ+ན།	ẽ
ཏ་ཐ་ཚ་ཏ་ཁ་ཁ་ཏ	t	ེ+ད་or་	eˀ	ཨ+ན།	ỹ
ཐ་ད་ན	tʰ	ི	i	+ག	k kz
ཀ་ཀ་ང་ཁ་ཁ་ཁ་ཁ་ཏ་	ʦ	ེ+ར་or་	i:	+མ	mmz
ཁ་ཀ་ཅ	ʦʰ	ེ+ད་or་	iˀ	+ང།	n nz
ཝ	w	ོ	o	+བ	p pz

(a) 无长音的音素集

Tibetan	IPA	Tibetan	IPA	Tibetan	IPA
ཀ་ཁ་ག་ང་ཅ	c	ཁ་ཀ་ཅ	ʦʰ	ོ	o
ཅ་ཆ	cʰ	ཝ	w	ཨ	u
ཏ་ཐ་ད་ན	h	ང་ཅ་ཇ་ཉ	ŋ	ཨ+ལ།	y:
ཨ	j	ཀ་ཁ་ག་ང་ཅ་ཆ་ཇ་	ɳ	ཨ+ས་or་།	yˀ
ཀ་ཁ་ག་ང་ཅ	k	ཁ་ག	ɕ	ོ+ན།	ø
ཁ་ཁ་ཀ	kʰ	ཏ་	ʂ	ོ+ད་or་	øˀ
ཆ་ལ་ཤ་ས་ཧ་ཀ	l	ཙ་ཚ་ཛ་ཝ	ʦ	ི+ན།	ĩ
ཨ་མ་ཡ་ཝ	m	ཆ་ཇ	ʦʰ	ཨ+ན།	ɛ
ན་ཅ་ཝ	n	ཅ་ཏ་ཐ་ད་ན་པ་ཕ	ʨ	ཨ+ད་or་	ɛˀ
བ་ཏ་ཐ་ཚ་ཛ་ཝ་ཟ	p	ཆ་ཇ་ཉ་ཏ་	ʨʰ	ེ+ན།	ẽ
ཁ་ཅ་ཏ	pʰ	ཨ	ˀ	ཨ+ན།	ỹ
ར	r	ཨ	a	+ག	k kz
ཤ་ས་ཧ་ཀ	s	ེ	e	+མ	mmz
ཏ་ཐ་ཚ་ཏ་ཁ་ཁ་ཏ	t	ེ+ས་or་	eˀ	+ང།	n nz
ཐ་ད་ན	tʰ	ི	i	+བ	p pz
ཀ་ཀ་ང་ཁ་ཁ་ཁ་ཁ་ཏ་	ʦ	ི+ད་or་	iˀ		

(b) 区分韵尾的辅音没有长音的音素集

图 3-3 藏语中无长音和区分韵尾的辅音没有长音的音素集

元音和韵尾合成一个音素（Combining vowels and consonants, VC）。由于设置的音素越少，声学模型的混乱程度就越低，而语言模型的混乱程度就越大。因此，我们需要在音素数量和字符音素区分之间进行权衡。在这种情况下，我们增加了19个由元音和辅音组合而成的音素，如图3-5所示。

3.1.2.4 实验验证

我们对藏语语料库进行音素识别实验，如下所述。实验分别使用GMM-HMM和DNN-HMM来基于不同音素集训练藏语声学模型。这两个实验都使用了从音素提示中得到的音素上的双轨语言模型。所有实验都是使用Kaldi（一种用于语音识别的免费开放源代码工具包）进行的[75]。

根据表3-1构建的音频语料库，实验中GMM-HMM参考系统是使用常规的13维MFCC特征以及一阶和二阶导数构建的。涉及到倒谱均值和方差归一化，并且通过线性判别分析（LDA）和最大似然线性变换（MLLT）将拼接9帧衍生的特征转换为40维。

在我们的实验性DNN设置中，有六个隐藏层，每层2048个单位。一个中心框架5个串联的先前帧和5帧后的背面形成11帧输入。通过预训练RBM[76]初始化DNN的权重，使用对比发散算法以分层的贪婪方式训练RBM。在预训练阶段之后，已经使用SGD算法进行了判别训练，标记的数据来自GMM-HMM系统。要

Tibetan	IPA	Tibetan	IPA	Tibetan	IPA
ཀ་ཁ་ག་ང་ཤ	c	ཐ་ད་ན	t^h	ཨ	a
ཆ་ཇ	c^h	ཏ་ཐ་ད་ན་པ་ཕ་བ་མ་ཙ་ཚ་ཛ་ཝ་ཟ་འ་ཡ་ར་ལ་ཤ་ས་ཧ་ཨ	ʦ	ེ	e
ཧ་ཧྲ་ཧྭ་ཧྲྭ	h	ཨ	w	ེ+ར or ཡ	e:
ཡ	j	ཇ་ཀྱ་ཁ	$ʦ^h$	ི	i
ཀ་ཁ་ག་ང་ཤ	k	ད་ངོ་ཚུ་ཞ	ŋ	ི+ར or ཡ or ཝ	i:
ཁ་ཀྲ་ཀྱ	k^h	ཨ་ཁྲ་རྐྱ་ཁྲྭ་རྒྱ	ɳ	ོ	o
ཨ་ལ་ལ་ཧ་ཤ་ཀྵ་ཪ	l	ཚ་ཤ	ɕ	ོ+ར	o:
མ་ཨ་ཨྱ་ཤྱ	m	ཏ་	ʂ	ུ	u
ན་ཨ་ཤྱ	n	ཅ་ཙ་ཇ་ཤྱ	ʦ	ུ+ར or ཪ	u:
པ་ཏ་ཐ་ཕ་བ་མ་ཡ	p	ཚ་ཁྱ	$ʦ^h$	ུ+ཡ	y:
ཕ་བ་ཤྱ	p^h	ཅ་ཆ་ཇ་ཉ་ཤྱ	ʨ	ོ+ན	ø
ར	r	ཚ་ཇ་ཐྱ་ཏྱ	$ʨ^h$	ོ+ལ	ø:
ཟ་ས་ཤ་ཧ	s	ཧ	ʔ	ེ+ན	ɛ
ཏ་ཏྲ་ཅ་ད་ཐ་ཙ་ཐྲ	t	ཨ+ར	a:	ེ+ལ	ɛ:
				ུ+ན	ỹ

图 3-4　　无鼻化元音的音素集

Tibetan	IPA	Tibetan	IPA	Tibetan	IPA	Tibetan	IPA	Tibetan	IPA
ཨ/ེ+ག	ak	ི+ག	ik	ོ+ང	on	ེ+མ	em	ོ+ག	ok
ཨ+མ	am	ི+མ	im	ོ+ཀ	op	ེ+ང	en	ོ+མ	om
ཨ+ང	an	ི+ང	in	ུ+ག	uk	ེ+ཀ	ep	ུ+ང	un
ཨ+བ	ap	ི+ཀ	ip	ུ+མ	um	ུ+ཀ	up		

图 3-5　　元音和韵尾合成一个音素中增加的19个音素

注意的是，说话人归一化也已使用特征空间最大似然线性回归（fMLLR）算法进行[77]。

为了评估DNN模型在实践中的表现，我们进行了9倍交叉验证。9次的结果示于表3-2。从表3-2可以看出，不同音素组之间通常没有太大的区别。无论应用范式GMM-HMM还是范式DNN-HMM，使用CTL音素集的AM建模都比使用其他五个音素集的AM建模效果更好。带有LV音素的声学模型表现最差。具体而言，在用于拉萨方言藏语语音识别的DNN-HMM范例中，使用CTL音素集的声学模型的音素错误率比使用BPH的音素错误率低2.46%，2.01%，2.19%，2.83%，0.55%，分别设置CT，NV，LV和VC。从CTL集的定义和表中可以看出，它使用的变化包括CT和LV中的变化。根据结果，它表明区分元音的四个辅音后缀很重要，而长元音期望/ y：/是没有用的。

表 3-2 HMM- GMM和HMM-DNN的音素识别结果

Model \ PER		BPH	CTS	CTL	NV	LV	VC
GMM - HMM	Mon	49.1	48.9	47.1	47.6	48.1	50.1
	Tri	30.4	29.9	28.8	30.8	30.6	31.2
DNN - HMM		23.5	23.1	21.1	23.3	23.9	21.6

此外，从表3-2中我们看出，在使用DNN时，音素编号在藏文语音识别中并不重要，因为带有VC的声学模型包含的音素数量最多，效果最好。但是，音素的数量确实影响了传统GMM-HMM的语音识别结果，因为在这种情况下CV表现最差。最后，如我们所料，结果表明，对于声学模型，范式DNN-HMM优于传统的GMM-HMM。

为了获得更好的藏语拉萨方言ASR准确性，对于6种不同音素集，研究使用DNN方法进行藏语声学建模的性能。确定这些不同音素集时，应考虑到藏语拉萨方言的语言知识和语音特征。使用二元语言模型，藏语语音识别结果表明，CTL的最佳音素集比基本音素集的准确度高出10.43%。这表明用于区分元音的四个辅音后缀很重要，而长元音期望/ y：/是没有用的。而且我们还观察到，在使用DNN时，音素号码在藏文语音识别中的作用并不重要。将来，我们将进一步调查CTL集表现良好的原因。我们还将建立拉萨藏语的语言模型，并根据基于不同音素集的声学模型观察语音识别性能。

3.2 藏语文本数据库构建和测试

3.2.1 藏语文本数据库构建

对于文本语料我们参照Mikolov等人构造的数据库，从网上抽取以藏语新闻为主的语料并进行预处理，其中包含了新闻、文化、教育和宗教等主题。在实验中，我们将Kneser-Ney平滑3-gram为KN3[78]。根据藏语的词汇频率特征，词汇表为2472个，其他未在词汇表中的都是以OOV符号替换表示。

在对所爬取的数据去除网上标签的噪声后，根据藏语的分句符对句子进行了

表 3-3　藏语语料数据

数据		#字数	%OOV
字典	-	2472	-
部件集合	-	57	-
组合部件集合	-	122	-
训练集	新闻	1.5m	1.08
	教育	1.2m	1.28
	法律	1.1m	1.35
	宗教	1.4m	1.61
	文化	1.2m	1.15
	文学	3.8m	1.35
	维基百科	11.1m	2.55
	全域	21.3m	1.48
验证集		125k	1.12
测试集		126k	1.11

分句。为了处理方便，我们把藏语中字(Character)之间分割符（“ tsheg”）替换为空格。经过去噪处理后，语料库分为25个部分（10：1：1）[5,22]。0-20部分是训练集（Train set），21-22和23-24部分分别是有效集（Valid set）和测试集（Test set），我们将数据库匿名为藏语新闻数据库（Tibetan News Corpus, TNC）。表3-3中是我们数据库的分布情况，其中基于字典的是按2472个字进行训练和测试。

表3-3中还有单个部件（Radical）和组合基字的统计个数，这些是为了方便以后进行结果对比。因为我们的测试集是以新闻为主题的，所以不同主题的数据会有不同的结果，由此，我们将训练集划分为新闻、教育、法律、宗教、文化和维基百科（因为数据主题混合）。以便于验证不同主题对模型的影响，以及我们提出的方法在不同主题上的实验结果。

为了验证数据库构建是否符合研究需要，我们做了一些预实验来验证。我们根据拉萨方言的特点，提出了一种新的音素集，这种音素集中包含了所有拉萨方言的音素，而且在我们提出的音频数据集上进行了验证。为了验证文本数据集符合我们的需求与否，我们提出了基于组合基字的语言模型，通过在文本数据集上

的验证，对传统的方法和现有的方法都进行了对比实验。

3.2.2 基于形态结构的组合基字藏语语言模型的测试

语言模型是自然语言处理（Natural Language Processing, NLP）和人工智能中一项基础性工作，在语音识别、机器翻译和手写体识别等领域广泛应用[4–6,79]。藏语作为低资源语言，数据稀疏是现存的一个问题，再加上藏语语言模型研究起步较晚，目前，主要是以字为粒度研究，基于N-gram的方法居多[37,65]。

近些年随着深度学习的发展，出现对藏语语言模型的研究。申彤彤等人根据形态结构出发，提出了基于循环神经网络语言模型(RNN)的藏语字符单元[80]研究方法，在一定意义上缓解了数据稀疏的问题，而且这种方法比传统的方法取得了较好的效果[29]。然而，这种方法还存在一些问题：1）首先，这种方法是基于字典，可以处理未登录词（Out-Of-Vocabulary, OOV）和罕见词（Rare Words）问题；2）其次，我们知道循环神经网络语言模型获取的是序列信息，但是在获取字信息时无法获取全局信息。基于这些问题，我们提出基于形态结构的组合基字的藏语语言模型。因为卷积神经网络(Convolutional Neural Networks, CNN)能够获取字的全局信息，所以我们利用CNN对藏文字进行了字的全局特征提取。

藏语在拼写字时具有一定的拼写规则。藏语字具有固定的搭配，例如：上加字和基字、上加字，基字和下加字、基字和下加字，以及一些特殊字符组合。在实际应用中，华光、同元、班智达和方正等输入法在早期的编码研究中也利用了藏语拼写中的这种组合[81]。这种方法的优势在于，它可以更加准确地获取到一些字中组合基字的信息，不需要判断每个字符是否是上加字、下加字和基字等信息。

在本研究中，我们根据藏语拼写规则并参照文章[5]的研究方法，将藏语字符作为输入进行卷积，发现一些字符的组合是固定形成的，不需要学习其规则，而且基字会影响字结构。因此我们提出了组合基字（Combination Tibetan Radical, CTR）的卷积神经网络（CNN）方法，这种方法可以确定字中基辅音，更加准确学习出其前缀和后缀等其它的全局信息。我们的方法比之前工作有一个优点就是不需要词典，我们知道之前的研究中都是基于字典来研究的，未在语料中出现的词都是以〈UNK〉符号来替换。而我们的方法在没有字典情况下解决了一些未登录词和稀有词问题，可以提高预测能力。

对于形态丰富的语言，它们主要研究单词的内部结构形成[32,82–84]。以英语形态学为例，目的是研究单词构成之间的关系，并梳理其构成规则。语言模型可以

应用这种类型的信息以获得具有可接受准确度的单词序列。

对于低资源语言，稀有词处理是一个常见问题。在[5]中，通过减少传统的单词嵌入级别到字符级别，大规模字符嵌入计算避免了稀有词，而且网络结构比较复杂，但是效果很好。该模型由两部分组成，即字符级作为CNN的输入，并通过输入到循环神经网络（Recurrent Neural Network Language Model, RNNLM），最终预测仍然是一个词。这种方法可以通过多次实验来获取一些语义和语法信息。

语言模型需要自适应允许使用多个其它域文本的技术资源[85,86]。使用其它域的文本资源构建语言模型灵活使用领域适应，开发了一种方法用于潜在变量空间中的模型合并[87,88]。在潜在的变量空间里面，一个单词被映射到一个潜在的变量空间，可以期望执行更灵活的状态下观察到的单词空间中分享的可能。正常类中的潜在变量，N-gram语言模型只是模型依赖指数，所以每个模型具有不同的潜在变量空间[89,90]。

3.2.2.1 藏文字符结构

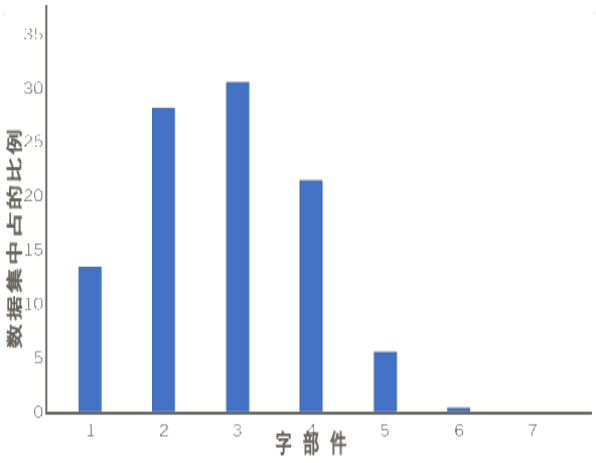

图 3-6　　藏文字符的分布情况

藏语中常用的字的构成一般都不超过7个字符(这里不包含梵文)，例如2-2中是藏语中常见的最长字，是藏语构字的基本结构。

图3-6是对我们的语料进行字符分布统计的结果，可以看出，字中的字符有1-4个组成的占了90%以上。字符为6-7的字占的比重少。这说明不是任意步长的字上都能够获取到有用的信息。所以，需要进一步研究字组合的方式及意义。

藏文字是由字符组合形成的。根据藏文字结构的接续规则，除了30个辅音和4个元音，还有一些上加字（T）、基字（B）和下加字（D）的组合，如上加字ra、

la和sa三种类型所能组合的基字。如图3-7所示，上加字和基字组合（T + B），基字和四个下加字组合（B + D）和上加字、基字和下加字的组合（T + B + D）。这里有一些焚文字符，这些文字在藏语中应用的较多，也将列入其中，以特殊字符（S）来处理。

组合方式	类型
T + B	ཀཁགངཅཆཇཉཏཐདནཔཕབམཙཚཛཝཞཟའཡརལཤསཧཨ ཀྱཁྱགྱངྱཅྱཆྱཇྱཉྱཏྱཐྱདྱནྱཔྱཕྱབྱམྱཙྱཚྱཛྱཝྱཞྱཟྱཡྱརྱལྱསྱཧྱ ཀྲཁྲགྲཅྲཆྲཇྲཏྲཐྲདྲནྲཔྲཕྲབྲམྲཙྲཚྲཛྲཞྲཟྲཤྲསྲཧྲ
B + D	ཀྭཁྭགྭངྭཅྭཆྭཇྭཉྭཏྭཐྭདྭནྭཔྭཕྭབྭམྭཙྭཚྭཛྭཞྭཟྭཡྭརྭལྭསྭཧྭ ཀྭྱཁྭྱགྭྱངྭྱཅྭྱཆྭྱཇྭྱཉྭྱཏྭྱཐྭྱདྭྱནྭྱཔྭྱཕྭྱབྭྱམྭྱཙྭྱཚྭྱཛྭྱཞྭྱཟྭྱཡྭྱརྭྱལྭྱསྭྱཧྭྱ ཀྭྲཁྭྲགྭྲངྭྲཅྭྲཆྭྲཇྭྲཉྭྲཏྭྲཐྭྲདྭྲནྭྲཔྭྲཕྭྲབྭྲམྭྲཙྭྲཚྭྲཛྭྲཞྭྲཟྭྲཤྲསྲཧྲ
T + B + D	ཀྭྱྲཁྭྱྲགྭྱྲཀྭྲཁྭྲགྭྲ ཅྭྱྲཆྭྱྲཇྭྱྲཉྭྱྲཏྭྱྲཐྭྱྲདྭྱྲནྭྱྲཔྭྱྲཕྭྱྲབྭྱྲམྭྱྲཙྭྱྲཚྭྱྲཛྭྱྲཞྭྱྲཟྭྱྲཤྲསྲཧྲ ཅྭྲཆྭྲཇྭྲཉྭྲཏྭྲཐྭྲདྭྲནྭྲཔྭྲཕྭྲབྭྲམྭྲཙྭྲཚྭྲཛྭྲཞྭྲཟྭྲཤྲསྲཧྲ
S	ཨ་ཨི་ཨུ་ཨེ་ཨོ།

图 3-7　　藏语组合基字规律

3.2.2.2　基于形态结构组合基字的藏语语言模型

对于基于形态结构的组合基字语言模型，研究主要是将字符嵌入作为CNN的输入，用CNN的输出经过一层Highway Network层处理表示字嵌入，然后作为RNNLM的输入。

Char-aware语言模型： CNNs 已经在计算机视觉方面取得了很好的成果[91]，并证明了对自然语言处理任务是有效的[92]。有研究以字符为粒度，以CNN的方法得到词的特征向量，将作为RNN的输入进行了应用，而且取得了好的结果[5]。

CNN: 基于字的语言模型是通过构建一个字将字表示成矩阵，输入到CNN模型中提取经过滤波器层和池化（max pooling）层得到一个输出表示。对于字符卷积，设C为字符的词汇量，d为字符嵌入的维数，$Q \in \mathbb{R}^{dx|c|}$为矩阵字符嵌入。假设单词由一系列字符组成$< c_1 \cdots c_l >$，其中l为单词k的长度，k的字符级表示由矩阵$C^k \in \mathbb{R}^{dxL}$给出，其中第j列对应于的字符嵌入。我们在和C^k宽度为w的滤波器之间应用一个卷积，然后我们添加一个偏置并应用非线性得到一个特征映射$f^k \in$

$\mathbb{R}^{l-w+1}$。具体地说，f^k的第i个元素由:

$$\mathbf{f}^k[i] = tanh(< \mathbf{C}^k[*, i:i+w-1], \mathbf{H} > +b), \tag{3-1}$$

$\mathbf{C}^k[*, i:i+w-1]$是从i到$i+m-1$列。最后我们计算池化（Max-over-time）

$$y^k = \max_i \mathbf{f}^k[i], \tag{3-2}$$

作为对应于过滤器H的特征，捕捉最重要的特征及给定过滤器的值最高的特征。过滤器本质上是选择一个字符N-gram，其中N-gram的大小对应于过滤器的宽度。

Highway Network: Srivastava等人提出的Highway Network[93]，Yoon Kim等人做了改进。而多层感知器（Multi-Layer Perception, MLP）的一层应用仿射变换后的非线性来获得一组新的特征：

$$z = g(Wy + b), \tag{3-3}$$

Highway的一层网络是：

$$z = t \odot g(W_H y + b_H) + (1 - t) \odot y, \tag{3-4}$$

g是非线性的$t = \sigma(W_T y + b_T)$，称为Transform gate，$\Phi 1 - t \Psi$被称为Carry gate。与LSTM网络中的存储单元类似，Highway层允许通过自适应地将输入的某些维度直接携带到输出来训练深度网络。通过构造y和z的维数必须匹配，因此W_T和W_H是方阵。

RNN(LSTM):循环神经网络(Recurrent Neural Network， RNN)是一种特别适合于序列现象建模的神经网络结构。在t时刻，RNN取输入向量和隐藏状态向量，通过循环操作生成下一个隐藏状态：

$$h_t = f(Wx_t + Uh_{t-1} + b), \tag{3-5}$$

从理论上讲，该神经网络可以用隐藏状态h_t概括出t之前的所有历史信息。然而，在实践中，RNN的无法获取长距离信息，因为会出现梯度消失和梯度爆炸问题。长短时记忆(Long Short-Term Memory ，LSTM)通过在RNN中增加一个记忆单元向量来解决学习长期依赖关系的问题：

$$i_t = \sigma(W^i x_t + U^i h_{t-1} + b^i), \tag{3-6}$$

$$f_t = \sigma(W^f x_t + U^f h_{t-1} + b^f), \tag{3-7}$$

$$o_t = \sigma(W^o x_t + U^o h_{t-1} + b^o), \tag{3-8}$$

$$g_t = tanh(W^g x_t + U^g h_{t-1} + b^g), \tag{3-9}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t, \tag{3-10}$$

$$h_t = o_t \odot tanh(c_t), \tag{3-11}$$

LSTM中的记忆是对时间的累加，缓解了梯度消失的问题。

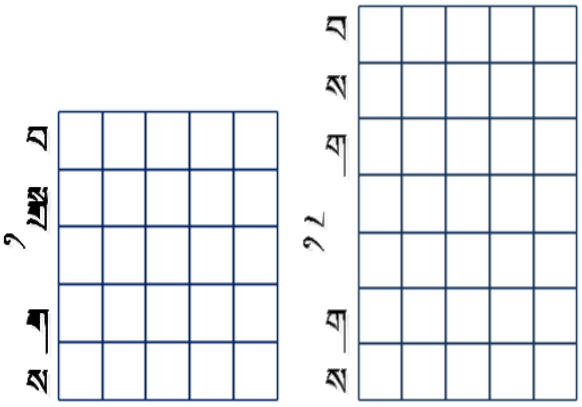

图 3-8　　藏文字符基本切分和组合基字切分

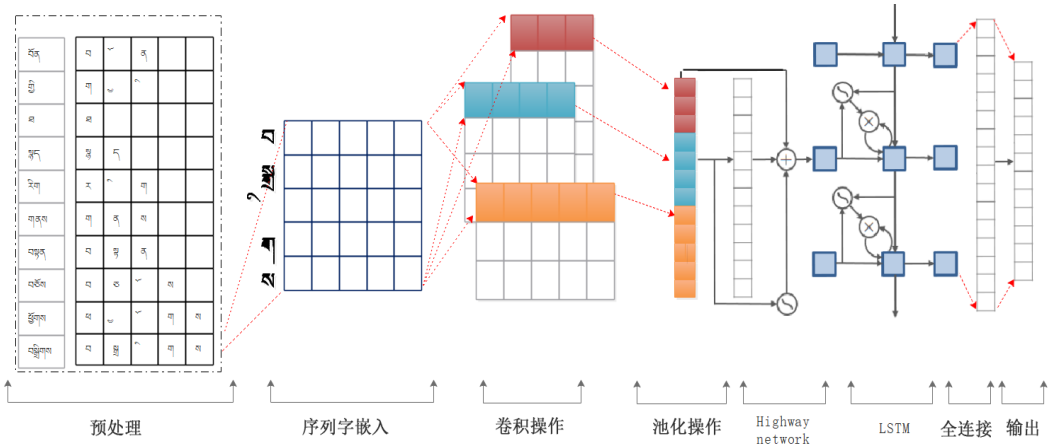

图 3-9　　基于CNN组合基字结构的藏语语言模型

基于形态结构组合基字的藏语语言模型：我们根据藏语的字符特点，首先对语料进行了预处理。基于传统的Char-CNN方法我们提出了基于组合基字卷积神经网络（Combination Tibetan Radical Convolutional Neural Networks, CTR_CNN）方法。如图3-8是藏语字的不同字符的结构，左图是基于组合基字的藏文字，我们可以看出基字很明显，而右图是基于字符的藏文字。可以看出，我们对字符长度为5，而基于全字符的长度为7。我们从图3-6中字符的分布中可以看出，6-7个字符构成的字并不多，这样在字符卷积中会出现参数过大，网络复杂等现象。

我们的方法和传统Char-CNN方法不同的是卷积的字符也不一样。传统方法是将每个字中的字符进行卷积，而这种方法的弊端是将一些规律性的字符组合进行了拆分，导致一些字基字无法确定。因为藏语中基字是每个字的核心，类似于英语中的词根。我们提出的CTR_CNN是将一些具有组合规律型的字符组合到一起，这样做的优势在于一方面，可以直接获取基字信息；另一方面，会更加准确获取字的结构信息，而且会减少模型参数和计算复杂度。

表 3-4　CNN参数设置

CNN	***d***	7
	w	[1,2,3]
	h	[25 · w]
	f	***tanh***
Highway	***l***	1
	g	***Relu***
LSTM	***l***	2
	m	300

图3-9是输入一个句子，对句子中的字进行预处理，通过卷积神经网络（CNN）和Highway Network对组合基字进行处理，输出为基于长短时记忆(LSTM)循环神经网络语言模型(RNN-LM)。我们利用字符级输入的神经语言模型，预测仍然是在文字层面进行的，尽管参数较少，但是我们的模型比字嵌入输入层的基线模型表现得更好。

3.2.2.3　实验结果与分析

语言模型任务中，通常将所有的数据集分成两个部分，训练集和测试集，用来衡量所训练的语言模型的质量。度量一个语言模型的好坏通常使用困惑度（Perplexity, PPL）作为评价指标。

数据：实验中的文本数据是我们构建的表3-3，利用我们提出的方法验证我们数据满足我们研究的要求。表3-4总结了模型的体系结构，我们使用200个隐层，*d*为字符嵌入的维数; *w* 为滤波器宽度; *h*为滤波器矩阵的数量，作为滤波器宽度的函数（宽度为[1，2，3]的滤波器[100，150，200]，总共450个滤波器）; *f*，*g* =非

表 3-5 我们的方法和最新方法进行PPL对比

Granularity	Model	News
Character	N-gram[23]	55.2
	RNNLM[23]	62.9
	CUED_RNNLM[94]	58.4
	CUED_RNNLM(LSTM)[94]	55.9
	Char-CNN[5]	60.7
Radical	Char-CNN[5]	55.2
	_TRU[78]	57.6
	_TRD[78]	54.3
	_TRC[78]	53.8
	CTR_CNN	**52.8**

线性函数; *l*为层数; *m* 为隐层数。选择这些以保持参数的数量类似于相应的字符级模型。

结果与分析：本研究主要是对藏语中的字符进行了重新组合，对这种组合实验中进行了实验验证。表3-5 中分为字和字符两种不同粒度上的结果，基于字的粒度上我们可以看出，传统的N-gram取得了较好的效果，而RNNLM[95]、CUED_RNNLM[26] CUED_RNNLM (LSTM)[94] 和Char-CNN[5]结果较差；而基于字符粒度上的Char-CNN取得了和N-gram相同的结果，_TRD和_TRC[27]结果比其他方法有了提升，而我们提出了组合基字的卷积神经网络（CTR_CNN）方法比其他方法都有提高。

表3-6我们将字粒度和字符粒度研究的结果和KN3插值计算，其中lambda取0.5。实验中，我们将粒度为字研究的方法和我们的方法对比，PPL下降了4.3%-16%左右；以字符为粒度和我们方法对比，PPL下降1.8%-8.3%左右。由此看出，本研究中提出的方法对PPL取得了很好的效果，加上同等条件下和KN3进行插值会出现较好的结果，说明我们提出的CTR_CNN方法是有效的。

综上所述，CTR_CNN 方法是对藏语字中字符的组合规律进行卷积，得出一个更加全面的结构信息，通过Highway Network来获取一个更好的字嵌入，把它作为LSTM的输入，获取了其序列信息的模型。

表 3-6　我们的方法和已有的方法插值以后进行PPL对比

Granularity	Model+KN3	News
Character	N-gram[23]+KN3	55.2
	RNNLM[23]+KN3	48.2
	CUED_RNNLM[94]+KN3	48.0
	CUED_RNNLM(LSTM)[94]+KN3	46.4
	Char-CNN[5]+KN3	48.1
Radical	Char-CNN[5]+KN3	47.3
	_TRU[78]+KN3	47.9
	_TRD[78]+KN3	47.0
	_TRC[78]+KN3	46.9
	CTR_CNN+KN3	**45.9**

在预实验中我们提出了一种藏语字符组合的CTR_CNN方法，模型增强了字符嵌入，并能够解决低资源语言的数据稀疏性问题。因为卷积特征的不同，它比传统基于Char-CNN方法取得了很好的效果。相比基于字典的方法，我们的方法不但可以预测出来字典中出现的字，还可以解决一些未登录字（OOV）和罕见词（Rare Words）。相对于语料丰富的语言，我们通过插值的方法，解决了藏语数据稀疏的问题。通过实验验证，我们提出的CTR_CNN方法具有较好的效果，困惑度比其他方法都（PPL）降低了。

3.3　本章小结

本章首先介绍了藏语音频数据集和文本数据集的构建。为了验证我们构建的数据集符合研究的需要，我们利用音频数据集和文本数据集进行了预实验，首先，对于音频语料的测试，提出了使用DNN方法对拉萨方言声学建模，利用6种不同音素集进行了验证。考虑到藏语拉萨方言的语言知识和语音特征，提出适合拉萨方言的CTL音素集，而且验证了该方法符合我们构建的音频数据库。其次，对于文本语料的测试，我们利用藏语形态结构组合基字的规律性，提出了基于组合基字的藏语语言模型，而且在我们的文本数据集上取得了良好的效果。除了以上研究

应用了此数据集，其他研究[37,38,43,78]也应用了我们的数据集，由此，我们可以验证我们所构建的数据集符合我们研究的需求。

第4章　基于静态形态结构的藏语语言模型

信息时代，信息交流和传递对加快地区经济和社会发展有着重要的意义。语言作为信息传递的载体，对信息发展具有重要作用。当前有很多语言模型研究方法如，N-gram，RNN等，针对的基本都是英语为主的资源丰富语言。而对于低资源语言的研究，基本都是应用现有技术直接移植的方式，并未考虑语言本身的特点。对藏语语言模型的研究，可以促进藏语机器翻译、藏语语音识别等相关领域的进步，进而还可以加快藏族地区的经济发展，促进藏族人民与其他民族和地区人民的沟通与交流，因而具有重要的社会意义和应用价值。

4.1　藏语虚词及相关研究

藏语作为低资源语言[34]，在获取数据方面存在困难，所以有必要在现有的语料中获取更多信息。藏语文字中存在特殊的形态结构关系，每个字由字符（Radical）组成。对于藏语语言模型的研究，利用形态结构的方法比传统的方法效果更好。然而这种方法取得了好的效果，但是这种方法未考虑对虚词的判定，即虚词的错误判定对句子语义的影响。

4.1.1　藏语虚词

藏语虚词不同于名词、动词等实词，为了满足句子不同的表义需求，通常出现在句子中的中、前、后。藏语虚词起着连接实词的作用，主要是连接名词和动词，在不同语境中表达不同的意思。一般情况下在句子中很少独立运用虚词，它通常和名词等实词搭配后才能表达出一定意义。藏族学者吉太加在《现代藏文语法通论》中引用了这么一段话“句子所要表达的意思三分之一是通过其他途径的，如手势和音高的变化来辅助完成”[96]。由此可以看出，生活在高原上的藏民族很早就把肢体语言和口头语言相结合使用。但是，毕竟口语语言是人类相互交流的主要工具，在外部环境和周围文化的相互交流和学习中，藏语形成具有独特语法

的语言，由此，产生了具有严谨的虚词语义关系。

公元七世纪，藏王松赞干布时期的吞米•桑布扎创造了文字，构建出了藏语语法理论，形成了研究藏语语法的重要基础，为以后更好地继承和弘扬民族文化做出了积极的贡献。自藏语语法理论创立以来，相继有很多研究藏语语法的论著不断问世，注解和分析了语法中深奥理论知识，为之后的研究者提供了很好的作用。对于藏语语法的研究，我们是以《三十颂》和《音势论》为主要研究对象进行研究。

《三十颂》共120句，以四句为一颂，全文正好30颂。主要论述了音节结构、正字法和虚词。其内容包含了四个部分：第一部分是介绍字母的分类和文字结构；第二部分是介绍格助词与虚词的应用；第三部分是介绍后加字及其缀联形式的重要性；第四部分是讲述语法理论的重要性[97]。《音势论》核心内容是字母的语音分类和动词。大致可分三大类：第一部分是字性分类，按发音方法将藏文字母分阴性、阳性和中性；第二部分是前加字和后加字的字性分类和约束规则；第三部分论述字性分类和正确缀联的重要性。

序号	藏语句子
A	སྒྲུ་མོས་དེའི་ཚར་ཡང་ཡང་ཤུ་བཞིན་འདུག
B	དེའི་ཚར་སྒྲུ་མོས་ཡང་ཡང་ཤུ་བཞིན་འདུག
C	李毛（人名）正在反反复复看书。
D	#书正在反反复复看李毛

图 4-1　藏语句子中虚词的作用

表中可以看出，句子A和句子B在藏语中都是一个意思，由于虚词在句子中充当接续关系，颠倒词语的顺序是不影响句子的语义。而句子C 和句子D是对照藏语句子A，B直接翻译的，可以看出，句子C能够准确表达原句的语义，而句子D不是一句正常的汉语句子，一般汉语中不会出现这样的句子，跟藏语句子的语义相差甚远。由此，可以看出，藏语虚词在句子中的作用，具有独特的连接词与词的作用。

藏语虚词（phrad）主要起连接作用，根据语境的不同所表达的意思也不同，通常和实词搭配才能表示一定的意义。并根据接续关系产生不同的含义。虚词本身不具备独立表意的功能，通过连接词语前后来表示特殊的语义功能。词语之间的搭配更加合适或更能够表示其意义（还有一些组合会对词性发生变化）[39–41]。虚词可以分为自由虚词（rang dbang can）和非自由虚词（gzhan dbang can）。

非自由虚词根据前一个字符的后缀或隐式的后缀确定接续关系。用于完结助

词（一类虚词）“ go”，“ ngo”，“ do”，“ no”，“ bo”，“ mo”，“ vo”，“ ro”，“ lo”，“ so”，“ to”，这类字后缀必须与虚词的词根相同（以便可以接续字和虚词）。对于属格助词（另一类虚词）“ gi”，“ kyi”，“ gyi”，“ vi”，“ yi”，它们根据的接续规则是（“ da”，“ ba”，“ sa”，“ kyi”；“ ga”，“ nga”，“ gi”；“ na”，“ ma”，“ ra”，“ la”，“ gyi”）[33]。换句话说，它是具有某些语法规则的接续关系。

自由虚词包括后缀和隐式后缀“ v”和“ da”。当没有后缀时，我们将“ v”和“ da”视为隐式后缀。但是，它们不是固定的。例如，与非自由词相比，“dang”，“nas”，“las”，“ma”，“mi”，“ni”具有相对自由的组合。它们根据句子的结构选择虚词。因此，值得注意的是，上述虚词并不完整，只是[39,40,57]中虚词的一部分，详见附件1 。

4.1.2 现有研究中问题及贡献

对于形态丰富的语言的研究，利用形态结构可以获取结构信息[32,82–84]。以英语形态学为例，其目的是研究单词构词方法，并理清其构词规则。然而，在之前的研究中，忽略了语法和语义信息在句子中的作用。语言模型是可以应用词的信息来获取准确的单词序列。

为了从不同层次获取更多信息，一些研究人员提出了许多结合词和字符信息的语言模型。其中，[98,99]提出了一种将词嵌入级别和字符嵌入级别模型知识相结合的新模型，该模型应用了选择词嵌入级别或字符嵌入级别的门控机制来获得词向量。该模型不仅利用字典中单词的信息，而且通过字符嵌入来利用低频和未登录单词的信息。实际上，该模型还包括其他级别的信息，例如词尾，前缀，后缀，以及包括单词级别和字符级别信息的形态结构形式。

藏语作为一种形态丰富的语言[33]，也具有静态形态结构信息，其中一个字最多包含七个字符（Radical）[80]。前缀和词根的字符对字或单词的含义有很大影响，而后缀会影响虚词，从而丰富藏语句子的语法规则和语义信息[39–41,97]。我们对于藏语语料的统计发现，藏语虚词在语料中的比重较大，而虚词对句子的语义有影响。虽然字和字符组合的方法补充了藏语语料不足，但是虚词对句子的语义影响很重要，而判定虚词的关键在于后缀。因为在藏语语法中有对于后缀的判定虚词的语法理论，由此，我们提出了基于显式后缀（TRSU-E）判定虚词的方法，解决了显式后缀在句子中非自由虚词的判定错误的问题。

我们进一步发现藏文中有一些后缀是隐含的字，这些隐含后缀经常存在于数

据中。由于书写方便省略了，但是这些隐式后缀将影响自由功能字（包括隐式“v”和“da”）的接续关系。因此，我们提出了一个藏语隐式后缀（TRSU-I）来判定虚词，它不仅可以解决显式后缀对非自由虚词的影响，而且还可以解决隐式后缀对自由虚词的影响，使句子的语义更准确。在藏语中，虚词是由后缀来判断的，后缀与虚词之间存在语法关系[39–41,97]。

本研究工作贡献概括如下：

1) 藏语句子中的虚词起着重要作用，而接续虚词的关键是后缀。据我们所知，这是首次提出后缀信息对虚词的影响藏语语言模型。
2) 根据藏语的形态结构关系，我们使用RNN方法基于字符信息的获取来集成藏语后缀信息。提出了TRSU-E和TRSU-I方法。 TRSU-E考虑显式后缀对虚词的影响，TRSU-I考虑了隐式后缀的影响。我们的方法可以考虑假想词在藏语中的接续关系。这使我们的方法能够解决句子中词汇词的连续错误的问题，并能够准确地表达句子的语义。
3) 作为一种低资源语言，藏语资源一直是人们关注的问题。我们的方法可以在语料有限的情况下获得更多的形态结构信息。

4.2 藏语后缀对虚词的影响

每种语言都有其自己的语法体系。藏语也不例外，是一种具有自己的拼写规则的字母书写文字。与英语相比，它在构词方式上有一些差异。英语用空格来分隔，而藏语需要进行分词。到目前为止，由于尚无标准的藏语分词准则，因此我们将字作为单元进行了切分。在本研究中，藏语字需要进行切分，字是指藏语分隔符切分的每个单元，而字符（Radical）则是组成字的各个构件。藏语有一套自己的语法系统[39–41]。藏语中最常见的句子与虚词是分不开的，虚词用于将词与词或短语联系起来以表达语义。

4.2.1 藏语字符形态结构

藏文字是由藏文字符组成的，由一个或几个字符拼写而成。藏文字通常由一到七个字符组成。图2-2显示了藏文字的每个字符的位置和名称。根（图中左侧的底部）是藏文字中必不可少的部分，而其他部分则取决于各个字符。通常，前缀与字根跟相邻的字没有直接关系，但是，后缀与虚词具有相关性，其中后缀接续

虚词会影响句子的语义。后缀可以包含后加字和后后加字，后后加字优先级高于后加字。

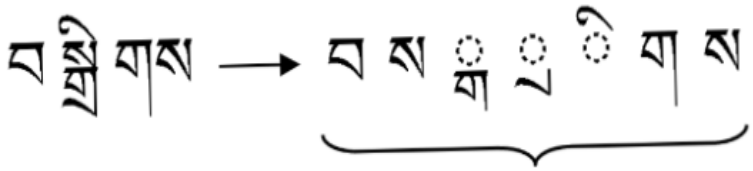

图 4-2　　藏语字的切分

图4-2 是藏语中一个字常用的字符（Radical）结构拆分，可以看出，我们把它分为七个部分，对应的分别是前加字、上加字、基字、下加字、元音、后加字和后后加字。每个字符都对字本身有一定的影响，所以，对于字本身，每个字符很重要。

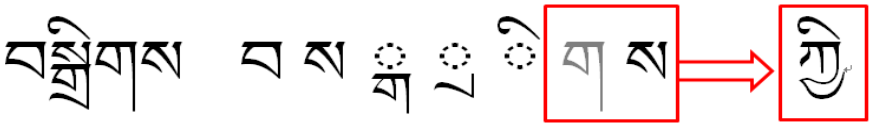

图 4-3　　藏语字后缀接续虚词

后缀主要影响虚词，起到连接的作用，并根据不同的上下文环境表达不同的语义。虽然虚词没有独立的表意功能，但它们可以通过将附近的词连接起来表达特殊的语义功能。根据藏文字符中不同的后缀（即，后缀出现在虚词的前面）来添加虚词。如图4-3是藏文字中后缀接续虚词的例子，左边红框中是后缀，包含了后加字和后后加字，箭头后的是后缀相应接续的虚词。藏语有两种形式的后缀，即后加字和后后加字，它们都包含在后缀中（“ ga”，“ nga”，“ da”，“ na”，“ ba”，“ ma”，“ v”，“ ra”，“ la”，“ sa”）[100]。

4.2.2　后缀的作用以及语义影响

图4-4是一些词拼接和添加虚词的对比。在表中的词A和C中，由于没有虚词，因此表达的含义不明确，句子的含义没有确切的解释。句子B和D中由于添加了虚词，句子可以流畅地表达含义。每个虚词在不同程度上增强了表意效果。例如，藏文“ nyi ma”的意思是太阳，接续不同虚词语义将发生变化。例如，句子E中虚

ID	Tibetan language/Transliteration in Latin	English translation
A	ལྷ་སა། འགྲོ། Lhasa , vkro	Lhasa, go/walk/take
B	ལྷ་སར་ འགྲོ Lhasar vkro	Go to Lhasa
C	ཁོ། སྒྲོལ་མ། གྲོགས་པོ། ཡིན kho, Sgrol Ma, grogs po, yin	He, Sgrol Ma, friend
D	ཁོ་ ནི་ སྒྲོལ་མའི་ གྲོགས་པོ་ ཞིག་ ཡིན་ ནོ། Kho ni Sgrol Mvi rogs po zhig yin no	He's a friend of Sgrol Ma.
E	ཉི་མའི་ འོད་ཟེར། Nyi Mvi vod zer	Sunlight
F	ཉི་མ་ ལ་ བཤད། Nyi Ma la bshad	Tell Nga Ma (Names).

图 4-4　　词拼接和添加虚词对句子语义的影响

Example	Type of functional words	Semantic
དཔེ་སྒྲོམ་ གྱི་ དུད་རྡུལ་ ལེགས་པར་ ཕྱིས་ སོ dpe sgrom gyi dud rdul legs par phyis so	End words(རྫོགས་ཚིག rdzogs tshig)	Tell others the bookcase has been cleaned up.
དཔེ་སྒྲོམ་ གྱི་ དུད་རྡུལ་ ལེགས་པར་ ཕྱིས་ སམ dpe sgrom gyi dud rdul legs par phyis sam	Detachable words(འབྱེད་སྡུད། vbyed sdus)	Ask others whether the bookcase clean up?
དཔེ་སྒྲོམ་ གྱི་ དུད་རྡུལ་ ལེགས་པར་ ཕྱིས་ ཤིག dpe sgrom gyi dud rdul legs par phyis shig	Indication function(ཚིག་ཕྱད་ tshig phyad)	Instruct others to clean up the bookcase!
དཔེ་སྒྲོམ་ གྱི་ དུད་རྡུལ་ ལེགས་པར ཕྱིས་ ཀྱང dpe sgrom gyi dud rdul legs par phyis kyang	Decorative words(རྒྱན་སྡུད། rgyan sdud)	Add the opposite sentence
དཔེ་སྒྲོམ་ གྱི་ དུད་རྡུལ་ ལེགས་པར་ ཕྱིས་ ཤིང dpe sgrom gyi dud rdul legs par phyis shing	Juxtaposition function(ཚིག་ཕྱད་ tshig phyad)	Play the role of juxtaposition, extend the next sentence.
དཔེ་སྒྲོམ་ གྱི་ དུད་རྡུལ་ ལེགས་པར་ ཕྱིས་ ཏེ dpe sgrom gyi dud rdul legs par phyis te	Pending words(ལྷག་བཅས lhag bcas)	Has the function of supplementing sentences.

图 4-5　　藏语后缀对显示虚词的影响导致语义信息的变化

词“ vi”的接续表示太阳“ vod zer”是光的意思，而句子F中虚词“ la”的延续语义发生了变化，其中“ nyi ma”表示一个人的名字（西藏名字是图腾，宗教，寄托和吉祥）。后缀虚词的添加可以增强句子的顺序关系，使句子更平滑并强调语义功能。

在实际情况中，我们知道不同的虚词对句子的语义影响不同。根据接续的关系不同，可以使用上下文和语境来判断虚词，以便选择能更好地表达句子含义的词。在图4-5中，相同的句子表达了不同的语义，因为它们后面接续不同的虚词。图4-6中列出的目的是表明同一句子中不同虚词的接续虚词存在语义差异。

图4-5中的“dpe sgrom gyi dud rdul legs par phyis”表示“清洁书架”。可以看出，不同虚词的语义效果是不同的。在明确存在后缀的情况下，可以基于后缀确定虚词。相同的后缀可以具有多种接续关系，因此有必要使用多个候选词来解决此类虚词。

Example	Type of functional words	Semantic
གྲོགས་པོ་(ཉ) ཀྱི་ སེབས་བྱུང་ groks po(v) ni phebs byung	Nominative function (རྩོལ་བྱེད)	Friends are coming
གྲོགས་པོ་(ཉ) ་ལ་ བཤད། groks po(v) la bshad	Dative word (ཉེར་གཞི)	Tell a friend
གྲོགས་པོ་(ཉ) ཡིས་ བྲིས། groks po(v) yis bris	Instrumental function (ཉེར་ཡ་པོ)	Written by a friend
གྲོགས་པོ་(ཉ) ་ལ་ བཟོ། groks po(v) la bzo	Dative word (ཕན་འདོགས་ཡུལ)	Do it for friends (clothes, etc.)
གྲོགས་པོ་(ཉ) · ལས་ སྐྱངས། groks po(v) las blangs	Ablative function (འཕྲོ་ཁུངས)	Taken from a friend
གྲོགས་པོ་(ཉ) · ་ལ་ ཡོད། groks po(v) la yod	Lovative function (གནས་གཞི)	(something) friends have

图 4-6 藏语后缀对隐式虚词的影响导致语义信息的变化

在图4-6中“groks po”中的“ v”没有明确的后缀，并且会有不同的虚词。此外，将不同的虚词应用于后续字，并且会发生语义变化，藏语中有多种类型的虚词，其比例很大。在藏文的实际应用中，这种字符后缀是隐含的。因此，出于这个原因，需要获得有关此类隐式后缀的信息以弥补显式后缀的不足，我们还对虚词进行了统计分析。本研究基于藏文语法理论，提出了一种考虑虚词关系的藏语语言模型。

4.3 考虑后缀的藏语建模

神经网络语言模型是当前流行的研究趋势。在一些资源丰富的语言中被研究和应用，例如英语，法语和中文等。在我们的研究中，我们发现语言除了共性之外，每种语言都有自己的特征，这是由于语言的差异所致。藏语也不例外，可以将语言中的虚词特征加到藏语语言模型中，以更准确地确定下一个字或单词。

4.3.1 标准的RNNLM

统计语言模型给出一系列单词，然后根据其概率值预测可能的句子。例如，RNNLM使用时间信息来保存上下文以获得隐藏层信息。标准循环神经网络的结构如图4-7所示。

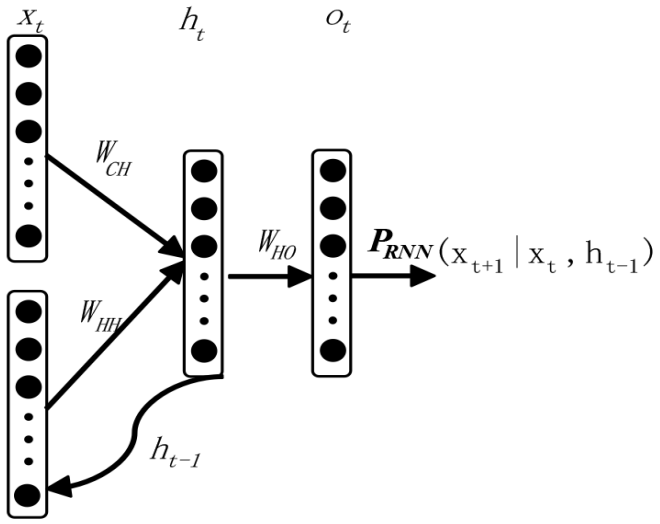

图 4-7　　循环神经网络语言模型的结构

在此，x_t表示时间t的输入层，它使用大小为V_{word}的词向量对当前字符w_t进行编码。用h_t表示的隐藏层通过使用激活函数来保留剩余上下文信息。RNNLM的输出层是o_t，这是在给定由softmax函数产生的历史字符序列$< w_t \cdots w_1 >$的情况下，词汇在时间$t+1$处每个词的概率。RNNLM传播过程中的输入层，隐藏层和输出层可以按以下方式计算：

$$h_t = f(W_{IH}x_t + W_{HH}h_{t-1}) \tag{4-1}$$

$$o_t = g(W_{HO}h_t) \tag{4-2}$$

$$P_{RNN}(x_{t+1} = k|x_t, h_{t-1}) = o_{t,k} \tag{4-3}$$

$$f(z) = \frac{1}{1 + e^{-z}} \tag{4-4}$$

$$g(z_m) = \frac{e^{z_m}}{\sum_k e^{z_k}} \tag{4-5}$$

除了RNNLM结构之外，其思想是在有关循环的隐含字符信息中充分利用序列信息，以便它可以模拟任何长范围的信息。但实际上，它仅限于前几个字符。但是，我们需要修改RNNLM和多模式功能（即某些结构信息和形态特征）的插值以丰富句子[101]。

4.3.2 藏文后缀特征融合

基于藏文语法[39,40,57]《音势论》“ Suncupa”理论，提出了考虑后缀接续虚词特征的藏语语言模型。我们为考虑藏语显式后缀对虚词（Tibetan Radical Suffix Unit-Explicitly, TRSU-E）和考虑藏语隐式后缀对虚词（Tibetan Radical Suffix Unit-Implicit, TRSU-I）提供了藏语语言模型。TRSU-E根据显式后缀添加虚词，将十个后缀（“ ga”，“ nga”，“ da”，“ na”，“ ba”，“ ma”，“ v”，“ ra”，“ la”，“ sa“）（以及更远的后后加字）功能（如图4-8所示）加入了模型，从而在与下一个虚词连接时改善了显式后缀的接续准确性和语义正确表达。以图4-7为例，使用后缀“ sa”作为后缀接续不同的虚词，语义会出现变化。TRSU-I在添加后缀“ v”之后将隐式关系添加到模型，由于藏语通常不会写隐性后缀“ v”，因此在实际应用中，TRSU-E主要针对显性后缀接续虚词，而TRSU-I考虑了隐性后缀接续虚词。

如图4-8所示，当使用RNN计算字概率时，我们添加从后缀中提取的特征信息，然后在解决序列中虚词的接续问题的同时预测潜在的语义关系。网络的输入层为c_t，隐层为h_t，输出层为o_t。输入层c_t是表示t时刻的输入，o_t是神经网络语言模型的输出，它是语言模型的概率计算每个词$t+1$时刻赋予历史的词序列，所得到的softmax函数。RNNLM计算过程中的输入层，隐藏层和输出层如下：

$$c_t = f(W_{IC}x_t + W_{SC}s_t) \tag{4-6}$$

$$h_t = f(W_{CH}c_t + W_{HH}h_{t-1}) \tag{4-7}$$

$$o_t = g(W_{HO}h_t) \tag{4-8}$$

公式4-6中，W_{IC}表示字权重，W_{SC}表示后缀特征权重，用激活函数输入的字和后缀特征权重融合。这里W_{SC}在TRSU-E和TRSU-I处理有点差别，处理TRSU-E时，

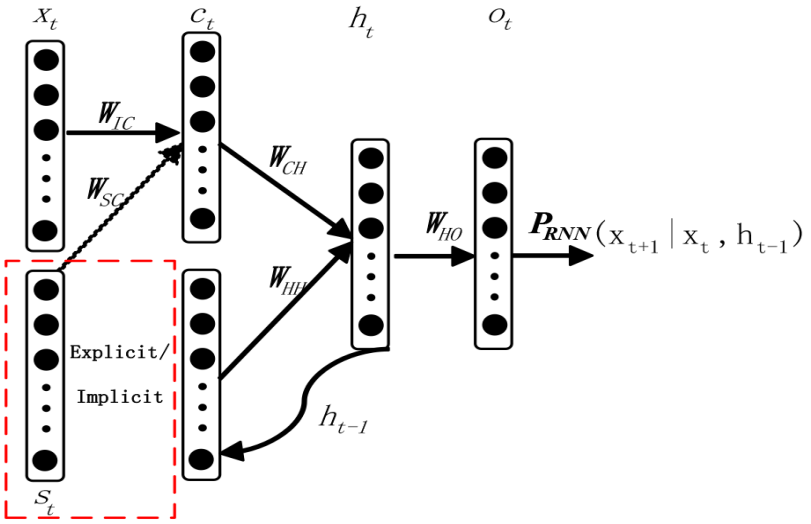

图 4-8　　考虑后缀虚词关系的藏语模型

只处理藏语语法中10个显式的后缀；而处理TRSU-I时，将所有后缀都进行处理，找出隐式后缀的接续关系和语义。公式4-7为隐藏层h_t，用激活函数对输入层c_t时刻W_{CH}权重和隐藏层h_{t-1}时刻的W_{HH}权重相加，用激活函数保留上下文信息，最后进行输出。

在基于语法的语言模型中，由于语法会影响藏语中的句子语义表达，因此后缀决定了虚词的接续。通过分析藏文语法关系，可以获得字的后缀信息，然后根据藏文语法的接续关系判断虚词。这种虚词在句子的语义中起着重要的作用，因此解决藏语虚词的接续关系可以进一步优化藏语句子的表示。

4.4　实验结果与分析

在本节中，我们首先介绍数据处理和分布情况，然后提取后缀特征，并将RNN方法应用于语言模型。我们选择困惑度（Perplexity, PPL）作为评价标准，PPL越小，语言模型的预测能力越好。最后，我们比较了以藏语语言模型为基准的最新方法[78]。先前的方法旨在获得字内部的形态结构信息，以补充低资源语言问题，而无需考虑语言本身的特征。

4.4.1 数据

藏语作为中国的一种少数民族语言，仅在藏族地区流行。因此，能在互联网上收集的语料库相对有限，而且研究人员相对稀少，因此获取语料存在困难。在本研究中数据均由我们构建的藏语新闻数据库（Tibetan News Corpus, TNC）作为实验数据。

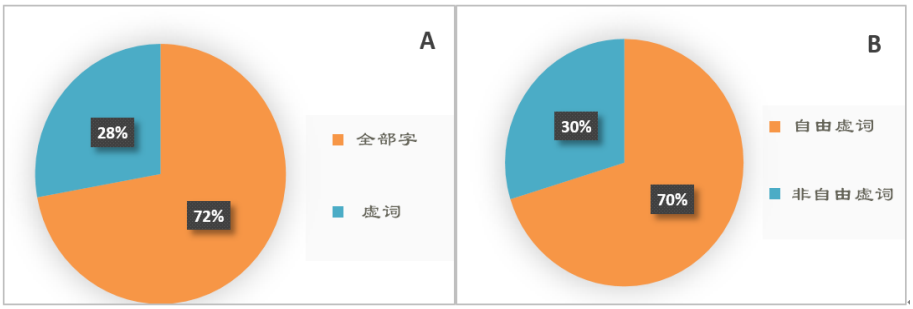

图 4-9 (A)语料中虚词的比例，(B)自由虚词与非自由虚词的比例

在藏语中，后缀特征对藏语虚词的接续关系非常重要，而虚词具有改变语义的功能。我们使用统计方法对数据进行分析，如图4-9（A）所示，我们发现在数据集中虚词的比例为28%，这充分说明了虚词的重要性。

为了更好地理解虚词中自由虚词和非自由虚词，我们根据数据集统计了它们的分布情况。从图4-9的B图中可以看出，自由虚词所占比例较大70%，也就是说，我们的任务不仅包含如何处理特定语法规则的接续关系，而且也能处理没有特定语法规则的接续关系。值得注意的是，非自由虚词接续关系是相对固定的，可以根据后缀来判断。而对于自由虚词，它们更灵活，没有固定的接续关系。我们方法可以从语料库中学习后缀特征，以产生一些接续的关系。

4.4.2 结果

本研究介绍了藏语后缀影响虚词的方法，即藏语显式后缀对虚词（TRSU-E）和藏语隐式后缀对虚词（TRSU-I）的语言模型。在实验中，我们把应用语言模型的传统方法与藏语模型研究的最新方法进行了比较。参数设置基于CUED-RNNLM方法[94,102–104]，N-gram中n=3，表示为KN3，且5次以上进行实验验证。由于我们数据集中的测试集与新闻相关，因此我们在实验中对相同类型，不同类型以及所有类型的数据集进行了实验，以验证我们方法的有效性。

在相同域上的数据集：由于测试集是以新闻为主题的，因此我们需要在实验中对相同类型的方法进行验证。在表4-1中，我们将实验分为三个部分。第一部分是现有语言模型研究的方法（传统方法和一些新方法），第二部分是藏语语言模型研究的最新方法，最后是我们提出的方法。我们首先将现有方法与我们的方法进行比较。可以看出，与现有语言模型相比，我们的方法困惑度（PPL）相对降低了约7.4%。与研究藏语模型的最新方法相比，我们的方法将困惑度（PPL）降低了4.8%。这表明我们所提出的方法可以有效缓解虚词错误的问题。

表 4-1　最新的方法和我们的方法同一域上的PPL对比

Model	News
N-gram[23]	55.2
RNNLM[23]	62.9
CUED_RNNLM[94]	58.4
CUED_RNNLM(LSTM)[94]	55.9
CharCNN[5]	55.2
_TRU[78]	57.6
_TRD[78]	54.3
_TRC[78]	53.8
TRSU-E	**52.7**
TRSU-I	**51.2**

为了验证我们提出的方法的有效性，不仅将我们的方法与其他方法进行了比较，而且还比较了我们提出的TRSU-E和TRSU-I。 TRSU-E解决了非自由词的接续关系，但是藏文中有些后缀是隐含的，这些隐式后缀将影响自由词的接续关系。因此，我们提出了TRSU-I来解决此类问题。实验结果表明，TRSU-I的PPL比TRSU-E低约3%，这也验证了我们提出的隐式后缀方法可以捕获后缀的隐式关系。

在不同域上的数据集：不同域上的结果显示在表4-2中。可以看出，在这种情况下，我们的方法比传统方法更准确。在教育、文化和文学类型上，我们的方法将PPL降低了18.6%至22.2%。与最新的藏语模型方法相比，我们的方法在教育、文化和文学方面的PPL降低了3.8%至10.4%。

表 4-2 已有的研究方法和我们的方法在不同域上PPL对比

Model	Edu	Law	Bud	Cul	Lit	Wikipedia
N-gram[23]	122.3	**288.8**	**497.5**	170.4	132.9	254.1
RNNLM[23]	147.9	374.1	698.9	196.2	155.9	218.7
CUED_RNNLM[94]	139.8	367.8	655.1	179.3	123.2	189.9
CUED_RNNLM(LSTM)[94]	127.2	364.1	618.4	169.8	118.2	160.2
CharCNN[5]	123.7	366.5	602.4	168.1	116.5	159.8
_TRU[78]	131.2	364.4	702.3	171.1	118.1	162.8
_TRD[78]	125.8	367.4	596.3	167.9	113.8	157.2
_TRC[78]	122.3	356.6	590.4	175.1	114.9	155.7
TRSU-E	118.5	358.9	562.8	165.8	110.4	152.4
TRSU-I	**117.6**	354.1	549.9	**165.2**	**109.8**	**152.1**

表4-2还显示，PPL的值在各个域上的差异，可以看出所有方法都在宗教和法律主题数据上没有传统N-gram方法好，这表明我们的测试集与此类数据类型有很大差异。尽管从上下文和形态结构信息上有一定的改进，但是没有N-gram方法有效。

在所有域上的数据集：表4-3汇总了所有类型的PPL值。在整个融合数据集上，我们的方法将PPL与传统方法相比降低了1.9%-16.2%左右，与最新藏语模型方法相比降低了4.7%-6.4%左右。再次证明了我们方法的有效性。

从表4-1的结果可以看出，新闻数据集的总体PPL值较低，这表明我们的测试集的域偏向新闻，因此在此数据集上已取得了不错的结果。所有域都是每种域的数据集成，这将产生噪声，而不是新闻域会影响结果。表4-2表明，我们的方法在包括所有领域的数据集上也取得了最佳结果，我们的方法对藏语语言模型是有效的。

由于数据量不足，为了公平地比较我们的方法和先前的方法，我们应用N-gram方法对每个类型的数据集所有模型进行了插值。我们使用线性插值方法，其中 λ 是比例插值权重模型。为了验证 λ 的范围，我们将隐藏层设置为700，并将 λ 从0设置为1，并比较RNNLM，CUED_RNNLM，LSTM，Char（CNN），TRU，TRD，TRC，TRSU-E，TRSU-I和Trigram。当 $\lambda = 0$ 时，我们采用RNN模型；当 $\lambda =$

表 4-3　所有域上的PPL对比结果

Model	ALL
N-gram[23]	98.6
RNNLM[23]	92.5
CUED_RNNLM[94]	89.2
LSTM[94]	84.1
CharCNN[5]	98.1
_TRU[78]	88.1
_TRD[78]	87.5
_TRC[78]	86.7
TRSU-E	**84.5**
TRSU-I	**82.6**

1时，使用N-gram方法，模型中一般取N = 3的三元模型。

表4-4进行了插值RNNLM，CUED_RNNLM，LSTM，Char（CNN），_TRU，_TRD，_TRC，TRSU-E，TRSU-I，和Trigram，其中lambda取0.5。从结果可以看出，我们的方法和其他方法都插值后，在我们的数据集上我们的方法仍然取得了良好的效果。在表4-2中，可以看到我们的方法在法律和佛教为主题数据集上的表现不如N-gram方法好。但是，插值后，与传统方法相比，我们的方法将PPL降低了3%至16.8%左右，尤其是在法律和佛教数据集上。由此可见，插值确实提高了预测的准确性。

4.4.3　分析

我们将藏语字符单元（Tibetan Radical Unit, _TRU）作为基准，_TRU是对藏文字中每个字符作为特征加入RNN中，但是不是每个字符都具有意义，会出现冗余信息，而且捕捉不到语法信息。本研究提出的方法是针对虚词，可以解决接续关系和句子语义准确表达。

TRSU-E方法是将藏语中后加字和后后加字提取出，作为特征加入RNN中。这种方法是可以将提取出后缀来判断虚词的接续关系，将解决非自由虚词和自由虚

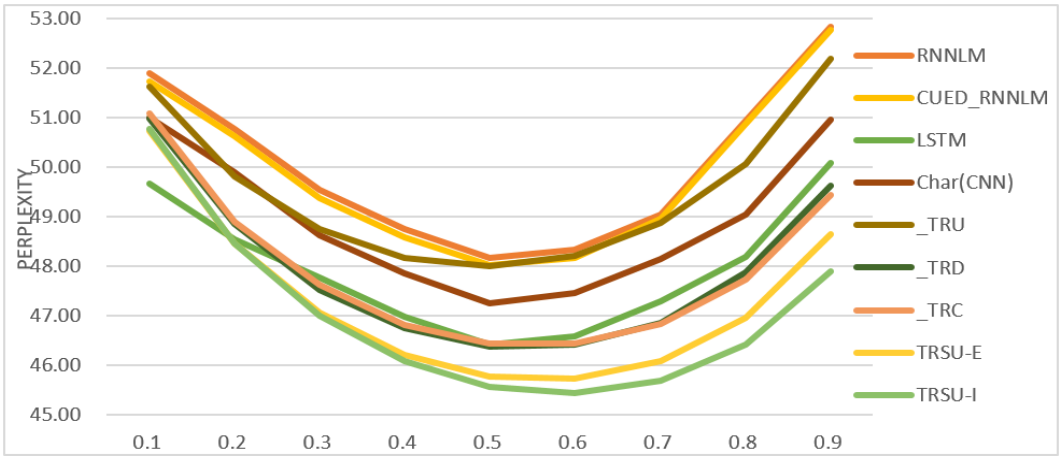

图 4-10 当隐藏层为700时，RNNLM，CUED_RNNLM，LSTM，Char（CNN），_TRU，_TRD，_TRC，TRSU-E，TRSU-I模型和Trigram插值

表 4-4 差值后的PPL结果对比

Model	News+ KN3	Edu+ KN3	Law+ KN3	Bud+ KN3m	Cul+ KN3	Lit+ KN3	Wikipedia+ KN3
N-gram[23]	55.2	122.3	288.8	497.7	170.4	132.9	254.1
RNNLM[23]	48.2	116.5	277.5	456.5	161.5	123.4	169.5
CUED_RNNLM[94]	48.0	110.4	272.5	427.2	153.2	112.6	151.2
LSTM[94]	46.4	107.3	274.8	419.5	143.3	107.8	139.1
CharCNN[5]	47.3	103.8	276.5	410.4	143.0	104.7	138.8
_TRU[78]	47.9	102.7	251.2	419.3	140.8	103.9	138.5
_TRD[78]	47.0	101.8	250.8	402.6	141.4	103.5	136.2
_TRC[78]	46.9	99.9	249.2	396.7	141.2	102.9	135.8
TRSU-E	45.7	98.1	249.1	387.3	139.8	102.0	134.5
TRSU-I	**45.1**	**97.3**	**247.1**	**383.1**	**137.9**	**100.4**	**133.9**

词接续问题。但是TRSU-E方法有个问题，RSU-E方法具有无法解决字符“ v”和“ da”之后的隐式问题。基于这个问题，我们提出了TRSU-I方法，此方法不但包含了TRSU-E方法，还对隐式后加字进行处理。TRSU-I方法不只是对后加字，只要是后缀，就把特征加入我们的模型。这种方法不但可以解决后加字对虚词的接续关系，还可以学习出隐式后缀“ v”和“ da”的隐式关系。这样就可以有效的解决非自由和自由虚词的接续关系以及准确预测句子语义表达。

表 4-5　　虚词对句子的语义影响

Method	Sentence
Original	ཁོང་གིས་ཐབས་འདི་ལ་བརྟེན་ནས་རིག་གཞུང་ཐད་དུ་བསམ་བློ་བརྗེ་རེས་བྱས་བ་དང
	Khong gis thabs vdi la brten nas rig gzhung thad du bsam blo brje res byas ba dng
N-gram	ཁོང་གི་ཐབས་འདི་ལ་བརྟེན་ནས་རིག་གཞུང་ཐད་འདུག་བསམ་བློ་བརྗེ་རེས་བྱེད་བ་དང
	Khong gi thabs vdi la brten nas rig gzhung thad vdug bsam blo brje res byas ba dng
RNNLM	ཁོང་གི་ཐག་དེ་ལ་བརྟེན་ནས་རིག་གཞུང་ཐད་དུ་བསམ་བློ་བརྗེ་རེས་བྱེད་བ་དང
	Khong gi thag de la brten nas rig gzhung thad du bsam blo brje res byas ba dng
Ours	ཁོང་གིས་ཐབས་འདི་ལ་བརྟེན་ནས་རིག་གཞུང་ཐད་དུ་བསམ་བློ་བརྗེ་རེས་བྱེད་བ་དང
	Khong gis thabs vdi la brten nas rig gzhung thad du bsam blo brje res byas ba dng

表4-5中的原始句子测试集中随意选择的句子。原始句子的意思是“他使用这种方法进行学术交流”。可以看出，N-gram方法产生的“thad vdug”语句中的虚词错误导致语义不清。 RNNLM方法还会在“khong gi”虚词中引起错误，从而导致后续字符和语义的变化。从句子的表达来看，“进行某种学术交流”可以用来确定这种虚假陈述。我们提出的方法很好地解决了这个问题，因此可以清楚地表达句子的语义。

4.5　本章小结

在本研究中，根据藏语字符的形态结构，我们提出了一种基于后缀特征的藏语语言模型，旨在预测语料库中的下一个字。通过引入后缀信息，提出的模型可以：1）从句子序列考虑虚词的接续关系，2）准确判断后缀相应虚词，3）区分字中的两种含义。这里，藏文中的虚词是根据接续和上下文来判断的。本文提出的方法不仅考虑规则的虚词，而且有效地利用了不规则的虚词。实验结果表明，该方法对藏语模型具有很好的辅助作用，其困惑度（PPL）比最佳基线低10%左右。为了更好地说明后缀方法的有效性，我们还将藏语字符单位（TRU）方法与

提出的方法进行了比较。此外，在和Trigram插值后本文的方法也取得了最佳效果，PPL比TRU插值的方法降低了5%左右。基于后缀特征融合的藏语模型用于支持藏语虚词的接续关系，可以解决句子中的语义关系。

第5章　基于动态形态结构的藏语语言模型

藏语语言模型（TLM）是藏语自然语言处理的关键。在本章中，我们首先观察到，与广泛使用的语言不同，藏语包含许多形态动词，这些动词很少出现在自然句子中，但在准确的文本预测中起关键作用。现有方法通常会忽略此属性，这会使传统的训练策略在构造准确而强大的藏语语言模型时效率较低。因此，我们提出了一种动态形态感知动词的藏语语言模型，它是通过字频率重加权策略进行离线学习并在线调整基于形态动词的判别权重来实现的。然而，由于形态动词对句子的时态和语义的影响，有必要考虑藏语中的形态动词。这里动态形态结构指的是藏语中形态结构变化的形态动词，是藏语中特有的一种动词类型。实验结果表明，与最新方法相比，我们的方法不仅减少了困惑，而且改善了文本预测和自动语音识别（ASR）任务中的字错误率。

5.1　引言

统计语言模型是描述一串文字序列成为句子的概率，在实践中应用于手写体识别、语音识别、机器翻译和信息检索等领域[4,8–10]。传统的N-gram一直是主要的语言模型[78,105,106]，因为该方法易于实现，训练速度快，泛化能力强。然而，有两个众所周知的问题：第一个问题是它不能长期捕获；二是数据稀疏性问题。随着技术的创新和改进，在传统的N-gram方法基础上，相继提出来了NNLM、RNNLM(LSTM)等最新方法，缓解了数据稀疏和长距离等问题[5,21,22,107,108]。

然而，RNNLM是目前最新的语言模型方法，需要大量的数据进行建模来参数估计，从而导致严重的数据稀疏性问题。对于低资源语言来说，语言模型无法有效地表示语言结构训练数据或关键词很少出现，这些关键词被称为稀有词。为了解决数据稀疏性问题，研究者们探索了许多不同的方法。一个是减少模型参数，包括基于语言模型的类和具有压缩层（Compression Layer）的语言模型[27,109]。然而，即使用较少的模型参数进行罕见词（Rare Words）的预测仍然很差。因此，另一种方法是利用更多的特征来丰富单词信息[27,28]，可以利用这些特征来增强

单词嵌入，并帮助RNN有效地学习更多的上下文信息[110]。另外，其他结构信息，例如子词(Subword)，字(Characters)，语素(Morphs)和词法，也已用于改进语言模型[29–32,111]。然而，这些方法主要集中于形态丰富的语言，例如英语或中文，而忽略了其他低资源语言，例如藏语等语言的特定属性。因此，通过仔细考虑与语言相关的属性，利用现有的通用的方法仍有很大的改进潜力。

关于藏语语言模型的研究很少，现有模型基本上使用N-gram方法[37]。最近，受[5,78]的启发，我们提出在基于静态形态结构上构建藏语语言模型，以从字中获取形态结构信息。已经提出了三种类型的嵌入式融合方法来增强模型，包括使用相同权重（UTibetan Radical Uniform Weight, TRU），不同权重（Tibetan Radical Different Weights, TRD）和自由基字组合（Tibetan Radical Combination Weight, TRC）。我们提出的模型的目的是使用特定的藏语基本单元和字嵌入，并通过引入不同的字嵌入因子来探索字嵌入的不同特征，以使模型更加灵活。每个部首嵌入都可以根据其对相应字整体含义的贡献进行内插。此外，藏语中存在一种激进的组合现象。自由基字入组合代码可以充分利用藏语基字嵌入的性质。由基本嵌入引入的字嵌入对于增强语义信息更有用，它可以帮助解决数据稀疏性问题。

在本章中，我们将重点放在藏语形态动词上来解决构建藏语原模型的问题。这里我们将动态形态结构定义为藏语特有的形态动词，因为藏语中形态动词的变化是根据形态进行变化的，是藏语特有的一类词。首先，我们观察到形态动词在藏语语言模型中的重要性，并观察到尽管形态动词很少出现在藏语词中，但它们占了藏语单词的很大比例，并在有效的藏语语言模型中起着关键作用。其次，我们通过对字频率进行加权来提出一种有效的离线学习方法，从而获得更强大的藏语语言模型。第三，我们通过在线调整其辨别权重以建立自适应离线学习模型，进一步增强了形态动词的重要性。通过上述努力，与基准模型相比，我们最新提出的藏语语言模型在研究预测和ASR任务上获得了更好的结果。

5.2 相关研究

在本节中，我们将调查和讨论与语言模型相关的工作。我们基于多个粒度的方法和传统的方法来调查，并讨论我们提出的方法对语言模型的推进作用。

不同粒度的输入。对于低资源语言而言，罕见词（Rare words）处理是常见的一个问题。在[5]中，通过将传统的词嵌入级别降低到字级别，避免了大规模的

嵌入计算和稀有词，并且通过Highway Network网络技术构建了更深的网络，从而获得了良好的效果。该模型由两部分组成：作为CNN输入的字级别，以及通过CNN和公路网的输出输入RNNLM的字，但是最终的预测仍然是一个单词。可以通过使用多种语言的语料库作为测试来获得语义和语法信息。此外，[4]还提出了一种将RNN与字嵌入级别的CNN结合的模型。

对于形态丰富的语言，研究词的形态结构形成是有帮助的[82–84]。英语形态（Morphology）学是研究词组成与试图整理词组成规则（但没有语法和语义信息）之间关系的研究。因此，语言模型可以应用这样的信息以获得高精度的词序列。

利用自适应的方法。在实际的ASR任务中，没有大量的域匹配数据集。语言模型需要使用域自适应技术，以允许使用多个域外文本资源[85,86]。在语言建模中最合适的领域自适应方法之一是基于混合模型[112,113]。可以通过组合语言模型和从域外文本资源分别构建的混合加权来构建适应模型[114–116]。在[117,118]中，由于存在域外数据，我们应用较大的数据来适应较小的数据，以解决数据量不足的问题。

为了使用从域外文本资源构建的语言模型进行灵活的域适应，[87]和[88]开发了一种在潜在变量空间中进行模型合并的方法。在潜在变量空间中，单词被映射到潜在变量空间，因此可以期望它比观察到的单词空间执行更灵活的状态共享。因此，本文将潜在词语言模型（latent word LMs, LWLMs）引入混合建模[119–122]。N-gram 方法的语言模型的普通类中的潜在变量只是与模型相关的索引，每个模型都有不同的潜在变量空间[123,124]。因此，传统的基于类的n元语法混合模型必须在观察到的文本空间中执行[90,118]。另外，LWLM中的潜在变量表示为特定的潜在词，多个LWLM可以共享一个公共的潜在变量空间，这使我们能够在考虑潜在变量空间的同时进行灵活的混合建模。

5.3 藏语中形态动词的作用

藏语属于汉藏语系，是一门历史悠久，在汉藏语系中具有重要影响的语言。传统上，国内的藏语分为拉萨语，康巴语和安多语。卫藏方言与安多方言差异较大互通有一定困难，康方言接近卫藏方言。藏语内部差异比较明显，卫藏和康方言有声调，安多方言没有声调。藏语语法偏孤立，主要靠语序和助词表达各种语法关系，有部分动词表现为屈折变化。

5.3.1 藏语形态动词

藏语动词是理解句子含义的关键，可以根据不同的性质分为不同的类。尤其是，藏语动词可以根据其形态特性分为形态变化和形态不变动词。特定的形态动词可能包含2-4个时态变体，包括将来式，现在式，过去式和命令式。

表 5-1　　形态变化动词/形态不变动词时态变化

verb	future tense	present tense	past tense	Command tense	Word meaning
Morphological verbs	ཟླེབ	ཟླེབ	ཟླེབས	ཟླེབ	"Finish"
	ཉོན	ཉོན	ཉོནས	ཉོན	"Listen"
	སྒྱུར	སྒྱུར	སྒྱུརས	སྒྱུར	"Translation"
Morphologically invariant	སྙན	སྙན	སྙན	སྙན	"Praise"

根据藏语动词变化的特点，我们可以从图5-1中可以看出可分为形态变化动词和形态不变动词。而形态变化动词又可以分为四种形态变化、三种形态变化和两种形态变化。四种形态变化指的是在将来式、现在式、完成式和命令式都发生字形体的变化；三种形态变化是指形态动词一般是将来式、现在式、完成式和命令式中某一个形态与其他三个里的一个相同，但是没有规律；两种形态变化是将来式、现在式、完成式和命令式中某一个形态与其他三个里的一个相同，或某两个形态与其他两个形态相同，也没有规律。

与英语的形态动词不同，藏语的形态动词是在字级别上定义的，并且包含附加命令，即命令式，这对于句子的理解非常重要[57,94,125,126]。表5-1列出了3种形态变化动词和1种形态不变动词。可以看出，不同时态动词在句子中有不同的变化，这些变化对句子的语义有影响。

5.3.2 基于类的藏语语言模型

我们可以使用RNNLM[22]来建模藏语。具体来说，要预测当前字w_i，我们可以将当前字的完整历史编码为$< w_{i-1}, \cdots, w_1 >$，并通过三层RNN计算概率：

$$P_{RNN}(w_i| < w_{i-1,\cdots,}w_1 >) = P_{RNN}(w_i|w_{i-1}, v_{i-2}), \tag{5-1}$$

其中v_{i-2}表示从 1 到 $i-2$的剩余历史上下文。我们进一步修改公式5-1通过添加基于类的分解输出层结构。基于频率计数，输出层词汇表中的每个单词都归因于唯一的类。然后我们得到

$$\begin{aligned} P_{RNN}(w_i|w_{i-1}, v_{i-2}) &= P_{RNN}(w_i|v_{i-1}) \\ &= P(w_i|c_i, v_{i-1})p(c_i|v_{i-1}), \end{aligned} \tag{5-2}$$

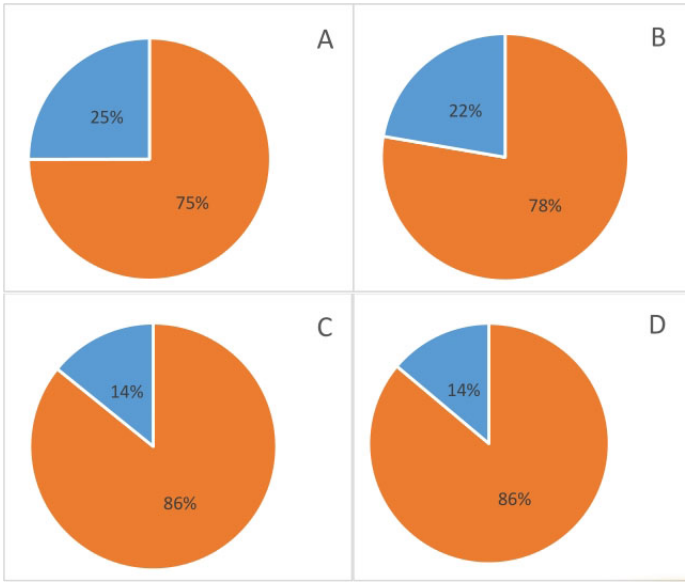

图 5-1　上图中A为藏语音频语料中形态动词比例；B为词典中形态动词所占的比例；C和D为训练语料和测试集中形态动词所占的比重。

c_i 表示类别标签。 $P(c_i|v_{i-1})$ 是基于条件概率的类别，并且根据训练语料库上的字频率定义为分组结果。我们还在图5-1和图5-2中显示了等式（2）的结构。已经证明具有等式（2）的RNNLM可以捕获英语中的长距离依赖项。 $P(c_i|v_{i-1})$在这种模型中起关键作用，并依赖于基于频率计数的字分类。该策略没有考虑特定语言属性的关键，例如形态动词在表示语言模型中的重要性。

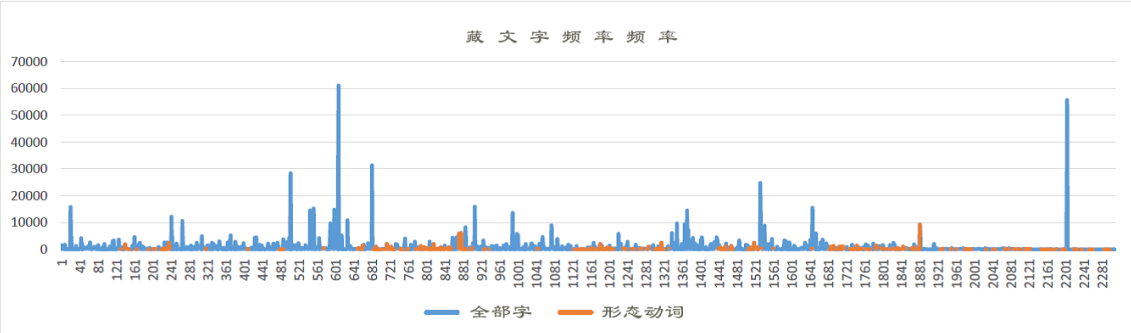

图 5-2　我们的语料中形态动词的词频分布

5.4 基于形态动词的藏语语言模型

自然语言中动词是不可或缺的部分，是句子表达的核心。因语言的差异性，对动词的划分是不同。由我们知道在英语中常用动词充当谓语，是句子表达语义的核心。在藏语中动词也很重要，不但跟语义相关，有一些跟时态也相关。因此，对于资源不足问题中形态动词的研究，对于探索更有效的藏语语言模型有很重要意义。

5.4.1 藏语语言模型中形态动词的重要性

字是藏文中最小和有意义的单元，类似于英语中的单词。如图5-1所示，我们发现藏语中的形态动词所占比例很大，在语料库和词典中分别占25%和22%左右（图5-1中的A和B），训练集和测试集中形态动词所占比例14%左右（图5-1中的C和D）。这些动词不仅影响句子的时态关系，而且影响句子的语义。

然而，如图5-2所示，那些重要的形态动词在我们的训练语料库中所占的比例很低，这导致受过训练的藏语语言模型容易错过形态动词的语义信息，并直接影响基于藏语语言模型的识别任务的准确性和速度。例如，我们在传统语料库上训练了两个基于传统RNNLM和形态动词感知藏语语言模型，并使用它们来执行文本预测。由于考虑到离线学习和在线调整中的形态动词，我们的方法获得的预测误差要低得多。

在表5-1和图5-2中，我们可以看到藏语中的形态动词在理解藏语句子中起着重要的作用。实际上，由于藏语是一种资源匮乏的语言，音频和文本数据是有限的，所以，我们在获取藏语形态动词信息时受到限制。因此，有必要增加形态动词的权重以更准确地预测句子的语义。

5.4.2 离线学习通过字频率重新调整

藏语形态动词是藏语动词中形态特点较明显的词类，藏语语言学家已经研究和总结出形态动词表[39,40,57]。这类词在我们的训练语料中有很重要的作用，它对句子时态关系和语义都会产生影响。由此，我们提出基于词频的形态动词加权方法。

从公式5-2中 $P(c_i|v_{i-1})$ 我们可以看出，在预测w_i词时需要从c_i词类和v_{i-1} 上下文

信息来获取，而c_i是根据训练语料中的词频获取。c_i是将训练语料中的w_i词进行分类，这样做的目的是为了能够快速准确获取w_i词。所以，我们根据藏语形态动词特点，对训练语料中的形态动词增加字频。我们这样做的目的是想让w_i词分到词较少的类中，这样在预测时获取的速度更快，而且增加了准确性。

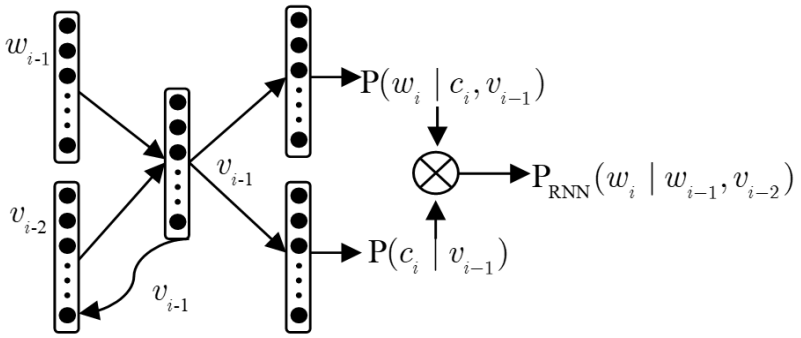

图 5-3　　RNNLM基本模型的结构

具体来说，在计算了语料库中所有字的字频后，我们通过将形态动词的值乘以权重 β 来对形态动词的字频进行加权。在本研究中，我们设置$\beta = 3$。然后根据新的加权字频率对所有字进行分类，从而得出新的 $P(c_i|v_{i-1})$。这样就可以将一些原来在低频率类形态动词划分到较高新类中，提高预测速度和精度。此方法称为RNNLM_字频加权法（Character Frequency Recounting, _CFR）。

5.4.3　在线调整权重

为进一步了解测试过程中的形态动词，我们提出了在线调整判别权重即$P(c_i|v_{i-1})$的方法，以提高预测精度，如图5-4所示。这种方法是把RNNLM的输出做一个反向目标词的加权，来影响对类的划分。具体的，是将RNNLM的输出给一个阈值，这里的阈值我们是根据输出的概率给定，生成一个二值化的输出。具体地说，RNNLM的输出被赋予阈值 ε，然后，生成二进制输出公式5-3；

$$\bar{P}_{RNN}(w_i|v_{i-1}) = P_{RNN}(w_i|v_{i-1}) > \varepsilon. \tag{5-3}$$

不属于形态动词的$\bar{P}_{RNN}w_i|v_{i-1}$的值设置为0，并获得$\bar{P}_{RNN}(w_i|v_{i-1})$。我们将$\tilde{P}_{RNN}(w_i|v_{i-1})$作为RNNLM的预测，并将其反向投影到$P(c_i|v_{i-1})$，并获得

$$\tilde{P}(c_i|v_{i-1}) = \frac{\tilde{P}_{RNN}(w_i|v_{i-1})}{P(w_i|c_i, v_{i-1})}. \tag{5-4}$$

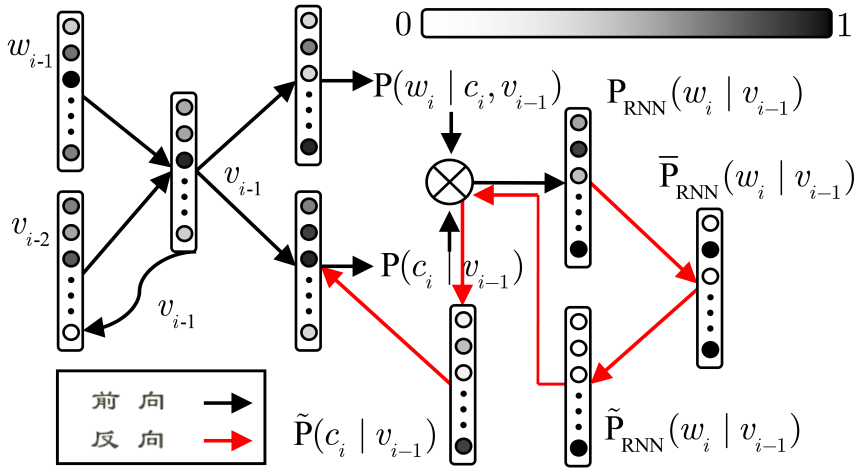

图 5-4　在线调整判别权重图

通过将其与 $\tilde{P}(c_i|v_{i-1})$组合，我们生成一个新的$P(c_i|v_{i-1})$

$$\bar{P}(c_i|v_{i-1}) = P(c_i|v_{i-1}) + \alpha\tilde{P}(c_i|v_{i-1}), \tag{5-5}$$

其中α表示 $\tilde{P}(c_i|v_{i-1})$的组合权重。我们将此方法表示为RNNLM _判别权重（tuning_TDW）。

表 5-2　藏语文本数据

数据	#字数	% OOV
字典	2472	-
STD	1.5m	1.08
LTD	21.3m	1.48
验证集	125k	1.12
测试集	126k	1.11

5.5　实验结果与分析

在实验部分，我们首先介绍本研究中应用的语料库，包括拉萨方言的语音语料库和文本语料库。然后给出了基于基线语言模型建模方法以及我们的提出的方法，对比了现有的方法和我们的方法在不同隐层时的结果。接下来，将我们的方

法和传统的N-gram方法进行插值，和其他插值的结果进行对比。在实验中，我们将N-gram方法中应用Kneser-Ney平滑，N取3记为KN3。最后，将我们的方法应用于ASR验证了方法的有效性。

5.5.1 实验准备

表 5-3 现有方法和我们的方法在同一域上的PPL比较

Granularity	Language Model	PPL
character	N-gram[23]	55.2
	RNNLM[23]	62.9
	CUED_RNNLM[94]	58.4
	LSTM[94]	55.9
	CharCNN[5]	55.2
radical	_TRU[78]	57.6
	_TRD[78]	54.3
	_TRC[78]	53.8
morph	**_CFR**	**50.6**
	_TDW	**49.8**

通过实验，我们验证了我们提出的方法性能。我们将字作为输入。在实验中，我们选择困惑度（Perplexity, PPL）和字错误率（Character Error Rates, CERs）作为评价标准。

语料库的选择直接影响藏语语言模型的质量，从而影响语音识别性能。对于语言模型，应该选择现实生活中常用的句子，以符合使用自然语言的人们的习惯。我们的数据集来自我们构建的数据集（表3-2）。其中文本数据是新闻为主域的小型藏语训练数据集（Tibetan Training Dataset, STD）在测试集的范围内。其他域的数据集合为藏语训练数据集（large Tibetan Training Dataset, LTD）中，根据音频数据的频率，单词限制在前2472个字内。词汇（Out-Of-Vocabulary, OOV）符号用于呈现不属于所选词汇的任何字。语料库的大小和OOV字的百分比如表5-2所示。

在实验中，我们有两个数据集，即STD和LTD，但有不同的域，包括：新闻，

表 5-4　我们的方法插值后相同域上的PPL进行比较

Granularity	Language Model + KN3	PPL
character	N-gram[23]+ KN3	55.2
	RNNLM[23]+KN3	48.2
	CUED_RNNLM[94]+KN3	48.0
	LSTM[94]+KN3	46.4
	CharCNN[5]+KN3	47.3
radical	_TRU[78]+KN3	47.9
	_TRD[78]+KN3	47.0
	_TRC[78]+KN3	46.9
morph	**_CFR+KN3**	**45.1**
	_TDW+KN3	**44.2**

政治，经济学，文化，佛教和格萨尔。而音频数据是之前构建的数据库（表3-1）。

5.5.2　结果

表5-3是我们在STD数据集上最新的方法做的结果，在RNNLM基础上，我们把基于部件（Radical）的Tibetan Radical Uniform Weight（TRU）方法作为基准。可以看出我们的方法比RNNLM困惑度（PPL）减少了15.5%左右，比基准TRU方法PPL减少了11.6%，相比Tibetan Radical Different weight（TRD）方法和Tibetan Radical Combination weight（TRC）方法PPL分别减少了6.8%和5.8%左右[78]，说明我们的方法有效性。比传统的N-gram方法我们的方法困惑度降低了8.3%。从实验结果我们可以看出，在藏语中形态动词对预测下一个字是有影响，我们在模型中对形态动词加权是可行的。

RNNLM和N-gram语言模型具有作为两个本质上不同的LM的建模特征。RNNLM通常将固定权重的线性插值与N-gram语言模型结合使用。表5-4显示了我们的方法和N-gram插值在STD数据集上的结果，参考[94,111,126]中的 λ 值（ $\lambda = 0.5$）。对于传统方法的插值，我们的提出的方法和N-gram进行插值取得了较好的效果。从结果可以看出我们的方法相对提高3.4%-18.8%不等。

在表5-4中，我们可以看到，我们所提出的方法比传统的N-gram方法具有更好的结果。另外，当我们的方法与N-gram方法结合使用时，可以实现一些改进。这证明了我们的方法和N-gram方法在解决稀有词问题方面具有互补的作用，这进一步证明了我们改进的方法在解决藏文数据稀疏问题方面的有效性。

表 5-5 我们对STD和LTD数据集进行N-gram插值的方法

# hidden	RNNLM(S)	RNNLM(L)	_TRU(S)	_TRU(L)	**_CFR(S)**	**_TDW(S)**	**_CFR(L)**	**_TDW(L)**
400	47.9	46.5	48.8	47.4	**46.2**	**45.7**	**44.6**	**43.8**
500	47.2	45.5	48.4	46.7	**45.9**	**44.9**	**44.2**	**42.9**
600	48.5	46.9	48.3	46.7	**45.7**	**44.5**	**44.1**	**42.3**
700	48.7	46.8	48.0	45.9	**45.1**	**44.2**	**43.2**	**42.1**

RNNLMS和N-gram语言模型作为两个本质上不同的语言模型具有各自的建模特点。RNNLMS通常与N-gramLM结合使用线性插值的固定权重。表5-5是我们对STD数据和LTD数据进行插值的结果。括号中的（S）是指Trigram模型在小数据（STD）训练中的应用，而括号中（L）是指Trigram模型用于较大数据（LTD）训练的模型。对于STD数据，与RNNLM方法和N-gram相比，我们的方法和N-gram分别减少了7.4%和9.3%。我们方法和N-gram比TRU 方法和N-gram困惑度减少了6.1%和7.9%左右，而N-gram在LTD数据上训练的模型和我们的方法进行插值，比RNNLM方法和N-gram方法困惑度减少了11.3%和13.6%，比TRU方法和N-gram方法困惑度减少了10%和12.3%。

可以看出，_CFR和N-gram（S）插值的方法比其他插值的方法产生更好的结果。特别是，我们提出的_TDW和N-gram（L）之间的插值方法不仅比现有方法获得了更好的结果，而且在PPL中相对_CFR方法降低了约4.5%左右。这个结果也表明我们的方法对改进藏语模型有很好的效果。

以上实验均以PPL为评估标准。我们可以看到，我们的方法比最新的藏语语言模型的方法取得了更好的结果。为了验证我们提出的方法的有效性，我们使用了不同粒度的实验结果，并将我们提出的方法应用于拉萨方言的ASR进行验证。结果表明，在基于字层面，RNNLM方法优于传统的N-gram方法。

TRC方法用于藏文的部件粒度达到了最佳效果。表5-6中的结果显示基于_CFR的RNNLM方法CER相对提高了3.1%。使用基于_TDW的RNNLM方法CER相对提高4.3%。与最新方法相比，我们提出的方法对藏语拉萨方言的语音识别有很好的效果。

我们知道，词格（Lattice）是在语音识别过程中一次解码的结构，其中包含大

表 5-6　最新方法和我们方法的%CER对比

Granularity	Language Model	%CER
character	N-gram[23]	35.20
	RNNLM[23]	34.60
	CUED_RNNLM[94]	34.25
	LSTM[94]	33.96
	CharCNN[5]	34.03
radical	_TRU[78]	34.09
	_TRD[78]	34.15
	_TRC[78]	33.94
morph	**_CFR**	**33.55**
	_TDW	**33.10**

量候选结果。由于神经网络使用历史信息来预测下一个单词，因此对词格重新定级将导致搜索速度降低。与词格的词结构相比，N-best更适合于长距离信息的模型扩展。本文使用N-best的中间结果进行评分[125]，如表5-7所示。

我们在藏语拉萨方言音频数据集的ASR实验中验证了我们的模型[38]。表5-7是我们对藏语拉萨方言识别%CER进行验证的结果。在N-best（n =100和n=1,000）评分中，我们的方法减少了约3.5%，这表明我们的方法对藏语语言模型中稀有词的加权方法有很好的效果。

表 5-7　具有N-best的字错误率评估结果

N	# hidden units	%CER with N-best rescoring						Original—
		RRNNLM	_TRU	_TRD	_TRC	**_CFR**	**_TDW**	
100	500	33.84	34.22	34.07	34.02	**33.65**	**33.20**	35.20
	600	34.03	33.99	34.06	34.05	**33.55**	**33.10**	
	700	33.97	34.09	34.15	33.94	**33.55**	**33.03**	
1000	500	33.73	34.15	33.92	34.02	**32.87**	**32.71**	
	600	33.88	33.87	34.06	33.82	**32.87**	**32.65**	
	700	33.83	34.05	34.08	33.78	**32.74**	**32.55**	

5.5.3 分析

以上实验结果表明，我们所提出的RNNLM_字频率重加权（_Character Frequency Reweighting, _CFR）和RNNLM_在线判别权重（Tuning Discriminative Weights, _CFR）是有效的同时，本研究发现一些有趣的现象如下：

对于稀有词的处理研究者都有不同的方法，有些学者提出基于稀有词的加权方法，有些学者提出自适应的方法。这些研究存在的问题是这些稀有词是否都有意义。有鉴于此我们根据藏语的形态动词这一特点，提出基于形态动词的藏语语言模型。结果表明，利用形态动词加权不但可以影响类型的变化，还可以提高预测能力。因此，增加形态动词的权重可以学习更多的语言特征，这种语言特点有助于提高藏语语言模型的性能。

表 5-8 对基线模型和我们的模型的输出句子进行主观评估

Number	1	2	3	4	5
Semantic influential number of sentences	46	49	45	48	47

在藏语中，形态动词对于句子的时态影响较大，而且在时态变化上没有统一的标准。表5-8 中我们就可以看出，藏语中的形态动词可分为形态变化词和形态不变词两种。前者是我们研究的重点，因为这类动词在句子中对时态变化起主要作用，对后者不作特殊处理。然而，这些词在训练语料中出现的频率少，这会影响句子的预测能力，尤其在语音识别任务中会影响识别结果。所以需要我们针对这类词进行研究分析，得出能够更加准确方法来解决这类问题。

为了验证我们的方法，我们在ASR上进行了实验。我们将测试集中基线输出语义不准确的100个句子，然后用我们提出的方法输出，再拿这些句子和原句进行对比。对比的方法是采用主观评价，我们找了5位藏语专家进行了打分。表5-8显示了中是我们提出的模型和基线模型比较的结果，评分的标准是以原句作为标准，100个句子中语义最相近原句句子。表中我们可以看出，我们提出的方法在语义理解上比基线方法好，在基线模型输出的100个句子中我们的模型平均可以表示47%的句子语义。

综上所述，我们提出的基于形态动词的加权方法，在一定程度上影响了句子的语义，而且也取得好的效果。由此，强化藏语中的形态动词是有意义的工作。

5.6 本章小结

在本节中，根据藏语的动态形态结构，我们表明藏语形态动词是稀有词，对于学习有效的藏语语言模型非常重要，并且我们提出了一种形态动词感知藏语语言模型。我们首先提出了一种通过字频率重新加权的离线学习藏语语言模型，以增强形态动词的权重。此外，我们建议在线调整形态动词的判别权重，以使离线学习的藏语语言模型在线自适应。结果表明，我们的方法在文本预测和ASR任务方面优于基线模型。

第6章　有效融合静态和动态形态结构的藏语语言模型

循环神经网络语言模型（RNNLM）在用于自动语音识别（ASR）时通常优于N-gram语言模型。对于藏语语言模型（Tibetan Language Model, TLM）的研究，当前的研究方法基本都是英语和中文中应用现有方法的移植。在我们的研究中，为了获取语言本身的特征对藏语语言模型的影响，我们分析了静态形态结构和动态型结构关系，发现藏语语法和形态动词对语言的重要性。在本文中，为了缓解静态形态结构和动态形态结构对藏语的的影响，我们提出了语法关系和形态动词感知藏语语言模型。我们的主要贡献有以下三个方面：1）利用形态结构关系，通过考虑藏语语法和形态动词对现有语言模型的影响，深入分析藏语语法和形态动词对语言理解的重要性。2）利用于静态形态结构关系，针对藏语中虚词对句子的影响，提出了加入后缀特征。 3）针对藏语稀有词问题，发现藏语中的形态动词基本上分布在低频词或稀有词中，由此，我们对形态动词进行了加权。后缀特征可以影响句子中虚词的接续关系，从而纠正句子中虚词的接续关系。对于稀有词我们提出了一种自适应加权方法，可以有效地增加此类单词的权重，从而增强此类单词在句子中的作用。实验证明，我们在考虑语法关系是比现有的最新方法Perplexity（PPL）相对提高了4.8%，而对形态动词的加权对最新的方法相对PPL提高了7.4%，而考虑语法和形态动词的方法相对PPL提高了9.8%，在ASR上也比现有的方法取得了提升。

6.1　引言

语言模型（LM）是各种自然语言处理任务所必需的，例如自动语音识别（ASR）、统计机器翻译和手写体识别等[127,128]。语言模型面临的最常见问题之一是数据稀疏性[127,129]。语言模型好坏取决于训练数据集的数量和质量。通过使用巨大的域匹配数据集来构建语言模型，通常可以获得优异的性能[18,21,95,130]。

对于低资源（Low Resource）语言来说，数据稀疏是常有的问题。为了减轻数据稀疏性问题，已经提出了几种技术。最传统的技术是在N-gram建模中平

滑，已经研究了改进语言模型概率估计的各种平滑方法用于N-gram 的语言模型[105,131]。另一种解决方案是基于降维，基于类的N-gram的语言模型[16]和决策树的语言模型[132]基于单词分类，基于神经网络的语言模型基于学习单词的分布式表示[22,108,126,133–136]。

对于藏语语言模型而言，因为是属于低资源语言[34]，目前研究的不多，常用模型基本都是用N-gram方法[37,65]。最近也对藏文部首（藏文部件）进行构建，提出了基于字符（Radical）的藏语语言模型。这种方法可以学习出藏语的字形形态结构关系，增强字的形态结构信息，可以帮助解决数据稀疏性问题[78]。但是未对语言自身的特点进行分析和应用，所以需要对语言特点进行分析说明。

在本文中，为了获取语言中语法对句子的影响，我们利用静态形态结构关系，根据藏语虚词接续关系，对藏语后缀接续虚词进行分析。虚词的接续关系是由后缀来判断，是判断是否接续准确的关键。这种语法关系的优势在于可以在句子中判断准确的虚词，因为当前虚词可能会影响下一个字或词，这样会提高句子的准确性。考虑后缀对虚词的影响不但可以辅助藏语虚词的接续关系，而且可以解决序列中句子语义关系。

对于藏语语言模型而言，低频词处理是常见的一个问题。我们对数据分析发现低频词对模型影响较大，而其中基于形态动词占了很大的比重，由此，我们根据藏语的形态动词这一特点，提出基于动态形态结构的藏语语言模型。利用形态动词加权不但可以影响类型的变化，这使得模型更具灵活性，有助于解决数据稀疏问题。其主要贡献在于对RNNLM中词类进行加权，改变一些罕见词或低频词的权重，还可以提高预测能力[137]。

为了获取到更好的信息，我们考虑了语法信息和形态动词加权方法，就是将藏语语法特征和藏语形态动词对词类的影响加入RNNLM模型。具体的，就是把后缀对虚词的信息加入输入，增加了语法信息，然后根据形态动词的加权方法，对低频词进行加权，使得一些低频词的词类发生变化，能够更加准确预测下一个词。这种方法不但可以获取模型中语法关系，还可以对罕见词和低频词进行加权，提高模型的性能。

6.2 语法关系和形态动词

藏语是一种有自己拼写规则的语言。与英语相比，构词法有些不同。英语使

用空格来划分单词，而藏语则需要进行分词。目前还没有现成的藏语分词标准，而且语料资源有限，我们的研究中常用的字为单位。构成藏文字一般都1-7个字符(这里不包含梵文)，藏语中常见的最长字长为7。其中后加字和后后加字对虚词接续关系相关，会影响句子的语义关系。

6.2.1 藏语语法关系

每种语言都有自己的语法体系，藏语也不例外。藏语有一套自己的语法体系，藏语中最常见的句子是与虚词不可分割的，虚词用来连接词与词或短语来表达语义。虚词本身不具备独立表意功能，通过连接词语前后来表示特殊的语义功能。虚词是起连接作用，根据语境的不同表达的意思也不一样。藏语虚词是根据后加字的不同来添加（虚词前出现的后缀），后缀有后加字和后后加字，即：ga，nga，da，na，ba，ma，v，ra，la，sa。虚词本身不具备独立表意的功能，通过连接词语前后来表示特殊的语义功能[39,40,57,100]。

为了更好地理解包含自由虚词和非自由虚词。从之前的研究中我们可以看出，自由虚词占据比例大，也就是说，我们的任务包含语法规则的接续关系，而且还没有特定规则的接续关系。值得注意的是，非自由功能的连续关系是相对固定的，可以根据后缀来判断。另一方面，对于自由虚词，它们更灵活，并且没有固定的连续关系。我们可以从语料库中学习后缀特征，以产生一些连续的关系。

6.2.2 形态动词对句子的影响

藏语动词是理解一个句子意义的关键，根据不同的性质可以分为不同的子集。藏语动词根据形态特征可分为形态不变动词和形态变化动词。具体来说，形态动词可能包含1到4个时态变体，包括将来式态、现在式、过去式和命令式[39,57]。

我们根据藏语动词形态屈折变化特点，可以看出可分为形态变化动词和形态不变动词。而形态变化动词又可以分为四种形态变化、三种形态变化、两种形态变化和没有变化。四种形态变化指的是在将来式、现在式、完成式和命令式都发生字（Character）形体的变化；三种形态变化是指形态动词一般是将来式、现在式、完成式和命令式中某一个形态与其他三个里的一个相同，但是没有规律；两种形态变化是将来式、现在式、完成式和命令式中某一个形态与其他三个里的一个相同，或某两个形态与其他两个形态相同，也没有规律，没有变化的是四个时态都形态不发生变化[39,57,100]。

与英语的形态动词不同，藏语的形态动词是在字面上定义的，这对理解句子非常重要。由于形态动词在理解藏语句子语义有重要性，我们有必要对形态动词进行分析。

6.3 考虑语法和形态动词的藏语语言模型

6.3.1 RNNLM

统计语言模型给出了一个词的序列，然后由其概率值预测可能得到一个句子。RNNLM可以利用时序信息，保存上下文信息来获取隐层信息。标准循环神经网络的结构是图6-1所示。

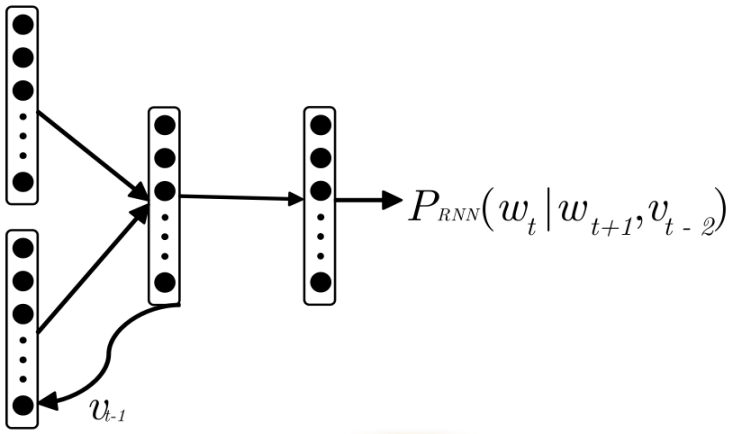

图 6-1 RNNLM基本模型

表示时间t的输入层，它使用大小为x_t 的一热向量对当前单词w_t进行编码。表示为h_t的隐藏层通过使用激活功能保留了其余的上下文信息。 RNNLM的输出层是o_t ，这是在给定历史单词序列$< w_t \cdots w_1 >$ 的情况下，词汇在时间$t+1$处每个单词的语言模型概率，这是通过softmax函数获得的。 RNNLM传播过程中的输入层，隐藏层和输出层可以按传统RNNLM公式进行建模。

我们可以使用RNNLM进行语言建模。具体来说，要预测当前字符w_i，我们可以将当前字符的全部和历史编码为 $< w_{i-1} \cdots w_1 >$，并通过三层RNN计算概率：

$$P_{RNN}(w_i| < w_{i-1,\cdots,}w_1 >) = P_{RNN}(w_i|w_{i-1}, v_{i-2}), \tag{6-1}$$

其中v_{i-2}表示从1到$i-2$ 的剩余历史记录上下文。我们进一步修改算法通过添

加基于类的分解输出层结构。基于频率计数，输出层词汇表中的每个单词都归于一个唯一的类。然后我们得到

$$P_{RNN}(w_i|w_{i-1}, v_{i-2}) = P(w_i|c_i, v_{i-1})p(c_i|v_{i-1}), \tag{6-2}$$

c_i 表示类标签。 $p(c_i|v_{i-1})$ 是基于条件概率的类，并且被定义为根据训练语料库上的字符频率的分组结果。$p(c_i|v_{i-1})$ 在这样的模型中起着关键作用，并依赖于它的分类。在RNNLM结构以外，这个想法是为了充分利用序列信息在循环中隐含的字信息，所以，它可以模拟任意长跨度的信息。但在实践中它是限于前几个字，获取的长度是有限的，虽然之后提出了LSTM（Long Short-Term Memory）、GRU（Gated Recurrent Unit）和Attention等方法[138]，获取的上下文信息长度和结构是有限的[108,139]。

6.3.2 语法关系影响藏语语言模型

藏语语法对句子语义影响较大，由此我们在RNN计算字的概率时加入后缀抽取出的特征信息，将字和后缀进行融合作为输入，解决序列中虚词的接续问题的同时预测出潜在的语义。

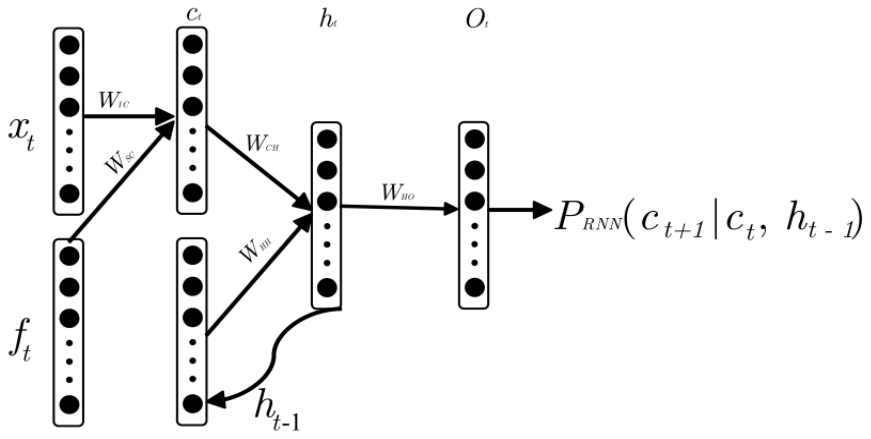

图 6-2 考虑后缀接续虚词的方法

具体地，是加入suffix 特征，网络的输入层为c_t，隐层为h_t，输出层为o_t。输入层c_t是表示t时刻的输入， o_t是RNNLM输出，它是语言模型的概率计算每个词$t+1$时刻赋予历史的词序列 $< w_{i-1} \cdots w_1 >$，所得到的softmax函数。输入层、隐层和输出层在RNNLM计算过程如下：

$$c_t = f(W_{IC}x_t + W_{SC}s_t) \tag{6-3}$$

$$h_t = f(W_{CH}c_t + W_{HH}h_{t-1}) \tag{6-4}$$

$$o_t = g(W_{HO}h_t) \tag{6-5}$$

公式6-3中，W_{IC} 表示字权重，W_{SC}表示后缀特征权重，用激活函数输入的字和后缀特征权重融合。这里W_{SC}是处理藏语语法中10个后缀。公式6-4为隐藏层h_t，用激活函数对输入层c_t时刻W_{CH}权重和隐藏层 h_{t-1}时刻的W_{HH}权重相加，用激活函数保留上下文信息，最后进行输出。这种方法我们称为RNNLM_ 藏语字符后缀单元（Tibetan Radical Suffix Unit，TRSU）。

6.3.3 形态动词相关的藏语语言模型

稀有词处理是语言模型研究中常见的一个问题。在[5]中通过将传统的词嵌入（Word Embedding）级降到Character-level，避免了大规模的嵌入计算和罕见词（Rare Words）的问题，通过Highway Network技术构建更深的网络，得到了不错的结果。这种方法虽然获取了更多字的形态结构信息，但是因为形态动词在语料中出现的较少，所以无法获取这类词。

在[32]中，通过利用形态结构在输入层和输出层的不同用途，一个明显的缺点是变形需要额外的工具来预处理。在数据匮乏时，因为存在数据上Out-Of-Domain，所以我们应用较大的数据对小数据进行自适应，解决数据量不足的缺陷[43,88]。但是因为形态动词对句子的时态和语义的影响，所以需要考虑藏语中形态动词。在形态上进行分析，可以根据藏语形态特点，切分出藏语中形态动词。

通过字频重新计算 藏语形态动词是藏语动词中形态特点较明显的词类，藏语语言学家已经研究和总结出形态动词表[39,57,100]。这类词在我们的训练语料中有很重要的作用，它对句子时态关系和语义都会产生影响。由此，我们提出基于词频的形态动词加权方法。

从公式6-2中 $P(c_i|v_{i-1})$ 我们可以看出，在预测w_i 词时需要从c_i词类和 v_{i-1} 上下文信息来获取，而c_i是根据训练语料中的词频获取。c_i是将训练语料中的w_i词进行分类，这样做的目的是为了能够快速准确获取w_i词。根据藏语形态动词特点，对训练语料中的形态动词增加字频，让w_i词分到词较少的类中，这样在预测时获取的速度更快，而且增加了准确性。

为了增加形态动词的词频，我们将生成藏语形态动词表，计算词频时根据形态动词表中出现的顺序，每出现形态动词都翻倍加，生成一个新类。统计字频时只对形态动词加，其它字根据数据库正常统计。这样就可以将一些原来在

低频率类形态动词划分到较高新类中，提高预测速度和精度。这种方法我们称为RNNLM_字频加权法（Character Frequency Recounting, _CFR）。

调整判别权重 这种方法是把RNNLM的输出做一个反向目标词的加权，来影响对类的划分。具体的，是将RNNLM的输出给一个阈值，这里的阈值我们是根据输出的概率给定，生成一个二值化的输出，公式6-6；

$$\bar{P}_{RNN}(w_i|v_{i-1}) = P_{RNN}(w_i|v_{i-1}) > \varepsilon. \tag{6-6}$$

根据藏语形态动词表，对公式 $\bar{P}_{RNN}(w_i|v_{i-1})$ 中不属于形态动词的值都设置0，保留形态动词值。然后，返回到RNNLM点乘生成一个新的类。最后，根据原来的和新生成的公式，重新生成新的类，因为我们只对形态动词对类进行操作，对一些低频词进行加权，将其映射到 $P(c_i|v_{i-1})$并获得

$$\tilde{P}(c_i|v_{i-1}) = \frac{\tilde{P}_{RNN}(w_i|v_{i-1})}{P(w_i|c_i, v_{i-1})}. \tag{6-7}$$

我们将其与 $P(c_i|v_{i-1})$ 结合生成一个新的$\tilde{P}(c_i|v_{i-1})$

$$\bar{P}(c_i|v_{i-1}) = P(c_i|v_{i-1}) + \alpha\tilde{P}(c_i|v_{i-1}), \tag{6-8}$$

这种方法是反向以形态动词来影响类，我们称为RNNLM_判别权重（Tuning Discriminative Weights, _TDW)

6.3.4 静态和动态结构相结合的语言模型

考虑语法的藏语语言模型主要考虑的是藏语语法对模型的影响，我们将语法信息融入RNN的输入，增加了藏语语法对预测下一个词的影响。而对于形态动词对词类的影响，我们提出了通过字频计算和调整判别权重两种方法。

为了融入语言特点，本文将语法和形态动词关系进行了融合，提出了考虑语法和形态动词的藏语语言模型。具体地，考虑语法关系基础上融合形态动词提出的通过词频调整权重和判别加权方法，在现有的预料中获取到更有用的信息，解决藏语中数据稀疏问题。

考虑语法关系融合词频调整权重的藏语语言模型，如图6-3中，我们在输入端加入了后缀信息，而且在预处理时将形态动词进行加权，不但融入了语法信息，而且提高形态动词在词类中的权重，改变形态动词在词类中的分类，解决低频词和罕见词在词类中的分布。

考虑语法关系融合调整判别权重的藏语语言模型，如图6-4中，这个方法在输入端加入语法信息，在加权时，我们根据藏语形态动词，除了形态动词值，其他

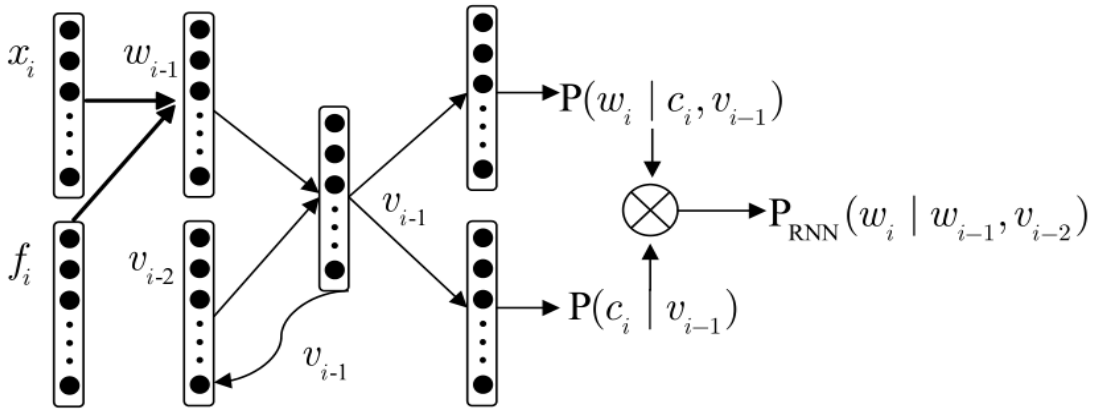

图 6-3　　基于词频考虑后缀接续虚词的方法

都设个阈值0，保留形态动词值，然后点乘生成一个新类，根据最新生成的类和原来输出的类进行融合生成新的词类。

$$P_{RNN}(w_i|w_{i-1}, v_{i-2}) = P(w_i|c_i, v_{i-1})\bar{P}(c_i|v_{i-1}). \tag{6-9}$$

公式6-9是对公式6-2的改进，具体地，在获取 w_i时，我们融入了基于后缀的信息，而对原有的 $P(c_i|v_{i-1})$进行了形态动词加权。对公式6-8中原有的词类和形态动词加权的词类进行了融合，生成新的词类，将这种方法应用到我们的模型。

6.4　实验结果与分析

在实验部分，数据我们应用了藏语拉萨方言的音频语料库和西藏新闻语料库（TNC）。然后给出了现有的最新方法和我们提出的考虑语法、形态动词加权和相结合的方法的对比，比较了现有的方法和我们的方法在同一域、不同域和多域的数据上进行Perplexity（PPL）对比。且将基准(Baseline)和我们的方法都和传统的N-gram方法进行插值进行对比。实验中，我们把kneser ney平滑的3-gram 表示为KN3。最后，将我们的方法应用到ASR中字符错误率（CER）作为评估标准，验证我们的方法的有效性。

本文介绍了藏语语法和形态动词加权对藏语语言模型的影响，提出了考虑语法和形态动词加权的语言模型。在实验中，我们比较了传统方法和最新的方法，我们将藏语模型研究方法_TRC为基准，提出了考虑语法和形态动词加权的模型，以及语法和形态动词加权相结合的藏语语言模型。参数设置基于CUED-

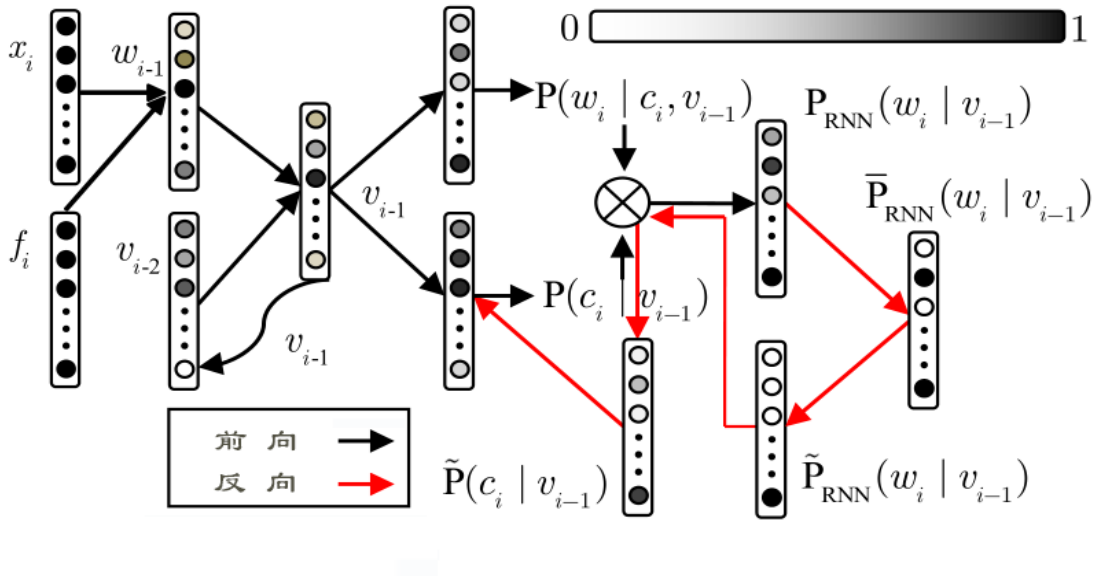

图 6-4 静态和和动态形态结构的藏语语言模型

表 6-1 已有的方法和我们的方法相同域上 PPL 对比

Granularity	Model	News
character	N-gram[23]	55.2
	RNNLM[23]	62.9
	CUED_RNNLM[94]	58.4
	LSTM[94]	55.9
	CharCNN[5]	55.2
radical	_TRU[78]	57.6
	_TRD[78]	54.3
	_TRC[78]	53.8
	TRSU	51.2
morph	_CFR[137]	50.6
	_TDW[137]	49.8
radical+morph	**TRSU +_CFR**	**49.5**
	TRSU +_TDW	**48.5**

RNNLM方法[94,103,140]。

由于我们的数据集中的测试集与新闻相关，因此我们对实验中的相同域，不同域和所有域的数据集进行了Perplexity, PPL）对比，以验证我们的方法的有效性。最后，在ASR上进行了CER评价，也取得了较好的效果。

6.4.1 困惑度评价

表 6-2 已有的方法和我们的方法差值后相同域上 PPL 对比

Granularity	Model + KN3	News
character	N-gram[23]+ KN3	55.2
	RNNLM[23]+ KN3	48.2
	CUED_RNNLM[94] + KN3	48.0
	LSTM[94]+ KN3	46.4
	CharCNN[5] + KN3	47.3
radical	_TRU[78] + KN3	47.9
	_TRD[78] + KN3	47.0
	_TRC[78] + KN3	46.9
	TRSU + KN3	45.1
morph	_CFR[137]+ KN3	45.1
	_TDW[137]+ KN3	44.2
radical+morph	**TRSU +_CFR + KN3**	**44.9**
	TRSU +_TDW + KN3	**43.8**

由于我们的数据集中的测试集与新闻相关，我们根据测试集在实验中使用相同域，不同域和所有域的数据集进行实验，以验证我们的方法的有效性。由于数据量不足，为了更加说明我们方法的有效性，使用线性插值方法对所有的方法和N-gram进行了插值。在本文中，我们设置lambda取0.5[78,94]。

相同域的数据集 因为测试集域偏向新闻，所以我们需要验证同一域下我们的方法。在表6-1中，我们将实验分为四个部分。第一部分是现有语言模型研究的方法（传统方法和一些新方法），第二部分是藏语模型研究的最新方法，第三部分是我们加入藏语语法信息的方法，最后是形态动词加权处理低频词的模型，以及语法和形态动词加权相结合的藏语语言模型。我们先将现有最新方法与我们提出的

方法进行比较，与现有语言模型相比，我们提出的语法、形态动词加权方法和二者相结合的方法PPL分别相对降低了约4.8%，8%和9.8%。

表 6-3 不同域上已有的方法和我们的方法的PPL对比

Granularity	Model	Edu	Law	Bud	Cul	Lit	Wikipedia
character	N-gram[23]	122.3	**288.8**	**497.5**	170.4	132.9	254.1
	RNNLM[23]	147.9	374.1	698.9	196.2	155.9	218.7
	CUED_RNNLM[94]	139.8	367.8	655.1	179.3	123.2	189.9
	LSTM[94]	127.2	364.1	618.4	169.8	118.2	160.2
	CharCNN[5]	123.7	366.5	602.4	168.1	116.5	159.8
radical	_TRU[78]	131.2	364.4	702.3	171.1	118.1	162.8
	_TRD[78]	125.8	367.4	596.3	167.9	113.8	157.2
	_TRC[78]	122.3	356.6	590.4	175.1	114.9	152.7
	RESU	117.6	354.1	549.9	165.2	109.8	152.1
morph	_CFR[137]	117.6	350.4	542.5	163.7	109.5	151.9
	_TDW[137]	116.9	348.6	540.2	161.8	109.5	150.5
radical+morph	**TRSU +_CFR**	**115.6**	344.9	529.8	**159.1**	**108.7**	**146.9**
	TRSU +_TDW	**114.1**	341.8	521.2	**157.9**	**105.2**	**140.5**

我们将最新方法与我们提出的方法在不同粒度进行比较，粒度为部件（Radical）提出的方法相比字（Character）方法PPL降低7.3%，而粒度为语素方法比字（Character）和部件（Radical）方法PPL分别相对降低了9.8%和7.4%。部件和语素相结合方法比字和部件方法PPL分别相对降低了约12.3%和9.9%。由此，可以看出，我们提出的方法在同一域的数据上具有较好的效果，获取了更多语言信息。

从表6-2中我们可以看出，我们的方法在插值后取得了良好的结果。由此可见，插值确实提高了我们的方法的准确性。基准方法和我们提出的方法对比（语法、形态动词和相结合）PPL分别相对降低了约3.8%，5.8%和6.6%。我们对不同粒度进行比较，我们粒度为部件和语素提出的方法相比字方法PPL分别降低2.8%和4.7%，部件和语素相结合方法比字和部件方法PPL分别相对降低了

约5.6%和7.1%。

由此可见，同一域的数据下我们提出的方法能够获取更加有用的信息，而且从实验结果可以看出，这些获取的语法和形态动词对藏语句子是有影响，能够提升模型的预测能力。

不同域的数据集 作为新闻为域的测试集，我们为了验证我们提出的方法的有效性，在各个域的数据上进行了验证，可以看出我们的方法仍然比传统方法都有提升。表6-3中，在教育，文化和文学方面，我们的方法比基准方法PPL分别相对降低了6.7%、9.8%和8.5%。在多域交织的维基百科数据上，我们的方法也取得了很好的效果，比基准方法PPL相对提高了9.8%。

表 6-4　不同域上已有的方法和我们的方法差值后PPL对比

Granularity	Model+KN3	Edu	Law	Bud	Cul	Lit	Wikipedia
character	N-gram[23]	122.3	288.8	497.5	170.4	132.9	254.1
	RNNLM[23]	116.5	277.5	456.5	161.5	123.4	169.5
	CUED_RNNLM[94]	110.4	272.5	427.2	153.2	112.6	151.2
	LSTM[94]	107.3	274.8	419.5	143.3	107.8	139.1
	CharCNN[5]	103.8	276.5	410.4	143.0	104.7	138.8
radical	_TRU[78]	102.7	251.2	419.3	140.8	103.9	138.5
	_TRD[78]	101.8	250.8	402.6	141.4	103.5	136.2
	_TRC[78]	99.9	249.2	396.7	141.2	102.9	135.8
	RESU	97.3	247.1	383.1	137.9	100.4	133.9
morph	_CFR[137]	98.6	248.2	385.4	139.4	101.8	133.8
	_TDW[137]	97.3	247.5	383.2	137.6	100.7	132.5
radical+morph	**TRSU +_CFR**	**96.6**	**246.6**	**381.7**	**137.3**	**99.5**	**131.6**
	TRSU +_TDW	**94.1**	**237.7**	**375.3**	**133.1**	**97.1**	**129.9**

表6-3还显示PPL因域而不一致导致的问题。在宗教和法律域数据上没有传统的N-gram方法，这表明我们的测试集与这些数据域不一致，而且在这些域上存在专业词汇和罕见词等现象。所以，虽然我们的方法从上下文和形态结构信息中有一定的改进，但也存在数据域问题。

表 6-5 综合数据集上的PPL对比

Granularity	Model	ALL
character	N-gram[23]	98.6
	RNNLM[23]	92.5
	CUED_RNNLM[94]	89.2
	LSTM[94]	84.1
	CharCNN[5]	98.1
radical	_TRU[78]	88.1
	_TRD[78]	87.5
	_TRC[78]	86.7
	TRSU	82.6
morph	_CFR[137]	83.3
	_TDW[137]	82.2
radical+morph	**TRSU +_CFR**	**81.7**
	TRSU +_TDW	**80.4**

表6-4是可以看出，我们的方法和N-gram方法插值后取得了好的结果。基准方法和我们提出的方法对比（语法、形态动词和相结合）PPL分别相对降低了约4.3%-5.9%不等。我们对不同粒度也进行了对比，PPL相对基准方法降低了9.3%-12.6%不等。总体上差值以后的结果都取得了提升，也说明差值补充了缺乏的信息。

综合数据集 在整个多域的数据集中，与基准方法相比，我们的方法PPL降低了7.2%-8.9%。从表6-5中的结果可以看出，我们的方法也在包括所有域的整个数据集上实现了最佳结果，证明了我们的方法对藏语模型有效。

由此可见，我们提出的考虑语法、形态动词和两者相结合的方法都有效补充语料不足的问题，不但解决了数据稀疏问题，而且还考虑了藏语语法对句子的影响，以及藏语形态动词对低频词作用。可以很好解决句子的中的时态关系和语义问题。

6.4.2 ASR evaluation

上述实验基于PPL作为验证结果的评估标准。我们可以看到，我们的方法取得了更好的结果。但是为了验证我们的方法能够在语音识别识别率上有提升，我们在西藏拉萨方言音频数据集中对ASR实验模型进行了验证[137]。

表 6-6 已有方法和我们的方法在%CER上的对比

Granularity	Model	%CER
character	N-gram[23]	35.20
	RNNLM[23]	34.60
	CUED_RNNLM[94]	34.25
	LSTM[94]	33.96
	CharCNN[5]	34.03
radical	_TRU[78]	34.09
	_TRD[78]	34.15
	_TRC[78]	33.94
	TRSU	33.35
morph	_CFR[137]	33.55
	_TDW[137]	33.10
radical+morph	**TRSU +_CFR**	**32.85**
	TRSU +_TDW	**32.55**

从表6-6中可以看出，在字粒度上RNNLM方法比传统的N-gram方法取得了提升。在部件粒度上针对藏语做的加入语法方法取得了最好的结果。使用语法的TRSU比RNNLM相对改善字错误率为3.6%。使用基于形态动词的_TDW比RNNLM相对改善从4.3%，使用TRSU +_TDW比RNNLM相对提高5.9%。所以我们提出的方法相对于最新的方法处理藏语语音识别取得较好的效果。

我们知道lattice是在语音识别过程中解码一次的结构，其包含大量候选结果。由于神经网络使用历史信息来预测下一个单词，因此对lattice进行重新调整会导致搜索速度变慢。与lattice的单词结构相比，N-best更适合于远程信息的模型扩展。本文使用N-best的中间结果进行重新校正[125]，如表6-7所示。

表6-7是N-best 重打分计算%CER的结果。我们的方法在N-best（n = 100和n =

表 6-7 N-best重打分的%CER对比

N	# hidden units	%CER with N-best rescoring									Original
		RNNLM	_TRU	_TRD	_TRC	TRSU	_CFR	_TDW	**TRSU+_CFR**	**TRSU+_TDW**	
100	500	33.84	34.22	34.07	34.02	33.55	33.65	33.20	**33.06**	**33.02**	35.20
	600	34.03	33.99	34.06	34.05	33.50	33.55	33.10	**32.85**	**32.68**	
	700	33.97	34.09	34.15	33.94	33.42	33.55	33.03	**32.65**	**32.45**	
1000	500	33.73	34.15	33.92	34.02	33.12	32.87	32.71	**32.75**	**32.65**	
	600	33.88	33.87	34.06	33.82	32.95	32.87	32.65	**32.58**	**32.23**	
	700	33.83	34.05	34.08	33.78	32.75	32.74	32.55	**32.42**	**32.12**	

1,000）重新调整中减少了约4.9%，表明我们的方法对藏语语言模型中语法和稀有词的加权方法有很好的效果。

6.4.3 分析

针对语法对模型的影响，本文将_TRC作为基准，_TRC是对Tibet 字中每个部件作为特征加入RNN中，但是不是每个部件都具有意义，会出现冗余信息，而且捕捉不到语法信息。由此可以看出不是所有的部件都对字有影响。本文提出的方法是针对虚词，可以解决接续关系和句子语义准确表达。

TRSU方法是将藏语中后加字和后后加字提取出，作为特征加入RNN中。这种方法是可以将提取出后缀来判断虚词的接续关系，将解决虚词接续问题。这样就可以有效的解决虚词的接续关系以及准确预测句子语义表达。实验中可以看出，考虑语法的TRSU方法比基准方法取得了更好的效果，说明语法信息能够解决语料匮乏带来的数据稀疏问题。

对于罕见词或低频词的处理，之前的研究都处理都不一样，有的是基于稀有词或罕见词的加权方法，有的是自适应的方法[137,141]。这些方法存在的问题是对所有Rate Word处理是否有意义。所以我们根据藏语的形态动词这一特点，提出形态动词藏语语言模型。利用形态动词加权不但可以影响类型的变化，还可以提高预测能力。实验中可以验证，_CFR和_TDW方法比基准方法都取得了提升，说明这类词影响模型的预测能力。

对于考虑语法信息和形态动词加权相结合的藏语语言模型，我们因为对词类加权方法不同，所以我们在俩个加权方法上融入语法信息提出TRSU

+_CFR和TRSU +_TDW方法。可以看出，这两种方法比考虑语法的TRSU和形态动词加权的_CFR和_TDW方法都取得了好的效果。

综上所述，我们提出考虑语法的TRSU和形态动词加权的_CFR和_TDW方法都取得了较好的效果。补充了语言本身的一些形态结构和语法信息，能够更加准确表达句子的语义表达。

6.5 本章小结

在本节中，根据形态结构关系，我们的工作对[78]基础上考虑了后缀接续虚词的特征，强化了语法对句子的影响，通过RNNLM方法，对考虑语法特征和形态动词加权方法进行了融合。发现这两种方法是互补的，两种类型的适应性的组合通常可以改善性能。对于考虑语法的适应，使用后缀特征作为RNNLM的增强特征输入有效，且取得了好的效果，还研究了形态动词加权方法的使用，使句子的语义更加准确。将语法和形态动词加权方法相结合的方法获取了更多信息，且具有更好的效果。实验表明，我们提出了方法，利用语法和形态动词加权方法的优势，在同一域、不同域和综合数据上获得最佳性能。此外，我们的方法比基准方法提高了ASR性能。

第7章 总结与展望

本文主要研究了藏语语言模型以及在藏语拉萨方言语音识别处理中的应用，其中介绍了藏语拉萨方言的音频库和文本语料库。本文从藏语特有的一些特点，针对目前低资源语言存在的数据稀疏问题，通过引入藏语语法，藏语形态动词，语法和形态动词相结合作为补充，实现精度高、鲁棒性好的藏语语言模型，也为在以后的藏语自然语言处理任务上研究提供了思路。下面对本文主要工作进行总结，并展望未来工作。

7.1 研究工作的总结

本文的主要研究内容是通过对藏语语言模型的改进，提高拉萨方言语音识别系统的识别率，从新闻语料为主题的文本语料库构建了藏语语言模型，分析了现有语言模型中的各种算法以及语音识别的识别率影响，提出了具有藏语语法和形态结构的藏语语言模型。具体包括了以下几方面的工作：

- 音频和文本语料库的构建。根据藏语音素平衡关系，我们选取了拉萨方言语音语料库和构建以新闻为域的藏文文本语料，并进行了验证。根据藏语音素平衡关系，我们选取了3123句藏语句子构建了拉萨方言语音语料库，并提出新的音素集来验证我们构建的音频数据库符合我们研究的需求；构建以新闻为域的藏文文本语料，提出了组合基字（Combination Tibetan Radical, CTR）的卷积神经网络（CNN）藏语语言模型，可以测试和验证我们构建的语料库。在拉萨方言语音识别过程中，语音信号是经过特征提取获取的，在声学模型和语言模型上进行识别和解码，最后得到识别的文本输出。
- 以静态形态结构对虚词的影响进行了深入的研究。由于低资源语言存在数据稀疏问题和语言的差异性，不是所有通用的方法都实用，由此，我们根据藏语特有的语法关系，提出了基于后缀对虚词判定的方法，补充了语法信息对于低资源语言的数据稀疏问题。

- 利用动态形态结构对句子的影响。稀有词是语言模型任务存在的一个问题，怎样有效利用或获取这类的信息是一个重要任务。藏语中形态动词的屈折变化对句子有影响，尤其是在语音识别中的同音字预测错误的可能性就较大，而且这种词的预测错误会导致句子语义的变化。我们统计发现在藏语中这些稀有词中含有形态动词，而实际语料中因语料资源缺乏，无法准确获取到这类词的概率分布，所以我们有必要对这类词进行加权来获取它的信息。因此，我们通过字频重新计算形态动词来调整判别权重，提出了形态特征的藏语语言模型。
- 融合静态和动态形态结构。通过对考虑特征和形态动词加权方法进行了融合。对于后缀对虚词的接续关系，使用RNNLM的增强特征输入有效，且加上形态动词加权方法的使用，有效补充和弥补了有用的信息。将语法特征和形态动词加权方法相结合互补了更多信息，且取得了更好的效果。

7.2 未来展望

本文虽然对藏语语言模型做出了一定的研究，但是其中还存在着许多的不足。语言模型虽然是一个基础性研究，但是对于语音识别、手写体识别、机器学习和其他自然语言处理等，都具有重要意义。由于藏文信息处理起步较晚，加上研究人员稀少，对其研究还处在起步阶段。下一步我们还需要做以下几个方面的工作：

- 由于语料的有限，现有的字无法表示所有字。虽然之前我们是以字作为单元进行研究的，但是由于字信息有限，因此，需要全面考虑藏语语言模型，以字、词、短语以及句子等信息。
- 对于一些最新的方法如希望在以后的工作中对ELMO[142]和BERT[143–145]等希望能够在今后的工作中加以融入，以能够进一步提高藏语语言模型的研究。
- 进一步优化藏语语言模型。藏语语言模型不是独立的存在，在语音识别任务中，要和声学模型、解码过程相结合，研究它们之间存在的联系。在实际任务中，则可以在自动语音识别（ASR）上取得更好的结果。尽管本文涉及到一些内容，但是我们仍然需要更多的研究以使其语法和语义效果标准化。我们希望在此基础上，将来藏语语言模型方法在其他相关任务也能

表现得更好。

近年来，虽然藏语的语音识别技术取得了一定的发展，但是还有许多的研究工作要做。希望在接下来的研究中，我们也多粒度以及语义相关信息进行研究，进一步提高藏语语言模型的性能。同时，我也希望信息技术能够影响一些学习者，有更多的科研人员加入藏文信息处理团队中，为藏语信息化做出一些贡献。

参考文献

[1] Ethnologue. OL [J]. http://www.ethnologue.coHL.

[2] summer Institute for Linguistics Ethnologue survey. OL [J]. https://afrobranding.wordpress.com/tag/summer-institute-for-linguistics-sil-ethnologue-survey, 1999.

[3] Besacier L, Barnard E, Karpov A, *et al*. Automatic speech recognition for under-resourced languages: A survey [J]. Speech Communication, 2014, 56: 85–100.

[4] Jozefowicz R, Vinyals O, Schuster M, *et al*. Exploring the limits of language modeling [J]. arXiv preprint arXiv:1602.02410, 2016.

[5] Kim Y, Jernite Y, Sontag D, *et al*. Character-Aware Neural Language Models. [C]. In AAAI, 2016: 2741–2749.

[6] Hwang K, Sung W. Character-level language modeling with hierarchical recurrent neural networks [C]. In Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on, 2017: 5720–5724.

[7] 韩纪庆. 语音信号处理 [M]. 清华大学出版社, 2004.

[8] Brown P F, Cocke J, Della Pietra S A, *et al*. A statistical approach to machine translation [J]. Computational linguistics, 1990, 16 (2): 79–85.

[9] Zhai C, Lafferty J. A study of smoothing methods for language models applied to information retrieval [J]. ACM Transactions on Information Systems (TOIS), 2004, 22 (2): 179–214.

[10] Kuhn R, De Mori R. A cache-based natural language model for speech recognition [J]. IEEE transactions on pattern analysis and machine intelligence, 1990, 12 (6): 570–583.

[11] Hasan T, Abdelwahab M, Parthasarathy S, *et al*. Automatic composition of broadcast news summaries using rank classifiers trained with acoustic and lexical features [C]. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016: 6080–6084.

[12] Witten I H, Bell T C. The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression [J]. Ieee transactions on information theory, 1991, 37 (4): 1085–1094.

[13] Good I J. The population frequencies of species and the estimation of population parameters [J]. Biometrika, 1953, 40 (3-4): 237–264.

[14] Kneser R, Ney H. Improved backing-off for m-gram language modeling [C]. In 1995 International Conference on Acoustics, Speech, and Signal Processing, 1995: 181–184.

[15] Kurland O, Krikon E. The opposite of smoothing: a language model approach to ranking query-specific document clusters [J]. Journal of Artificial Intelligence Research, 2011, 41: 367–395.

[16] Graves A, Mohamed A-r, Hinton G. Speech recognition with deep recurrent neural networks [C]. In 2013 IEEE international conference on acoustics, speech and signal processing, 2013: 6645–6649.

[17] Srivastava N, Hinton G, Krizhevsky A, *et al*. Dropout: a simple way to prevent neural networks from overfitting [J]. The journal of machine learning research, 2014, 15 (1): 1929–1958.

[18] Bengio Y, Ducharme R, Vincent P, *et al*. A neural probabilistic language model [J]. Journal of machine learning research, 2003, 3 (Feb): 1137–1155.

[19] Lukoševičius M, Jaeger H. Reservoir computing approaches to recurrent neural network training [J]. Computer Science Review, 2009, 3 (3): 127–149.

[20] Kombrink S, Mikolov T, Karafiát M, *et al*. Recurrent neural network based language modeling in meeting recognition [C]. In Twelfth annual conference of the international speech communication association, 2011.

[21] Mikolov T, Karafiát M, Burget L, *et al*. Recurrent neural network based language model [C]. In Eleventh annual conference of the international speech communication association, 2010.

[22] Mikolov T, Kombrink S, Burget L, *et al*. Extensions of recurrent neural network language model [C]. In 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011: 5528–5531.

[23] Mikolov T. Statistical language models based on neural networks [J]. Presentation at Google, Mountain View, 2nd April, 2012, 80.

[24] Hochreiter S, Schmidhuber J. Long short-term memory [J]. Neural computation, 1997, 9 (8): 1735–1780.

[25] Cho K, Van Merriënboer B, Bahdanau D, *et al*. On the properties of neural machine translation: Encoder-decoder approaches [J]. arXiv preprint arXiv:1409.1259, 2014.

[26] Chung J, Gulcehre C, Cho K, *et al*. Empirical evaluation of gated recurrent neural networks on sequence modeling [J]. arXiv preprint arXiv:1412.3555, 2014.

[27] Mousa A E-D, Kuo H-K J, Mangu L, *et al*. Morpheme-based feature-rich language models using deep neural networks for LVCSR of Egyptian Arabic [C]. In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013: 8435–8439.

[28] Shi Y, Wiggers P, Jonker C M. Towards recurrent neural networks language models with linguistic and contextual features [C]. In Thirteenth Annual Conference of the International Speech Communication Association, 2012.

[29] Mousa A E-D, Schlüter R, Ney H. Investigations on the use of morpheme level features in language models for Arabic LVCSR [C]. In 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012: 5021–5024.

[30] He Y, Hutchinson B, Baumann P, *et al*. Subword-based modeling for handling oov words inkey-word spotting [C]. In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014: 7864–7868.

[31] He T, Xiang X, Qian Y, *et al*. Recurrent neural network language model with structured word embeddings for speech recognition [C]. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015: 5396–5400.

[32] Fang H, Ostendorf M, Baumann P, *et al*. Exponential language modeling using morphological features and multi-task learning [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2015, 23 (12): 2410–2421.

[33] Miyamoto Y, Cho K. Gated word-character recurrent language model [J]. arXiv preprint arXiv:1606.01700, 2016.

[34] Li G, Yu H, Zheng T F, *et al*. Free linguistic and speech resources for tibetan [C]. In 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2017: 733–736.

[35] Rinchenskid r t. A study of Tibetan language model [J]. China Computer & Communication, 2015.

[36] Wang W, *et al*. Formation of standard Tibetan syllables and comparison as well as analysis of the statistical results [C]. In 2008 IEEE Conference on Cybernetics and Intelligent Systems, 2008: 379–384.

[37] Li J, Wang H, Wang L, *et al*. Exploring tonal information for lhasa dialect acoustic modeling [C]. In 2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP), 2016: 1–5.

[38] Shen T, Wang L, Chen X, *et al*. Tibetan language model based on recurrent neural network [C]. In ISSP, 2017.

[39] Xiarong T. Detailed Explanation about Tibetan grammar [M]. 1954.

[40] Tharrgyal L. A Study of Tibetan Grammar [M]. 2008.

[41] Tournadre N. The Classical Tibetan cases and their transcategoriality: From sacred grammar to modern linguistics [J]. Himalayan Linguistics, 2010, 9 (2).

[42] 宗成庆. 统计自然语言处理 [M]. 清华大学出版社, 2013.

[43] Shen T, Wang L, Chen X, *et al*. Investigation of long short-term memory for tibetan language model [C]. In NCMMSC, 2017.

[44] Clarkson P R. Adaptation of statistical language models for automatic speech recognition [D]. [S. l.]: Citeseer, 1999.

[45] Raju A, Filimonov D, Tiwari G, *et al*. Scalable Multi Corpora Neural Language Models for ASR [J]. arXiv preprint arXiv:1907.01677, 2019.

[46] Kneser R, Steinbiss V. On the dynamic adaptation of stochastic language models [C]. In 1993 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1993: 586–589.

[47] Rosenfeld R. A maximum entropy approach to adaptive statistical language modeling [J], 1996.

[48] Berger A L, Pietra V J D, Pietra S A D. A maximum entropy approach to natural language processing [J]. Computational linguistics, 1996, 22 (1): 39–71.

[49] 黄非. 面向中文语音识别的自适应语言模型研究 [J], 1999.

[50] 曲卫民，张俊林，孙乐，等. 基于记忆的自适应汉语语言模型的研究 [J]. 中文信息学报，2003，17 (5)： 14–19.

[51] 张俊林，孙乐，孙玉芳. 一种改进的基于记忆的自适应汉语语言模型 [J]. 中文信息学报，2005，19 (1)： 9–14.

[52] Beck E, Zhou W, Schlüter R, *et al*. LSTM Language Models for LVCSR in First-Pass Decoding and Lattice-Rescoring [J]. arXiv preprint arXiv:1907.01030, 2019.

[53] Gyumsto M, Ney H. Open vocabulary speech recognition with flat hybrid models [C]. In Ninth European Conference on Speech Communication and Technology, 2005.

[54] Gerz D, Vulić I, Ponti E M, *et al*. Language Modeling for Morphologically Rich Languages: Character-Aware Modeling for Word-Level Prediction [J]. Transactions of the Association for Computational Linguistics, 2018, 6: 451–465.

[55] Matthews A, Neubig G, Dyer C. using morphological knowledge in open vocabulary neural language models [C]. In NAACL HLT 2018: 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018: 1435–1445.

[56] 陈玉忠，俞士汶. 藏文信息处理技术的研究现状与展望 [J]. 中国藏学，2003 (4)： 97–107.

[57] Gyumsto S L T. Tibetan grammatical theories by Seduo [M]. 1957.

[58] Garofalo J S, Lamel L F, Fisher W M, *et al*. The DARPA TIMIT acoustic-phonetic continuous speech corpus cdrom [J]. Linguistic Data Consortium, 1993.

[59] Paul D B, Baker J M. The design for the Wall Street Journal-based CSR corpus [C]. In Proceedings of the workshop on Speech and Natural Language, 1992: 357–362.

[60] Godfrey J J, Holliman E C, McDaniel J. SWITCHBOARD: Telephone speech corpus for research and development [C]. In [Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1992: 517–520.

[61] 张金溪，李照耀，肖俊生，等. 面向语音合成的藏语音素切分算法研究 [J]. 西北民族大学学报（自然科学版），2016.

[62] 刘晓凤. 藏语语音深度特征提取及语音识别研究 [D]. [S. l.]: 中央民族大学, 2016.

[63] 黄晓辉，李京．基于循环神经网络的藏语语音识别声学模型 [D]．[出版地不详]：中文信息学报，2018.

[64] Zhang X, Wang B, Wu Q, *et al*. Prosodic Realization of Focus in Statement and Question in Tibetan (Lhasa Dialect) [C]. In Thirteenth Annual Conference of the International Speech Communication Association, 2012.

[65] Li G Y, Yu H Z. Large-Vocabulary Continuous Speech Recognition of Lhasa Tibetan [C]. In Applied Mechanics and Materials, 2014: 802–806.

[66] Zhao Y, Cao Y, Pan X. Tibetan Language Continuous Speech Recognition Based on Dynamic Bayesian Network [C]. In 2009 Fifth International Conference on Natural Computation, 2009: 91–94.

[67] Li J, Mohamed A, Zweig G, *et al*. Exploring multidimensional LSTMs for large vocabulary ASR [C]. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016: 4940–4944.

[68] Dahl G E, Yu D, Deng L, *et al*. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition [J]. IEEE Transactions on audio, speech, and language processing, 2011, 20 (1): 30–42.

[69] Hinton G, Deng L, Yu D, *et al*. Deep neural networks for acoustic modeling in speech recognition [J]. IEEE Signal processing magazine, 2012, 29.

[70] Seide F, Li G, Yu D. Conversational speech transcription using context-dependent deep neural networks [C]. In Twelfth annual conference of the international speech communication association, 2011.

[71] Yu D, Seltzer M L, Li J, *et al*. Feature learning in deep neural networks-studies on speech recognition tasks [J]. arXiv preprint arXiv:1301.3605, 2013.

[72] Cho E, Kumar S. A Conversational Neural Language Model for Speech Recognition in Digital Assistants [C]. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018: 5784–5788.

[73] Wang H, Khyuru K, Li J, *et al*. Investigation on acoustic modeling with different phoneme set for continuous Lhasa Tibetan recognition based on DNN method [C]. In 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2016: 1–4.

[74] 才旦夏茸. 藏文文法 [M]. 2005.

[75] Povey D, Ghoshal A, Boulianne G, *et al*. The Kaldi speech recognition toolkit [C]. In IEEE 2011 workshop on automatic speech recognition and understanding, 2011.

[76] Hinton G E, Osindero S, Teh Y-W. A fast learning algorithm for deep belief nets [J]. Neural computation, 2006, 18 (7): 1527–1554.

[77] Rath S P, Povey D, Veselỳ K, *et al*. Improved feature processing for deep neural networks. [C]. In Interspeech, 2013: 109–113.

[78] Shen T, Wang L, Chen X, *et al*. Exploiting the Tibetan Radicals in Recurrent Neural Network for Low-Resource Language Models [C]. In International Conference on Neural Information Processing, 2017: 266–275.

[79] Hermans M, Schrauwen B. Training and analysing deep recurrent neural networks [C]. In Advances in neural information processing systems, 2013: 190–198.

[80] Yeh E T. Tibet and the problem of radical reductionism [J]. Antipode, 2009, 41 (5): 983–1010.

[81] LI Y-h, HE X-z, AI J-y, *et al*. Tibetan coding method and mutual conversion [J]. Journal of Computer Applications, 2009, 29 (7): 2016–2018.

[82] Lazaridou A, Marelli M, Zamparelli R, *et al*. Compositionally derived representations of morphologically complex words in distributional semantics [C]. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2013: 1517–1526.

[83] Luong T, Socher R, Manning C. Better word representations with recursive neural networks for morphology [C]. In Proceedings of the Seventeenth Conference on Computational Natural Language Learning, 2013: 104–113.

[84] Yildiz E, Tirkaz C, Sahin H B, *et al*. A morphology-aware network for morphological disambiguation [C]. In Thirtieth AAAI Conference on Artificial Intelligence, 2016.

[85] Bellegarda J R. Statistical language model adaptation: review and perspectives [J]. Speech communication, 2004, 42 (1): 93–108.

[86] Koehn P, Schroeder J. Experiments in domain adaptation for statistical machine translation [C]. In Proceedings of the second workshop on statistical machine translation, 2007: 224–227.

[87] Masumura R, Asami T, Oba T, *et al*. Domain adaptation based on mixture of latent words language models for automatic speech recognition [J]. IEICE TRANSACTIONS on Information and Systems, 2018, 101 (6): 1581–1590.

[88] Masumura R, Asami T, Oba T, *et al*. Viterbi Approximation of Latent Words Language Models for Automatic Speech Recognition [J]. Journal of Information Processing, 2019, 27: 168–176.

[89] Blunsom P, Cohn T. A hierarchical Pitman-Yor process HMM for unsupervised part of speech induction [C]. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, 2011: 865–874.

[90] Moore R C, Lewis W. Intelligent selection of language model training data [C]. In Proceedings of the ACL 2010 conference short papers, 2010: 220–224.

[91] Hinton G E, Srivastava N, Krizhevsky A, *et al*. Improving neural networks by preventing co-adaptation of feature detectors [J]. arXiv preprint arXiv:1207.0580, 2012.

[92] Zhang Y, Wallace B. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification [J]. arXiv preprint arXiv:1510.03820, 2015.

[93] Srivastava R K, Greff K, Schmidhuber J. Training very deep networks [C]. In Advances in neural information processing systems, 2015: 2377–2385.

[94] Chen X, Liu X, Qian Y, *et al*. CUED-RNNLM—An open-source toolkit for efficient training and evaluation of recurrent neural network language models [C]. In 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP), 2016: 6000–6004.

[95] Mikolov T, Zweig G. Context dependent recurrent neural network language model [C]. In 2012 IEEE Spoken Language Technology Workshop (SLT), 2012: 234–239.

[96] 吉太加. 现代藏语语法通论 [M]. 兰州:甘肃民族出版社, 2000.

[97] 钦饶威色. 三十颂详解 [M]. 1986.

[98] Bojanowski P, Joulin A, Mikolov T. Alternative structures for character-level RNNs [J]. arXiv preprint arXiv:1511.06303, 2015.

[99] Nuo M, Liu H, Ma L, *et al*. Automatic acquisition of Chinese-Tibetan multi-word equivalent pair from bilingual corpora [C]. In 2011 International Conference on Asian Language Processing, 2011: 177–180.

[100] Zhabsdrung T. Sumrtags kyi bshadpa Thonmivi Zhallung [M]. 1989.

[101] Pusateri E, Gysel C V, Botros R, *et al*. Connecting and Comparing Language Model Interpolation Techniques [J]. arXiv preprint arXiv:1908.09738, 2019.

[102] Chen T, Caseiro D, Rondon P. Entropy based pruning of backoff maxent language models with contextual features [C]. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018: 6129–6133.

[103] Chen X, Liu X, Gales M J, *et al*. Improving the training and evaluation efficiency of recurrent neural network language models [C]. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015: 5401–5405.

[104] Chen X, Liu X, Gales M J, *et al*. Recurrent neural network language model training with noise contrastive estimation for speech recognition [C]. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015: 5411–5415.

[105] Chen S F, Goodman J. An empirical study of smoothing techniques for language modeling [J]. Computer Speech & Language, 1999, 13 (4): 359–394.

[106] Roark B, Saraclar M, Collins M. Discriminative n-gram language modeling [J]. Computer Speech & Language, 2007, 21 (2): 373–392.

[107] Pascanu R, Mikolov T, Bengio Y. On the difficulty of training recurrent neural networks [C]. In International conference on machine learning, 2013: 1310–1318.

[108] Sundermeyer M, Ney H, Schlüter R. From feedforward to recurrent LSTM neural networks for language modeling [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2015, 23 (3): 517–529.

[109] Brown P F, Desouza P V, Mercer R L, *et al*. Class-based n-gram models of natural language [J]. Computational linguistics, 1992, 18 (4): 467–479.

[110] Irie K, Lei Z, Schlüter R, *et al*. Prediction of LSTM-RNN full context states as a subtask for n-gram feedforward language models [C]. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018: 6104–6108.

[111] Chen X, Liu X, Ragni A, *et al*. Future word contexts in neural network language models [C]. In 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 2017: 97–103.

[112] Katz S. Estimation of probabilities from sparse data for the language model component of a speech recognizer [J]. IEEE transactions on acoustics, speech, and signal processing, 1987, 35 (3): 400–401.

[113] Iyer R M, Ostendorf M. Modeling long distance dependence in language: Topic mixtures versus dynamic cache models [J]. IEEE Transactions on speech and audio processing, 1999, 7 (1): 30–39.

[114] Lee K, Park C, Kim N, *et al*. Accelerating recurrent neural network language model based online speech recognition system [C]. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018: 5904–5908.

[115] Liu X, Liu S, Sha J, *et al*. Limited-memory bfgs optimization of recurrent neural network language models for speech recognition [C]. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018: 6114–6118.

[116] 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.

[117] Dalmia S, Li X, Metze F, *et al*. Domain robust feature extraction for rapid low resource asr development [C]. In 2018 IEEE Spoken Language Technology Workshop (SLT), 2018: 258–265.

[118] Niesler T, Woodland P C. Combination of word-based and category-based language models [C]. In Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96, 1996: 220–223.

[119] Deschacht K, De Belder J, Moens M-F. The latent words language model [J]. Computer Speech & Language, 2012, 26 (5): 384–409.

[120] Masumura R, Masataki H, Oba T, *et al*. Use of latent words language models in ASR: a sampling-based implementation [C]. In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013: 8445–8449.

[121] Masumura R, Masataki H, Oba T, *et al*. Viterbi decoding for latent words language models using gibbs sampling. [C]. In INTERSPEECH, 2013: 3429–3433.

[122] Masumura R, Asami T, Oba T, *et al*. N-gram approximation of latent words language models for domain robust automatic speech recognition [J]. IEICE TRANSACTIONS on Information and Systems, 2016, 99 (10): 2462–2470.

[123] Goldwater S, Griffiths T. A fully Bayesian approach to unsupervised part-of-speech tagging [C]. In Proceedings of the 45th annual meeting of the association of computational linguistics, 2007: 744–751.

[124] Blunsom P, Cohn T. A hierarchical Pitman-Yor process HMM for unsupervised part of speech induction [C]. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, 2011: 865–874.

[125] Liu X, Chen X, Wang Y, *et al*. Two efficient lattice rescoring methods using recurrent neural network language models [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2016, 24 (8): 1438–1449.

[126] Chen X, Wang Y, Liu X, *et al*. Efficient GPU-based training of recurrent neural network language models using spliced sentence bunch [C]. In Fifteenth Annual Conference of the International Speech Communication Association, 2014.

[127] Goodman J T. A bit of progress in language modeling [J]. Computer Speech & Language, 2001, 15 (4): 403–434.

[128] Masumura R, Asami T, Oba T, *et al*. Combinations of various language model technologies including data expansion and adaptation in spontaneous speech recognition [C]. In Sixteenth Annual Conference of the International Speech Communication Association, 2015.

[129] Rosenfeld R. Two decades of statistical language modeling: Where do we go from here? [J]. Proceedings of the IEEE, 2000, 88 (8): 1270–1278.

[130] Brants T, Popat A C, Xu P, *et al*. Large language models in machine translation [C]. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), 2007: 858–867.

[131] Teh Y W. A hierarchical Bayesian language model based on Pitman-Yor processes [C]. In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, 2006: 985–992.

[132] Potamianos G, Jelinek F. A study of n-gram and decision tree letter language modeling methods [J]. Speech Communication, 1998, 24 (3): 171–192.

[133] Alumäe T. Multi-domain neural network language model [C]. In INTERSPEECH, 2013: 2182–2186.

[134] Chen X, Tan T, Liu X, *et al*. Recurrent neural network language model adaptation for multi-genre broadcast speech recognition [C]. In Sixteenth Annual Conference of the International Speech Communication Association, 2015.

[135] Park J, Liu X, Gales M J, *et al*. Improved neural network based language modelling and adaptation [C]. In Eleventh Annual Conference of the International Speech Communication Association, 2010.

[136] Tilk O, Alumäe T. Multi-Domain Recurrent Neural Network Language Model for Medical Speech Recognition. [C]. In Baltic HLT, 2014: 149–152.

[137] Khyuru K, Jin D, Dang J. Morphological verb-aware Tibetan language model [J], 2019.

[138] Mei H, Bansal M, Walter M R. Coherent dialogue witbased lah attention-nguage models [C]. In Thirty-First AAAI Conference on Artificial Intelligence, 2017.

[139] Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need [C]. In Advances in neural information processing systems, 2017: 5998–6008.

[140] Chen X, Liu X, Gales M J, *et al*. Recurrent neural network language model training with noise contrastive estimation for speech recognition [C]. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015: 5411–5415.

[141] Masumura R, Asami T, Oba T, *et al*. Domain adaptation based on mixture of latent words language models for automatic speech recognition [J]. IEICE TRANSACTIONS on Information and Systems, 2018, 101 (6): 1581–1590.

[142] Peters M E, Neumann M, Iyyer M, *et al*. Deep Contextualized Word Representations [C]. In NAACL HLT 2018: 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018: 2227–2237.

[143] Devlin J, Chang M-W, Lee K, *et al*. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [C]. In NAACL-HLT 2019: Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2019: 4171–4186.

[144] Sun Y, Wang S, Li Y, *et al*. ERNIE 2.0: A Continual Pre-training Framework for Language Understanding [J]. arXiv preprint arXiv:1907.12412, 2019.

[145] Wu C-S, Hoi S C H, Socher R, *et al*. ToD-BERT: Pre-trained Natural Language Understanding for Task-Oriented Dialogues [J]. arXiv preprint arXiv:2004.06871, 2020.

附 录

附件1

前音节的最后一个字母 / 助词的名称、略语及形式		后加字								后加字或后后加字	无后加字母	后后加字
		ག(ga)	ང(nga)	ད(da)	ན(na)	བ(ba)	མ(ma)	འ('a)	ར(ra)	ལ(la)	ས(sa)	ཨ(mchava med)	ད(da)
逻义格助词 (Ld)	འདོན(la don)	ལ(la)及ན(na)											
自由 从格助词 (Ab)	འཕྲལ་སྒྲ(vbyung khungs)	ནས(nas)及ལས(las)											
不自由格助词(c p) 属格助词(G)	འབྲེལ་སྒྲ(vbrel sgra)	ག(gi)	ང(gi)	ད(kyi)	ན(gyi)	བ(kyi)	མ(gyi)	འ(r)或 ཡ(y)	ར(gyi)	ལ(kyi)	ས(su)	ཤ(r)或ཡ(y)	ད(kyi)
逻义格助词 (Ld)	འདོན(la don)	ར(du)	ར(du)	ར(du)	ར(du)	ར(tu)	ར(du)	ར(ra) 或 ར(ru)	ར(du)	ར(du)	ས(su)	ར(ra)或ར(ru)	ར(tu)
作具格助词 (I)	ཀྱིས(byed sgra)	གིས(gis)	གིས(gis	ཀྱིས(kyis)	ཀྱིས(gyis	ཀྱིས(kyis	ཀྱིས(gyis)	ཡིས(yis)	ཀྱི(gyi	ཀྱིས(gyis	ཀྱིས(kyis)	ཡ(y)或ཡིས(yis)	ཀྱིས(kyis)
副助词(advp)	ཅིའང་ཞེ་སྟོན་པའི	ཅ(tsam) ཉིད(nyid) ཁ་ན(kho na) འབའ་ཞིག(vbv zhig) རུང(zaad) ཤས་ཙམ(sha stag)等											
助													

词	一	自由		tshig phrad)												
			多数助词 (mp)	མིང་ཚིག(ming tsgig)	དག(dag) རྣམས(rnams) ཚོ(tsho) ཅག(cag) --等											
			强调助词 (nip)	ནི་སྒྲ(ni sgra)	ནི(ni)											
			连接助词(dp)	དང་སྒྲ	དང(dang)											
	般		语尾助词 (mthp)	མིང་མཐའི་སྒྲ (ming mthavi sgra)	པ(pa)	བ(ba)	པ(pa)	པ(pa)	པ(pa)	བ(ba)	བ(ba)	བ(ba)	བ(ba)	པ(pa)	བ(ba)	པ(pa)
	助		指小助词 (dimp)	ཆུང་ངུ་སྟོན་པའི་ཚིག་འཇུག(chung ngu ston pavi tshig vjug)	གུ (gu) ནུ(nu) པུ(pu) བུ(bu) འུ(vu) རུ(ru) ལུ(lu) --等											
			不定助词 (infp)	ངེས་མེད་ཀྱི་ཚིག་ཕྲད(nges med kyi tshig phrad)	ཅིག(cig)	ཞིག(zhi g)	ཅིག(cig)	ཞིག(zhig)	ཅིག(cig)	ཞིག(zhig)	ཞིག(zhig)	ཞིག(zhi g)	ཞིག(zhig)	ཤིག(shig)	ཞིག(zhig)	ཅིག(cig)
	词 (g	不自	完结助词 (dzp)	རྫོགས་ཚིག(rdzogs tshig)	གོ(go)	ངོ(ngo)	དོ(do)	ནོ(no)	བོ(bo)	མོ(mo)	འོ(*vo*)	རོ(*ro*)	ལོ(*lo*)	སོ(*so*)	འོ(*vo*)	ཏོ(to)

(p)	p)	由	分合助词(bdp)	འབྱེད་སྡུད(vbyed sdud)	གམ(gam)	ངམ(ngam)	དམ(dam)	ནམ(nam)	བམ(bam)	མམ(mam)	འམ(*vm*)	རམ(*ram*)	ལམ(*lam*)	སམ(*sam*)	འམ(*vm*)	ཏམ(tam)
			疑问助词(qp)	ཚིག་ཕྲད(tshig phrad)	ཅེ་ན(ce na)	ཞེ་ན(zhe na)	ཅེ་ན(ce na)	ཞེ་ན(zhe na)	ཅེ་ན(ce na)	ཞེ་ན(zhe na)	ཞེ་ན(zhe na)	ཞེ་ན(zhe na)	ཞེ་ན(zhe na)	ཤེ་ན(she na)	ཞེ་ན(she na)	ཅེ་ན(ce na)
			命令助词(ipvp)		ཅིག(cig)	ཞིག(zhig)	ཅིག(cig)	ཞིག(zhig)	ཅིག(cig)	ཞིག(zhig)	ཞིག(zhig)	ཞིག(zhig)	ཞིག(zhig)	ཤིག(shig)	ཞིག(zhig)	ཅིག(cig)
			嚴摄助词(gdp)	རྒྱན་སྡུད(rgyan sdud)	ཀྱང(kyang)	ཡང(yang)	ཀྱང(kyang)	ཡང(yang)	ཀྱང(kyang)	ཡང(yang)	འང(vang)或ཡང(yang)	ཡང(yang)	ཡང(yang)	ཀྱང(kyang)	འང(vang)或ཡང(yang)	ཀྱང(kyang)
			带余助词(hcp)	ལྷག་བཅས(lhag bcas)	སྟེ(ste)	སྟེ(ste)	དེ(de)	ཏེ(te)	སྟེ(ste)	སྟེ(ste)	སྟེ(ste)	ཏེ(te)	ཏེ(te)	ཏེ(te)	སྟེ(ste)	ཏེ(te)
	接续	自由	逻义格助词(Ld)ལ 的转用	ལ(la)												
			逻义格助词(Ld)ན 的转用	ན(na)												
			从格助词(Ab) 的转用	ནས(nas)及ལས(las)												
			连接助词(dp) 的转用	དང												

	助词(conjp)	不自由	对等助词(cop)	ཚིགས་ཕྲད(tshig phrad)	ཅིང(cing)	ཞིང(zhing)	ཅིང(cing)	ཞིང(zhing)	ཅིང(cing)	ཞིང(zhing)	ཞིང(zhing)	ཞིང(zhing)	ཞིང(zhing)	ཤིང(shing)	ཞིང(zhing)	ཅིང(cing)
			时态助词(dsp)	དབྱ་གསལ་བྱེད(da lta gsal byed)	གིན(gin)	གིན(gin)	ཀྱིན(kyin)	གྱིན(gyin)	ཀྱིན(kyin)	གྱིན(gyin)	ཀྱིན(kyin)	གྱིན(gyin)	གྱིན(gyin)	ཀྱིན(kyin)	ཡིན(yin)	ཀྱིན(kyin)
			属格助词(G)的转用		གི(gi)	གི(gi)	ཀྱི(kyi)	གྱི(gyi)	ཀྱི(kyi)	གྱི(gyi)	འི(*vi*) 或 ཡི(*yi*)	གྱི(gyi)	གྱི(gyi)	ཀྱི(kyi)	འི(*vi*)或 ཡི(*yi*)	ཀྱི(kyi)
			作具格助词(I)的转用		གིས(gis)	གིས(gis)	ཀྱིས(kyis)	གྱིས(gyis)	ཀྱིས(kyis)	གྱིས(gyis)	ས(*s*) 或 ཡིས(*yis*)	གྱིས(gyis)	གྱིས(gyis)	ཀྱིས(kyis)	ས(*s*)或 ཡིས(*yis*)	ཀྱིས(kyis)
			逻义格助词(Ld)- ར(ra)或 ཟ(ru)的转用		只能接在ད 或无后加字母音节之后											
			嚴摄助词(gdp)的转用		ཀྱང(kyang)	ཡང(yang)	ཀྱང(kyang)	ཡང(yang)	ཀྱང(kyang)	ཡང(yang)	འང(vang) 或 ཡང(yang)	ཡང(yang)	ཡང(yang)	ཀྱང(kyang)	འང(vang) 或 ཡང(yang)	ཀྱང(kyang)
			带余助词(hcp)的转用		སྟེ(ste)	སྟེ(ste)	དེ(de)	ཏེ(te)	སྟེ(ste)	སྟེ(ste)	སྟེ(ste)	ཏེ(te)	ཏེ(te)	ཏེ(te)	སྟེ(ste)	ཏེ(te)
	引号助词(quop)		不自由	ཚིགས་ཕྲད(tshig phrad)	ཅེས(ces)	ཞེས(zhes)	ཅེས(ces)	ཞེས(zhes)	ཅེས(ces)	ཞེས(zhes)	ཞེས(zhes)	ཞེས(zhes)	ཞེས(zhes)	ཤེས(shes)	ཞེས(zhes)	ཅེས(ces)
	传闻助词(cep)		不自由	ཚིགས་ཕྲད(tshig phrad)	ཅེའོ(cevo)	ཞེའོ(zhevo)	ཅེའོ(cevo)	ཞེའོ(zhevo)	ཅེའོ(cevo)	ཞེའོ(zhevo)	ཞེའོ(zhevo)	ཞེའོ(zhevo)	ཞེའོ(zhevo)	ཤེའོ(shevo)	ཞེའོ(zhevo)	ཅེའོ(cevo)

发表论文和参加科研情况说明

（一）发表的学术论文

[1] **Kuntharrgyal Khysru**, Di Jin, and Jianwu Dang. Tibetan Language Model Based on Combination Tibetan Radicals. NCMMSC, 2019. (对应本文第三章)

[2] Wang, Hongcui., **Kuntharrgyal Khysru**, Li Jian., Li, Guanyu., Dang, Jianwu.,& Huang, Lixia. Investigation on acoustic modeling with different phoneme set for continuous Lhasa Tibetan recognition based on DNN method. In Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2016 Asia-Pacific (pp. 1-4). IEEE.(EI会议,对应本文第三章) Accession number: 16602992 WOS:000393591800123

[3] **Kuntharrgyal Khysru**, Di Jin, Yuxiao Huang, Hui Feng, Longbiao Wang and Jianwu Dang. Tibetan Language Model Considering Successive Relationships Between Suffixes and Semantic Functional Words [J]. IEEE Signal Processing Letters . (SCI,审稿中, 对应本文第四章)

[4] **Kuntharrgyal Khysru**, Di Jin, and Jianwu Dang. Morphological verb-aware Tibetan language model[J]. IEEE access,2019. (SCI, IF:4.098,对应本文第五章) WOS:000472009400001

[5] **Kuntharrgyal Khysru**, Qing Guo, Di Jin, and Jianwu Dang. Grammar relationship and morphological verb-aware Tibetan language model[J].Language Resources and Evaluation.(SCI,审稿中, 对应本文第六章)

（二）其他论文

[1] Li, Jian and Wang, Hongcui and Wang, Longbiao and Dang, Jianwu and **Kuntharrgyal Khysru**, and Lobsang, Gyaltsen. Exploring tonal information for Lhasa dialect acoustic modeling. In Chinese Spoken Language Processing (ISCSLP), 2016 10th International Symposium on (pp. 1-5). IEEE.2016, October.

[2] Shen, Tongtong and Wang, Longbiao and Chen, Xie , **Kuntharrgyal Khysru**, and Dang, Jianwu. Tibetan language model based on recurrent neural network. ISSP 2017.

[3] Shen, Tongtong and Wang, Longbiao and Chen, Xie ,**Kuntharrgyal Khysru**, and Dang, Jianwu. Exploiting the Tibetan Radicals in Recurrent Neural Network for Low-Resource Language Models. International Conference on Neural Information

Processing, 2017, pp. 266–275.

[4] Shen, Tongtong and Wang, Longbiao and Chen, Xie ,**Kuntharrgyal Khysru**, and Dang, Jianwu . Investigation of Long Short-Term Memory for Tibetan Language Model, NCMMSC 2017.

（三）参与的科研项目

[1] 国家重点基础研究发展计划（ 973项目）“互联网环境中文言语信息处理与深度计算的基础理论和方法”。

致　谢

转眼间，已渡过五年的博士研究生生涯，感概和感恩之情涌上心头，我要感谢这一路摸爬滚打中伴我成长的每一位老师、同学、好友，尤其是家人，为我的学业付出一切。

首先我要由衷感谢我的导师党建武教授，在这五年的学习和论文写作期间给予我的指导、鼓励和帮助。党老师国际的视野、先锋的科研抱负、以及掌控全局的学术思维，从他的言传身教中都让我受益匪浅。回首这几年来自己在科研领域的进展，每一次的进步无不凝结了老师的汗水。回首读博这些年，自己有很多欠缺和不足的地方，党老师都给予了我最大程度上的包容。我深知，自己的学术研究水平还未达到老师的要求，惟有在今后的工作和学习中加倍努力，方能不辜负老师的谆谆教诲和悉心栽培。

其次，感谢金弟老师和龙从军老师对我科研上的帮助。在我每次科研遇到瓶颈的时候，金老师都会给予耐心的指导和方向上的把控；龙老师在藏语知识和理论上给予很好的帮准，也多次进行了探讨和研究；在科研论文撰写期间，他们和我一起打磨论文思路，指导论文写作，点点滴滴难以忘怀，特此向他们表达我衷心的感谢。

同时，感谢魏建国教授、王龙标教授、冯卉老师、贺瑞芳老师、陈彧老师和王洪翠老师，还要感谢曹金鑫、郭青、薛万利、郭凤羽、张句、郭丽丽、王英奎、王晓宝、李健、刘瑞琳、申彤彤、潘立馨等同学和朋友，我们在生活上互相帮助和扶持，在科研上互相讨论和互相学习。很荣幸这一路有你们相伴左右，愿你们心想事成。

最后，感谢我的父母、妻子和俩孩子，为了支持我的求学之路，他们默默地牺牲了很多，在我最困难的日子里不离不弃，并为我打气给我鼓励。感谢我的女儿阿央措姆和儿子丹桑尼玛，无论我如何疲惫，一听到他们稚嫩的声音和看到他们纯真的笑脸就会令我充满动力。家始终是我这些年求学生涯最坚强的后盾，家人们健康是我最大的心愿!

更太加

2019年7月　于北洋园