Computer Science and Application 计算机科学与应用, 2020, 10(10), 1714-1727
Published Online October 2020 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2020.1010182

Hans汉斯

基于循环神经网络的西班牙语词汇发音预测模型研究

赵皎谷，马延周，黄晓辉

战略支援部队信息工程大学洛阳校区，河南 洛阳
Email: zjg0307zjg@163.com, myz827@126.com, 531895671@qq.com

收稿日期：2020年10月1日；录用日期：2020年10月6日；发布日期：2020年10月23日

摘　要

依据西班牙语词汇和音素的特征以及词汇标音过程的特点，将西班牙语词汇标音过程建模为序列标注任务，提出基于字符嵌入 + 循环神经网络 + 连接时序分类的端到端词汇标音模型。首先利用word2vec框架在自建的西班牙语词库上训练字符嵌入向量，从而形成西班牙语字符的分布式向量编码表示；之后基于循环神经网络和连接时序分类算法构建了西班牙语词汇标音模型，并在自建的发音词典语料上进行了训练与测试。试验结果显示，基于字符嵌入 + 循环神经网络 + 连接时序分类的词汇标音模型可以获得较其他统计模型或是神经网络模型更高的标音准确率，同时较传统标音模型有更简单的标注流程，对数据集的要求也要低得多，可有效实现端到端的西班牙语词汇标音任务。

关键词

西班牙语，发音词典，字音转换，循环神经网络

Research on Predictive Model of Spanish Vocabulary Pronunciation Based on Recurrent Neural Network

Jiaogu Zhao, Yanzhou Ma, Xiaohui Huang

Information Engineering University, Luoyang Henan
Email: zjg0307zjg@163.com, myz827@126.com, 531895671@qq.com

Received: Oct. 1st, 2020; accepted: Oct. 16th, 2020; published: Oct. 23rd, 2020

Abstract

According to the characteristics of these vocabularies and phonemes and the characteristics of the

文章引用: 赵皎谷, 马延周, 黄晓辉. 基于循环神经网络的西班牙语词汇发音预测模型研究[J]. 计算机科学与应用, 2020, 10(10): 1714-1727. DOI: 10.12677/csa.2020.1010182

vocabulary transcription process, the word vocabulary transcription process is modeled as a sequence labeling task, and an end-to-end vocabulary transcription model method based on character embedding + recurrent neural network + connection arrangement classification is proposed. First, this paper uses the word2vec framework to train the character embedding vector on the self-built serial thesaurus to form a distributed encoding representation of the character; then based on the recurrent neural network and the connection classification algorithm, a model called vocabulary transcription is constructed. The test results show that the word transcription model of string embedding + cyclic neural network + connection order classification can use higher transcription accuracy than other statistical models or neural network models. At the same time, it has a simpler labeling process than traditional phonetic models. The requirements of the phonetic transcription should also be reduced, that can effectively realize the end-to-end task called phonetic transcription.

Keywords

Spanish, Pronunciation Prediction, Grapheme-to-Phoneme Conversion, Recurrent Neural Network

1. 引言

西班牙语发音词典是西班牙语语音识别以及语言合成的基础。西班牙语词形的复杂变化会导致西班牙语发音词典词汇量巨大，同时随着时代的发展，西班牙语中不断涌入新的词汇，固定的发音词典难以满足西班牙语语言处理任务的要求，实现西班牙语词汇发音的机器自动预测功能对于西班牙语的自然语言处理研究具有重要意义。

本文对西班牙语发音字典的构建方法进行了简单的回顾，分析了目前存在的问题与主流的解决方案。最后选择将西班牙语发音词典的构建作为一个序列标注任务来建模，利用 word2vec 框架在自建的西班牙语词库上训练字符嵌入向量，从而形成分布式的字符编码向量表示。之后，基于循环神经网络和连接时序分类算法构建了西班牙语字——音标注模型，并在自建的发音词典语料上进行了训练与测试，取得了理想的效果。

2. 发音词典概述

发音词典的主要功能是实现文本单词到语音发音基元的映射，在文字与语音的转换过程中起着关键性的作用，也是语音识别、语音合成等高层应用必备的基础数据资源。西班牙语发音词典的传统构建方法多采用穷举法，即列举出词表中所有的西班牙语单词，并以人工方式为每个单词标上相应的发音基元，从而构建一个以键－值对形式存储的单词——发音对照词典。然而由于西班牙语属于屈折型语言，是典型的字符拼写型文字，具有数量极其庞大的单词形式，单纯采用人工方式来标记发音词典不仅需要大量的时间和精力，同时还需要召集足够多的西班牙语专业人员来进行发音基元的标注，其低下的构建效率和高昂的人力成本对于实际应用来说都是难以承受的。并且随着经济社会的发展，不断的会有新的词汇出现，基于键—值对的发音词典根本无法应对词表以外单词的发音预测任务。因此，业界开始致力于研究基于数据驱动的词汇标音模型，即以人工标记的小规模、高精度发音词典作为原始训练数据，通过构建机器学习模型并基于原始数据进行模型训练，从而使模型具备一定的标音能力，再结合机器标音和人

工校验来不断的补充、扩展原始数据集以形成更大规模的发音词典数据，然后再重复机器学习－机器标音－人工校验的过程，最终实现词汇标音模型的构建以及大规模发音词典的生成。随着词典数据的持续积累，机器学习模型通过不断的训练来提升其标音准确率，直至实现完全自动化的词汇标音。与人工标记相比，训练好的机器学习模型不受工作时间、工作环境、工作状态的影响，具有较人工标注更高的稳定性和持续性，其标音速度也要快的多，因此已经成为当前解决大规模发音词典构建问题的有效手段。值得注意的是，要实现高质量的字音转换，机器学习模型的选择是极其关键的。恰当、合适的机器学习模型可以快速、准确的学习到西班牙语单词与发音之间的关联模式，从而高效、准确的实现单词到音素的自动转换。不合适的模型就难以有效建模单词与发音之间的关系，不仅无法实现精准的字－音转换，还会干扰到正常的发音词典构建，反而起到事倍功半的反作用。

3. 字音转换问题分析与任务建模

早期发音词典大多采用键－值对形式存储，即一个单词对应一个音素序列。词表中的所有单词都会被纳入发音词典中，并以人工标注方式为每个单词标上对应的音素。由于这种发音词典只能包含词表中的词，对于词表以外的词就无法处理，因此其开放性较差，并且对人力成本的要求极高。随着人工智能技术的发展，语言智能处理相关应用对大规模成熟标注语料的需求越来越旺盛，大规模、开放式的发音词典越来越受到关注。就目前来讲，将发音词典模型化、参数化是解决大规模发音词典构建问题的有效方式。发音词典模型化、参数化的基本理念是采用带有参数的数学模型作为转换器，以构成单词的字符序列作为输入，由模型将其自动转换为对应的音素序列。模型的参数是在已有的小规模发音词典上通过机器学习方式训练获得的，并且可以通过不断的迭代训练进行优化调整。这种形式的发音词典只需要存储模型的参数，再根据模型的计算原理即可实现单词到发音的转换功能，具有自动化程度高、可扩展性强的特点，并且能够有效应对未登录词的标音任务。因此，本文的主要研究内容就是通过试验来遴选合适的模型，以构建模型化、参数化的西班牙语发音词典。

从形式上来看，发音词典的主要作用是实现单词到发音的转换，即为一个文本形式的单词标注上正确的发音基元(一般定义为音素)。对于西班牙语来讲，词是由多个字符按照一定的规则构成的序列，而词的发音也是由多个音素构成的序列。因此，构建发音词典的过程实际上是一个序列标注的过程，即输入序列是构成西班牙语单词的字符序列，而输出则是对应的音标序列，其形式如图 1 所示。

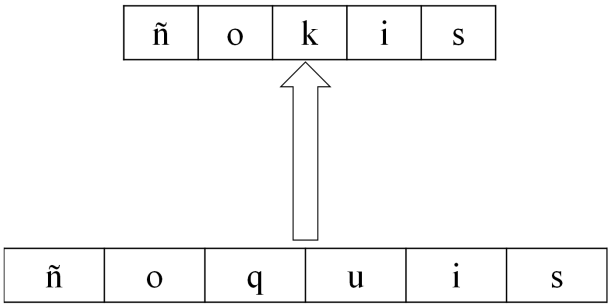

Figure 1. Modeling the best budget dictionary reconstruction task
图 1. 西班牙语发音词典构建任务建模

在西班牙语单词中通常会有连续多个字符对应一个发音基元的现象，也就是说在一个单词中既存在一个字符对应一个发音的现象，也存在连续多个字符对应一个发音的现象(如图 1 中的 qu 两个字符对应一个音标 k)。总体来看，在大部分的西班牙语单词中，构成单词字符个数会多于其音素标记个数。这也就是说在西班牙语的字－音转换任务中，输入序列的长度通常会比输出序列要长，而现有的序列标注模

型都是一对一的标注模式，因此通常还要进行一些预处理或是后处理才能实现最终的字音转换。现有用于解决字符个数与音素个数不一致的方法主要有两种思路，一种思路是在单词层面先对词进行字素切分，使一个字素对应一个发音基元，这样即可使用标注模型进行一对一的序列标注；另一种思路是先用序列标注模型为每个字符都标注一个发音基元，再用语言模型来对标注的音素进行修正，最终形成正确的发音序列。第一种方法中字素切分结果的准确性将直接影响发音序列的标注，因此存在误差传播问题。第二种方法则需要构建专门的声学基元语言模型，大大增加了数据准备的难度。在这两种方式下，模型的训练和解码过程都是流水线式的流程，存在不可避免的误差传播问题。因此本文本将探索将连接时序分类算法引入字音转换过程中，通过合理的设置音素标记集使字音转换任务转变为较长字符序列到较短音素序列的标注任务，从而连接时序分类算法实现端到端的字－音转换过程。

从模型选择上来看，能够实现序列标注的模型有很多，目前常用的是基于统计的序列标注模型和基于神经网络的序列标注模型。其中基于统计的序列标注模型从概率分布的角度来建模输入序列与标签序列之间的关联关系，典型的如隐含马尔可夫模型(Hidden Markov Model, HMM)、条件随机场模型(Conditional Random Field, CRF)等。然而由于统计模型为了能够在合理的时间内求解，通常对序列内部以及序列之间的依赖关系做了简化处理，因而忽略了许多实际存在的关联因素，最终导致模型对数据特征分布的拟合能力不足，需要依靠专家设计的特征模板才能发挥作用。相比之下，基于神经网络的序列标注模型则主要基于连接主义的思想，通过向量化的表示学习来实现对输入序列的特征提取。同时，神经网络的连接结构使其具有强大的数据分布拟合能力，从而可以建模复杂的映射关系。随着深度学习技术的发展，基于神经网络的序列标注模型已经在语音识别、信息抽取、机器翻译等多个领域取得突出的成果。大量实验结果也已证明，只要给予足够的训练数据，基于循环神经网络的序列标注模型要明显优于传统的统计模型。这其中，循环神经网络(Recurrent Neural Network, RNN)对序列数据具有天然的上下文依赖特征建模能力，在许多应用上都取得了显著的成果，因而成为序列标注任务的首选模型[1]。

基于以上分析，本文拟将字－音转换过程作为序列标注问题进行建模，并探索合适的循环神经网络作为序列标注模型，结合连接时序分类算法来实现端到端的训练与标注，以实现西班牙语发音字典的模型化和参数化，提升西班牙字－音转换的效果。

4. 基于循环神经网络的字音转换模型

本文设计的字音转换标注模型总体架构如图 2 所示。

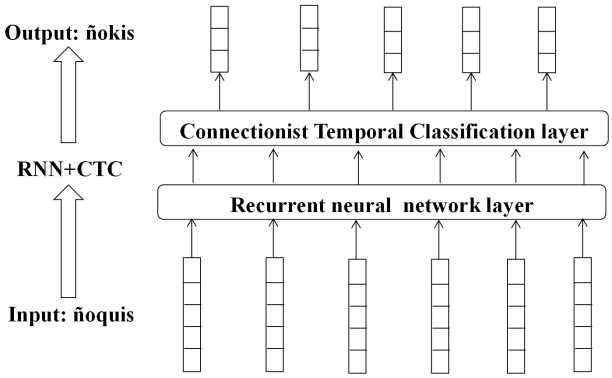

Figure 2. Spanish phonetic labeling model based on character embedding + RNN + CTC
图 2. 基于字符嵌入 + RNN + CTC 的西班牙语字音标注模型

从图中可以看出，该框架主要包括三部分内容，一是基于字符嵌入的序列输入层，二是基于循环神

经网络的特征提取层，三是基于连接时序分类构建的标签输出层。其中，输入层是以向量形式构成的输入序列，代表顺序输入的字符。本文用实数向量的方式来对西班牙语字符进行编码，将字符自身特征及其所处的上下文特征进行向量化表示，以作为神经网络特征提取器的输入数据；之后，基于循环神经网络的特征提取层对输入的字符向量序列进行线性的加权和非线性映射，从而提取出能够区分不同音素的字符特征作为输出层类别判断的依据[2]。由于循环神经网络具有天然的时序结构，因此能够有效建模字符序列的上下文特征，从而为最终的音素标注提供高质量的特征依据。最后，基于连接时序分类的输出层依据循环神经网络提供的数据特征来为每个时刻的输入向量预测一个音素类别，最终形成音素类别的序列，从而实现对单词发音基元的标注。连接时序分类层从结构上来讲实际是一个 Softmax 分类层，在每个时刻为每个输入的向量判断一个音素类别。但其特殊之处在于一个特别的损失函数，该函数用于衡量两个序列之间的差异，从而用于模型的训练和优化。最重要的是，连接时序分类层可以实现长序列到短序列的自动映射。由于构成西班牙语单词的字符个数一般多于其音素个数，而循环神经网络的输出序列与其输入序列是等长的，因此通过采用连接时序分类可以有效的实现输入序列到真实音素序列的映射关系，从而实现端到端训练与解码。

4.1. 基于向量表示的字符嵌入输入层

由于神经网络的输入只能是以数字形式表示的向量或张量，因此对符号形式的西班牙语字符进行编码是构建基于神经网络的字－音转换模型首先要解决的问题。传统模式下，字符编码通常采用 one-hot 向量形式将所有字符编码到一个正交的向量空间中，向量的维度就是字符的个数。这种编码方式无法体现字符与字符之间在构词上的相似度，忽略了不同字符在构成一个单词时存在的上下文关联关系。因此，本文基于 NLP 中常用的词嵌入思想[3] [4]，以字符作为编码对象，采用词向量训练工具来训练字符向量，从而实现西班牙语字符的嵌入式表示。在这种模式下，表示各个字符的向量不再是正交的 one-hot 向量，而是可计算相似度的多维实数向量，其形式如图 3 所示。

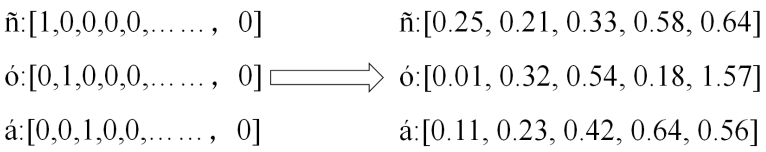

Figure 3. One-hot vector encoding and embedded vector encoding of Spanish characters
图 3. 西班牙语字符的 one-hot 向量编码与嵌入向量编码

目前，可用于训练词向量的工具有很多，其中典型的一个是由 Google 发布的 Word2vec 框架。该框架的本质是基于神经网络的语言模型计算工具，即利用神经网络来计算语言模型，其结构原理如图 4 所示。

可以看出，基于神经网络的语言模型计算框架包含三个神经网络层，第一层是输入层，表示采用 one-hot 编码的词序列；第二层是一个映射层，用于将 one-hot 向量映射成一个维度较低的实数向量；第三层是一个神经网络隐含层，对实数向量进行非线性变换；最后是 softmax 层，用于预测输入词序列后面下一个词的可能性。等模型训练收敛后，存储在映射层中的向量就是对应输入词的嵌入式表示，因此词向量实际上是神经网络语言模型的副产物。由于图 7 的神经网络模型要经过隐含层的非线性变换，计算量过大，因此 word2vec 又设计了 CBoW 模型和 Skip-gram 模型，两种模型从结构上都摒弃了隐含层的非线性变换，只保留了分类层。其中 CboW 模型以单词两边的上下文作为输入来预测当前位置的单词，而 Skip-gram 模型则当前单词为输入来预测其左边和右边的其他单词，两种模型的计算原理分别如图 5、图 6 所示。

Word2vec 词向量训练工具可以在百万数量级的词典上进行高效的训练。因此，本文采用 Word2vec 作为西班牙字符向量的训练工具，以单词的字符作为模型的输入，以 CBoW 模型作为字符嵌入的训练模型，训练语料则为本文自建的西班牙语词库。

4.2. 基于循环神经网络的特征提取层

循环神经网络(Recurrent Neural Network, RNN)是一种特殊的深层神经网络，其显著特点就是隐含层自连接结构使其具有天然的时序特征，因此能够有效建模序列数据的上下文依赖关系[5] [6]。RNN 的结构特点可如图 7 所示。

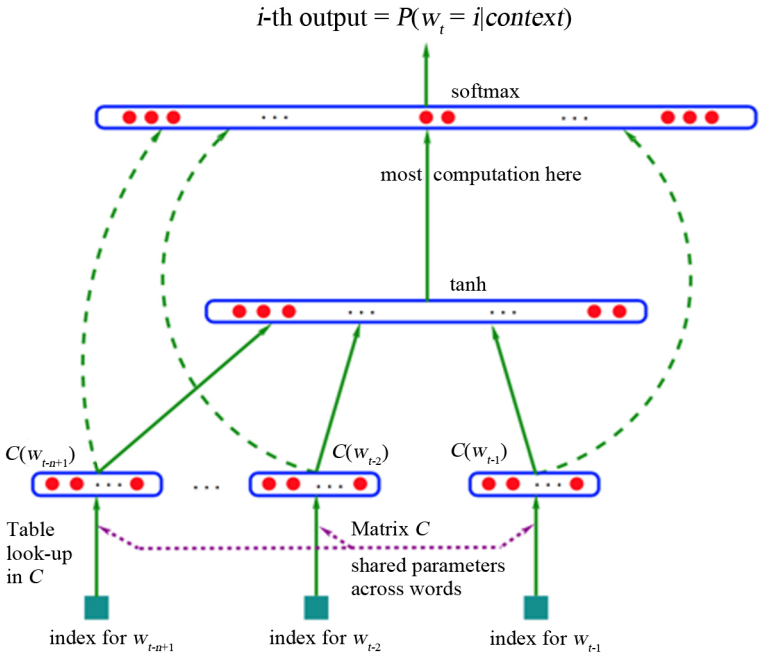

Figure 4. The architecture of word2vec
图 4. Word2vec 的架构

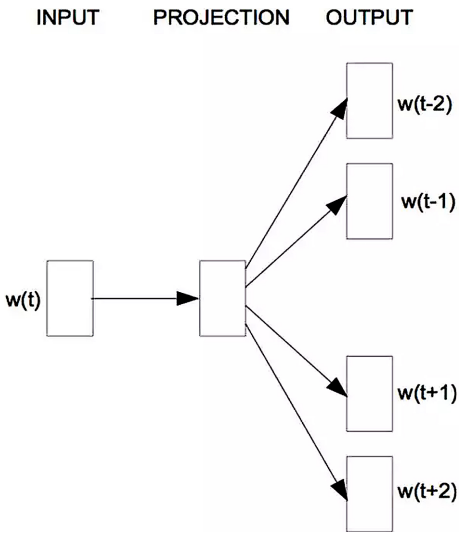

Figure 5. CBoW model
图 5. CBoW 模型

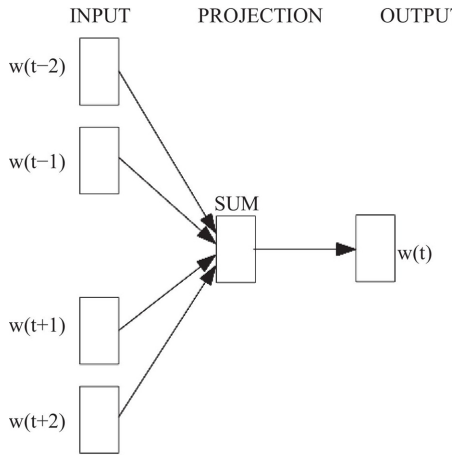

Figure 6. Skip-gram model
图 6. Skip-gram 模型

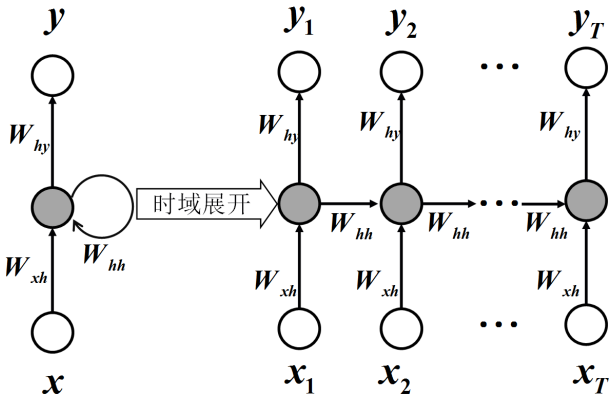

Figure 7. Structural characteristics of recurrent neural network
图 7. 循环神经网络的结构特点

RNN 的输入层表示输入样本的编码向量，输入层神经元个数即输入数据编码向量的维数。编码向量经加权后会进入中间的隐含层；RNN 的隐含层具有自连接特性，即将上一时刻隐含层的状态向量进行加权计算后，与输入层的加权向量一起送入激活函数，经计算得到隐状态向量。该隐状态向量再进入输出层进行分类预测，同时会进入下一时刻的隐含层进行时序迭代。RNN 的输出层实际是一个分类层，其中的每个神经元代表一个类别，每个神经元的输出值表示输入样本属于该类别的后验条件概率值。可以看出，RNN 实际上是一个三维的网络结构，其中时间维度上的展开步数等于输入序列的长度，但在任意时刻，网络连续的权值是恒定不变的。序列数据进入 RNN 隐含层的计算过程可如公式 1 所示：

$$\begin{aligned} h_t &= H\left(W_{xh}x_t + W_{hh}h_{t-1} + b_h\right) \\ y_t &= O\left(W_{hy}h_t + b_y\right) \end{aligned} \tag{1}$$

其中 x_t 为 t 时刻的网络输入，H 为隐含层激活函数(通常为非线性函数，如 Tanh 函数)，O 代表输出层的分类函数(通常是 Softmax 函数)，y_t 为 t 时刻的输出向量，通常表示一个类别的编码向量。权值 W_{xh}、W_{hh}、W_{hy} 以及偏置项 b_h、b_y 都是需要通过训练才能确定的网络参数。

RNN 通常采用基于梯度的训练方法，在输入序列较长时，其时域展开步数也随之线性增加。由于只有简单的激活函数结构，因此在模型训练时产生的梯度并不能按照理想的方式持续回传，普遍存在梯度消失问题，这也导致普通的 RNN 在实际应用中无法捕获序列内部的长距离时序依赖关系。为了改善循环

连接对时序关联特征的建模能力，人们提出了门控机制，即通过门控函数的方式来提升其对长距离时序特征的记忆、强化及弱化能力，典型的如长短时记忆单元(Long-Short Term Memory, LSTM)和门循环单元(Gated Recurrent Unit, GRU)等，两者的内部结构及运算过程分别如图 8 和图 9 所示。

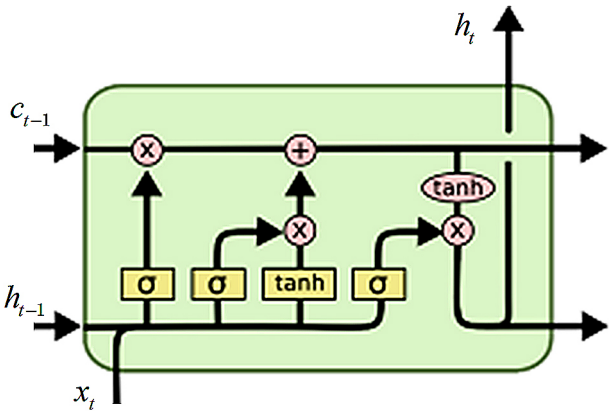

Figure 8. The internal structure and operation process of LSTM
图 8. LSTM 的内部结构及运算过程

从图中可以看出，LSTM 实际上是加入门控函数的 RNN 激活单元。其神经元内部共有三个门函数和一个存储单元。三个门函数通常称为输入门、遗忘门和输出门，分别控制神经元的输入、存储以及输出。存储单元是 LSTM 神经元的记忆组件，受遗忘门的控制，同时接收新的输入，并在达到激活条件时产生输出值。LSTM 的内部计算过程可如公式 2 所示：

$$\begin{aligned} i_t &= \sigma\left(W_i \cdot \left[h_{t-1}, x_t\right] + b_i\right) \\ f_t &= \sigma\left(W_f \cdot \left[h_{t-1}, x_t\right] + b_f\right) \\ o_t &= \sigma\left(W_o \cdot \left[h_{t-1}, x_t\right] + b_o\right) \\ c_t &= f_t c_{t-1} + i_t \tanh\left(W_c \cdot \left[h_{t-1}, x_t\right] + b_c\right) \\ h_t &= o_t \tanh\left(c_t\right) \end{aligned} \tag{2}$$

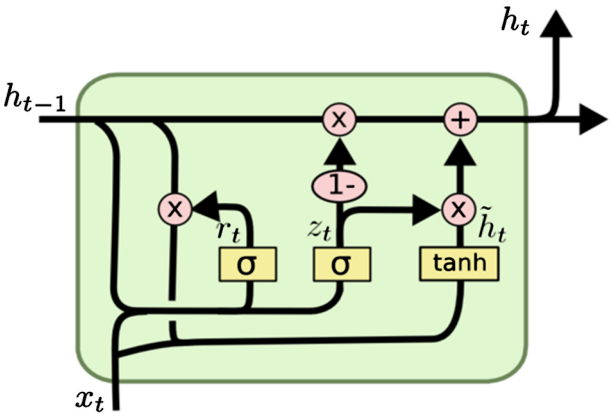

Figure 9. The internal structure and calculation process of GRU
图 9. GRU 的内部结构及运算过程

与 LSTM 相比，GRU 同样使用了门控机制。但是 GRU 只有两个门函数，即更新门和重置门。其中，更新门用于控制前一时刻状态信息对神经元激活的有效程度，而重置门则用于控制经过激活值对当前时

刻状态信息的有效程度。GRU cell 的内部结构及运算过程如图 9 和公式 3 所示：

$$\begin{aligned} z_t &= \sigma\left(W_z \cdot \left[h_{t-1}, x_t\right]\right) \\ r_t &= \sigma\left(W_r \cdot \left[h_{t-1}, x_t\right]\right) \\ \tilde{h}_t &= \tanh\left(W\left[r_t h_{t-1}, x_t\right]\right) \\ h_t &= \left(1-z_t\right) h_{t-1} + z_t \tilde{h}_t \end{aligned} \tag{3}$$

门控机制能够实现对输入信息流的控制，从而缓解训练过程中的梯度消失现象，但也带来了参数量成倍增加的问题。相对 LSTM 来讲，GRU 的参数要少，但在应用中的效果却并不比 LSTM 差，甚至在某些情况下还要优于 LSTM。因此，GRU 是目前循环神经网络的首选变体。

另外，传统 RNN 是单向展开，因此只能利用之前的历史信息来预测当前时刻的输出，而序列标注任务是对整个输入序列的转写，当前元素后面的上下文信息对当前元素也有直接的影响。因此双向 RNN (Bidirectional RNN, Bi-RNN)可以使用两个独立的隐含层分别从前往后和从后往前处理数据，之后再同时进入输出层，可有效解决两个方向上的时序特征建模该问题[7]。Bi-RNN 隐含层的结构可如图 10 所示。

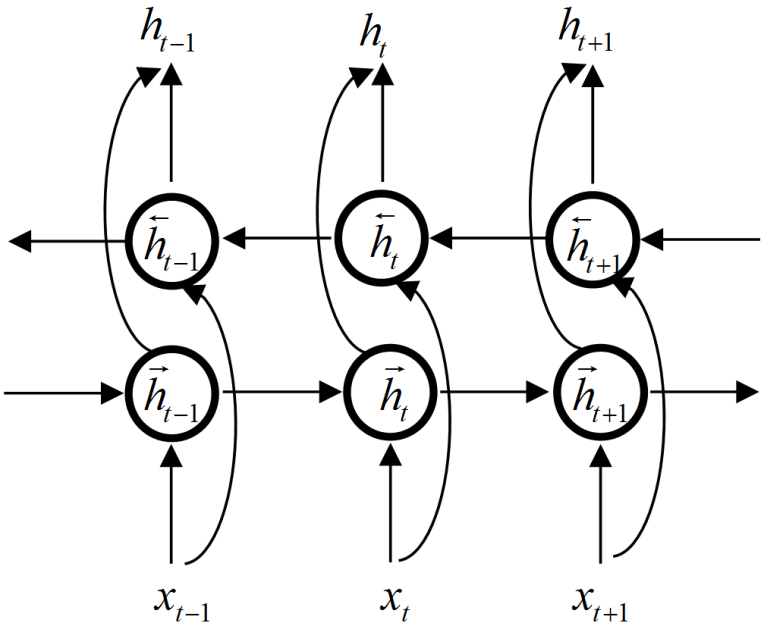

Figure 10. Hidden layer structure of bidirectional recurrent neural network
图 10. 双向循环神经网络的隐含层结构

因此，双向 RNN 隐含层的计算过程可如公式 4 所示：

$$h_t = W_{\vec{h}}\vec{h}_t + W_{\overleftarrow{h}}\overleftarrow{h}_t \tag{4}$$

组合两个方向的 LSTM(GRU)单元就可以构成双向的 LSTM(GRU)网络。

4.3. 适用于端到端训练的连接时序分类层

由于构成西班牙语单词的字符个数与其对应的音素个数并非一一对应，并且通常情况下其发音音素的个数会小于字符的个数。因此，字－音转换模型还要实现较长字符序列到较短音素序列的自动映射。连接时序分类(Connectionist Temporal Classification, CTC)作为一种在语音识别领域比较经典的序列标注算法，其本身可以实现从长序列到短序列的映射[8]。从结构上来讲，连接时序分类实际上是一个经过改

进的 Softmax 分类器。在传统 Softmax 分类器中，每个输出节点代表一个真实有效的类别标签，但在 CTC 中，Softmax 层除了设置有效的类别标签节点外，还增加了一个表示“空输出”的节点。相应的，其标签集中也会增加一个空(blank)标签，表示某一时刻的输入元素没有实际的类别标签。同时，从功能上来讲来讲，CTC 本身又是一个用于模型训练的误差函数，可以建模长序列与短序列的差异。具体的，假设循环神经网络的输出是一个长度为 T 的向量序列 π，则 T 实际也是最后循环神经网络在时域上的展开长度，而序列 π 中的每个向量都表示一个标签类别的概率。标签类别来自于真实有效的标签类别集合(记为 K)以及一个“空”(blank)类别。神经网络输出层生成序列 π 的计算过程如公式 5 所示：

$$\begin{aligned} p\left(k,t\middle|x\right) &= soft\max\left(out_t^k\right), k \in K \bigcup \text{blank} \\ p\left(\pi\middle|x\right) &= \prod_{t=1}^{T} p\left(k,t\middle|x\right), k \in K \bigcup \text{blank} \end{aligned} \tag{5}$$

式中 x 表示输入序列，$p(k, t|x)$表示在 t 时刻模型预测类别为 k 的概率值，是由 Softmax 分类器根据前一层神经网络输出的特征值计算得到的；$p(\pi|x)$则为整个输出序列的概率值，是由每个时刻的类别概率值 $p(k, t|x)$相乘得到的。此时，使用一个多对一映射 $\beta(\pi)$即可将模型输出序列 π 转换成真实标签序列 y，其映射过程可如公式 6 所示：

$$y = \beta\left(\pi\right) \Leftrightarrow \left(a,b,c\right) = \begin{cases} \beta\left(a,b,-,-c\right) \\ \beta\left(-,a,b,-,c\right) \\ \beta\left(-,a,a,b,-,c,c\right) \\ \quad\vdots \\ \beta\left(-,a,-,b,-,c,c\right) \end{cases} \tag{6}$$

可以看出，该函数首先将序列 π 中两个彼此相邻并且值相同的标签合并为一个类别标签，然后再删除 π 中的空(blank)标签，即可形成最终的标签序列 y。因此，CTC 分类层的序列转换过程可如图 11 所示。

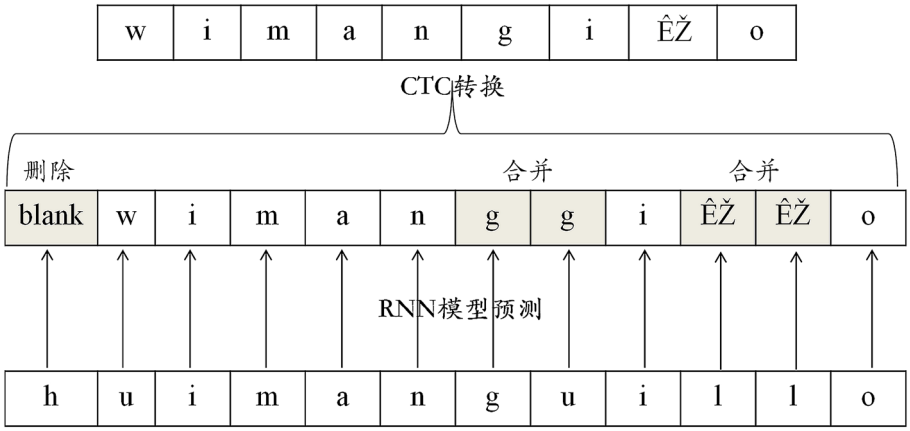

Figure 11. CTC layer sequence conversion process schematic
图 11. CTC 层序列转换过程示意

不难看出，可以存在多种形式的序列 π 对应着同一个标签序列 y。因此真实标签序列 y 的概率实际上是所有能够转换成 y 的序列 π 的概率之和，如公式 7 所示：

$$p\left(y\middle|x\right) = \sum_{\pi \in \beta^{-1}(y)} p\left(\pi\middle|x\right) \tag{7}$$

CTC 训练的目标即是让得到真实标签序列 y 的概率达到最大值(极大似然估计)。通常我们会将上述目标函数转换为负对数似然的形式，以便使用梯度下降法使该目标函数达到最小值(等价于最大化真实标签序列的概率)。因此，目标函数的定义如公式 8 所示：

$$\text{CTCLoss} = -\log p\left(y|x\right) \tag{8}$$

公式 8 的计算可用类似于 HMM 前向算法的动态规划算法有效解决，并且该算法的中间值还可用于计算公式 7 的梯度值，从而加速模型的训练过程。当模型训练完成之后，即可用于对西班牙单词发音基元的预测。

由于 CTC 可有效实现较长序列到较短序列的映射转换，因此对训练数据不再要求单词内字素与发音基元的对齐关系，在实现端到端序列转换的同时，大大减轻训练集的建设成本。

5. 实验设置与结果分析

5.1. 实验数据准备

为验证所提西班牙语字—音转换模型的性能，本文以自建的西班牙语发音词典语料为实验数据进行了一系列验证实验。具体的，本文首先以自建的西班牙词库作为训练语料，基于 word2vec 来实现实现西班牙字符嵌入向量的训练。该词库共包含西班牙单词 9 万余条，单词的长度从 1 到 20 分布不均。构成西班牙语单词的字符种类有 27 种，其中五个元音字符{a, e, i, o, u}在作为重音元音发音时其书写形式为{á, é, í, ó, ú}，因此如果从书写形式上来区分的话，构成西班牙语单词的字符种类实际上有 32 种。训练字符嵌入所用词库的详细统计信息如表 1 所示。

Table 1. Spanish thesaurus statistics for character embedding training
表 1. 用于字符嵌入训练的西班牙语词库统计信息

	0 ≤ L ≤ 4	5 ≤ L ≤ 8	9 ≤ L ≤ 12	13 ≤ L ≤ 16	17 ≤ L ≤ 20
单词长度分布	2865	41216	39804	6798	357
所占比例	3.1%	45.3%	43.7%	7.5%	0.4%
字符集	á, é, í, ó, ú, a, e, i, o, u, l, m, n, p, s, t, c, q, d, b, v, r, z, f, j, g, ñ, h, y, x, k, w				

在完成字符的嵌入表示训练之后，本文在自建的西班牙语发音词典语料上进行字—音转换模型的训练与性能验证。该语料共包含了 91,040 条西班牙语单词，每个单词都进行了音素序列的人工标注，其中共设置了 44 个音素。因为存在部分不同音素对应同一个音标，以及同一个音素有不同音标的情况，所以总共有 25 个不同的音素标记符号。由于本文研究设置了新的音素类别集合，因此该语料中单词的长度全部大于或等于音素序列的长度，具体的统计信息如表 2 所示。

Table 2. Statistics of pronunciation dictionary corpus
表 2. 发音词典语料的统计信息

	词长小于音标长	词长等于音标长	词长大于音标长
单词与音标长度	0	79,200	11,840
所占比例	0%	87.0%	13.0%
音素标记集	a, É>, i, o, w, u, l, m, n, p, s, t, k, d, b, r, É3/4, f, x, g, É2, j, ks, ÊŽ, Ê§		

我们将发音词典语料按照 70%、10%和 20%的比例进行训练集、验证集和测试集的划分，其中训练

集用于模型参数的调整与优化，而验证集则用于记录模型的训练效果，防止产生过拟合，测试集则用于评估模型的标注性能。

5.2. 模型的构建与训练

依据本文提出的解决方案，本文采用 python 3.5 + Tensorflow 1.4 来构建西班牙语字－音转换序列标注模型，并在自建的西班牙语数据集上完成模型的训练与测试。具体来讲，本文分别基于 word2vec 中的 CBOW 模型和 skip-gram 模型来实现字符嵌入向量的训练，基于 Tensorflow 来构建基于 RNN + CTC 的字音转换模型。在训练过程中，采用网络搜索的方式，通过设置多组模型结构并行独立实验来实现模型结构的探索与验证，依据最终的标注效果来科学的选择嵌入向量的维度、神经网络隐含层的节点个数、隐含层的节点类型等，并基于充分、全面的交叉验证实验来保证结果的可靠性和稳定性。需要强调的是，对于 CTC 输出层的设置，本文依据常用的西班牙音素集共设置 26 个节点，分别对应 25 个音素类别和一个空类别。

模型在训练集上训练时的目标函数如 4.3 节公式 8 所示。训练方法采用基于 Minibatch 的随机梯度下降算法，Minibatch 大小设置为 512，即一次从训练集中随机选取 512 个单词作为输入，逐个 Minibathch 的进行模型训练。每个 Minibatch 计算一次目标函数，进行一次梯度更新。另外，在训练过程中我们还对单词进行填充处理，即将同一 Mini-batch 内的单词以补 0 方式填充到统一的长度，以便于利用并行计算来提升模型训练速度。当训练集中每个 Minibatch 的数据都逐个完成训练后，便在验证集上进行一次模型标注效果的验证。在验证集上的标注效果采用单词的标注精确率作为评估指标，即正确标注的单词数占验证集总体单词数的比例。该指标的计算相对较容易，能够明确反映模型对西班牙单词发音的标注效果。当验证集上的标注准确率不再上升或是开始下降时，即停止模型的训练。同时将模型在测试集上进行单词标音测试，最后，以测试集上获得最高标音精确率的模型作为最终模型。

需要说明的是，本文所有实验均在同一台安装 Linux 系统的高性能服务器上运行，该服务器内置 8 颗至强 CPU，64G DDR 内存，同时配置 4 块 NVIDIA 的 Tesla K80 GPU 卡来进行训练加速。

5.3. 实验结果分析

循环神经网络(Recurrent Neural Network, RNN)是一种特殊的深层神经网络，其显著特点就是隐含层自连接结构使其具有天然的时序特征，因此能够有效建模序列数据的上下文依赖关系[5] [6]。RNN 的结构特点可如图 7 所示。

在参考一些经典神经网络模型配置的基础之上，我们对模型的各种结构参数都进行了不低于 10 次的训练和测试，并记录了数次实验的平均值作为最终的衡量指标，以在测试集上标音性能最好的网络结构作为最终的网络模型。最终，在多次实验之后，我们确定了最优的字音转换模型结构，具体如表 3 所示。

Table 3. Specific parameters of the optimal grapheme-to-phoneme conversion model
表 3. 最优字音转换模型的具体参数

层数	结构参数
第一层(输入层)	10 个节点，表示 10 维字符嵌入向量，采用 CBOW 模型训练获得
第二层(隐含层)	双向循环神经网络，各 128 个 GRU 单元
第三层(输出层)	Softmax 分类层，设置 26 个输出节点

采用表 3 所示的模型结构，在经过多轮训练之后，该模型在测试语料上取到了最高的标音准确率，

证明了本文所提模型的可行性。同时，为了衡量本文所提模型的标音性能处于何种水平，我们在实验中还构建了几个目前常用的字音转换模型作为对比。这些模型与我们设计的模型都是在同样的硬件条件下、并且采用相同的语料资源完成了训练与测试。我们同样记录了各个模型最优的标音准确率(标音正确的单词占测试集单词的比例)，具体结果如表 4 所示。

Table 4. Comparison of transcription results of different grapheme-to-phoneme conversion models
表 4. 不同字音转换模型标音结果的对比

模型	辅助处理	测试集标音准确率
HMM-GMM	字素切分并与音素对齐，数据平滑	81.24%
CRF	字素切分并与音素对齐，特征模板	88.66%
RNN(tanh)	字素切分并与音素对齐	87.54%
LSTM + CTC	端到端	88.63%
GRU + CTC(ours)	端到端	91.88%

从表中可以看出，本文所提出的模型在西班牙语字－音转换任务上取得了最优的标音结果，证明了本文所提方法的优越性。同时需要强调的是，除了标音准确率最高之外，基于 RNN + CTC 的方式可以实现端到端的训练与解码，即模型的输入是构成单词的原始字符序列，而输出直接就是对应的音素序列，中间不再需要进行单词字素的切分以及与音素间的对齐操作，不仅大大简化了训练数据的构建复杂度，还可以直接实现较长字符序列到较短音素序列的自动转换，避免标音过程中额外的预处理或是后处理操作。这是 HMM、CRF 以及普通 RNN 无法实现的功能。同时与基于统计方法的 HMM、CRF 模型相比，神经网络能够自动的提取字符序列的特征，从而避免了不同语种特征工程的制约，让不具备西班牙语背景的计算机从业人员专注于发音词典模型的构建，而让西班牙语专业人员致力于高质量词库的建设，从而实现更优化的人力资源配置。

6. 总结

文章首先对西班牙语发音字典的基本构建办法进行了介绍，分析了该方法存在的问题与目前主流的解决方案，最终采取基于神经网络模型建模的方式完成了西班牙语词汇的发音预测。总体上，我们将西班牙语发音词典的构建作为一个序列标注任务来建模，首先利用 word2vec 框架在最初的西班牙语词库上训练字符嵌入向量并形成分布式的字符编码向量表示。之后，基于循环神经网络和连接时序分类算法构建了西班牙语字——音标注模型，并在自建的发音词典语料上进行了训练与测试。试验结果显示，基于字符嵌入 + 循环神经网络 + 连接时序分类的字－音标注模型可以获得更高的标音准确率，同时较传统的模型有更简单的标注流程，对数据集的要求也要低得多，可有效实现端到端的西班牙语单词标音任务，对于西班牙语语音信息处理任务具有重要意义。

致　谢

最后，特别感谢马延周老师和黄晓辉老师的指导与帮助，感谢面向特定领域连续语音识别的多语种发音词典构建研究基金项目的支持，另外，对所有给予转载和引用权的资料、图片、文献、研究思想和设想的所有者表示感谢。

基金项目

面向特定领域连续语音识别的多语种发音词典构建研究(编号：2019JZSY0002)。

参考文献

[1] 唐美丽, 胡琼, 马廷淮. 基于循环神经网络的语音识别研究[J]. 现代电子技术, 2019, 42(14): 152-156.

[2] Veena, P.V., Anand Kumar, M. and Soman, K.P. (2018) Character Embedding for Language Identification in Hindi-English Code-Mixed Social Media Text. *Computación y Sistemas*, **22**, 65-74. https://doi.org/10.13053/cys-22-1-2775

[3] Fang, C., Moriwaki, Y., Li, C.H. and Shimizu, K. (2019) Prediction of Antifungal Peptides by Deep Learning with Character Embedding. *IPSJ Transactions on Bioinformatics*, **12**, 21-29.

[4] 方春, 孙福振, 李彩虹, 邢林林. 基于深度学习和字符嵌入的细胞穿透肽预测[J]. 计算机仿真, 2019, 36(10): 353-358.

[5] 杨丽, 吴雨茜, 王俊丽, 刘义理. 循环神经网络研究综述[J]. 计算机应用, 2018, 38(S2): 1-6+26.

[6] Graves, A. and Jaitly, N. (2014) Towards End-to-End Speech Recognition with Recurrent Neural Networks. *Proceedings of the* 31*st International Conference on Machine Learning* (*ICML*-14), United States, 1764-1772.

[7] Schuster, M. and Paliwal, K.K. (1997) Bidirectional Recurrent Neural Networks. *IEEE Transactions on Signal Processing*, **45**, 2673-2681. https://doi.org/10.1109/78.650093

[8] Graves, A., Fern´andez, S., Gomez, F. and Schmidhuber, J. (2006) Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. *ICML*'6: *Proceedings of the* 23*rd International Conference on Machine learning*, Pittsburgh, PA, 369-376. https://doi.org/10.1145/1143844.1143891