严泉勇,吴跃成. 基于深度学习的汉盲转换系统[J]. 智能计算机与应用,2024,14(6):183-187. DOI:10.20169/j.issn.2095-2163.240626

基于深度学习的汉盲转换系统

严泉勇, 吴跃成
(浙江理工大学 机械工程学院, 杭州 310018)

摘 要: 汉盲转换是将中文电子文本信息自动转换成盲文数字化信息,这将会对视障人士的学习方式和生活水准产生巨大的影响。本文为解决汉盲转换中的分词连写问题,基于深度学习技术构建了一种双向门限循环单元条件随机场(BIGRU-CRF)网络模型。该模型通过双向门限循环单元获取上下文信息,再结合条件随机场模型序列标注特性,从而得到较为准确的分词连写结果,极大地提升了汉盲转换的准确率。测试结果表明,所设计的汉盲转换系统能够精确地将汉字转换成盲文,具有一定的实用性。

关键词: 汉盲转换系统; BIGRU-CRF; 分词连写; 深度学习
中图分类号: TP391 **文献标志码**: A **文章编号**: 2095-2163(2024)06-0183-05

Chinese-braille conversion system based on deep learning

YAN Quanyong, WU Yuecheng
(School of Mechanical Engineering, Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract: Chinese-braille translation is the automatic translation of Chinese electronic text information into Braille digital information, which will have a huge impact on the learning style and living standards of the visually impaired. In this paper, a Bidirectional Gated Recurrent Unit Conditional Random Field (BIGRU-CRF) network model is constructed based on deep learning technology to solve the problem of Chinese-blind translation. The model obtains the context information through the bidirectional gated recurrent unit, and combines with the characteristics of conditional random field model sequence labeling, so as to obtain more correct word segmentation and sequential writing results, and greatly improves the accuracy of Chinese-braille translation. The experimental results show that the Chinese-blind translation system designed in this paper can accurately convert Chinese characters into braille, and has certain practicability.

Key words: Chinese-braille conversion system; BIGRU-CRF; segmentation and sequential writing; deep learning

0 引 言

研究指出,中国的视障人士总群体目前多达1700万[1]。虽然这部分人存在视力障碍,但是却也有学习能力和阅读需求[2]。盲文是盲人获取信息的主要途径,如今也出版了多种纸质盲文书籍。盲人通过触摸盲文书籍获取信息,但是盲文书籍成本高、制作困难,同时盲文书籍多以医学为主,覆盖领域狭窄、且内容更新有所滞后,盲人无法获取多领域和最新的信息。在当前数字化、信息化技术不断发展的时代背景下,许多高校和研究所均已研发推出了显示盲文的盲文点显器[3-4],将盲文用机械设备显示出来,但目前着重于研究盲文点显器的机械部分和电路设计部分,对于汉盲转换的研究仍较为匮乏。基于此,本文提出一种基于深度学习的汉盲转换系统。

1 汉盲转换的相关研究

汉语不同于字母文字,词与词之间没有空格作为分隔符,并且汉语有很强的语义和语义规则,汉盲转换必须要先对文本进行中文分词处理。然而中文分词规范与盲文分词规范有很大的不同。例如,在盲文中,'不'与单音节中的动词、形容词、介词中的程度副词应当进行连写[5]。另外,盲人是通过摸读的形式获取盲文信息,无法像正常人一样一目十行。如果盲文的词串太长,会对盲人造成压力,使其产生

基金项目: 浙江省基础公益研究计划(LGF19E050005)。
作者简介: 严泉勇(1999-),男,硕士研究生,主要研究方向:人机交互,Email:3565796809@qq.com; 吴跃成(1966-),男,博士,副教授,主要研究方向:人机交互。
收稿日期: 2023-04-06

疲劳和混淆。比如像“工匠精神”要切分为“工匠 | 精神”两个词。为解决这一问题,有 2 种处理方案。一种是先对文本进行中文分词,再根据中国盲文标准中给出的分词连写的词法、语法和语义规则,对中文分词的结果进行调整,转换成符合盲文规则的词串。黄河燕等学者[6]在汉盲转换中最先设计了基于 SC 文法的多知识一体化的规则。李宏乔等学者[7]定义了一系列形式化的连写规则。庄丽等学者[8]在分词时尝试融合语义知识和语言模型以提高准确率。还有一种思路是直接建立分词连写词库,其数据源是已经出现过的分词连写组合,然后基于词库直接对汉语文本分词[9]。然而,分词连写组合的数量是十分庞大的,直接建立一个分词连写库的工作量极大。最好是在目前已经建设完成的中文数据集的基础上按照盲文规则进行加工处理。

综合前文论述可知,汉盲转换的难点在于文本按照盲文规则分词连写。一般的分词方法按原理可分为 3 类:词典匹配法、统计学习法、深度神经网络法。其中,词典匹配法原理简单,但该方法难以解决分词过程中的歧义切分和未登录词的识别。统计学习法常见的有隐马尔可夫模型(HMM)、最大熵隐马尔可夫模型(MEMM)、条件随机场模型(CRF)等。具体地,HMM 在多特征描述观察序列的情况时不适用,MEMM 借助最大熵框架进行特征选取解决这一问题,但是 MEMM 会出现标识偏置问题。CRF 具有 MEMM 的一切优点,同时 CRF 解决了 MEMM 的标记偏置问题。基于统计学习的分词方法可以减少歧义和未登录词的影响,分词性能有所提升。

目前,统计学习技术已经在人们常见的领域之中得到充分利用,例如语音识别、图像处理、中文分词等。传统的统计学习提取一些原始数据的特征时,需要人工构造复杂的特征提取器,比如在汉语分词连写时,往往要人为从语料库提取特征,这样就提升模型复杂度,人工提取特征工作量极大。为此就需要一种可以自己从原始数据中自动识别并提取那些所需的特征的算法,而这种算法就是近年来流行的深度学习算法。深度神经网络模型已经在语音识别、目标检测等领域取得了丰硕成果。同时,在自然语言处理(NLP)领域,也凸显其优势。Zheng 等学者[10]较早将神经网络模型应用到中文分词领域。Chen 等学者为解决分词中长序列信息难以学习的难题,先后引入了 GRNN 模型[11]和 LSTM 模型[12]。传统的 RNN 网络模型在处理中长序列信息时,隐藏节点简单使用了 tanh 函数,网络模型内部进行非线性的循环信息传递,这将导致模型训练过程中会出现梯度爆炸和梯度消失的问题。为解决 RNN 模型存在的问题,LSTM 模型单元则被提出,LSTM 模型利用门单元对历史信息的自动选择和记忆,解决了长序列信息中的依赖性问题。LSTM 模型拥有输入门、遗忘门和输出门三个门结构,结构相对比较复杂,模型训练时间较长。Cho 等学者[13]提出一种门限循环单元(Gated Recurrent Unit,GRU)神经网络分词方法,将 LSTM 中的输入门和遗忘门合并成更新门,所以 GRU 只具有重置门和更新门两个门结构。LSTM 模型和 GRU 模型都是由 RNN 模型发展而来,所以两者有相似的分词效果,但 GRU 模型的结构更为简洁,为此 GRU 模型的分词速度更快,Jozefowicz 等学者[14]在对比训练了 LSTM 模型和 GRU 模型后,已经验证了这一点。但是 LSTM 模型和 GRU 模型都只有单向处理信息的能力。为此,金宸等学者[15]提出了双向的 LSTM (BI-LSTM)模型,能够同时捕获上下文两个方向的信息。

本文构建双向门限循环单元条件随机场(BIGRU-CRF)网络模型,此模型结合了神经网络模型,考虑长远的上下文信息和条件随机场(CRF)联合考虑序列内部信息和外部观测信息的特性,得到了较好的分词连写效果。

2 汉盲转换的设计与实现

2.1 汉盲转换整体流程

本文的汉盲自动转换采用多步转换方法,即对输入文本采用 BIGRU-CRF 模型进行分词连写处理,再将文本的汉字和非汉字字符进行识别区分,其中非汉字字符主要包括数字、标点符号、英文字母等。对于非汉字字符可以直接根据字符与盲文对应的词典进行盲文转换,而汉字字符则先要进行汉语到拼音的拼音化处理转换,然后再进行拼音到盲文的转换,最后输出为盲文 ASCII 码和盲文点序,同时还有盲文显示设备可以直接读取的数据。总体流程如图 1 所示。

2.2 分词连写的数据库的构建

盲文分词连写规则的设计是为了让盲人摸读时更加方便以减少学习盲文和阅读盲文的困难,其本质还是符合汉语语法的逻辑性和规则性,其中很多可以看作在中文分词的基础上进行单音节的连写和对音节较长的词进行分词。对于这些简单的分词连写处理,可以在中文分词的基础上使用正则表达式替换处理即可,而对于一些无法用计算机处理的内

容可以使用人工替换。总之，构建汉语盲文分词连写库完全可以在按照标准中文分词规则的数据库的基础上进行改进，进而得到按照盲文分词连写规则的汉字文本数据库。

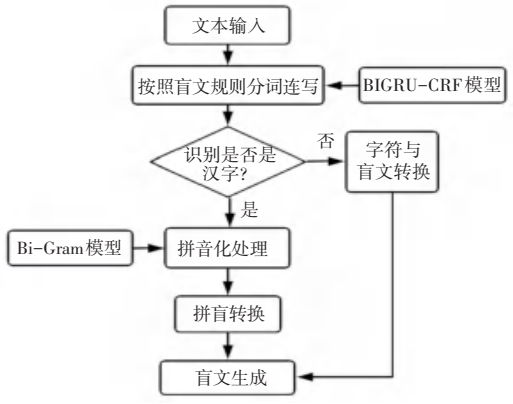

图 1　汉盲转换的整体流程

Fig. 1　The overall process of Braille conversion for Chinese

2.3　BIGRU-CRF 汉语盲文分词连写模型

本文所采用的分词模型的功能是将汉字文本按照盲文规则进行分词连写，这需要在前文所构建的盲文分词连写规则的数据库训练分词模型。汉语盲文分词连写问题可以看作是自然语言处理中的序列标注问题。序列标注分词中常用的方法是四位序列标注法（B，M，E，S），其中 B 表示一个词的开头位置，M 表示一个词的中间位置，E 表示一个词的末位置，S 表示单独成词。所以分词的过程可以通过分词模型得到每个字的位置类别，然后合并成词再输出即可。传统的统计学习解决序列标注问题最好的方法是采用条件随机场模型，本文在传统的统计学习算法的基础上引入深度神经网络模型，最终构造了 BIGRU-CRF 分词模型。在 BIGRU-CRF 算法中，BIGRU 具有利用句中长远的上下文信息进行中文分词的能力，这对处理容易产生歧义的地方非常有帮助；CRF 在序列标注的过程中，能够联合考虑到每个字前后相邻的标签特征，在最后的解码计算最大路径时，可以得到全局最优的序列标注的结果。

模型的结构如图 2 所示。由图 2 可看到，该模型一共有 5 层网络结构。第 1 层是 Word embedding 层，将输入的文本转换为字向量。第 2 和第 3 层是 BIGRU 网络层。第 4 层是 Dropout 层，是为了防止 BIGRU 模型出现过拟合，而在 BIGRU 网络层添加 Dropout 层，每次按一定比例舍去一部分网络节点。第 5 层是 CRF 模型层。

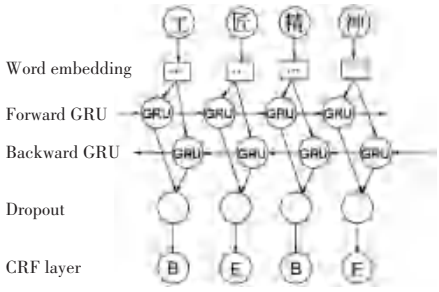

图 2　BIGRU-CRF 网络模型结构

Fig. 2　BIGRU-CRF network model structure

2.4　拼音化处理

拼音化处理就是对汉字进行拼音转换和添加声调，通常可以采用汉语-拼音词典匹配的方式。然而汉字存在一字多音和一音多调的歧义现象，对于这些词的拼音化处理非常复杂。如果只采用汉语-拼音词典匹配的方式将受歧义问题影响较大，本文将多音字的汉语 - 拼音语料库作为训练语料构建 n 元语法模型（n - gram），对于 n 元语法模型，可以近似认为一个词的概率只依赖其前面 n - 1 个词，n 的值通常不会取太大，本文取 n = 2，此时模型被称为二元语法模型（Bi - Gram）。例如，在拼音标调处理时，一个词假如有 2 种声调形式设为 w_1（上声）和 w_2（去声），在该词的前一个词拼音形式为 w_0，此时要比较 $P(w_1 \mid w_0)$ 和 $P(w_2 \mid w_0)$ 值的大小，选取概率值最大作为多音字标调形式的结果。

2.5　盲文生成

盲文的转换格式有不带调现行盲文、全带调现行盲文、国家通用盲文、双拼盲文四种，而每种盲文有盲文 ASCII 码和盲文字符两种固定表达形式。本文设计的汉盲转换系统还将盲文转换成盲文点序数据形式，这是一种盲文显示设备可以读取的数据形式，这样就可以使用户通过盲文显示设备学习和阅读中文文本。从盲文的点序图来看，一方盲文主要有 6 个点，从上至下左列点位记为“123”、右列点位记为“456”。如果将点位处的凸点记为 1，点位处的凹点记为 0，那么空方表示为“000000”，满方表示为“111111”，一方共有 64 种不同的表示形式。其中，盲文字符图、盲文 ASCII 码和盲文点序三者是存在映射关系的。盲文之间的转换示意如图 3 所示。中文文本通常包括汉字、标点符号、阿拉伯数字等内容。其中，汉字需要经过分词、拼音化处理再转换成盲文，这里以空方来表示分词的间隔。而其他非汉字字符则可直接转换成盲文。汉盲转换的演示如图 4 所示。本文以“2020 年，中国全面奔小康。”为例句，对汉盲转换的流程进行演示。

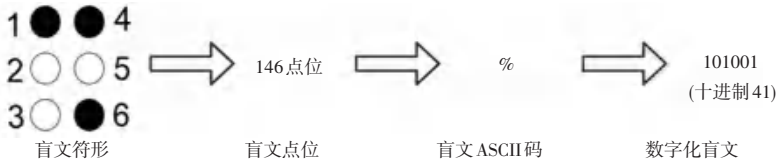

图 3 盲文转换示意

Fig. 3 Braille conversion diagram

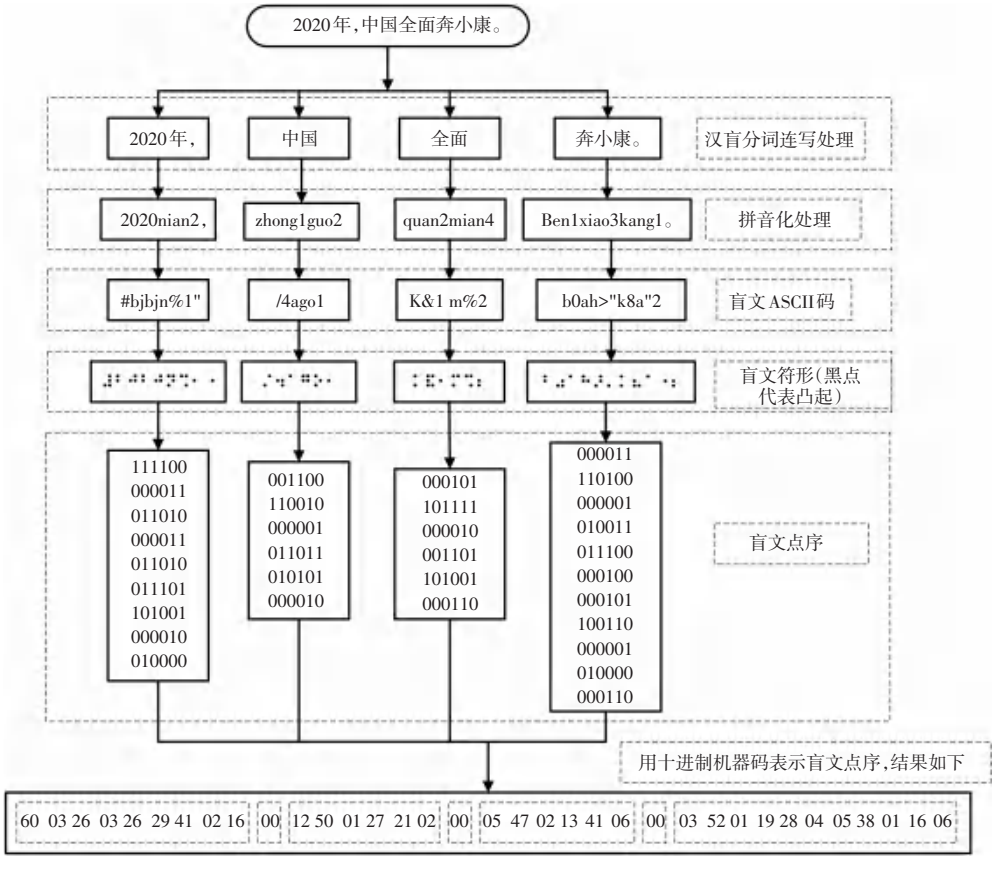

图 4 汉盲转换的演示图

Fig. 4 Demonstration diagram of Chinese-braille conversion

3 实验测试与结果分析

汉盲转换系统生成的盲文点序最终将应用到自主研制的盲文点显设备中,盲文点显设备的展示如图 5 所示。盲文点显设备的工作原理是将盲文点序的二进制序列作为控制机械本体凸凹显示的序列,最终实现中文到盲文的转换。非常有利于盲人学习和阅读盲文。

图 5 盲文点显设备的展示图

Fig. 5 Display diagram of Braille dot display device

3.1 汉盲分词连写算法实验测试

3.1.1 数据集

本文选取来自北京大学语言研究所提供的 Peking University 数据集(简称 PKU 数据集)和 2014 年人民日报数据集(简称 PFR 数据集),由于 PKU 和 PFR 数据集是按照中文分词规则进行分词,所以要将数据集按照汉盲分词连写规则进行改进,同时采用 BMES 四词位标注法对数据标注。将改进后的数据按照 9∶1 的比例分割为训练集和测试集。

3.1.2 评价指标

汉盲分词连写算法采用标准评估指标:准确率 $P(precision)$,召回率 $R(recall)$ 及 $F1$ 值对模型进行评估。其对应的计算公式如下:

$$P = \frac{\text{正确分词的数量}}{\text{模型切分的词语数量}} \times 100\% \qquad (1)$$

$$R = \frac{\text{正确分词的数量}}{\text{测试集中的总词语数量}} \times 100\% \quad (2)$$

$$F1 = \frac{2 \times P \times R}{P + R} \times 100\% \quad (3)$$

3.1.3　实验环境

汉盲分词连写模型训练环境配置见表 1，实验测试结果见表 2。

表 1　实验环境配置表

Table 1　Experimental environment configuration table

实验环境	参数值
CPU	Intel(R) Core(TM) i7-11800H
GPU	NVIDIA GeForce RTX 3050 Ti
操作系统	Windows 10
编程语言	Python
深度学习框架	TensorFlow

表 2　实验测试结果

Table 2　Experimental test results　%

训练语料	准确率 P	召回率 R	$F1$ 值
PKU	96.72	96.58	96.65
PFR	96.69	96.55	96.62

3.2　汉盲转换系统的测试

为了验证多步转换的汉盲转换系统的可行性，本文选取人民日报数据集 2014 年中部分的新闻文本作为中文语料，以此构建汉盲对照库。为方便语料库的使用，将超过 50 字的句子在标点处分割为短句，对于不适合拆分的句子依然保留。选取约 18.6 万条句子语料，数据库不同长度句子的数量见表 3。将中文语料按上述方式切分后，由阳光专业盲文编辑排版系统转换成现行盲文的 ASCII 码，将其作为汉盲对照库。将新闻文本语料在本系统得到的盲文 ASCII 码和汉盲对照库中的盲文 ASCII 码进行比较，在忽略标调和分词连写的影响时，最终得到该系统的准确率为 96.17%。

表 3　汉盲对照库中的中文句子数量统计

Table 3　The statistics of the number of Chinese sentences in Chinese-braille parallel corpus

句子长度	(0,10]	(10,20]	(20,30]	(30,40]	(40,50]	>50	总计
数量	6 188	29 937	44 617	50 927	49 877	4 787	186 333

4　结束语

面对传统纸质盲文书籍成本高且制作困难的难题，同时现如今的盲文书籍也多是以医学为主，覆盖领域狭窄且内容陈旧，盲人无法获取多领域和新兴的信息，难以满足盲人的需求。随着信息数字化时代的到来，发展数字化盲文则尤显重要，而汉盲转换是盲文数字化建设的关键一环。为解决这一问题，本文基于深度学习技术设计了一种汉盲自动转换系统。实验结果表明，该系统能够有效地实现汉盲转换，为盲人学习和阅读盲文提供了便利和支持。

参考文献

[1] 毛志伟，傅悦，崔瑶. 视障群体的信息无障碍应用现状分析[J]. 信息记录材料，2019，20(7)：51-53.

[2] 赵英. 针对残障人士的信息无障碍影响因素研究[J]. 四川大学学报(哲学社会科学版)，2018(5)：84-93.

[3] LI Zhipeng, WANG Rui, ZHANG Tianchi, et al. Intelligent braille conversion system of Chinese characters based on Markov model[C]// IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference. Chengdu, China: IEEE, 2019: 1283-1287

[4] 吴新丽，祝盼飞，杨文珍，等. 触觉再现的分层电磁式盲文点显器[J]. 系统仿真学报，2016，28(9)：2220-2226，2234.

[5] 滕伟民，李伟洪. 中国盲文[M]. 北京：华夏出版社，2006.

[6] 黄河燕，陈肇雄，黄静. 基于多知识分析的汉盲转换算法[C]// 全国第七届计算语言学联合学术会议. 哈尔滨：中国中文信息学会，2003：617-623.

[7] 李宏乔，樊孝忠，李良富，等. 汉语-盲文机器翻译系统的研究与实现[J]. 计算机应用，2002，22(11)：3-6.

[8] 庄丽，包塔，朱小燕. 盲人用计算机软件系统中的语音和自然语言处理技术[J]. 中文信息学报，2004，18(4)：72-78.

[9] 杨潮，车磊. 汉字盲文转换系统的设计[J]. 北京印刷学院学报，2011，19(6)：36-38.

[10] ZHENG Xiaoqing, CHEN Hanyang, XU Tianyu. Deep learning for chinese word segmentation and POS tagging[C]//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle, USA: Association for Computational Linguistics, 2013: 647-657.

[11] CHEN Xinchi, QIU Xipeng, ZHU Chenxi, et al. Gated recursive neural network for Chinese word segmentation [C]// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Beijing: Association for Computational Linguistics, 2015: 567-572.

[12] CHEN Xinchi, QIU Xipeng, ZHU Chenxi, et al. Long short-term memory neural networks for Chinese word segmentation [C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal: Association for Computational Linguistics, 2015: 1385-1394.

[13] CHO K, VAN M B, GULCEGRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation [C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar: Association for Computational Linguistics, 2014: 1724-1734.

[14] JOZEFOWICZ R, ZAREMBA W, SUTSKEVER I. An empirical exploration of recurrent network architectures [J]. Journal of Machine Learning Research, 2015, 37(1): 2342-2350.

[15] 金宸，李维华，姬晨，等. 基于双向 LSTM 神经网络模型的中文分词[J]. 中文信息学报，2018，32(2)：29-37.