文章编号：1003-0077(2019)04-0060-08

基于汉盲对照语料库和深度学习的汉盲自动转换

蔡 佳[1,2]，王向东[1]，唐李真[3]，崔晓娟[1,2]，刘 宏[1]，钱跃良[1]

(1. 中国科学院 计算技术研究所 移动计算与新型终端北京市重点实验室，北京 100190；
2. 中国科学院大学，北京 100049；
3. 中国盲文出版社，北京 100142)

摘 要：汉盲转换是指将汉字文本自动转换为对应的盲文文本，其在盲文出版、盲人教育等领域具有重要应用价值，但当前已有系统性能难以满足实用需求。该文提出一种基于汉盲对照语料库和深度学习的汉盲自动转换方法，首次将深度学习技术引入该领域，采用按照盲文规则分词的汉字文本训练双向 LSTM 模型，从而实现准确度高的盲文分词。为支持模型训练，提出了从不精确对照的汉字和盲文文本中自动匹配抽取语料的方法，构建了规模为 27 万句、234 万字、448 万方盲文的篇章、句子、词语多级对照的汉盲语料库。实验结果表明，该文所提出的基于汉盲对照语料库和深度学习的汉盲转换方法准确率明显优于基于纯盲文语料库和传统机器学习模型的方法。

关键词：汉盲转换；中国盲文；盲文语料库；深度学习

中图分类号：TP391 **文献标识码**：A

A Deep Learning Method for Chinese-Braille Conversion Based on Parallel Corpora

CAI Jia[1,2], WANG Xiangdong[1], TANG Lizhen[3], CUI Xiaojuan[1,2], LIU Hong[1], QIAN Yueliang[1]

(1. Beijing Key Laboratory of Mobile Computing and Pervasive Device, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China;
2. University of Chinese Academy of Sciences, Beijing 100049, China;
3. China Braille Press, Beijing 100142, China)

Abstract: The Chinese-Braille conversion can be applied to fields such as Braille publication, education for the blind, etc. This paper presents a deep learning solution to automatic Chinese-Braille conversion based on parallel corpora. A Bi-directional LSTM model is trained using segmented Chinese texts according to the Braille segmentation rules and achieves high accuracy of Braille word segmentation. In order to support the model training, this paper also presents a strategy of automatically generating a corpus from Chinese and braille texts with the same content, with alignments at article-level, sentence-level and word-level, totaling 270 000 sentences, 2.34 million Chinese characters, and 4.48 million Braille symbols. The experimental results show that the proposed method outperforms the existing models.

Keywords: Chinese-Braille conversion; Chinese Braille; Braille corpus; deep learning

0 引言

盲文是盲人阅读和获取信息的重要方式。它是一种触觉符号系统，印刷在纸张或显示在点显器上，通过触摸进行阅读。盲文的基本单位称作“方”，一方包含 6 个点位，通过设置每个点位是否有点共可形成 64 种组合，这些组合构成了最基本的盲文符号。图 1(a)给出了一个盲文符号的示例。

为了生成盲文内容，需将普通人使用的文字内容转换为盲文。不同语言对应的盲文是不同的。对于字母文字，其对应的盲文往往直接定义了从字母到盲文符号的唯一映射，因此转换相对简单。当前，英语、葡萄牙语、丹麦语、西班牙语、印地语等语言的

收稿日期：2018-06-14 **定稿日期**：2018-10-17
基金项目：国家科技支撑计划课题(2014BAK15B02)

文本到其相应的盲文文本的自动转换，都已有可用的计算机系统[1-5]。而在汉语中，由于不可能将汉字唯一映射到盲文符号，汉语盲文被定义为一种拼音文字，并且还定义了分词连写和标调等规则。汉语盲文的这些特点为汉盲转换，即汉字到盲文的转换带来了很大困难。现有的汉盲自动转换系统准确率较低，难以实用。在盲文出版、盲人教育等行业中，目前仍主要采用人工进行汉盲转换，效率低、成本高，导致盲文读物匮乏、盲人获取信息困难，严重限制了盲人在信息社会的发展。

汉语盲文有 3 种相近的方案，分别称为现行盲文、双拼盲文[6-7]和通用盲文[8]，其中现行盲文使用最广，当前占据主导地位，双拼盲文使用较少，通用盲文是对现行盲文的改进和规范，目前正在进行推广。汉语盲文一般用 2～3 方表示一个汉字，其中一方表示声母，一方表示韵母，现行盲文和通用盲文中有些情况需要再增加一方表示声调。图 1(b)给出了一个盲文词(“中国”)的现行盲文表示。汉语盲文与汉字文本最大的区别在于盲文的“分词连写”规则，即要求词与词之间用空方分隔。但盲文分词与常用的汉语分词不同，为减少单音节词可能带来的歧义，许多汉语中的短语在盲文中需要连写，例如，“王老师”“大红花”“不能”等都需要连写。针对分词连写，中国盲文标准中给出了 100 多条基于词法、语法和语义的细则，如“‘不’与动词、能愿动词、形容词、介词、单音节程度副词均应连写”[6]。另一方面，为减少同音字造成的歧义，盲文还制定了标调规则。双拼盲文和通用盲文中几乎每个字都可确定声调。而在现行盲文中，为节省阅读时间和印刷成本，规定只对易混淆的词语、生疏词语、古汉语实词、非常用的单音节词等标调。一般认为现行盲文的标调率大约在 5%。

可以看出，汉盲转换的关键在于分词和标调。当前研究大多集中在分词方面，主要遵循两种思路。一是按照盲文分词连写本身的逻辑，首先对文本进行汉语分词，然后使用预定义的规则对汉语分词结果进行调整，将汉语词串转换为盲文词串[9-13]。当前大多数研究都基于这一思路，例如，黄河燕等[9]最先提出和采用了基于 SC 文法的规则；李宏乔等[13]定义了 183 条形式化连写规则，其中包含 41 条准短语性规则；朱小燕团队[10-12]尝试融合语义知识和语言模型以进一步提高盲文分词的准确率。但是，盲文分词连写涉及主观性很强的语法和语义规则，计算机定义和处理都很困难，导致这种方法的性能存在瓶颈，难以进一步提升。第二种思路是从盲文语料中提取出现过的连写组合，建立分词连写库，然后基于分词连写库进行文本分词[14]或对汉语分词结果进行后处理。但是，盲文将汉语中的许多短语连写，所形成的连写组合是无限的，无法通过分词连写库穷举。因此这一方法性能有限，目前主要和第一种方法结合，作为一种补充式的后处理操作使用[14-15]。

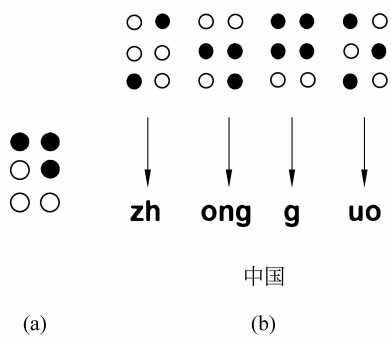

图 1　盲文示例

注：(a)为一个盲文符号示例，对应英语盲文中的字母 D 或汉语盲文中的声母 d；(b)为“中国”的现行盲文表示。

最近几年，中科院计算所的 Wang 等[16]提出了基于机器学习的盲文直接分词框架，不再基于汉语分词结果进行后处理，而是利用训练好的盲文分词模型直接对盲文串进行分词。这种方法采用机器学习模型隐含地刻画盲文分词连写规范，避免了计算机直接处理复杂的语法和语义规则，实验结果表明，此方法可大大提升汉盲转换的准确率。但这一方法也存在不足：一方面，该方法基于感知机模型，而近年来，深度学习技术在很多领域已逐步替代感知机和统计机器学习等传统方法；另一方面，模型训练基于盲文语料，而盲文只表示汉字的读音(且大多数不加声调)，导致可能因同音产生歧义，进而影响最终的分词结果。如果采用按照盲文规则分词的汉字文本作为训练语料，则可以避免上述问题。

要得到按照盲文规则分词的汉字文本语料，相当于将汉字文本及与其对应的盲文文本进行词语级对齐，即需要建设一个词语级对照的汉盲语料库。目前尚无可用的此类语料库。2014 年，国家社科基金启动了重大项目“汉语盲文语料库建设研究”，计划建成约 1 000 万方的汉语盲文语料库，保持“盲文—拼音—汉字”的对照形式。该项目在语料库构建中采用信息技术进行处理，但人工校对工作量仍然极大，语料库目前仍在建设中[17]。

在盲文自动标调方面，由于现行盲文的标调规

则极为主观，计算机难以有效判定生僻词和易混淆词，因此已有系统大多只支持全标调或不标调等简单模式。Wang 等[16]提出了一种基于 n-gram 语法的标调方法，采用机器学习模型从盲文语料中自动学习标调规律，取得了较好的效果。

本文提出了一种基于汉盲对照语料库和深度学习的汉盲自动转换方法，首次将深度学习技术引入该领域，采用按照盲文规则分词的汉字文本训练双向 LSTM 模型，从而实现高准确度的盲文分词。为支持模型训练，本文提出了从不精确对照的汉字和盲文文本中自动匹配抽取语料的方法，利用 126 种盲文书籍构建了规模为 27 万句、234 万字、448 万方的篇章、句子、词语多级对照的汉盲语料库。实验结果表明，本文提出的基于汉盲对照语料库和深度学习的汉盲转换方法准确率，明显优于基于纯盲文语料库和传统机器学习模型的方法。

1 汉盲对照语料库

1.1 语料库设计

在进行语料库设计时，首先根据当前已有的盲文语料的问题及汉盲转换系统的应用特点，明确了语料库需要满足的若干需求，具体包括规模、内容、形式三个方面。

① 规模：为支持机器学习算法，特别是当前主流的基于深度神经网络的模型，语料库应具有较大规模，预期首期建成 20 万句以上。

② 内容：针对当前汉盲转换系统的需求，语料内容涉及的领域应兼顾通用性与若干重点领域。为保证通用性，内容应覆盖多个领域，文本来自多个作者的多种书籍；另一方面，对于盲文出版较为集中的特定领域，如中医推拿按摩等，应给予重点关注。语料库应尽可能地按领域划分为子语料库。

③ 形式：为生成按照盲文规则分词的汉字文本，语料库应在汉字和盲文文本之间实现篇章、句子、词语等的多级对照。语料应采用计算机方便读取的编码和存储格式。

根据上述需求，在内容方面，选用了中国盲文出版社编辑的 126 种书籍，划分为通用与文学、科学、医学三个子类，具体情况如表 1 所示。之所以将通用与文学并列为一类，一是由于两类别的内容较为相似，二是因为两类别的书籍种数较少，作为两类略嫌不足。

表 1 语料库领域子类划分

类别	书籍种数	内容	书籍名称
通用与文学	9	文学和盲文月刊类	《将苦难转化成祝福》《幸福在哪儿》《我的精神家园》《爱的教育》等
科学	100	新视野百科全书	《浩瀚宇宙》《太阳系》《地球与月球》《星座与观星》《天气与气候》等
医学	17	内科、外科等医学相关类	《中医基础理论》《中医诊断学》《中药学》《头痛 151 个怎么办》等

在编码方面，盲文领域一直存在多种计算机内编码，常用的有 Unicode 盲文编码、ASCII 编码及使用 Unicode 扩展域的自定义编码等。本文构建的语料库中盲文符号采用 ASCII 编码，这是由于 ASCII 编码更为简单，相对于 Unicode 等多字节编码更节省存储空间，且无需安装任何插件或字体即具备一定的可读性。另一方面，由于只需简单的字节映射，ASCII 编码可方便地转换为其他编码。

在存储格式方面，为了简单、方便，语料库设计为直接采用 txt 文件存储。为每个类别构建两个文件夹，每个文件夹中分别是每一篇文章对应的汉字和盲文 txt 文件，文件中每个句子占一行，汉字和盲文句子都按盲文规则分词。同名的汉字和盲文 txt 文件对应相同的篇章，对应篇章中相同行的文本对应同一句子，对应句子中相同位置的词对应同一词语（或含按盲文分词连写规则连写的词串）。这样，就以最简单的方式实现了汉字和盲文文本之间的篇章、句子和词语级对照。文件夹目录参见图 2，txt 文件中的内容如图 3 所示。

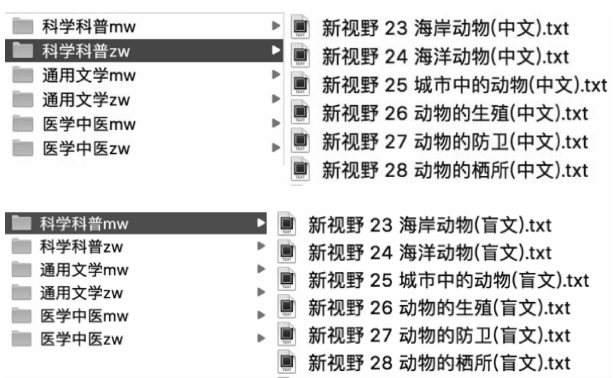

图 2 汉盲语料库存储目录示意图

我 的 一贯 做法	O D IG] ZOF9
是 每部 译作 出版 后	: MIBU I2ZO QUBV H(
并不 将 其	B*BU GX KI
束 之 高阁	:U / G6G5
而 是 时常 保持	R : :1Q8 B6Q1
亲密 接触	K<MI GEQU
会 常常 拿 原著 进行 比较 对照	HW Q8Q8 N91 &/U G<H* BIG> DW/6
有时 还 请 高手 阅读	\:1 H[K* G6:( )2DU
校对	G>DW
挑出 一些 毛病 来	T>QU IHE M6B* L[
这样	/5X
那些 译得 不够 准确	N9HE I2D5 BUG(/3K)
不够 满意 之 处 都 能 在 每次 再版 时 及时 改正 过来	BUG(MVI2 / QU D(N# Z[M!C2 Z[2BV :1 GI:1 G[/# GOL[
使得 译作 日臻 完善	:'D5 I2ZO J2/0]:V
做到 不愧对 作者 和 读者	ZOD6 BUKWDW ZO/5' H5 DU/5'
当然	D8JV
爱 的 教育	[2 D G>+2
也 不例外	E BULIY

图 3　对应的汉字与盲文 txt 文件内容示例

1.2　语料库构建

本文中，语料库构建主要采用自动方式，从内容相同的汉字和盲文书籍文件中自动对齐并抽取文本从而形成语料库。每本盲文书籍存储为一个阳光盲文编辑软件所用的 bdo 文件，每本汉字书籍存储为一个 Microsoft Word 文件。

语料库构建的主要难点在于实现汉字和盲文文本的句子级和词语级对应，原因有以下几点：第一，汉字和盲文的内容并不完全对应。为了便于盲人理解，盲文编辑会对内容进行适当的修改，比如文本增删、段落拆分和合并等。第二，盲文会增加目录、页码等内容，且都作为文本，不能通过特定的格式标记去除。第三，bdo 文件中合并了一些非标准的格式标记，有可能和文本内容混淆。因此，很难通过计算机自动化处理实现所有句子和词语的完全对应，只能抽取能够对应成功的部分、丢弃匹配失败的部分。由于本文目标是构建训练机器学习模型所需的语料库，所以这种处理是可以接受的。

语料库构建的主要流程如图 4 所示。从汉字文件和盲文文件中分别抽取文本，将盲文文本转换为 ASCII 编码，在各自进行句子切分等预处理后，利用匹配算法进行汉字和盲文的字符对齐，根据对齐结果输出多级对照的汉盲对照语料，形成汉盲对照语料库。

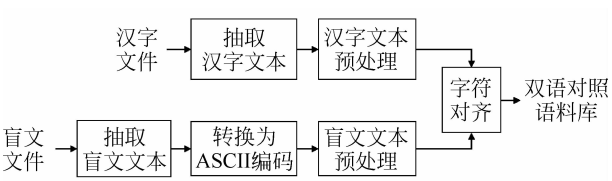

图 4　语料库构建流程图

1.2.1　预处理

在进行匹配和对齐之前，需要将汉字和盲文文本切分为句子。本文采用的方法为检测标点符号。采用的标点集如表 2 所示。

表 2　汉语-盲文 ASCII 码标点符号对照表

汉语	，	：	《	》	。	“	”	、	？	！	％
盲文	”	—	”—	—1	“2	^	^	@	”’	;1	⌗JO

汉语	……	（	）	·	——	‘	’	；	—	—
盲文	”"”	;’	,2	,’	,—	^^	^^	;	,”	—

汉字文本中标点的检测相对简单，直接搜索相应字符即可。盲文 ASCII 文件中的标点符号的形式相对复杂，标点符号之间存在包含关系，所以在预处理时需要添加规则判定以确认标点符号。首先使用 KMP 算法获得盲文标点的位置列表，然后对比具有包含关系的标点符号的位置信息，如果存在相同的位置信息，则删掉被包含的短字符串的位置信息。

1.2.2　字符对齐

在预处理阶段，对于同一篇章，汉字和盲文文本都已被切分为句子，形成两个句子集合。但是由于上文所述原因，两个句子集合并不能精确对应；更为重要的是，汉字文本中的句子（以下称为“汉字句子”）是不分词的，无法与盲文形成词语级别的对照。因此，字符对齐的任务就是：第一，匹配并且保留内容精确对应的汉字和盲文句子，丢弃无法建立对应关系的句子。第二，在句子中，将每个汉字与盲文建立对应关系，从而把汉字句子也按盲文的分词形式分词，形成如图 3 所示的对照。

设预处理后得到的汉字句子集合为$\{A, B, C, \cdots\}$，盲文句子集合为$\{A', B', C', \cdots\}$。首先，将每个汉语句子通过汉盲字典转换为对应的盲文句子集合。汉盲字典中列出了每个汉字对应的盲文符号串。由于汉字句子不分词，因此此时生成的盲文句子也并不分词。之所以是盲文句子集合，是因为汉语句子中的多音字可以对应多个不同的盲文符号串，因此根据句中多音字的所有读音进行全部组合，得到所有可能的盲文句子的集合。此时，汉语句子集合$\{A, B, C, \cdots\}$被转化为盲文句子集合的集合$\{\{a_1, a_2, \cdots\}, \{b_1, b_2, \cdots\}, \{c_1, c_2, ..\}, \cdots\}$，其中$\{a_1, a_2, \cdots\}$为汉语句子 A 对应的盲文句子的集合，其他依此类推。

对于每一个由汉语句子生成的盲文句子集合$\{a_1, a_2, \cdots\}$，检查其中的每个句子，判断是否与$\{A', B', C', \cdots\}$中的句子匹配。所谓匹配，是指两

个盲文句子在不考虑分词(即忽略空方)和不考虑标调(即忽略声调符号)的情况下完全相同。

若找到 A' 与 a_i 匹配,则将 a_i 按照 A' 分词,并进一步将 a_i 对应的汉字句子 A 按相同的方式分词,得到按照盲文规则分词的汉字句子 A''。这样就得到了词语级对照的汉语句子 A'' 和盲文句子 A'。保存 A'' 和 A'。如果没有找到 $\{A', B', C', \cdots\}$ 中的盲文句子能够与 $\{a_1, a_2, \cdots\}$ 中的任意一个句子匹配,则丢弃 $\{a_1, a_2, \cdots\}$ 及其对应的汉语句子 A,继续处理下一个汉语句子及其生成的盲文句子集合。整个流程如图 5 所示。

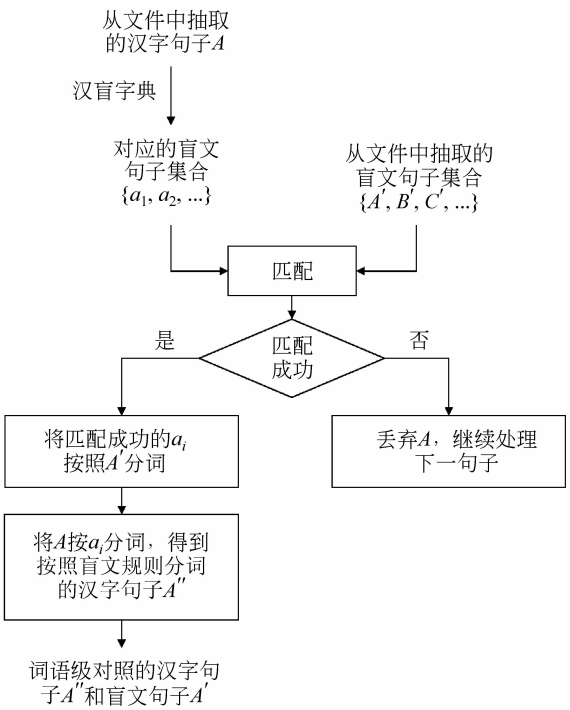

图 5 字符对齐算法流程图

1.3 WLCBC 语料库

经过上文所述的语料库构建步骤,我们利用 126 种书籍,成功构建了 WLCBC(Word Level Chinese-Braille Corpus)语料库。其规模如表 3 所示。

语料库的汉字部分编码为 UTF-8,盲文编码为 ASCII,语料库设计为直接采用 txt 文件存储。为每个类别构建两个文件夹,每个文件夹中分别是每一篇文章的中文和盲文的 txt 文件,每个句子占一行,汉字和盲文句子都按盲文规则分词,如图 2、图 3 所示。

2 基于深度学习的汉盲转换方法

基于本文构建的汉盲对照语料库,本文提出了一种基于深度学习的汉盲转换方法。该方法的核心是利用与盲文分词连写对应的汉字文本语料,训练符合盲文分词规范的深度神经网络分词模型。这种方法通过机器学习模型一次性地将汉字文本按盲文分词规范进行切分,相对于传统的先按汉语分词规范分词再利用盲文规则进行合并的方法更为简单、直接,避免了计算机处理人工定义的语义和语法规则时存在的困难。相对于利用纯盲文语料库训练盲文分词模型的方法[16],本文方法充分利用了汉盲对照语料库的优势,直接训练面向汉字文本的分词模型,可避免盲文因同音字词带来的歧义性。

表 3 汉盲对照语料库统计结果

句子数	中文字符数	盲文字符数
273 279	2 343 762	4 480 766

本文提出的基于深度学习的汉盲转换方法的主要流程如图 6 所示。首先将汉字文本按照盲文的规则分词,其中分词部分使用基于深度学习的双向 LSTM 模型。然后使用 n-gram 模型对分词后的汉字标调。最后将已经分词和标调的汉字文本转换为盲文,生成盲文文本。

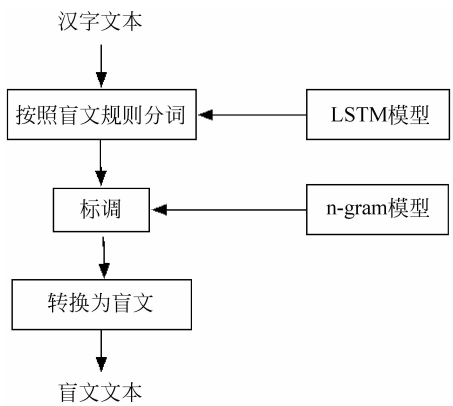

图 6 基于深度学习的汉盲转换方法流程

2.1 基于盲文规则的汉字文本分词

如上文所述,本文方法的核心在于直接将汉字文本按盲文规则分词,这是通过直接采用按盲文规则分词的汉字文本训练分词模型实现的。本文所构建的 WLCBC 语料库,通过在汉字和盲文文本间进行句子和字符匹配,实现了汉字和盲文在词语级的对照,获取了按照盲文规则分词的汉字文本语料(图 3)。利用这一语料训练的分词模型,即可用于将汉字文本直接按盲文规则分词。在分词模型方面,本文尝试采用深度学习模型,该模型近年来在汉

语分词等许多领域均得到了广泛应用，被证实效果优于传统的神经网络及统计机器学习模型。

通过深度学习进行分词属于分类问题：将每个字的位置分为 4 种，即 B、E、M、S，其中 B 代表词的开头，M 代表词的中间，E 代表词的末位，S 代表单独成词，分词的目的就是通过模型得到每个字的位置类别，然后合并成词。

本文选取了最近分词领域普遍采用的 LSTM 神经网络模型[18]，尝试将其用于基于盲文规则的汉字文本分词。本文采用的网络结构如图 7 所示。该模型共有 6 层网络，第 1 层是 Word embeddings 层，基于词向量模型，将训练语料中的字由 one-hot 编码映射为低维稠密的字向量。第 2 和第 5 层是 Bi-LSTM 网络层，共有两层 Bi-LSTM 层，为了防止过拟合，Bi-LSTM 网络层之后添加 Dropout 层，每次随机丢弃一定比例的神经网络节点。第 6 层输出层是一个全连接层，因为是多分类问题。设置全连接层的激活函数为 Softmax，它将多个神经元的输出映射到 0 到 1 之间的数值，选择概率最大的类别作为该字的类别。

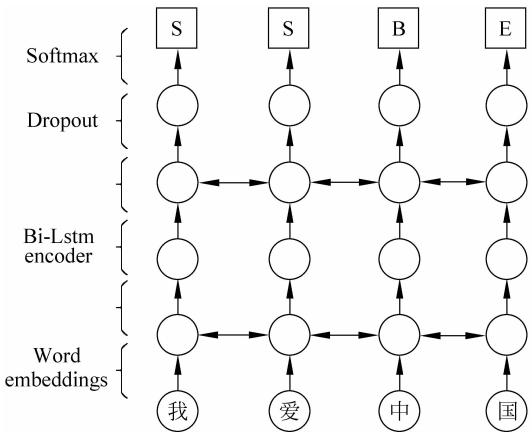

图 7　本文采用的深度神经网络结构示意图

模型训练前，需要将语料句子中的每个词以字为单位进行标记。另外，由于分词模型的输入是向量形式，因此需要训练训词向量模型，将语料转为向量表示。经过多轮训练，可生成所需的分词模型。分词模型的训练流程如图 8 所示。

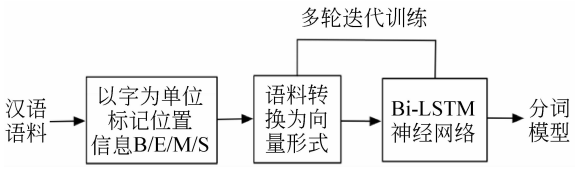

图 8　分词模型的训练流程

对一句话进行分词时，将文本转换为词向量，输入分词模型，通过模型计算得到每个字（向量）属于四种状态的概率，选择概率最大的作为该字的状态，最后合并得到分词结果。

2.2　基于统计学习的自动标调

本文基于构建的汉盲对照语料库，采用统计机器学习方法训练标调模型，从语料中学习隐含的标调模式，从而实现自动标调。本文采用的方法与文献[16]相似，区别在于文献[16]中的方法采用纯盲文语料库，其 n-gram 模型构建的对象为盲文词（含连写的词串），而本文方法采用汉盲对照语料库，n-gram 模型构建的对象为汉字词（含连写的词串）。由于多个同音的汉字词可对应同一个盲文词，因此本文方法更为精确。

对于构建的汉盲语料库，将其中所有的盲文词对应的汉字词的集合作为使用的词表。将语料中出现的同一词的不同标调形式（如不标调，首字标调，第二字标调……）作为不同词添加至词表。采用此词表和语料库训练一个 n-gram 语言模型。本文中，采用简单的 bi-gram 模型，训练时采用了 Kneser-Ney 平滑策略。

标调时，对于每一个待标调的词，根据其前 $n-1$ 个词的语言模型概率确定该采用哪种标调形式。例如，若某个两字词存在不标调形式 w_0 及两种标调形式 w_1（首字标调）和 w_2（第二字标调），此时，比较 $P(w_0|w)$，$P(w_1|w)$，$P(w_2|w_1)$ 的大小（其中，w 为该词之前的一个词），取概率最大的标调形式为最终选择。

2.3　汉一盲字符转换及特殊处理

在词语级对照的汉盲对照语料库的支持下，上文中的分词和标调两个步骤都是针对汉字文本进行的，相对于针对盲文文本进行分词和标调的方法[16]，避免了因盲文只表示读音而导致的信息丢失和歧义增加。本文方法中，在进行分词和标调之后，利用发音词典和发音-盲文映射表将分词和标调的汉字文本转换为盲文，转换过程中保留并复制其中的标调信息。

在文本转换时，会遇到一些特殊情况，如汉语文本中有时会夹杂阿拉伯数字、英文字母及一些特殊符号，盲文在“数字＋量词”和采用数字形式的年月日时会需要特殊处理（在数字后增加一个连接符）。针对这些情况，本文采用文献[16]中的方法进行必

要的处理。

3 实验

3.1 实验设置

基于第2节所述方法,本文搭建了一个用于实验的原型系统,其代码框架基于Python的Keras库。

为测试系统性能,将WLCBC语料库随机分为训练集和测试集,训练集规模约为21万句,测试集规模约为6万句。训练集和测试集的数据不重合,且不源于相同的书籍。训练集和测试集保留了通用文学、科学、医学的分类。具体情况如表4所示。

表4 实验数据情况

		句子数	词数
训练集	科普	45 172	1 007 802
	医学	119 375	
	通用文学	43 752	
	总计	208 299	
测试集	科普	16 012	79 171
	医学	42 023	167 472
	通用文学	6 945	39 810
	总计	64 980	286 453

实验的任务为汉字文本到盲文文本的转换。在训练时,使用句子、词语级对照的汉字和盲文文本按第2节介绍的方法训练分词模型和标调模型。在测试时,将测试集中的汉字文本去除分词(即删除词与词之间的空格字符)后得到的文本作为输入,系统输出转换后的盲文结果。将语料库中测试数据的盲文文本作为标准答案,与输出结果进行比较,计算转换准确率。准确率的计算方法为将输出结果与标准答案以词为单位进行编辑距离对齐,然后统计正确的词的个数,将正确的词的个数与标准答案总词数的比值作为准确率。实验同时统计了考虑标调和不考虑标调的准确率,前者代表最终的汉盲转换性能,后者可基本代表分词性能。

实验中,训练了用于分词的LSTM模型和用于标调的bi-gram模型。LSTM为两层双向网络,维度为512。bi-gram模型采用SRILM工具包训练而成。为进行比较,获取了文献[16]中基于盲文语料库和感知机模型的系统进行对比实验。同时,为验证深度学习方法的优越性,还训练了一个多层感知机(MLP)模型,其结构为两层Dense网络,7×100个神经元为输入层,隐藏层单元数为100,输出层单元数为4。词向量模型的训练语料为Sogou语料库,向量维度为200,迭代50次,使用Python的Gensim库训练模型。实验时,采用本文构建的汉盲对照语料库训练MLP模型,然后采用该模型实现按照盲文分词连写规则的汉字文本分词,后续的标调等处理与上文所述相同。

3.2 实验结果

汉盲转换的实验结果如表5和表6所示。可以看出,无论是考虑标调还是不考虑标调,对于所有领域,基于汉盲对照语料库的MLP模型和LSTM模型效果均优于采用纯盲文语料库的方法(文献[16]系统),LSTM模型的结果优于MLP模型,由此可以看出采用汉盲对照语料库和更复杂的机器学习模型的重要性。在不考虑标调时,本文提出的基于汉盲对照语料库和深度学习的分词算法可达到94.42%的准确率,已经达到实用水平。从各领域来看,科学科普的准确率最高,但这可能是由于训练语料和测试语料来自同一套丛书相似性较高造成的。而医学领域性能相对较低,这可能是因为其中与中医相关的测试语料包含一定的古文内容和中医专用词汇,而训练语料主要为现代汉语,只有一部分为医学领域语料,总量规模不是很大,导致训练尚不充分。

表5 汉盲转换准确率(不考虑标调)(%)

模型	全部数据	科学科普	医学医用	通用文学
文献[16]系统	91.84	95.49	85.90	90.79
MLP	92.22	95.58	91.98	92.45
Bi-LSTM	94.42	96.17	93.53	92.88

表6 汉盲转换准确率(考虑标调)(%)

模型	全部数据	科学科普	医学医用	通用文学
文献[16]系统	79.63	82.49	70.80	81.55
MLP	83.30	87.83	79.15	82.60
Bi-LSTM	85.11	92.44	83.50	85.91

4 结论

本文提出了一种基于汉盲对照语料库和深度学习的汉盲自动转换方法,首次将深度学习技术引入

该领域，采用按照盲文规则分词的汉字文本训练双向 LSTM 模型，从而实现高准确的盲文分词。为支持模型训练，采用从汉字和盲文文本中自动匹配抽取语料的方法构建了篇章、句子、词语多级对照的汉盲对照语料库，其规模为 27 万句、234 万字、448 万方盲文。实验结果表明，本文提出的基于汉盲对照语料库和深度学习的汉盲转换方法准确率明显优于基于纯盲文语料库和传统机器学习的方法，也优于基于汉盲对照语料库和多层感知器模型的方法。

参考文献

[1] Christensen L B, Keegan S J, Stevns T. SCRIBE: A model for implementing robobraille in a higher education institution[C]//Proceedings of International Conference on Computers Helping People with Special Needs. Springer-Verlag, 2012:77-83.

[2] Christensen L B, Chourasia A. Document transformation infrastructure[C]//Proceedings of 8th International Conference on Universal Access in Human-Computer Interaction (UAHCI 2014), Springer International Publishing, 2014:93-100.

[3] Christensen L B, Stevns T. Universal access to alternate media[C]//Proceedings of 9th International Conference on Universal Access in Human-Computer Interaction. Springer International Publishing, 2015: 406-414.

[4] Coutinho L R R, Girao A M, Frota J B B, et al. Device to assist the visually impaired in reading printed or scanned documents[C]//Proceedings of Brazilian Symposium on Computing System Engineering. IEEE Computer Society, 2012:25-30.

[5] Bodale F, Bhide U, Gore D, et al. Braille translation [J]. International Journal of Research in Advent Technology, E-ISSN, 2014, 57(20):2321-9637

[6] GB/T 15720—2008 中国盲文[S], 2008.

[7] 滕伟民，李伟洪. 中国盲文[M]. 北京：华夏出版社，2006.

[8] 钟经华. 汉语盲文规范化的新起点[J]. 现代特殊教育，2017，5(3):25-26

[9] 黄河燕，陈肇雄，黄静. 基于多知识分析的汉盲转换算法[C]. 全国计算语言学联合学术会议，2003.

[10] Xiaoyan Zhu, Ta Bao. EasyBraille: a translation system for Mandarin and Braille[C]//Proceedings of Natural Language Understanding and Machine Translation Proceedings of the 6th Joint Symposium on Computational Linguistics in China (JSCL-2001), Tsinghua University Press, Beijing, China, 2001: 326-331.

[11] Minghu Jiang, Xiaoyan Zhu. Segmentation of Mandarin Braille word and Braille translation based on multi-knowledge[C]//Proceedings of International Conference on Signal Processing, Publishing House of Electronics Industry, Beijing, China, 2000: 2070-2074.

[12] 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):73-79.

[13] 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.

[14] 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.

[15] 吕先超. 视障汉语转换软件 SunBraille 的设计实现[D]. 兰州：兰州大学硕士学位论文，2016.

[16] Wang X, Yang Y, Liu H, et al. Chinese-Braille translation based on Braille corpus[J]. International Journal of Advanced Pervasive and Ubiquitous Computing (IJAPUC), 2016, 8(2):56-63.

[17] 肖航，钟经华. 汉语盲文语料库建设方案[J]. 语言文字应用，2015，3(3):109-118.

[18] Chen X, Qiu X, Zhu C, et al. Long short-term memory Neural Networks for Chinese word segmentation [C]//Proceedings of Conference on Empirical Methods in Natural Language Processing, 2015: 1197-1206.

蔡佳(1991—)，硕士研究生，主要研究领域为自然人机交互。

E-mail: caijia@ict.ac.cn

唐李真(1977—)，硕士，工程师，主要研究领域为盲人信息无障碍技术及盲人辅具。

E-mail: tanglizhen@blc.org.cn

王向东(1979—)，通信作者，博士，高级工程师，主要研究领域为人机交互、语音识别、残疾人信息无障碍技术等。

E-mail: xdwang@ict.ac.cn