计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210100025

基于 Transformer 的汉字到盲文端到端自动转换

蒋 琪[1] 苏 伟[1] 谢 莹[2] 周弘安平[2] 张久文[1] 蔡 川[1]
1 兰州大学信息科学与工程学院 兰州 730000
2 中国盲文出版社 北京 100142
(jiangq2018@lzu.edu.cn)

摘 要 汉字到盲文自动转换是改善我国1700万视障人群生活学习和贯彻落实国家信息无障碍建设的重要问题。现有汉盲转换方法均采用多步转换方法,先对汉字文本进行盲文分词连写,再对汉字进行标调,最后结合分词和标调信息合成盲文文本。该文提出一种基于编码器-解码器模型 Transformer 的端到端汉盲转换方法,利用汉字-盲文对照语料库训练 Transformer 模型。基于《人民日报》六个月约1200万字中文语料,该文构建了国家通用盲文、现行盲文、双拼盲文三种对照汉盲语料库。实验结果表明,该文提出的方法可将汉字一步转换为盲文,并在国家通用盲文、现行盲文、双拼盲文分别有80.25%,79.08%和79.29%的 BLEU 值。相比现有汉盲转换方法,该方法所需语料库的建设难度较小,且工程复杂度较低。

关键词:汉盲转换;端到端深度学习;编码器-解码器模型;Transformer
中图法分类号 TP391

End-to-End Chinese-Braille Automatic Conversion Based on Transformer

JIANG Qi[1],SU Wei[1],XIE Ying[2],ZHOUHONG An-ping[2],ZHANG Jiu-wen[1] and CAI Chuan[1]
1 School of Information Science & Engineering,Lanzhou University,Lanzhou 730000,China
2 China Braille Press,Beijing 100142,China

Abstract Chinese-Braille automatic conversion concerns the life and learning of 17 million visually impaired people in China and the national information accessibility construction. All existing Chinese-Braille conversion methods adopt multi-step process, which firstly segment Chinese text according to Braille word segmentation rules,then mark tone for Chinese characters. This paper studies end-to-end deep learning system that directly converts Chinese into Braille. The encoder-decoder model transformer is trained on Chinese-Braille corpus. Based on six-month data of People's Daily,totaling about 12 million characters,this paper builds three Chinese-Braille corpora of Chinese common Braille,current Braille and Chinese double-phonic Braille systems. The experimental results show that the method proposed in this paper can convert Chinese into Braille in one step,and reaches BLEU score of 80.25%,79.08% and 79.29% in Chinese common Braille,current Braille and Chinese double-phonic Braille. Compared with the existing methods,this method requires a corpus which is less difficult to construct and the engineering complexity is lower.

Keywords Chinese-braille conversion,End-to-end deep learning,Encoder-decoder model,Transformer

1 引言

盲文(Braille)是指专为盲人设计、供盲人使用的、靠触觉感知的文字,是我国语言文字的重要组成部分。汉字到盲文的转换系统,是将已有的汉字资源转换为盲文资源,最终生成盲文文档,可以供各种盲文系统使用。

盲文最基本的符号为盲文点字(盲符),以6个位置固定的凸点为基本结构,根据点的凸起与否可形成64种变化,即64个盲符。一个盲符所占的长方形位置,称作"方"。计算机中可用盲文 ASCII 码表示盲文,64个盲符对应64个盲文 ASCII 码(单个 ASCII 字符)。盲文 ASCII 码极大地方便了盲文处理,是盲人点显器、盲文刻印机等设备的标准输入。

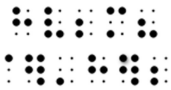
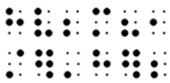
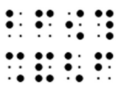
我国目前使用3种盲文:现行盲文[1]、双拼盲文、国家通用盲文[2]。现行盲文是以普通话为基础,以词为单位,以声、韵、调三方表示一个完整音节,采用盲文分词连写规则记录汉字的一套盲文方案。现行盲文使用最广,具有易学易用的优点,但存在标调不够规范的缺陷,且现行盲文"需要时标调"对计算机来说是无法实现的,因此已有汉盲转换系统大多只支持不带调或全带调现行盲文。双拼盲文旨在克服现行盲文的缺点,曾在部分盲校试行,但因符形类别多、规则繁难,较现行盲文难学,只被部分盲人接受。2018年,教育部、国家语委、中国残联共同发布《国家通用盲文方案》。国家通用盲文完全

基金项目:国家自然科学基金项目(61772006);中国残联-中国盲人协会专项项目((14)0218);广西科技项目(桂科 AA17204096,桂科 AB17129012);广西"八桂学者"专项资助
This work was supported by the National Natural Science Foundation of China(61772006),Research Program of China Disabled Persons' Federation and China Association of the Blind((14)0218),Guangxi Province Science and Technology Project(AA17204096,AB17129012) and Fund for Guangxi Province "Baguischolars".
通信作者:苏伟(suwei@lzu.edu.cn)

沿用了现行盲文的声母、韵母、声调、标点符号,没有改变任何一个符号,没有删减、增加任何一个符号,确立了全部音节标调的总原则,废止了需要时标调的旧体系,其核心是在字字标调基础上,提出了按声母省写声调符号的规范,解决了现行盲文的声调"猜谜"问题,消除了盲文阅读中依靠上下文猜测语音和语义的障碍。国家通用盲文具有易学易用、读音准确、省时省方、利于信息化等优点[3],处在全国推广阶段。短语"汉盲转换"的 3 种盲文表示如表 1 所列。

表 1　"汉盲转换"盲文表示

Table 1　Braille of "Chinese-Braille conversion"

Braille Type	Braille	BrailleASCII Code
Chinese common Braille		HV2M8　/]'H]2
Chinese current Braille		HV2M81　/]'H]2
Chinese double-phonic Braille		BF[?　%&.<F

注:表中现行盲文为全带调现行盲文,即字字标调

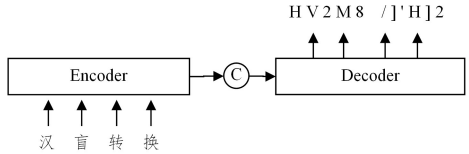
我国盲人阅读物存在数量稀缺、种类稀少等诸多问题[4]。我国视力残疾人总数约有 1 700 万人,每位明眼人每年平均占有 10 种出版物,盲人只平均占有 0.36 种[5]。盲文读物存在专业制作人士数量少、制作难度大、资金有限、盲文出版物成本高的问题。由于汉盲转换软件欠缺、技术落后,如今的网络盲文数字资源存在无法满足视障者学习、生活、工作需求的问题,盲文出版、盲人教育事业也受到限制。

汉盲转换技术的一大难点在于盲文分词与中文标准分词有着显著差异,现有中文标准分词软件与工具不适用于盲文分词连写。中国盲文的分词连写机制涵盖了分词和连写两个步骤,其中分词是将一个汉字序列切分成一个个单独的词。所谓连写,即是按照盲文的特殊性,避免音节结构过于松散,便于摸读和理解,使词意迅速形成概念,将意义上结合得较为紧密的一些词连写在一起。中文标准分词中四字成语或习惯用语为一个切分单位,而中国盲文中四字成语,能独立分写时,应按词分写,如"对\牛\弹琴""肆\无\忌弹""危\在\旦夕"。

汉盲转换技术的另一大难点是转换规则繁多。3 种盲文通用一套基于词性、词组、语义的分词连写规则,约有 145 条。书写格式约有 20 条,其中大多与标点符号相关,如逗号、顿号、分号、冒号前面均不空方("空方"指 6 点均不凸起的盲符),后面空一方,句号、问号、叹号前后均不空方。此外,中国盲文还有标调、拼写和简写规则,且大部分规则的正确执行依赖于准确的标调,可见多音字标调也是一个需要解决的问题。以国家通用盲文为例,标调规则中有一条:声母为 p、m、t、n、h、q、ch、r、c 的音节,省写阳平符号,音节 tóu 的声调符号不省写。

很多学习系统需要多个阶段的流水线处理,而端到端学习系统忽略所有的这些阶段,并用单个神经网络替代。端到端语音识别系统[6]忽略提取一些人工设计特征(音位等)、根据这些特征合成词和文本等阶段,直接学习音频到文本的映射函数。端到端学习系统,最大的挑战就是它需要大量的数据用以起端到端地学习一个能从输入直接映射到输出的函数。

鉴于机器翻译的复杂性和应用前景,学术界和产业界都把该领域作为重点研究方向,其已成为当前自然语言处理最活跃的研究领域之一[7]。机器翻译的定义是将一种自然语言(源语言)转换为另一种自然语言(目标语言)的过程。盲文是一种文字,不是一种语言,因此,汉盲转换问题本质上不是一个机器翻译问题。但是,机器翻译领域最常见的模型——编码器-解码器模型[8],作为一种序列生成序列(Sequence to Sequence,Seq2Seq)模型,非常适合用于实现端到端汉盲转换,其读取汉字序列生成盲文 ASCII 码序列的示意图如图 1 所示。编码器将输入汉字序列转化成语义向量 $\boldsymbol{C}$,解码器解码语义向量 $\boldsymbol{C}$ 和已生成的盲文 ASCII 码序列继续生成下一个盲文 ASCII 码,直至生成完整的盲文 ASCII 码序列。



注:图中盲文 ASCII 码为国家通用盲文转换结果

图 1　编码器-解码器模型

Fig. 1　Encoder-decoder model

盲文的数字化和信息化可以有效缓解我国盲人读物资源匮乏和内容较为陈旧的问题,更好地满足视障者日常学习生活工作需求。人工智能在盲文信息处理领域应用日益广泛,作用日益显著。中国盲文数字平台[1)]为我国首个面向全国 1 700 万视障者学习、生活和娱乐为一体的综合性公共服务平台,具有开放、合法、共享、免费的特点,为全国视障者提供盲文转换及盲文数字资源服务。2018 年 6 月 25 日,国家通用盲文标准正式通过,国家五部委将在全国推广,该平台为推广数字化平台之一。该平台原汉盲转换系统采用"汉字-分词-标调-盲文"的多步转换方法,转换准确率为 90%～95%。本文的研究和成果属于并应用于该平台现有汉盲转换系统。

本文基于"端到端深度学习"思想提出一种基于编码器-解码器模型 Transformer[9] 的端到端汉盲转换方法,利用汉字-盲文对照语料库训练 Transformer 模型。Transformer 模型解决了长距离依赖的问题,更擅长捕捉句子内部重要的语义信息,在许多自然语言处理领域中特别是机器翻译任务中表现突出。本文基于《人民日报》6 个月约 1 200 万字中文语料,构建了国家通用盲文、现行盲文、双拼盲文 3 种汉盲对照语料库。该方法忽略通过规则或多个模型提取分词和标调等特征、根据这些特征和特殊规则合成盲文文本的多个处理阶段,直接学习汉字到盲文的映射函数,首次实现将汉字一步转换为盲文,并首次实现将汉字转换为国家通用盲文。

2　现有汉盲转换方法

现有汉盲转换方法均采用"汉字-分词-标调-盲文"的多步转换方法:首先通过分词模型或盲文分词连写规则将汉字文本分词成汉字串;再对汉字进行标调;最后结合分词和标调信息将汉字串转换为盲文。该方法将汉盲转换问题分解为分词

1) www.braille.org.cn

问题和标调问题。标调问题的主要难点是多音字问题。

2003 年，Huang 等[10] 根据汉语特征与盲文特征的内在联系，设计了多种知识的统一形式化描述和相应的规则处理机制，有效地解决了转换过程中的分词歧义和连写问题，实现了汉字到盲文的自动转换。2016 年，中科院计算所的 Wang 等[11] 提出基于统计机器学习的盲文分词连写算法。2017 年，Wang 等[12] 基于统计机器学习模型，并结合中文标准分词和盲文分词连写信息，实现 92.81% 的盲文分词连写准确率。2019 年，李志鹏等[13] 根据自然语言服从马尔可夫源，提出一种基于马尔可夫模型的智能汉字盲文转换系统。2019 年，Wang 等[14] 提出基于汉盲对照语料库和深度学习的汉盲转换方法，其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，其多步转换方法示意图如图 2 所示。首先建立盲文分词连写语料库，利用 BiLSTM 模型将每个汉字标注为 BMES 4 个词位（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）。接着建立汉字标调语料库，需要将汉字与标点符号分割开来，然后利用 bi-gram 语言模型对汉字进行标调。完成分词和标调后，利用发音词典和发音-盲文映射表将完成盲文分词连写和标调的汉字文本转换为盲文，在文本转换时，会遇到一些特殊情况，如汉字文本中有时夹杂阿拉伯数字、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）。

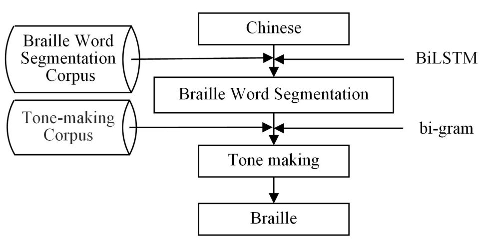


图 2　文献[14]中的多步转换方法

Fig. 2　Multi-step conversion method in reference 14

现有多步转换方法的不足之处：1）需要同时建立两个语料库——盲文分词连写语料库和汉字标调语料库，为了提取分词和标调信息得到上述两个语料库，需要将汉字语料和与其对应的盲文语料进行词语级对齐，即需要建设一个词语级精确对照的汉盲语料库；2）多步骤转换，转换过程至少包括分词、标调、合成等阶段，合成阶段可能还需特殊处理，算法复杂，转换效率也受影响。

3　端到端汉盲转换方法 ECB

本文提出的端到端汉盲转换方法（End-to-End Chinese-Braille Conversion Method，ECB），利用单个端到端编码器-解码器模型直接将汉字转换为盲文。模型及方法中所举转换示例为表 1 中短语“汉盲转换”的国家通用盲文转换结果，其盲文 ASCII 码为“HV2M8 /]′H]2”，分词信息为“汉盲\\转换”，盲文点字为“⠓⠧⠆⠍⠦ ⠌⠻⠄⠓⠻⠆”。本文基于编码器-解码器模型 Transformer 的 ECB 如图 3 所示。

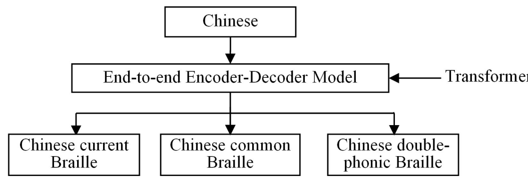


图 3　基于 Transformer 的 ECB 方法

Fig. 3　ECB based on Transformer

Transformer 模型是一种编码器-解码器（Encoder-Decoder）模型，完全摈弃了 RNN 和 CNN，使用了注意力机制 Attention 实现编码-解码。Encoder 由 N 个结构相同的 encoder 连接而成。encoder 由 Multi-head Self Attention 层（多头自注意力层）、FFN 层（Feed-Forward Network，包含两层全连接 Dense 层）构成。Decoder 由 M 个结构相同的 decoder 连接而成。decoder 相对 encoder，将多头自注意力层添加 mask 形成 Masked Multi-head Self Attention 层，确保预测位置 i 时仅仅依赖于位置小于 i 的输出，且在中间插入一层由传统 Attention 层结合多头注意力机制形成的 Multi-head Attention 层。

本文实验中 $N=M=3$，基于 Transformer 的 ECB 模型结构如图 4 所示。

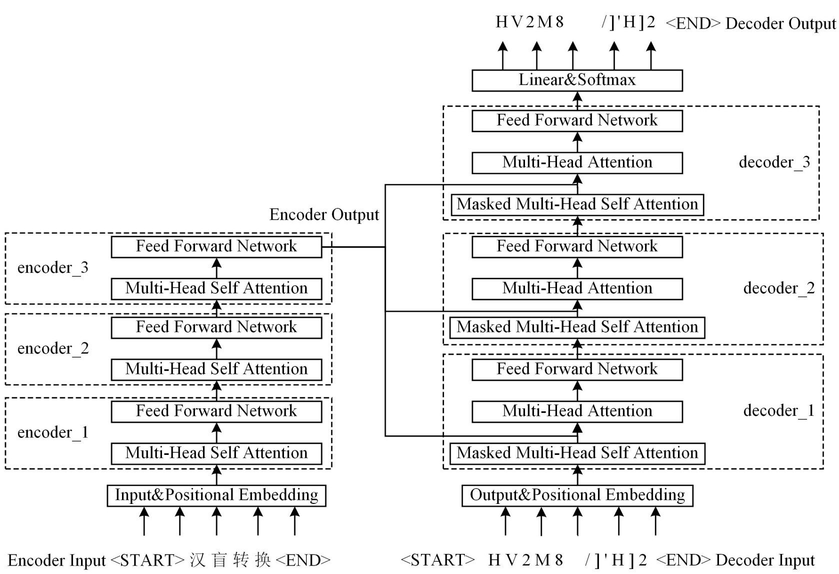


图 4　基于 Transformer 的 ECB 模型结构图

Fig. 4　Model structure of ECB based on Transformer

设模型输入为 $X=(x_{\text{start}},x_1,x_2,\cdots,x_i,\cdots,x_n,x_{\text{end}},x_{\text{pad}},\cdots)$，其中 x_{start} 标记句子开始，x_{end} 标记句子结束，输入句子过短用 x_{pad} 补齐，$x_1,x_2,\cdots,x_i,\cdots,x_n$ 为输入汉字句子，如 $(x_1,x_2,x_3,x_4)=$(汉，盲，转，换)。

设模型输出序列为 $Y=(y_1,y_2,\cdots,y_i,\cdots,y_n,y_{\text{end}})$，其中 y_{end} 标志模型应停止生成盲文 ASCII 码，$y_1,y_2,\cdots,y_n$ 为依次生成的盲文 ASCII 码，如 $(y_1,y_2,\cdots,y_{11},y_{12})=$(H，V，2，M，8，，/，]，'，H，]，2)。

Transformer 中注意力(Scaled Dot-Product Attention，缩放点积注意力)计算有 3 个输入 Q(Query，查询向量)、K(Key，键向量)、V(Value，值向量)，计算公式为：

$$Attention(Q,K,V)=softmax(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}})V$$

其中，$\sqrt{d_k}$ 为缩放因子，d_k 是 K 的维度。

Transformer 使用了多头注意力机制 Multi-head，将 Q，K 和 V 分别线性投射到 h 个不同子空间 head 上，最后将利用上述公式计算出的 h 个结果拼接在一起得到最终的注意力向量。

$$head_i=Attention(Q\,W_i^Q,K\,W_i^K,V\,W_i^V)$$

$$MultiHead(Q,K,V)=Concat(head_1,\cdots,head_i)W^O$$

其中，$Q\,W_i^Q$，$K\,W_i^K$，$V\,W_i^V$ 为 Q，K，V 对应第 i 个头的线性变换矩阵，W^O 是线性变换参数矩阵。

Transformer 中有 3 处需要 Multi-head Attention 计算。第一处位于图 4 encoder 间，用于计算输入句字符间的关联，其 $Q=K=V=$前一个 encoder 的输出。第二处位于图 4 decoder 间，用于计算目标句字符间关联，其 $Q=K=V=$前一个 decoder 的输出，记为 decoder self attention。第三处位于图 4 Encoder Output 与 decoder 间，用于计算输入句、目标句间关联，其 $K=V=$ Encoder Output，$Q=$ decoder self attention 的输出。多头注意力计算的好处是可以允许模型在不同的表示子空间里学习到位置、语义等相关信息。

ECB 相比多步转换方法的特点及优点：1)相比多步转换方法需要同时建立两个语料库，ECB 只需建设单一语料库——汉字-盲文对照语料库，可直接利用句子级对照盲文数字资源，无需通过专业知识和对齐算法实现词语级精确对照，语料库建设难度较小、耗时短；2)相比多步转换方法的复杂过程，ECB 只需一步，输入为汉字，直接输出盲文，无需特殊处理，工作量较少，工程复杂度较低。

4　实验与分析

4.1　语料

中文语料大数据集为 1998 年《人民日报》1－6 月数据，大小为 34.5 MB，约 1 200 万字。其中将 1，2 两个月数据作为小数据集，大小为 10.7 MB，约 372 万字。该语料包含中文、英文、数字、标点符号等多种字符类型。表 2 给出了大小数据集的各字符类型字符数统计信息。中文语料根据逗号、句号、感叹号、问号等中文标点符号和设定的截断长度切分成中文句子。

国家通用盲文、现行盲文、双拼盲文的盲文语料是将上述切分后中文句子以句子为单位由中国盲文数字平台原汉盲转换系统转换得到的盲文 ASCII 码，并由中国盲文出版社专家校改该转换结果。其中，现行盲文为全带调现行盲文。

最终，我们得到 3 种盲文的句子级汉盲对照语料库。每种汉盲对照语料库包含中文文本和盲文文本，两个文本均以句子为单位，每个句子占一行。

表 2　中文语料各字符类型统计信息

Table 2　Statistical information of various character in Chinese corpus

语料大小	中文	英文	中文标点	英文标点	数字	特殊字符
小数据集	3 251 756	3 544	354 912	251	107 940	4 175
大数据集	10 471 535	11 612	1 131 230	3 530	355 473	14 906

4.2　数据预处理

数据预处理分为字典生成、最大长度计算、标记添加、字典转换 4 个步骤，最终将中文文本、盲文文本转换为 Transformer 模型训练和测试所需数据。

首先，遍历中文文本和盲文文本生成汉字字典和盲文 ASCII 码字典。两个字典的 3 个键 0，1，2 均分别对应于补零标记'〈PAD〉'、开始标记'〈START〉'、结束标记'〈END〉'。逆盲文 ASCII 码字典为盲文 ASCII 码字典键值对交换后字典，用于模型转换时将输出转为盲文 ASCII 码。

计算中文句子最大长度 *maxlen* 和盲文句子最大长度 *maxlen_ascii*。经计算，中文句子的最大长度为截断长度，盲文句子的最大长度因语料大小和盲文类别的不同而不同。

每个中文句子在首添加开始标记'〈START〉'，在尾添加结束标记'〈END〉'，长度小于 *maxlen* 时添加补零标记'〈PAD〉'。通过汉字字典将其转为数字表示，作为编码器输入，记为 Encoder Input。

每个盲文句子在首添加开始标记'〈START〉'，在尾添加结束标记'〈END〉'，长度小于 *maxlen_ascii* 时添加补零标记'〈PAD〉'。通过盲文 ASCII 码字典将其转为数字表示，作为解码器附加输入 Decoder Input。解码器主要输入为编码器输出 Encoder Output。

Decoder Input 去除开始标记'〈START〉'，作为解码器输出 Decoder Output。

最后，随机抽取 10%数据作为测试集 test_set，剩余 90%的数据作为训练集 train_set。数据预处理相关信息统计如表 3 所列。

表 3　数据预处理信息

Table 3　Data preprocessing information

语料大小	语料	汉字字典大小	截断长度	训练集	测试集
小数据集	252 494	5 120	48	227 244	25 250
大数据集	811 354	5 706	48	730 218	81 136

注：语料、训练集和测试集的数据为中文句子数或盲文句子数

4.3　实验设置

本文训练 Transformer 模型的实验环境：CentOSLinux release 7.7.1908 系统，Tesla P100 PCIe 16GB 显卡，256 GB 内存，深度学习框架 Keras。

深度学习模型的性能与模型超参数有着直接的关系。良好的模型结构及超参数设置甚至可使结构相对简单的模型实现比相对复杂的模型更好的效果[15]。本文构建的 Transformer 模型超参数如表 4 所列。

表 4　Transformer 模型超参数

Table 4　Transformer model parameter

encoder 层数 N	decoder 层数 M	Multi-head head 数	词向量维度	FFN 层隐藏层维度	Dropout 层舍弃率
3	3	4	100	150	0.05

本文使用 BLEU[16] 作为衡量参考盲文句子和生成盲文句子相似度的指标。BLEU 常用于评价机器翻译模型的性能，但不仅仅局限于机器翻译，也可应用于文本摘要[17]等领域。汉盲转换问题虽然不是机器翻译问题，但是 BLEU 值的计算更侧重于两个句子的相似度，也适合用于本文中参考盲文句子和生成盲文句子相似度的计算，其值的高低代表生成盲文句子相对参考盲文句子的匹配度和通顺性。BLEU 取值范围为 0～1，1 代表完全匹配，越靠近 1 转换质量越好。本文使用 NLTK 工具计算 BLEU 值。

端到端汉盲转换直接将汉字一步转换为盲文，跳过了拼音这一中间步骤，对汉字一字多音的消歧效果是一个需要注意的问题。本文统计了部分高频多音字在现行盲文大数据集的测试集的标调准确率。

计算在现行盲文大数据集的测试集的标调准确率，并与多步转换方法进行对比。声调所对应的盲文 ASCII 码为：阴平(A)、阳平(1)、上声(‘)、去声(2)、轻声(不标调)。标调准确率等于标调正确的汉字数与总汉字数的比值，不包括标点符号、数字、英文等其他字符。

计算大数据集的测试集的汉盲转换准确率，并与多步转换方法进行对比。汉盲转换准确率计算方法为：将转换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。

4.4 实验结果

4.4.1 模型性能与语料大小的关系

深度学习模型的性能往往和语料大小有很大关系，训练数据量也是神经机器翻译的一大挑战[18]。更多数据能让模型更好地学习输入到输出的映射函数，也能在一定程度上减小过拟合。

表 5 给出了本文构建训练的 Transformer 模型分别在大小数据集的测试集的性能表现，大数据集括号内数字为相对小数据集提高的 BLEU 值。计算 BLEU 时，每个生成的盲文句子只有一个参考句子。(ω_1，ω_2，ω_3，ω_4)为 1-gram，2-gram，3-gram，4-gram 在计算 BLEU 值的权重，满足 $\omega_1+\omega_2+\omega_3+\omega_4=1$。实验结果表明，基于 Transformer 的 ECB 可实现汉字一步转换为盲文。在 NLTK 工具计算 BLEU 值的默认权重 $\omega_1=\omega_2=\omega_3=\omega_4=0.25$ 下，该方法在国家通用盲文，现行盲文，双拼盲文分别有 80.25%、79.08%、79.29%的 BLEU 值，生成的盲文文本质量较高。大数据集下该方法的 BLEU 值相对小数据集：在国家通用盲文和现行盲文下提高较大(2%～7%)，且在单 n-gram($\omega_n=1$)计算方式下 n 取值越大，提高越大，说明生成的盲文句子不仅在词的层面更匹配，在句子的层面也更通顺；在双拼盲文下提高不大。

表 5 基于 Transformer 的 ECB 在小数据集、大数据集的 BLEU 值(测试集)

Table 5 Bleu score of ECB based on Transformer in small and large test data sets

(单位：%)

Weight of BLEU	Chinese common Braille		Chinese current Braille		Chinese double-phonic Braille	
	small data set	large data set	small data set	large data set	small data set	large data set
(0.25,0.25,0.25,0.25)	75.03	80.25(+5.22)	73.68	79.08(+5.40)	77.24	79.29(+2.05)
(1,0,0,0)	85.25	88.02(+2.77)	84.06	87.50(+3.44)	86.34	87.64(+1.30)
(0,1,0,0)	78.42	82.96(+4.54)	77.05	81.94(+4.89)	80.28	82.16(+1.88)
(0,0,1,0)	72.07	77.97(+5.90)	70.67	76.60(+5.93)	74.57	76.80(+2.23)
(0,0,0,1)	65.77	72.84(+7.07)	64.39	71.21(+6.82)	68.88	71.47(+2.59)

4.4.2 高频多音字消歧效果分析

通过对测试集各个多音字出现次数进行统计，本文选择了部分高频多音字并给出在现行盲文大数据集的测试集的标调准确率，如表 6 所列。

表 6 高频多音字(部分)标调准确率

Table 6 Tone-making accuracy of some high frequency polyphonic characters

Polyphonic Character	*Frequency*	*Accuracy*/%
中	10061	99.46
为	5993	96.63
长	3389	96.13
重	2937	97.87
种	1728	98.28
教	1617	99.82

从表 6 可以看出，本文的端到端方法在所使用的现行盲文测试集，对表中多音字的消歧效果比较好。其中，“中”和“教”因在语料中某一种拼音占绝大多数，其标调准确率可超过 99%。

4.4.3 ECB 与多步转换方法对比

ECB 与多步转换方法在现行盲文的标调准确率对比如表 7 所列，多步转换方法数据为文献[11]王向东团队的实验结果。在汉盲转换领域并无公开数据集，第 1 节中所提黄河燕、王向东等采用各自团队收集的数据集。因此表 7 中准确率高低不能完全代表两种方法的标调性能高低。文献[11]训练集为涵盖教育、历史、医学、小说等领域的盲文文本，大小为 62.7 MB，测试集为《读者》杂志盲文版的 11 篇文章，由 704 个句子组成。本文数据集基于《人民日报》6 个月约 1200 万字中文语料构建了现行盲文、国家通用盲文、双拼盲文 3 种汉盲对照语料库，测试集由随机抽取的约 8 万条句子组成。

表 7 标调准确率

Table 7 Tone-making accuracy

Method	*Accuracy*/%
Reference 11	99.80
ECB	99.70

ECB 与多步转换方法在现行盲文的转换准确率对比如表 8 所列，多步转换方法数据为文献[14]王向东团队的实验结果。同样地，由于数据集的不同，表 8 中准确率高低也不能完全代表两种方法的性能高低。文献[14]数据集基于 234 万字中文语料构建了现行盲文的汉盲对照语料库，包含科学科普、医学医用、通用文学 3 种类型，测试集约为 6 万句。本文方法在国家通用盲文、双拼盲文的转换准确率见表 9。

表 8 汉盲转换准确率(1)

Table 8 Accuracy of Chinese-braille conversion (1)

Method	*Accuracy*/%
Reference 14	85.11
ECB	86.46

表 9　汉盲转换准确率(2)

Table 9　Accuracy of Chinese-Braille conversion (2)

Braille Type	*Accuracy*/%
Chinese common Braille	86.89
Chinese double-phonic Braille	87.87

从标调准确率和对部分多音字消歧效果的分析可以看出，本文端到端转换方法的标调性能较好，均在 96% 以上。但是，转换准确率和 BLEU 值均在 88% 以下，可以得出的结论是，非标调的盲文 ASCII 码(包括载有盲文分词连写信息的空格)的生成准确率有待提高。

结束语　本文基于“端到端深度学习”提出了一种基于编码器-解码器模型 Transformer 的端到端汉盲转换方法 ECB，利用汉字-盲文对照语料库训练 Transformer 模型。本文基于《人民日报》6 个月约 1 200 万字中文语料，构建了国家通用盲文、现行盲文、双拼盲文 3 种汉盲对照语料库。实验结果表明，本文提出的方法首次实现了将汉字一步转换为盲文，在国家通用盲文、现行盲文、双拼盲文分别有 80.25%，79.08%，79.29% 的 BLEU 值，并首次实现了将汉字转为国家通用盲文。相比现有多步转换方法，ECB 所需语料库的建设难度较小，且工作量较少，工程复杂度较低。

未来，我们将实践基于其他端到端编码器-解码器模型[19]的 ECB 并进行性能比较，如基于 RNN 的 Seq2Seq、基于 CNN 的 ConvS2S，并不断提高汉盲转换准确率。

参 考 文 献

[1] GB/T 15720-2008 中国盲文[S]. 北京，2008.

[2] GF 0019-2018 国家通用盲文方案[S]. 北京，2018.

[3] ZHONG J H. Analysis of the characteristics of Chinese common Braille Scheme[J]. Modern Special Education，2018(23)：23-25.

[4] GUO L H. Research on the current situation and development trend of Braille Publishing[J]. Media Forum，2019，2(11)：121-122.

[5] LI N. The current situation and trend of Braille Publishing[J]. Modern Publishing，2016，(5)：30-33.

[6] ZEGHIDOUR N，USUNIER N，SYNNAEVE G，et al. End-to-End speech recognition from the raw waveform[C]// Interspeech. 2018：781-785.

[7] DABRE R，CHU C，KUNCHUKUTTAN A. A Survey of Multilingual Neural Machine Translation[J]. ACM Computing Surveys，2020，53(5)：1-38.

[8] SUTSKEVER I，VINYALS O，LE Q V. Sequence to sequence learning with neural networks[C]// Advancesin Neural Information Processing Systems. 2014：3104-3112.

[9] VASWANI A，SHAZEER N，PARMAR N，et al. Attention is all you need[C]// Proceedings of Advances in Neural Information Processing Systems. 2017：6000-6010.

[10] HUANG H Y，CHEN Z X，HUANG J. Chinese-Braille Translation Approach Based on Multi-Knowledge Analysis[C]// The 7th China Joint Conference on Computational Linguistics. 2003：607-613.

[11] WANG X，YANG Y，LIU H，et al. Chinese-Braille translation based on Braille corpus[J]. International Journal of Advanced Pervasive & Ubiquitous Computing，2016，8(2)：56-63.

[12] WANG X，YANG Y，ZHANG J，et al. Chinese to Braille translation based on Braille word segmentation using statistical model [J]. Journal of Shanghai Jiaotong University (Science)，2017，22(1)：82-86.

[13] LI Z，WANG R，ZHANG T，et al. Intelligent Braille conversion system of Chinese characters based on Markov model[C]// Proceedings of IEEE 3rd Information Technology，Networking，Electronic and Automation Control Conference (ITNEC). 2019：1283-1287.

[14] CAI J，WANG X D，TANG L Z，et al. A Deep Learning Method for Chinese-Braille Conversion Based on Parallel Corpora[J]. Journal of Chinese Information Processing，2019，33(4)：60-67.

[15] MA J，GANCHEV K，WEISS D. State-of-the-art Chinese word segmentation with BiLSTMs[C]// The 2018 Conference on Empirical Methods in Natural Language Processing. 2018：4902-4908.

[16] PAPINENI K，ROUKOS S，WARD T，et al. BLEU：a method for automatic evaluation of machine translation[C]// ACL. 2002：311-318.

[17] GAMBHIR M，GUPTA V. Recent automatic text summarization techniques：a survey[J]. Artificial Intelligence Review，2017，47(1)：1-66.

[18] KOEHN P，KNOWLES R. Six challenges for neural machine translation[C]// The First Workshop on Neural Machine Translation. 2017：28-39.

[19] YANG S H，WANG Y X，CHU X W. A Survey of Deep Learning Techniques for Neural Machine Translation[J]. arXiv：2002.07526，2020.

JIANG Qi，born in 1995，postgraduate. His main research interests include natural language processing and Chinese-Braille conversion technology.

SU Wei，born in 1977，associate professor. His main research interests include natural language processing and information accessibility technology.