基于多知识分析的汉盲转换算法*

黄河燕 陈肇雄 黄静

(中国科学院计算语言信息工程研究中心，北京 100083)
Email: heyan.huang@hjtek.com

摘要：本文，提出了一种基于多知识一体化分析的汉盲翻译转换算法，该算法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汉语分词歧义和连写问题，实现了汉语到盲文的高效自动翻译转换。
关键词：多知识分析，盲文信息处理，汉盲转换

Chinese-Braille Translation Approach Based on Multi-Knowledge Analysis

Huang Heyan Chen Zhaoxiong and Huang Jing
Research Center of Computer & Language Information Engineering, CAS, Beijing, 10083
Email: heyan.huang@hjtek.com

Abstract In this paper, an approach of Chinese-Braille translation is presented, which is based on multi-knowledge analysis. According to the internal relation between Chinese and Braille, a unified representation form of multiple kinds of knowledge and the corresponding rule processing mechanism is designed, so it is possible to efficiently solve the problem of Chinese segmentation and joining of Braille words. Also, the automatic translation from Chinese to Braille is implemented.
Keywords Multi-Knowledge Unified Analysis, Braille Information Processing, Chinese-Braille Translation

1. 前言

中国的汉语盲文，自借鉴莱尔字符体系创造出康熙盲字，又“心目克明”盲字，又现行盲文，又汉语双拼盲文，经过不断的实践和完善已形成一套门类基本齐全，且具有民族特色的汉语盲文符号系统，为广大盲人学习文化和掌握科学技术知识发挥了极其重要的作用。然而，现有盲文读物种类少、信息来源狭窄，远远无法满足盲人对信息的需求。因此，如何实现对多元信息的汉盲自动翻译转换，让广大盲人能跟上信息时代的步伐，提高全社会信息应用的能力和水平，是计算机研究工作者的一项艰巨而光荣的任务，具有重大的科学研究意义和广阔的应用前景。

本文，我们在多年实用化多语种机器翻译系统研究开发工作的基础上，提出了一种基于多种知识一体化分析的汉盲翻译转换算法，该算法根据汉语特征与汉语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汉语分词歧义和连写问题，设计并实现了一种汉语到盲文的高效自动翻译转换机制。

*基金项目：国家自然科学基金资助项目（60272088）

2. 总体设计

2.1 特点和难点分析

汉语盲文是用汉语拼音的方式来表示汉字的，应该说是汉语的另一种形式化表示方法，实质上是一种实行分词连写的拼音盲文。但汉语盲文也有自身的特殊性，首先，盲文是一种触摸文字，盲人通过指间触摸，感知文字的读音，同时理解文字的含义。其次，盲人特有的默读心理和摸读习惯，制约着他们不能象明眼人阅读时那样，一目十行，而只能是一个盲符一个盲符地感知和摸读。

由于盲文实行分词连写规则，因此将汉语翻译成盲文，首先要对汉语句子进行分词，并对分词结果进行词性标注，然后按照盲文的分词连写规则对分词以后的结果重新进行连写，再根据词的拼音查找盲文拼音编码表，进行汉字到盲字的转换。因此，汉盲翻译转换的主要工作和需要解决的难点问题有：

（1） 汉语分词

汉语不同于英语等其它西方曲折型语言，在表层形式上汉语的词与词之间不存在空格等分隔符。而由于词是计算机进行语言信息处理的基本要素，是进行句法分析和理解的基础。要进行汉盲的翻译转换首先必须对汉语进行分词。针对汉语分词中歧义切分字段与未登录词识别困难的特点，我们采用了逆向全切分与规则统计相结合的方法，有效地解决了歧义切分字段与未登录词识别的问题。从而使得分词算法具有较高的通用性和良好的可扩展性，为深层次的中文信息处理打下了很好的基础。

（2） 连写处理

由于盲人是通过摸读来阅读理解的，为了让一些意义上结合得较为紧密的短语连写起来以便于盲人理解，在进行汉盲翻译转换时，还需要对盲文进行分词连写。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。如句子：*我们学会了解答问题。*

经过汉语分词，结果为：*我们 学会 了 解答 问题 。*

经过连写处理，结果为：*我们 学会了 解答 问题 。*

显然连写处理结果更便于盲人的理解。自 1953 年许多语言文字学家就开始了分词连写规则的研究，直到 1993 年才最终审定了《汉语盲文分词连写规则》。分词连写规则的制定主要是根据汉语的语法特征，语言的逻辑性和习惯性，并在一定程度上考虑音节长度。现有的分词连写规则主要对名词、动词、形容词、数词、量词、代词、副词、介词、连词、助词、叹词等十几类通用词的连写做出了规定。如动词连写规则有：

① 单音节动词重叠式连写：中间插入“一”和“了”也是连写。如：

看看 说说 看一看 说一说 看了看 说了说

② 动词跟时态助词“着”、“了”、“过”连写；如果出现两个以上的动词，则时态助词跟最后一个动词连写。如：

学习着 看见了 思考过 参观访问了

为了让计算机自动进行分词连写，需要将这些规则进行形式化的描述并在进行汉盲翻译转换时进行相应的处理。

（3） 汉拼转换

汉盲机器翻译系统的另一个主要任务是将汉语串与拼音串对应起来，按照盲文规范生成盲文。汉字和拼音的对应可以在字、词、短语三个层次上建立对应关系。由于汉语中有大量的多音字的存在，直接在汉语字集合与盲文字集合上建立映射是不可能的，如重

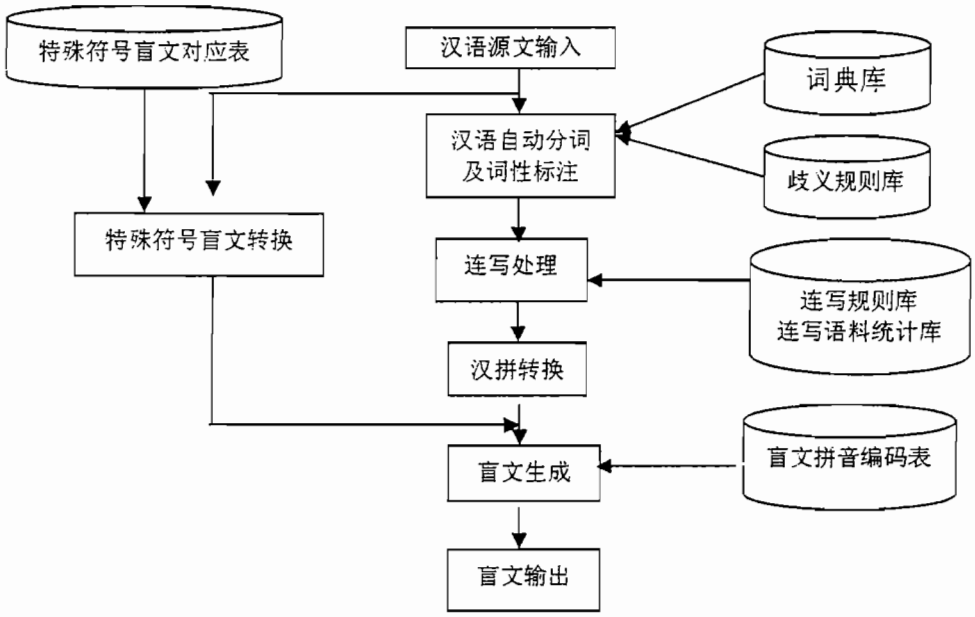
(zhong4)奖，重(chong2)逢，好(hao3)人，爱好(hao4)，（其中的数字表示声调），然而通过汉语分词，以词为单位标音就可以基本上消除多音字。对于词一级无法消除的多音，可以通过句中前后词的语境来消除，如播种(bo1zhong3)造林，播种(bo1zhong4)季节。由此通过词一级的标音可以在汉语词集合与盲文词集合上建立对应关系，从而可以提高汉盲翻译转换的准确率。

（4） 盲文生成

根据盲文的拼音编码表将汉语词的拼音转换成盲文。

2.2 总体方案

针对汉盲翻译转换中的主要特点和难点问题，我们设计的基于多知识一体化分析的汉盲翻译转换算法的总体功能模块结构如下图所示：



图一 汉盲翻译转换算法总体结构图

在汉盲转换算法中，首先对输入的原文进行扫描，如果是标点或 ASCII 字符就调用特殊符号盲文转换模块，根据特殊符号盲文对应表直接产生盲文，否则调用自动分词及词性标注模块，输出词切分结果；对词切分结果调用连写处理模块，根据连写规则库和连写语料统计库对词切分结果进行合并，输出新的词切分结果；对连写后的词切分结果调用汉拼转换模块，输出拼音串；对拼音串调用盲文生成模块，根据盲文拼音编码表将拼音转换成盲文，输出盲文串。

3. 算法设计

在本节中，给出多知识汉盲翻译转换算法中多种知识表示形式和总体算法流程的描述。

3.1 多语言知识表示

在多知识汉盲翻译转换算法中涉及的知识包括：词典知识、分词歧义处理规则、连写规则和盲文拼音转换规则等，下面给出各种知识的表示形式：

（1） 词典知识表示

汉盲翻译系统中的词典一方面要为分词和词性标注提供词条和词性信息；另一方面要为拼音生成提供相应的拼音串信息；除此之外还要为连写提供词音节数信息。因此，词典库中各词条应具有以下特征信息：词条串、音节数、拼音串、词性标记集、词性优先级等。其中，词条串、词性标记集、词性优先级是与分词和词性标注相关的，音节数是与采用连写规则进行局部组合相关的，而拼音串项则为拼音生成服务的。词典中每个词的词性优先级获取是通过对大量语料的统计动态得到的。

词典知识库中，词的表示形式为：

$汉字串＝｛分类｛属性｝｝ {拼音|音节数}

其中{}表示可重复 0 到任意次

如字典中“北京”表示为：

$北京 NP(SPLA) "bei1jing1|2"

如字典中“爱好”表示为：

$爱好 NP(NCGEN);VP(VT) "ai4hao4|2"

（2） 规则知识表示

在多知识汉盲翻译转换算法中涉及的规则包括：分词歧义处理规则、连写规则和盲文拼音转换规则。为了便于计算机的自动处理，需要将这些规则进行形式化。同时，为了综合利用多种知识，提高翻译转换的准确率，需要在规则中将各种语言知识进行一体化的有机表示。为此，我们设计了一种基于 SC 文法的多知识一体化的规则表示形式，为：

<头部> －> <条件函数>|<操作函数>

其中，<头部>用于定义当前词的组合结构形式，其具体内容为句子中各个词的语法类、词长等信息；<条件函数>用于定义一些需要满足的条件信息；<操作函数>为当前规则满足时所进行的操作；如切分规则：

NUL(会诊) NUL([断;疗;脉;治]) －>|CECUT(会;诊+N2) *(1)*

这条规则的意思是：如果当前词的组合结构是“会诊”，“断或疗或脉或治”时， 则要重新切分，将“会”单独切开为一个词，诊与第 2 项的字合成一个词；

VP() NUL(了解) NUL([答;毒;说;决;围;救;气;手;放]) －>|CECUT（N1.了,解+N3）* *(2)*

这条规则的意思是：如果当前词的组合结构是“动词”，“了解”，“答或毒或说或决或*救或气或手或以放开始的词*”时， 则要重新切分，将“了”单独切开为一个词，解与第 3 项合成一个词；

如连写规则

VP() NUL([着;了;过]) －>|MERGE(N1,N2) *(1)*

这条规则的意思是：如果当前词的组合结构是“动词”，“着或了或过”时， 则要合并第 1 项（动词）与第 2 项（着或了或过）；

连写规则按长头部优先和确定规则优先的原则进行排序，如规则

规则 1：AP（1）AP(1) AP（1）AP(1)－>SAME(N1,N2),SAME(N3,N4)|MERGE(N1,N2,N3,N4)

规则 2：AP(1) NUL(不) AP(1) －>SAME(N1,N3)|MERGE（N1,N2,N3）

规则 3： NUL（[副；总；非；反；超；老；阿；可；无]）NP()－> SAME(N1,N2), SAME(N3,N4) |MERGE(N1,N2,N3,N4)
规则 4： NP(1) NP(1) －>SAME(N1,N2)|MERGE（N1,N2）
规则 1 比规则 2 优先，规则 3 比规则 4 优先；

连写语料库：记录了一些从语料中统计出来的，而无法用连写规则进行连写的连写块，库基本结构如下图所示：

大中小学
贫下中农
…
连词块 m
…
连词块 n

盲文拼音编码表中盲文的表示形式为：

拼音	盲文
如	men

3.2 算法总体流程描述

基于上述的知识表示形式，我们设计了相应的汉盲转换算法。算法输入为汉语文章，输出为盲文文章，算法总体流程描述如下：

（1）将原文 T_c 切分为句子 $T_c = \{S_1S_2...S_n\}$；
（2）对句子 S_i 进行成分（如标点，ASCII 字符，汉字串(包括数字)等）的切分 $Si = \{C_1C_2...C_n\}$；
（3）对句子成分 C_i，如果是标点或 ASCII 字符，根据特殊符号盲文对应表直接产生盲文，转（12）；
（4）如果是汉字串，将汉字串转换成字结点(一个汉字就是一个字结点，数字串为一字结点)，形成字结点序列 $Ci = \{N_1N_2...N_n\}$；
（5）采用逆向最大匹配算法，从右到左进行词典匹配，得到所有可能的有效边；
（6）若结点 N_i 包括两条或两条以上的边，则出现歧义，调歧义切分规则，推理消除歧义；
（7）遍历词切分路径得到正确的词切分结果 $Si = \{W_{11}W_{12}...W_{1n}\}$；
（8）对词切分结果 $Si = \{W_{11}W_{12}...W_{1n}\}$，依次取词结点 W_{1i}，匹配连写规则，若成功，根据连写规则所要求的连写长度，将若干词连写，形成新的词切分结果 $Si = \{W_{21}W_{22}...W_{2n}\}$；
（9）对词切分结果 $Si = \{W_{21}W_{22}...W_{2n}\}$，检索连写语料库，形成新的词切分结果 $Si = \{W_{31}W_{32}...W_{3n}\}$；
（10） 依次取词结点 W_{3i}，用 W_{3i} 的汉语拼音检索盲文拼音编码表，产生相应的盲文 M_i，形成盲文词结果 $Si = \{M_1M_2...M_n\}$；
（11） 判别句子是否结束，若不是，取下一句子成分 C_{i+1}，转（3）；
（12） 判别原文处理是否结束，若不是，取下一句子成分 S_{i+1}，转（2）；
（13） 输出盲文 Tm，算法结束；

例如，对于句子“我们学会了解答问题”，通过前面给出的汉语歧义切分规则(2)，分词结

果为“我们 学会 了 解答 问题 ”，通过前面给出的连写规则(2)，得到连写结果为“我们 学会了 解答 问题”，再经过盲文编码表，产生盲文“
”。

4．模块实现

4．1 分词及词性标注

汉语盲文对分词有很高的要求，常见的分词算法如正向最大匹配、逆向最大匹配和双向最大匹配，这些算法因很多歧义问题无法解决，最终分词质量均不能完全满足要求。为此，我们采用逆向全切分与规则统计相结合的算法，依据词典进行切分，调用歧义规则进行消歧，根据语料库进行评估，选取最优的切分结果，并在此基础上进一步对地名、人名、常见的缩略语进行自动识别。

词性标注就是给句子中的每个词赋一个合适的词类标记，由于汉语词的兼类现象较为频繁，这给词性标注带来了很大困难，词性标注也是自然语言处理中的一个重要的课题。我们采用规则和统计相结合的处理方法进行词性标注。

4．2 盲文连写处理

盲文连写处理主要是利用连写规则（85 条规则）和盲文连写统计库进行词组合处理，通过一些实际的盲文语料库，对一些常见的连写词组合建立一个连写统计库。处理中先应用人工建立的形式化连写规则，对分词和词性标注后的各个词进行分析，符合连写规则的将这些词组合，然后按照这个连写统计库对词与词进行再组合。

盲文连写规则之间基本上是独立的，但这些规则有优先级别的，如准短语的规则要优先于短语规则，处理单音节词的要优先于处理多音节词的，处理具体的词的要优先处理词类的，因此，需要对这些规则进行分等级，优先级高的规则先执行。具体实现策略为：自底向上、依据规则局部识别，对句子中的每个切分块，及对连写规则库中的每条规则依据优先级依次进行匹配检索，如果当前块和其邻接块与所检索的规则匹配，则将这些相应的块连接起来形成一新块。

在此基础上再使用连写统计库，进一步识别可能的连写，盲文连写统计库中存储有一些常见的连写词，采用最大匹配策略对词进行再组合，具体算法为：对句子中的每个切分块，从连写语料库中查找所有以当前块开头的记录，如果当前块与其相邻的块完全匹配该记录则组合这些块。

4．3 拼音生成与盲文生成

通过前面的分词和词性标注及连写处理后，对于给定词性的词，除了极少数词以外，其读音就基本确定了。这些极少数的词都只是音调不同，区别不是非常明显。因此，要实现汉语词的拼音生成和盲文生成，只要根据汉语词及其词性建立一个拼音对应表，再根据这个对应表生成相应的拼音串，实现汉语词到拼音的转换。对少数不能确定其读音的多音

词再进行特殊的处理，如采用不标调，或者采用常见的音调，也可以建立一些专门的规则进行处理。

根据盲文的拼音编码表、拼法规则和标调规则将汉语词的拼音转换成盲文，对汉语中的标点符号、阿拉伯数字单独根据盲文编码表生成相应盲文标点符号和数字，然后根据盲文书写规则组合这些盲文块形成盲文。

可以看出，分词连写是汉盲机器翻译过程中的关键，分词连写又直接与分词和词性标记的正确性密切相关，另外词性标注的正确性还直接影响多音词的择音，因此分词、词性标记和连写分析的质量直接关系到汉盲机器翻译的质量。

5. 实验结果

在上述多知识汉盲翻译转换算法的实现中，我们搜录了 38090 个基本词，293 条歧义规则，97 条人名识别规则，85 条连写规则。为了评估算法的性能，我们用随机获取的一万字的常规语料进行了测试，结果表明，其翻译转换的处理速度可以达到 3850 汉字/秒，汉字转换的准确率达到了 98.4%，分词连写的准确率达到了 95%以上，翻译转换的准确率达到了 95.6%，已经具备实用性。

目前，我们还在不断扩展和完善相应的汉语词典库和各种规则库，其翻译转换的准确度应该有进一步的提高。该算法现已具体应用于本中心承担的中国盲文出版社委托的中国盲文计算机系统项目中，其阶段开发成果已投入市场实际应用。

6. 结束语

上面，我们给出了一种基于多种知识一体化分析的汉盲翻译转换算法，该算法通过设计一种规则表示形式和相应的规则处理机制，很好的解决了汉盲转换中存在的分词连写问题，实现了汉盲转换算法的高效性和正确率。

在进一步的工作中，需要对汉盲翻译中的关键难点多音字处理，即对多音字规律作归纳总结，并制定相应规则进行特殊处理，以降低多音字的出错概率，提高系统的准确率。

最后，作者诚挚希望本文研究工作对中文信息处理同行的研究工作有参考借鉴作用。

参 考 文 献

[1] 滕伟民、李伟洪、杨忠诚、高旭，《中国盲文》，北京：华夏出版社，1996.12。

[2] 刘开瑛，《中文文本自动分词和标注》，北京：商务印书馆，2000。

[3] 陈肇雄，《SC 文法功能体系》，计算机学报第 11 期，1992.11。

[4] 黄河燕，《智能型机器翻译研究论文集》，中国科学院计算技术研究所智能机器翻译研究开发中心，1995.10。

[5] 孙茂松，《汉语自动分词研究的若干最新进展》，辉煌二十年——中国中文信息学会二十周年学术会议，2001.11。

[6] 黄河燕，陈肇雄，集成化盲文信息处理平台的总体设计及关键技术研究，《第二届中日自然语言处理专家研讨会论文集》，CJNLP2002，北京，2002。

[7] 唐英杰、伍春洪，盲文阅读机的设计及实现，北京印刷学院学报，1998。

[8] 朱双六、宋文兰，盲文信息处理研究，上海铁道大学学报（自然科学版），1996.12。