EasyBraille:
A Translation System for Mandarin and Braille

Zhu Xiaoyan and Bao Ta
State Key Lab of Intelligent Tech. & Systems,
Department of Computer, Tsinghua University, Beijing 100084, China
(Email:zxy-dcs@tsinghua.edu.cn)

EasyBraille:中文汉语盲文自动转换系统

朱小燕 包塔
智能技术与系统国家重点实验室,清华大学计算机系，北京 100084

摘要：本文提出并实现了一个中文汉语盲文自动转换系统。该系统不仅能够将打印的盲文文档通过扫描仪输入成为图像文件，进一步进行自动识别转换。同时可以将中文汉字文档自动转换成为盲文文档便于盲人阅读。本系统包含两个关键技术：汉语自动分词；盲文拼音到汉字的转换。汉语分词是在词典与规则的基础上，根据双向最大匹配方法进行的。同时在统计模型的基础上，根据盲文得分词连写规则进行排岐。另一方面，盲文到汉字的转换过程中，一音多字的歧义问题是在大规模预料库统计基础上，采用 Viterbi 搜索算法完成的。系统给出 N-Best 的结果，提供用户选择。实验结果表明，普通文本的汉到盲的正确转换率接近 98%，盲到汉的正确转换率约 97%。
关键词：盲文转换、Virterbi 算法、平滑算法、多知识

Abstract: This paper will present a translation system---EasyBraille between Braille and Mandarin. It can not only process printed Mandarin Braille document by scanner and translate it into Mandarin automatically, but also translate Mandarin into Chinese Braille for the blind to read usual documents. In this paper two key technologies are involved. One is word segmentation of Mandarin document when translate Chinese into Mandarin Braille. The other is transformation from *pinyin* to Chinese words when translate Braille into Mandarin. Chinese word segmentation module consists of a word dictionary, a rule base and a knowledge base for disambiguation by using adjacency constraints and bi-directional maximum matching with a dictionary. Our system's segmentation precision achieved is better than 98% for the common text. By developing a statistical language model, we perfectly solved the problem of ambiguity in the translation Braille into Mandarin. By using a multi-knowledge base to carry out the disambiguation process for each Braille sentence, we built a multi-level graph and used Viterbi search to find the sequence of Chinese characters with maximum likelihood, and used an N-Best algorithm to get the N most likely character sequences. The experimental results show that the system's overall precision for translation from Braille codes to Chinese characters is 97.38%.
Keywords: Braille translation, Virterbi algorithm, smoothing method, multi-knowledge.

I. Introduction

Braille is a kind of tactile writing, each character consisting of six raised dots as the basic structure. Mandarin Braille is a phonemic alphabetic representation [1]. In one case of translation from Braille to Mandarin, Braille texts were firstly scanned into images. By image recognition module, Braille images were transformed into Braille code files, then into *pinyin* code files. In transformation from *pinyin* to Mandarin, we developed a N-Gram language modal with Katz's discount smoothing to reduce errors from *pinyin* to Chinese words.

In the other case of translation from standard written Chinese into Braille, two steps are divided: the first is segmentation of characters into words, and then according to the dictionary translate word to Braille. Because segmentation of Chinese words and translation of Chinese Braille involves ambiguity and errors, we affiliated understanding function to solve illegibility and mistakes.

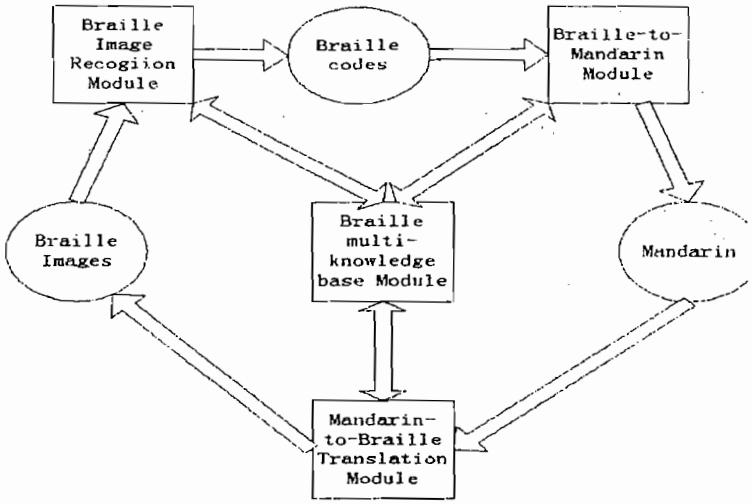

Figure 1 System Architecure of EasyBraille

II. System Architecture of EasyBraille

As shown in Fig.1, there are four modules in EasyBraille. The first is called Braille image recognition module, which performs the function of tansforming Braille images to Braille code files. The next two modules are two translation modules between Mandarin and Braille. The last module is an center-controlled module which interplays with the other three modules and gives them intelligent information.

2.1 Braille Image Recognition Module

Recognizing Braille images to Braille code files is a problem of OCR. As Braille codes have uniform interior gray level and rest upon a background of different, but uniform, gray level, we decided to choose global threshold method to identify the Braille codes from its background. The practical effect is better than 99%.

2.2 Braille multi-knowledge base Module

Braille multi-knowledge base is as follows:

1). The Braille electronic dictionary consists of a dictionary of word-class tags and phrase analyses, Braille to character correspondingly [2], and a dictionary of Braille abbreviations for high-frequency words, and several special forms of Braille codes. The knowledge dictionary records all the legal combinations of Mandarin consonants and vowels.

2). The rule base includes words, phrases, syntactic information and information on the segmentation of Braille words [2,3,4]. The knowledge-acquisition module takes grammar as its basis, which decides what information can be obtained from the training corpus and what information will be the responsibility of the understanding system.

3). The statistical base includes 7,5000 words' bi-gram and tri-gram statistics gathered from a 400-megabyte corpus, depending on statistical knowledge and the relationship between tag sequences [4,5].

2.3 Mandarin-to-Braille Translation Module

2.3.1 Mandarin word segmentation

The first step in translation from standard written Chinese characters to Braille is segmentation of sentences into words. Word segmentation separates a sentence into words and phrases in ways that are particular to Braille, which serves to avoid overly sparse syllable structure and ensure convenience for the blind to touch by connecting some words together. Braille word segmentation includes general rules, the general principle of which is shown as follows [1]:

1). Basically, Standard Written Chinese is made up of words, morphemes and syllables of a word put together.

2). Names of countries, social units, books and periodicals are divided as words.

3). Set sequences of two or more characters which denote a unitary concept are considered one word.

4). Some short phrases containing fewer syllables combined more tightly were put together in order to read or understand more easily and form concept quickly and reduce some scattered syllable. Braille word segmentation combines the work of several sub modules:

2.3.2 Bi-directional max-matching algorithm with a dictionary

Bi-directional maximum matching used was carried out to segment words using a dictionary, and emphasized particularly on checking and correcting the error. By comparing the maximum matches from a forward and a backward, the non-matching parts are segmented again by combining with the contextual information.

2.3.3 Knowledge base for disambiguation

In our word-token dictionary, every word marked with *pinyin* information to avoid ambiguity. Tagging ambiguity brought about by the changing of part-of-speech belongs to tags with multiple meanings. Adverbs, function words, and adjectives may be responsible for a large proportion of the ambiguous tokens. These categories contain relatively few words, while the frequency of the words tends to be high. However, these words play only the modifying, complementary and correlative roles in Chinese syntax, so one disambiguation method is to use a separate segmentation process for this kind of words. Its segmentation precision was better than 98% for the common text.

2.4 Braille-to-Mandarin Translation Module

2.4.1 Chinese Language Models with Katz's Discount Smoothing

Assuming a sentence consists of a sequence of words $W = w_1 w_2 \cdots w_n$, its probability can be decomposed into a conditional form [4]:

$$P(w_1 w_2 \cdots w_n) \approx P(w_1)P(w_2/w_1)\cdots P(w_n/w_{n-N}w_{n-N-1}\cdots w_{n-1}) \qquad (1)$$

$P(w_n / w_{n-N}w_{n-N-1}\cdots w_{n-1})$ is called an N-gram probability.

Chinese language N-gram models: Supposed a N-1 order Markov process generated a sequence of words; the probability of the i-th word w_i given in a conditional form depending on the last N-1 words, that is: $P(w_i / w_1 w_2 \cdots w_{i-1}) \approx P(w_i / w_{i-(N-1)} \cdots w_{i-2} w_{i-1})$

The smoothing method of bi-gram was to use Katz 's Discounting Model. That is to cover events that were not seen in the training text but could be observed with nonzero probability in the future texts by a redistribution of a part probability mass among unseen events by means of count dependent discount factors $\lambda_{N(h,w)}$. Thus, we obtain the language model with Katz's discount smoothing and conditional probability $p(w|h)$.

We assume count dependent discounting factors λr and use discounting only for counts $r \leqslant k$, say k=7.

$$p(w \mid h) = \begin{cases} \dfrac{N(h,w)}{N(h)} & \text{if } k < N(h, w) \\ [1 - \lambda_{N(h,w)}] \cdot \dfrac{N(h,w)}{N(h)} & \text{if } 1 \le N(h, w) \le k \\ \left(\sum_{w':1 \le N(h,w') \le k} [\lambda_{N(h,w')} \cdot \dfrac{N(h,w')}{N(h)}] \right) \cdot \dfrac{\beta(w \mid \bar{h})}{\sum_{w':N(h,w')=0} \beta(w' \mid \bar{h})} & \text{ifN(h, w) = 0} \end{cases} \quad (2)$$

$$\beta(w \mid \bar{h}) = \frac{N(\bar{h}, w)}{N(h)} \quad (3)$$

$$\lambda_r \cong 1 - \frac{r^*}{r} \quad (4)$$

$$r^* = \frac{(r+1)n_{r+1}}{n_r} \quad (5)$$

Where $\beta(w \mid \bar{h})$ is a more general distribution which is conditioned on backoff history h. It can be obtained by relative frequency as formula (3).Where λr is the discount factor, which we get by (4).Where r^* is referred to as Turing-Good count and n_r is total number of distinct joint events (h,w) occurred exactly r times.

2.4.2 Disambiguation from Braille into Chinese

In the case of translation from Braille into Chinese characters, every Braille character has several candidate Chinese characters, consisting of all nodes of multi-level graph. There are adjacent probabilities between two words, which are weighted between two nodes. Thus, a multi-level graph of transform Braille to Chinese characters was built. Our aim is to search a best path of maximum probability in the graph. Viterbi algorithm was used in the transformation which is suitable to search fast for a best path from a weighted multi-level graph.

Assuming N was the length of a Chinese character string: $Y=\#Y_1Y_2\ldots Y_N\#$ ('#' was starting or terminal symbols between a character string). Then the string Y corresponding to Chinese language candidates is: $w_{i1}w_{i2}\ldots w_{i,u_i}$, the weights of multi-level graph are formed. After the multi-level graph was built, a Viterbi algorithm was adopted to search for a best path.

The recursion formula of Viterbi dynamic programming is shown as follows:

$$G_1(w_{1j}) = \log \hat{P}(w_{1j} / \#) \qquad j=1,2,\ldots,u_1$$

.....

$$G_i(w_{ij}) = \max_{C_{i-1,k}} \{G_{i-1}(w_{i-1,k}) + \log \hat{P}(w_{ij} / w_{i-1,k})\} \qquad j=1,2,\ldots,u_I$$

.....

$$G_{N+1}(\#) = \max_{C_{N,k}} \{G_N(w_{N,k}) + \log \hat{P}(\# / w_{N,k})\} \qquad j=1,2,\ldots,u_N$$

By calculating $G_{N+1}(\#)$, w^* is gotten:

$$w^* = \arg Max_w \{P(w \mid Y)\}$$

Where, w^* was a character string corresponding to Chinese words.

2.4.3 N-Best searching algorithm

We used an N-Best algorithm to search for the N sequences of Chinese characters with highest estimated likelihood. When several paths lead to the same node in the word graph, the Viterbi criterion expands only the best scored path, and abandons the rest. The best path can be determined simply by comparing the cumulative scores of all possible paths leading to the terminal node of the word graph. Assuming that the single best hypothesis sentence was found by the Viterbi algorithm through a given word graph, the second best path is one of the paths which competed with the best one but was recombined at some node of the best path. Thus, in order to find the second best hypothesis sentence, we should consider all of the paths which share with the best path. That is, we should expand all of the nodes of the best path in the search for the second best path. The second best hypothesis sentence can be found by taking the path with the best score among the candidate paths, which might share a remaining section of the best path. By using the above rules repeatedly, we can find the N best paths. After comparing the costs of all paths the expansion of all the nodes in N-Best tree, that part of the second best path which is different from the best path is copied into the N-Best tree. Then the cumulative score of the new entry node is calculated in the backwards direction. Suppose the N best paths have been found. The (N+1)-th best path can be determined by examining all the existing nodes and comparing the paths with the expanded paths in N-Best tree as shown in Fig.2. The method can perform a comprehensive and efficient search of all the paths through the word graph structure.

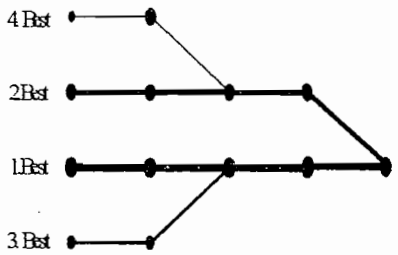

Figure 2. A expanded N-Best tree

III Realization of the System

Recognition module of Braille image, segmentation module of Mandarin Braille words, transforming module of Braille to *pinyin*, and transforming module of *pinyin* to Chinese character were integrated and a viewable interface was generated. The statistic information base of Bi-gram and Tri-gram is based on 75,000 thousand Chinese words and phrases. The natural language post-processing based on syntax rule to build Braille knowledge base. The translation of Braille or Chinese language to *pinyin*, *pinyin* to Chinese language used the most of homophone and polyphone knowledge of Chinese phonetics, syntax and language understanding to carry out multi-level rectifying mistakes and disambiguation processing. Braille knowledge base, rule base and statistic information base were built to realize the composite multi-level information processing and auto-correcting mistakes by the recognition modules of Braille scanning and codes, Braille electronic dictionary, statistic information base, language rule base. A N-gram model was adopted from a *pinyin* string to the weighted multi-level graph, the system searched for a best path in the graph which the algorithm was based on a sentence.

The test step is shown as follows: Braille texts were scanned into images,147 pages. Braille image recognition was transformed into Braille code files and the recognition mistakes were mended artificially. Braille code file was transformed into *pinyin* code file and then into Chinese character file. Obtained Chinese language text file was compared with the file of standard Chinese character and the compared result obtained. The test results are shown in table 1 and 2.

Table1: The test results of best and N-Best searching algorithm (N=5) in Braille translation into Chinese characters. NOFTW(NUMBER OF TOTAL WORDS), NOCW(NUMBER OF CORRECT WORDS), CR(CORRECT RATE), NOCWOT5C(NUMBER OF CORRECT WORDS OF TOP-5 CANDIDATES), CROT5C(CORR RATE OF TOP-5 CANDIDATES)

NOFT	NOCW	CR%	NOCWOT5C	CROT5C %
5216	4829	92.58	4988	95.63

Table2. The test results of Braille translation into Chinese characters (containing proper nouns and common ones) BFC(BRAILLE FILES CODE), NOTW(NUMBER OF TOTAL WORDS) NOCWC (NUMBER OF CORRECT WORDS,[CONTAINING]),CRC (CORRECT RATE,[CONTAINING]), NOCWNC (NUMBER OF CORRECT WORDS,[NO CONTAINING]),CRNC(CORRECT RATE ,[NO CONTAINING])

BFC	NOTW	NOCWC	CRC%	NOCWNC	CRNC%
p1-10	1950	1912	98.05	1918	98.36
p11-20	2714	2609	96.13	2615	96.35
p40 -49	1022	965	94.42	969	94.81
p55 -64	1935	1780	91.99	1840	95.09
p90 -99	2295	2140	93.24	2221	96.78
p110-119	2604	2410	92.55	2481	95.28
Total	**12520**	**11816**	**94.38**	**12044**	**96.20**

The test result of translation system from Braille to Chinese characters was 94.38% which included all kinds of types: P1-10, and P11-20 was illuminating articles; p40-49, and p110-119 was essays, p55-64 was comic dialogues; p90-99 was illuminating articles which introduced products. The experiments show if proper nouns were not considered we can get a 2% improvement further. The results on N-Best algorithm show the correct rate for top-5 hypothesis was 3% higher than that of the best hypothesis.

IV. Conclusion

Braille word's segmentation system consists of rules base, the signs base of segmentation and knowledge base for disambiguation and mistakes, by using adjacency constraints and bi-directional maximum matching with a dictionary, the system's segmentation precision is better than 99% for the common text. By the knowledge dictionary and the statistical language model, we perfectly solved the ambiguity in the translation from Braille to Chinese. A smoothing method was used to overcome sparse data in the model to be consistent with the phrase-level N-gram model in our system. For each *pinyin* sentence, we built a multi-level graph and used Viterbi algorithm to search the best Chinese character sentence, and used N-Best algorithm to get N Best Chinese character sentences, all the these made the transforming rate of the system inceased obviously.

References

[1] Teng Weimin, Li Weihong. China Braille, China Press, 1996

[2] Yu Shiwen, Zhu Xuefeng, Wang Hui et al. The grammatical knowledge-base of contemporary Chinese——A complete specification. Tsinghua University Press, 1998.

[3] Liu Yuan, Tan Qiang, Shen Xukun. Modern Chinese Language words segmentation criterion and automatic words segmentation methods in information processing", Tsinghua University Press, 1994

[4] Jiang Minghu, Zhu Xizoyan, Yuan Baozong, "Chinese Corpus N-gram Probability Statistic and Its Smoothing Algorithm" Journal of Tsinghua University. 1999, No.9. p99-102.

[5] Frank K. Soong, Eng-Fong Huang. A Tree-Trellis Based Fast Search for Finding the N Best Sentence Hypotheses in Continuous Speech Recognition. Proceedings of IEEE ICASSP 1991, Vol.1 pp705-708, Toronto, Canada.