Chinese-Braille Translation Based on Braille Corpus

Xiangdong Wang, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
Yang Yang, Jiangsu Enterprise Information Operation Center, China Telecom Corporation Limited, Beijing, China
Hong Liu, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
Yueliang Qian, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China

ABSTRACT

For people with visual disabilities, reading Braille text is an important way to acquire information. There are great challenges for Chinese-Braille translation due to the characteristics of word segmentation and tone marking in Chinese Braille. In this paper, a novel scheme of Chinese-Braille translation is proposed. Unlike current methods which use heuristic rules defined by experts for Braille word segmentation, the proposed method performs Chinese-Braille translation based on a Braille Corpus without experts on Braille. Under the scheme, a Braille word segmentation model based on statistical machine learning is trained on a Braille corpus, and Braille word segmentation is carried out using the statistical model directly without the stage of Chinese word segmentation. Tone marking and some special treatment are also performed based on word and rule mining on the Corpus. This method avoids manually establishment of rules concerning syntactic and semantic information and uses statistical model to learn the rules by stealthily and automatically. Experimental results show the effectiveness of the proposed approach.

KEYWORDS

Braille, Braille Word Segmentation, Chinese Braille, Chinese-Braille Translation

1. INTRODUCTION

For people with visual disabilities, reading Braille text is an important way to acquire information. Braille is a tactile writing and reading system used by the blind. Blind people read Braille by touching and recognizing the dots on the paper or a refreshable Braille display connected to computers or other terminals. Many systems have been developed to convert text in languages such as English, Danish, Spanish, Portuguese and Devanagari into corresponding Braille text (Christensen et al., 2012; Christensen and Chourasia, 2014; Christensen and Stevns, 2015; Coutinho et al., 2012; Bodale et al., 2014). The conversion from the text of the above-mentioned languages to corresponding Braille is relatively easy and simple, since there exists direct mapping from letters or words to Braille characters (known as cells). However, when concerning the language of Chinese, there are great challenges due to the characteristics of word segmentation and tone marking in Chinese Braille (Jiang and Zhu, 2006).
Unlike alphabetic languages, e.g. English, the basic unit of Chinese is character. And there are tens of thousands of characters used in the Chinese language. Therefore, it is impossible to map Chinese characters to Braille cells. In China, the most widely-used Braille system (called prevailing Mandarin Braille) maps the pronunciation of characters to Braille cells. Each syllable is written with up to three cells, representing the initial, final, and tone, respectively. To reduce ambiguity, Braille words are separated by spaces in writing, which is different from Chinese text. Furthermore, there is no direct mapping from Chinese word to Braille word, because phrases with relatively complete meaning are defined as words in Braille to further reduce ambiguity. In the standard of Chinese Braille, hundreds of rules are given to define Braille words, most of which are syntactic or semantic rules (AQSIQ, 2009; Teng and Li, 1996). This poses great challenge to automatic Chinese-Braille conversion, since the vocabulary of Braille words can be infinite according to the rules and the syntactic or semantic rules are difficult to be understood and processed accurately by computers. Another challenge of automatic Chinese-Braille conversion is the tone marking problem. To reduce printing cost and avoid information overload, in prevailing Mandarin Braille, an average of about 5% syllables need to be marked with tones. Besides some simple situations, most situation of tone marking depend on subjective judgment of experts, e.g., unfamiliar words and words may cause ambiguity, which makes it difficult to be automated by computers.

Currently, most research work on Chinese-Braille translation focuses on Braille word segmentation. Most researchers used expert-defined rules to obtain Braille words. The main idea is to firstly segment the Chinese text into Chinese words and then use pre-defined rules to convert Chinese words into Braille words, mainly by merging adjacent Chinese words into Braille words according to POS (part-of-speech) tags of words. For example, Huang et al. (2003) and Chen (2014) used rules based on SC grammar; Zhu and Bao (2001) and Jiang and Zhu (2000) adopted rules based on Chinese syntax; Li et al. (2002) defined 183 formal rules based on POS tags. However, as the rules of Braille words are defined according to syntactic and semantic information, it is quite difficult to convert those rules into formal representation which can be understood and executed by computers. Therefore, the performance of those methods is quite limited. Furthermore, the performance of state-of-art syntactic analysis and POS tagging techniques remains immature, resulting in cascade error in Braille word segmentation. Considering the problems of rule-based approaches, some researchers proposed corpus-based methods, the main idea of which is to extract Braille words from a pre-constructed Braille corpus and perform Braille word matching in the state of Chinese word merging (Yang and Che, 2011), sometimes in conjunction with human-defined rules (Chen, 2014; Li et al., 2002). However, as mentioned above, the vocabulary of Braille is infinite. Therefore, it is difficult to extract enough rules for accurate Braille word segmentation.

Currently, there is little work on tone marking of Braille words. Most current systems return Braille words without tone marking, which may cause ambiguity and make the Braille text difficult to understand. Also, current work only considers translating pure Chinese text to Braille. However, in practice, Chinese text always includes punctuations, and sometimes includes numbers, letters, and other symbols. Special treatment is need for these situations. For example, some punctuation may be separated from Braille words and some are not; some letters and numbers may be concatenated to a Braille word by a hyphen and some are not. Some rules considering these situations are given in the standard of Chinese Braille (AQSIQ, 2009; Teng and Li, 1996), but the rules are also dependent on semantic information and are difficult to be formalized.

In this paper, a novel scheme of Chinese-Braille translation is proposed. Unlike current methods which use heuristic rules defined by experts for Braille word segmentation, the proposed method performs Chinese-Braille translation based on a Braille Corpus without experts on Braille. Under the scheme, a Braille word segmentation model based on statistical machine learning is trained on
Handhelds for Digital Libraries
www.igi-global.com/chapter/handhelds-digital-libraries/37828?camid=4v1a

A Trust Model for Detecting Device Attacks in Mobile Ad Hoc Ambient Home Network
www.igi-global.com/article/a-trust-model-for-detecting-device-attacks-in-mobile-ad-hoc-ambient-home-network/179244?camid=4v1a