Good Spelling of Vietnamese Texts, one aspect of computational linguistics in Vietnam

PHAN Huy Khanh
Department of Information Technology
DaNang University
17, Le Duan Street, DaNang City, Vietnam
phanhuykhanh@dng.vnn.vn

Abstract

There are many challenging problems for Vietnamese language processing. It will be a long time before these challenges are met. Even some apparently simple problems such as spelling correction are quite difficult and have not been approached systematically yet. In this paper, we will discuss one aspect of this type of work: designing the so-called Vietools to detect and correct spelling of Vietnamese texts by using a spelling database based on TELEX code. Vietools is also extended to serve many purposes in Vietnamese language processing.

Introduction

For the past two decades computational linguistics (CL) has progressed substantially in Vietnam, mainly in these basic aspects: data acquisition from the keyboard, encoding, and restitution through an output device for Vietnamese diacritic characters, updates on the fonts in Microsoft DOS/Windows, standardization for Vietnamese (James Do, Ngo Thanh Nhan), automatic translation of English documents into Vietnamese and vice versa (Phan Thi Tuoi, Dinh Dien), recognition of handwriting (Hoang Kiem, Nguyen Van Khuong), speech processing (Nguyen Thanh Phuc, Quach Tuan Ngoc), building bilingual dictionaries such as English-Vietnamese and V-E, French-Vietnamese and V-F dictionaries (Lac Viet), archives of old Sino-Vietnamese documents (Ngo Trung Viet, Cong Tam), etc. Some of these works have been presented in Informatics and IT workshops organized in Vietnam. These efforts are modest and do not yet show our full potential. There are many reasons for this weakness. The major reasons that the different efforts are quite isolated and there is not enough coordination. Some coordinated workshops held from time to time would be very helpful.

At the IT Dept. DaNang University we are building a lexical database based on TELEX code for accomplishing the following tasks:

- Converting Vietnamese texts from any font to any other font.
- Putting texts in alphabetical order independently of the font in use.
- Looking up words up in the monolingual and / or multilingual dictionary.
- Building specialized monolingual dictionaries.

At present, we are taking part in the GETA, CLIPS, IMAG, France, in the FEV project: for a multilingual dictionary: French-Vietnamese via English.

In fact, inputting Vietnamese texts still encounters many problems, not yet solved properly. The most common mistakes in detecting and correcting spelling errors are:

- wrong intonation or misspelling,
- not following spelling specialization, not using syllables systematically in the same texts, etc.

Winword, a commercial text processor, is not able to detect and correct spelling mistakes. The program designed by Ngo Thanh Nhan (without an associated spelling dictionary) and other software packages for Vietnamese still do not offer adequate solutions.

We propose here a general solution for building the so-called Vietools for detecting and correcting spelling errors. Vietools is designed for office application such as Winword, Excel, Acess, PowerPoint, etc. in Microsoft Windows. Vietools has also been extended for converting and rearranging Vietnamese words in the dictionaries and consulting the Vietnamese dictionaries, including multilingual dictionaries.

1 Building spelling database

In the spelling dictionary by Hoang Phe (1995), there are 6760 syllables in the writing

system (6616 syllables in the phonology system) to compose single words or complex words. Each syllable has two parts: initial consonant (optional) and rhyme pattern (including rhyme and tone). Altogether, there are 27 initial consonants, and 1160 rhyme patterns (including 6 tones).

Based on Vietnamese syllable structure, the spelling database is built in a tabular form. Each element of the table helps to check the correction of a syllable based on the column position of initial consonants and the row position of rhyme patterns, for example, the syllable *lamf* (work) in the TELEX form, is composed of the initial consonant *l* and rhyme pattern *am* with by low falling tone (or grave accent) *f*. Each element of the table can be understood as:

- syllables used in Vietnamese.
- elements between tone sign positions (on *o*: *oja* or on *a*: *oaj*), pronunciation or dialect with spelling (*z* is equivalent to *d* or *gi*, *y* is equivalent to *i*...) and borrowings such as *karaoke*, *photocopy*, *fax*...
- Sino-Vietnamese word: *coongj* (addition) → *congj*, *quoocs* (country) → *nuwowcs*...
- being unable to form syllables: *quts*, *quoon*, *coan* , *cuee*...

Techniques have been developed to recognize the compound words from two syllables, such as *baor damr* or *damr baor* (guarantee), *chung chung* (vague), etc., from three syllables, such as *howpj tacs xax* (cooperative), etc., from four syllables, such as *coong awn vieecj lamf* (work, job), etc.

2 Designing Vietools

The error detecting program reads one syllable at a time from the text. The syllable is divided into an initial consonant and a rhyme pattern, paying attention to solving initial consonants such as: *gi* containing vowel *i*; the consonant *qu* has vowel *u*, but it is easy to separate it from the syllable for it does not have the consonant *q*; the other combined initial consonants have the length of 2, or 3. The error-correcting unit checks the conformity of initial consonants (if present) and the rhyme pattern.

3 Code converting

At present, there are many Vietnamese fonts built on different codes (different in number of bytes used: 1 byte or 2 bytes, order of tones, letter arrangements, etc.). Because there has not been a unified code for Vietnamese text, we selected a pivot code and TELEX code. There are many codes to convert from such as IBM-CP01129, Microsoft-CP1258, VISCII, VietKey, VietWare, VNI, TCVN3, Unicode, etc. Vietools works on syllables converted to TELEX. Vietools analyses syllables to detect initial consonants and rhyme pattern in TELEX code.

Conclusion

The main advantage of our method is that the tool operates independently of the Vietnamese font used. The design of Vietools is open: one can add new functions such as text or data conversion Spelling data base structure design helps building multi-functional dictionaries, which are essential for natural language processing.

Acknowledgements

My thanks go to my students for the realization of Vietools and my colleagues for their opinions. In particular, I thank Professor Aravind Joshi, University of Pennsylvania, Philadelphia, USA, for his helpful suggestions I am grateful to Christian Boitet, Professor, Joseph Fourier University, GETA, CLIPS, IMAG, France, for his comments on this paper.

References

1. Hoang Phe (1995) *Dictionary of Orthography*. Center of Lexicography, DaNang Publishing House, 509 p.
2. Hoang Phe (1997) *Vietnamese Dictionary*. Center of Lexicography, DaNang Publishing House, 1130 p.