**Automatic Insertion of Vietnamese Diacritics in Mobile Message**

Nguyen Cong Chinh – Nguyen Vu Long – Bui The Huy – Nguyen Quang Truc

Student at FPT University

**Abstract – This paper presents a solution for the insertion of diacritics into a text where they are missing. Missing diacritics on text are making difficult for the repient to understand the content being transmitted or wasting time re-reading the content. This paper is a result of our group for the purpose to solve the problem of automatic insertion of diacritics into a text, especially Vietnamese text. Due to language specificity, Vietnamese has many diacritical marks and more particularly, white space is not often seen as the basis for determining the world. This paper also provide the proposal a point wise approach for automatically recovering missing the diacritics, using three conceptual bases: n-gram of syllable, n-gram of syllable types and dictionary word. Building corresponding feature for the classification through this approach, our team’s experiments have shown that resilience markers have an approximate rate of 88~92%.**

**Keyword- automatic insertion of Vietnamese Diacritics, classification, natural language processing, Vietnamese.**

1. Introduction

Text Classification, detecting spelling error and correct spelling are problems that related to natural language processing. The problem of detecing spelling error and correct spelling error are the most commonly encountered in the problem of spelling correction for text. However, there is one very common problem in correcting a misspelling that is to recover the lost mark for word. Due to linguistic characteristics of each country, this problem only occurs in countries like Southeast Asia, some countries in Africa, and some European countries such as France, Romani,… Recovering tone mark offers many benefits: avoid misunderstanding the content is transmitted, for situations requiring rapid recording, bring convenience, still express the full meaning… In addition, it is a necessity in today’s generation with a larger number of people that tend to ignore the signs… missing diacritics pose a serious problem not only for automatic text processing and information retrieval, but also for human readers.

Typically, there are 2 ways to approach the problems that recovery tone mark for word: word-based and character-based

Word- based: Using techniques such as parsing, word-type analysis, etc... to guess the sign of a word. This method can’t be applied to Vietnamese, because Vietnamese must conduct word analysis, and the accuracy of this process will be very low in Vietnamese without accent. In addition, this method requires a large enough data (dictionary). In contrast, character-based systems use language-independent algorithms based on statistical information that has been learned from training data. For languages in which diacritics signal grammatical or semantic roles, word-based systems are much more reliable than character-based systems

Compared with other languages, Vietnamese is much more complex. We can see through the following statistics:

+ More than 95% of Vietnamese words contain accents, while French is only 15%, roman is 35%.

+ More than 80% of Vietnamese syllables are missing in duplicate and unclear words (for example: “sach” may be “sách” or maybe “sạch”, etc…). Normally, the choice between two approaches will depend on some factors: the technology, the user request, language specify,…

Vietnamese text uses the Latin alphabet. However, Vietnamese has letters that are modified with diacritics: đ, ă, â, ô, ơ, ư and letters in Vietnam are also have accents: ắ, ạ, ố, ồ, ổ, ộ,… Due to the complexity language, many Vietnamese choose to use non-diacritical Vietnamese, which can be entered using any computer, phone and is easier and quicker to type. However, non-diacritical Vietnamese is difficult to understand and can make confusing for repients or readers.

The abudant of diacritics along with the absence of a word seperator also make Vietnamese a difficult language for recovering character-based. Therefore, we need an approach to restoring diacritics for the Vietnamese language.

The rest of this paper is organized as follow. Section II present briefcase overview of Vietnamese orthography, section III describies our approach to recovering diacritics on text and the section IV is the result of an experimental evaluation of our approach. And the last section V, presents our conclusion.

1. Vietnamese Orthorgraphy

written Vietnamese uses the Latin alphabet, known as “quoc ngu” (quốc ngữ) in Vietnamese, include:

* 26 letters in English alphabet except f, j, z, w
* 7 letters are used with the diarictics: đ, ă, â,ê, ô, ơ, ư (Including upper case ).

In addition, diacritics are used to indicate different tones in Vietnamese. The tone markings (using the letter a as an example) are a (no mark), à, á, ả, ã, and ạ. Letters modified with diacritics and tone markings can be combined to produce complex diacritics such as ằ, ở, ể, and ẫ.

Moreover, Vietnamese language has a problem that is white space is not used to separate consecutive words. In Vietnamese, spaces may used to separate syllables within words . Therefore, white space does not always serve as a word separator.

Non-diacritical Vietnamese means diacritical marks will be removed from a word . For example, we consider “ố”, “ồ”, “ộ”, “ổ” will becom “o” when the diacritical marks are removed . All letters that involve diacritics are show in Table 1.

|  |  |
| --- | --- |
| Diacritical Character | Non-diacritical character |
| "a", "à", "ả", "ã", "á", "ạ", "ă", "ằ", "ẳ", "ẵ", "ắ", "ặ", "â", "ầ", "ẩ", "ẫ", "ấ", "ậ" | “a” |
| "e", "è", "ẻ", "ẽ", "é", "ẹ", "ê", "ề", "ể", "ễ", "ế", "ệ" | “e” |
| "i", "ì", "ỉ", "ĩ", "í", "ị" | “i” |
| "o", "ò", "ỏ", "õ", "ó", "ọ", "ơ", "ờ", "ở", "ỡ", "ớ", "ợ", "ô", "ồ", "ổ", "ỗ", "ố", "ộ" | “o” |
| "u", "ù", "ủ", "ũ", "ú", "ụ", "ư", "ừ", "ử", "ữ", "ứ", "ự" | “u” |
| "y", "ỳ", "ỷ", "ỹ", "ý", "ỵ" | “y” |
| "đ", "d" | “d” |

Through Table I, we can see that it is easy to transform diacritical characters to non-diacritical character in Vietnamese, but the inverse tranformation is more complicated. For example, the syllable “bàn” (desk) in diacritics Vietnamese become “ban” in non-diacrictical Vietnamese. However, the word “ban” in non-diacritical Vietnamese can have mean of words “bận” (busy), “ban” (department), “bắn” (shoot). We refer word that like “ban” in so far can have more than one diacritical or non-diacritical syllable as “ambigous syllable”.

According to Simard (1998) [5], approximately 85% of the words in arbitrary French texts contain no accents, while more than approximately 85% of the syllables in Vietnamese and 95% of the words contain letters that use diacritics.

Vietnamese uses diacritical marks more extensively than other languages, and when they are missing it is much more difficult to restore them.

1. Point wise approach

The approach focuses on recovering diacrtics for each word in the sentence independently, for each syllable in sentence form information: features.

For example, give non-diacritical sentence s = s1s2s3…sm ( si is syllable with 1<= i <= m). The diacritic restoration whether a syllable si depend on features that form on the surrounding non-diacritical syllables. We are using three conceptual bases : n-gram of syllable, n-gram of syllable types and dictionary word for the purpose forming the corresponding feature for syllable:

N-gram of syllable describes which syllable arround the given syllable. With the meaning of window size W , specify a particular number and only syllables within this window are used. Approximately 70% of the words in the Vietnamese language are composed of two syllables, and approximately 14% are composed of at least three syllables. The high frequency of twosyllable compounds suggested using the window sizes W = 2 and W = 3.

N-gram of syllable types using symbols to characterize surrounding syllables:

* + - Lower case syllable (L): syllable contains only lower case letters.
    - Number (N) syllable is a number.
    - Other (O): syllable is something else (other), such as a symbol

And the last, dictionary word are words that contain the given syllable.

For example, consider the first occurrence “ban” in the sentence “cai ban nay duoc dat o phong ban ke toan” (this table is place in financal department).

Syllable and Syllable type 1-gram and 2-gram feature with the window size = 2:

cai ban nay duoc dat o phong ban ke toan

|  |  |  |
| --- | --- | --- |
|  | Syllable | Syllable types |
| 1-gram | “cai”, “ban”, “nay”, “duoc” | “L” |
| 2-gram | “cai ban”, “ban nay”, “nay duoc” | “LL” |

And dictionary words are also “cai ban” and “ban nay”.

The second occurrence “ban”:

cai ban nay duoc dat o phong ban ke toan

|  |  |  |
| --- | --- | --- |
|  | Syllable | Syllable types |
| 1-gram | “o”, “phong”, “ban”, “ke”, “toan” | “L” |
| 2-gram | “o phong”, “phong ban”, “ban ke”, “ke toan” | “LL” |

With the first occurrence “ban” we have features (“cai”, “ban”, “nay”, “duoc”, “cai ban”, “ban nay”, “nay duoc”, “L”, “LL”) and the second occurence “ban” (“o”, “phong”, “ban”, “ke”, “toan”, “o phong”, “phong ban”, “ban ke”, “ke toan”, “L”, “LL”). These features are created for each syllable and are used to classify the diacritical marks.

1. Experiments

We write a simple application that crawl data from the journalism pages to create a corpus of the text it’s looking for. With this raw data, we removed all diacritical marks, follow on the rule in table I. Using two data file (raw data and data without tone mark) to train model data.

We used linear support vector machine, implemented in SCIKIT-LEARN library [2] to solve the classification task. With the chosen features, the accuracy of pointwise approach depends on the window size W and the size of training data. The above table shows the result of our experiments with window size = 2 and the training data size of 16, 32 and 64 mb.

|  |  |
| --- | --- |
| The training data size | Accuracy (%) |
| 16 MB | 75% |
| 32 MB | 88% |
| 64 MB | 92% |

Moreover, because data we used which are crawled from the journalism websites, so that it may contain unknow words and may errors. Therefore, the accuracy of diacritics restoration in some cases is low. The highest obtained was 93.7% with window size = 2. For a complex language such as Vietnamese, this is an acceptable result.

1. Conclusion

The automatic insertion of diacritics into written Vietnamese texts is important for many applications including mobile message reading, search engines, …

This paper has presented a solution for recovering diacritics in Vietnamese texts. In our experiments, testing with the window size = 2, the pointwise approach show the acceptable result. We belive that larger window size may show the better accuracy result. Moreover, larger training data size may impore the accuracy.

The negative consequence of this approach is required a large data, with data size does not large enough, the accuracy is very low. Moreover, with the large data, the training data processing take long time to train, and some situation can affect on the training data process including “not enough memory”, crash computer, … Finally, with the window size = 2 and the training data size is MB, the total size of the model’s file was GB. In the near future, we hope can find the better feature so that the model’s file can made much smaller.

REFERENCE