|  |  |  |  |
| --- | --- | --- | --- |
| ***Thinh, Le Phuc***  FPT University  Ho Chi Minh City, Vietnam | ***Nguyen, Vo Thach***  FPT University  Ho Chi Minh City, Vietnam | ***Hung, Pham Chanh***  FPT University  Ho Chi Minh City, Vietnam | ***Anh, Dang The***  FPT University  Ho Chi Minh City, Vietnam |

**Abstract** – Statistics prove that companies across the globe invest heavily on employee training and development. For example, employee training and education spends in the United States alone are growing significantly by 14% every year. Besides enhancing knowledge and skills, corporate training has proven to be an important tool to boost employee engagement and retention. However, in Vietnam, corporations are underestimating the importance of training and education. Therefore, companies are having limited and uneffective ways of training their employees. Seeing this problem, we build a system that may change corporations in Vietnam for a better and more efficient in training employees.

1. **INTRODUCTION**

Nowadays, online learning turns out to be more and more practiced. Many corporations start to post their training programs and courses online for their employees. It represents an easy and comfortable method to achieve knowledge in almost every field, from law and accounting, to even human sciences, such as psychology and physics. It represents a great way to study many fields and to boost the level of self-motivation. We can see that e-learning is a great alternative for corporations, especially for companies which can’t afford the time, money, human resources to train their employees. Currently, many corporations are having problems to find a suitable system that can provide them an environment to train staffs, manage training plans and view employees report easier and more effective. We build a system called TRAISY, which allows any corporates to post their training program with ease and privacy on our website. Through our system, staffs can access to training anywhere and anytime, making learning and training more effective, faster and easier. For corporations’ admin and manager, they can manage their programs effortlessly, and view their employees’ statistics.

TRAISY aims to provide all corporations in Vietnam with a professional environment for training staffs. To achieve that goal, the application has to, not only perform its business activities accurately such as course management, learner statistics, and user management, but also assist user with good UX (User experience) and beautiful UI (User interface).

1. **PROBLEM AND SOLUTION PLAN**

When implement the TRAISY system, we faced several problems:

1.Buy or build?

Video streaming is one of our mandatory functions, so the system must implement this function with these requirement:

* Pro-quality for mobile view.
* Low-latency and wide bandwidth for smooth streaming.
* Flexible and customizable for developer to maintain and upgrade.
* Professional-grade streaming

At beginning, we found some platforms for streaming video with open source and free such as YouTube. We did integrate to our system and everything worked well. But soon, we found out that these free systems are hard to modified and not so great experience.

After that

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 files (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 shows 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