An Optimization Text Summarization Method

Based on Naïve Bayes and Topic Word

for Single Syllable Language

Ha Nguyen Thi Thu

Department of Computer Science
Vietnam Electric Power University
235 Hoang Quoc Viet, Tu Liem, Hanoi, Vietnam

Abstract

Text summarization since the late 50’s of the 20th century by the simple technical based on term frequency and it applied for technical text summarization at IBM institute. During more than 50 years of development, text summarization is still a hot topic that attracting many researchers, scholars in the field of data mining and natural language processing proposals development of the text summarization system. For the English, there are some automatic text summary systems was built as SUMARIST, SWESUM,… But for single syllable languages like Chinese, Vietnamese, Japanese, Thai, Mongolian and other "native" languages in Southeast and East Asia. Amount people use single syllable language more than 60% of all language on the world. So that, processing of single syllable language is very necessary. However, it is very complex for language processing problem because it’s very hard to determine word or term based on white space and all word segmentation tools not reach 100% accuracy currently. In this paper, we propose a text summarization method based on Naïve Bayes algorithm and topic words set. We’ve experimented with 320 Vietnamese texts (equivalent to 11,670 Vietnamese sentences) show that our method is really effective; text summary is readable, understandable and closer with summary of the human.

Keywords: Single syllable language, text summarization, dimensionality reduction, Naïve Bayes, supervised learning, natural language processing, topic word
1 INTRODUCTION

The rapid increase in the amount of data on Internet has brought to users the convenience of retrieving and searching information but also difficult to read because texts is often too long. Text Summarization is a subfield of natural language processing, it return a short text (shorter than original text) and maintaining the gist of information from original text [4], [5].

During more than 50 years of text summarization development, there are many methods proposed for the purpose of building the automatic text summarization that satisfying the user's requirements. There are many approaches but has two main approaches: supervised learning and unsupervised learning. In the recent research’s often approached by unsupervised learning methods. With this method, often reduce the cost when build data set, reduce time, costs in training phrase and reduce the computation complexity through learning model. However, this approach quality of text summarization is not high, because of calculating sentences score by linear combination of features and select sentences that has high score [4], [5], [6], [19], [26].

Text summarization based on Naïve Bayes was Kupiec and colleagues using with English text in 1995 with simple features such as length of sentence, word frequency [21]. In 1999, Aone gives a method based on word frequency (tf) and inverse document frequency (idf). In his method, he use vocabulary and WordNet to increases in natural language processing in summary system, and this method also used to build the automatic text summarization system (call Dimsum) and is considered as the first foundation method that developed system automatic text summarization successful [1].

For some single syllable languages as Chinese, Japanese, Vietnamese, Thai ... when we build automatic summarization system, automatic classification system, retrieval system… we usually word segmentation tool as a phase in process of system building. Because words in single syllable language cannot determine based on whitespace, for example: preparation (English) is translated into Vietnamese: chuẩn bị, and Chinese is: 准备(this is a compound). On the other hand, word segmentation tools are not high accuracy: With Chinese, accuracy rate is only about 85% [17], [19], [34], [38], with Vietnamese exact rate at approximately 80% [37]. Therefore, if this problem resolved the quality of summary text can be better.

Because of enhance quality of text summary, accuracy and reduce time for computing in single syllable language, we have chosen a approach based on supervised learning method using Naïve Bayes with the goal of improving the quality of summaries through the selection of training data and training data by experts, and we also use tagger tool for extracting topic words set in the training phase. In summary phase, without segmentation tool will reduce computational complexity in terms of time (Figure 1).
The rest of this paper is structured as follows: Section 2 introduces the related works in the field of text summarization. Section 3 introduces construction method of topic word and how to reduce the computational complexity. Methodology of summary base on Naïve Bayes and topic word will be presented in section 4, section 5 is the experimental and results and we use Vietnamese for experimental, finally is conclusion.

2 RELATE WORKS

The first research text summarization is usually based on some simple characteristics to identify and extract the important sentences in a document, features are often used: sentence length, sentence position, the title appears, the frequency of keywords (or terms). While some text summarization methods based on unsupervised learning techniques [2], [7], [10], [24], [38] still has a number of published works that based on supervised learning techniques like Edmundson, he implement data set includes 400 technical documents for training [6].

Since 1990, along with the development of computer science and machine learning techniques in natural language processing has a new look at text summarization, the researchers used machine learning techniques to summarize text. The machine learning models are mentioned as: Naive Bayes, artificial neural networks, decision trees, SVM models, hidden Markov models.

Kupiec et al (1995) have described a method of text summary based on the ideas that has inherited from Edmunson by training data set. Naive Bayes classifier function used to divide into two sentences class with meanings extracted or not. Kupiec used some features : heading word, length of sentences and words frequency [21]. In 1999, Aone also propose a method using Naïve Bayesian classification to develop an automatic text summarization system called Dimsum, in his method, tf×idf be used as a main characteristic and he also used WordNet dictionary and vocabulary to increase the quality of summaries [1].

Text summarization using neural network is presented since 2000’s, with the typical study by Lin, Nenkova and Svore. In this study, they have used the standard database with more than 1000 documents labeled by experts [23], [33], [36].
Text summarization based on Hidden Markov Model (HMM) were Conroy and O’Leary announced in 2001, this method based on idea: features is dependence. There are three features were used: position of the sentence in the document (built into the state structure of the HMM), number of terms in the sentence, and likeliness of the sentence terms given the document terms [18].

For single-syllable language, the proposed methods is mainly used unsupervised learning techniques, word segmentation tools often mentioned in the proposal. With Chinese text, some automatic text summarization systems based on surface information of context likes: Jun-Jie Li and Key-Sun Choi proposed a method that based on corpus, in this method they use segmentation tool and designed a new full text indexing method called Natural Hierarchical Network (includes 6 levels) for calculating sentence score through 2 key features: important words and few number of sub-sentences [19]. Yanmin Chen, Xiaolong Wang, and Yi Guan proposed a chinese text summarization by computes lexical chains based on the HowNet knowledge database [38]. And with Vietnamese text, some text summarization method was proposed likes: Thanh Le Ha in his article used include word frequency, sentence position, sentence length, title words. In his method also mentioned word segmentation tool for separate words from document and calculate tfxidf [37] ... Nguyen Le Minh proposed method Vietnamese text summary based on SVM [30]...

3 FEATURE REDUCTION

3.1 Text Representation

Text representation methods that describe the content or characteristics of the text. When performing vector text as it is common to use the vector components represent corresponding features are the words (terms). With the value of each characteristic is called the weight of the word (term), describe the frequency of words (terms) appear in the text.

Suppose that we have a document d, figure 1 illustrates representing d. In which, \( w_{ij} \) is the weight of word (or term) \( t_{ij} \), i is sentence \( i^{th} \) in document d, j is position of word (term) in the sentence \( i^{th} \) from left to right.

\[
T = \begin{bmatrix}
W_{1,1} & W_{1,2} & \cdots & W_{1,n} \\
W_{2,1} & W_{2,2} & \cdots & W_{2,n} \\
\vdots & \vdots & \ddots & \vdots \\
W_{m,1} & W_{m,2} & \cdots & W_{m,n}
\end{bmatrix}
\]

Figure 2. Matrix of text representation.

Example 1: Have an original Vietnamese text include 34 words.

“Các nhà nghiên cứu thuộc trường Đại học Michigan vừa tạo ra một nguyên mẫu đầu tiên cho hệ thống tính toán quy mô nhỏ, có thể chứa dữ liệu một tuần khi tích hợp chúng vào trong những bộ phận rất nhỏ như mắt người.”
Optimization text summarization method

"密歇根大学的研究者已经创建了一个小规模计算的第一个原型系统，它可以存储一个星期的资料，还可以集成在很小的部分如人眼睛。"

Pinyin:"michigen da xue de yan jiu zhe yi jing chuang jian le yi ge gui mo ji suan de di yi ge yuan xing xi tong, ta ke yi cun chu yi ge xing qi de zi liao hai ke yi ji cheng zai hen xiao de bu fen ru ren yan jing"

Translate in to English:
"Researchers at the University of Michigan have created a first prototype system for small-scale computing, which can contain data for a week while integrating them into very small parts as the human eyes."

Like this document, we must calculate weight for 34 words. And representation matrix with 1 row and 34 columns like below

\[ T = \{t_{1,1}, t_{1,2}, \ldots, t_{1,34}\} \]  

In the general case, the original text often includes multiple sentences, each sentence consists of many words, so if expressed in the above, the weight matrix representation of the text will have some very large, requiring more time for computation. With common Vietnamese text, first we need separate all the words (use word segmentation tool) in the text and the compute weight of each word will require more and more time treatment.

3.2 Methodology of feature reduction.

Feature selection is one of the key topics in machine learning and other related fields. Real-life datasets are often characterized by a large number of irrelevant or redundant features that may significantly hamper model accuracy and learning speed if they are not properly excluded. Feature selection involves finding a subset of features to improve prediction accuracy or decrease the size of the structure without significantly decreasing prediction accuracy of the classifier built using only the selected features.

To overcome the disadvantages of large feature vector and the cost to build a large set of terms for each topic. In this paper, we propose a new method which can reduce complexity computing of large feature set by using a word segmentation tool for separating word into two word sets: nouns set (called topic word) and other words set. In a any text, nouns contain information of text. So, when we extract nouns from text, remarkable reduction of large feature set.

In example 1, we separate document d into two sets, one set include nouns, called nouns set, if we use Vietnamese text, nouns set is below:

Nouns set $T' = \{nha_nghienv_cutu, truong, dai_hoc, Michigan, nguyen_mau, he_thong, quy_mot, dur_lietu, tuan, bo_phan, mat, nguoi\}$.

If we use Chinese text to separate words, nouns set is $T''$ like below

Noun set $T'' = \{密歇根, 大学, 研究者, 规模, 原型, 系统, 星期, 资料, 部分, 人, 眼睛\}$. 
Pin yin T’’={mi_ke_gen, da_xue, yan_jiu_zhe, gui_mo, yuan_xing, xi_tong, xing_qi, zi_liao, bu_fen, ren, yan_jing}.

Use text separation technique in two sets, the size of the matrix T will be reduced, for example, with the original text in example 1, instead of using the T matrix contains one row and 34 columns, we only need the matrix T’ consists of one row and 14 columns:

\[ T = \{ t_{1,1}, t_{1,2}, \ldots, t_{1,14} \} \]  

So now, we reduced from 34 feature vectors to 12 feature vectors.

4 SINGLE SYLLABLE TEXT SUMMARIZATION

4.1 Features Selection

In previous studies, we also use this features reduction method with purpose: reduce processing time, reduce computational complexity and enhance quality of summary. In this method, we use three key features for calculating weight of sentences as: Information significant, amount of information and position of sentence.

- Information Significant (topic word information Significant): It express important of topic word in sentences and it is calculated by number of sentences that appear this word divide by all sentences in training set.

- Amount of information in a sentence: Other method usually mentions sentence length as a key feature, but in our method, we use amount of information feature instead of sentence length feature. It is calculated by total number of topic word in sentence.

- Position of sentence: Sentence in a text has an each difference of important. Commonly, position of sentence at first paragraph or title of text has high value. Position of sentence score is calculated by equation \( \text{Pos} (s_i) = \frac{1}{i} \). i = 1 when si in the top of paragraph.

4.2 Naïve Bayes classification

Naïve Bayes classification used to calculate the probability of a sentence s with k features like F1, F2, ..., Fk be the set S or not? by the following formula

\[
P(s \in S | F_1, F_2, \ldots, F_k) = \frac{P(F_1, F_2, \ldots, F_k | s \in S) \times P(s \in S)}{P(F_1, F_2, \ldots, F_k)}
\]

Assuming that the feature is independent formula (3) converted to

\[
P(s \in S | F_1, F_2, \ldots, F_k) = \prod P(F_j | s \in S)P(s \in S) / \prod P(F_j)
\]

Using logarithmic rule (4) into:

\[
P(s \in S | F_1, F_2, \ldots, F_k) = \log(P(s)) + \sum \log P(F_j | s)
\]

In which:
P(s) = C(s)/C(w) where C(s) is the number of sentences in training set and C(s) are in class C, C(w) is total of sentences in training set.

P(Fj|s) = C(Fj,s)/C(s). Where C(Fj,s) is the number of ocularity Fj in sentence of class C.

We use Naïve Bayes to classify in to two classes (extracted class and non-extracted class). So, we compute the probability corresponding to each case \( P(s \in S | F_j) \) and \( P(s \notin S | F_j) \). Sentence will be selected for extraction when \( P(s \in S | F_j) > P(s \notin S | F_j) \).

4.3 Using Naïve Bayes classification for single syllable text summarization

We use Naïve Bayes classification for single syllable text with two phases: Training and summarization. In Training phase, we trained from data and with the participation of human to build a set of extracted sentences.

- Training Phase:

  - Step 1: Build the training data set D consists of n documents D = \{d_1, d_2, ..., d_n\}

  - Step 2: Use the POS tagging tool to create the set of nouns V.

  - Step 3: Human will select out the most important sentences in each ds to creat text summary.

  - Step 4: Division into two classes: Selected sentences labeled (+), and not-select sentence is labeled (-). Save position of sentences (including the labeled (+) sentences and the labeled (-) sentences in the text ds.

  - Step 5: Calculate the value of topic word information significant in from both labeled (+) sentences and (-) sentences. Save this values to database table.

  - The figure 3 below illustrated 5 steps in training phase.
**FEATURE SCORE (WEIGHT)**

**Input**

S: set of sentences; V: dictionary of nouns;

**Output**

W’: list of sentences and value corresponding of it

1. **Initialization**

   \[ W' = \phi; \ I = \phi; \ P = \phi; \ M = \phi; \ x = 0; y = 0; \]

2. **For i=1 to count(S) do**

   2.1 **For k=1 to length(si) do**

      2.1.1 If (s_i[k] is noun)

      2.1.1.1 x=x+I(s_i[k]);

      2.1.2 \[ I[k] \leftarrow x; \]

      2.1.3 x=0;

   2.2 **For k=1 to length(s_i) do**

      2.2.1 If (s_i[k] is noun) calculate

      2.2.1.1 y++;

      2.2.2 \[ M[k] \leftarrow y; \]

      2.2.3 y=0;

   2.3 \[ P[k] = \leftarrow MATCH(s_i,d); \]

**Figure 3.** Calculating value of topic word in S and S’.

**-Extraction phase**

- Step 1: For T original text (the text to be summarized).

- Step 2: Splitting T into a set of sentences \( S = \{s_1, s_2, ..., s_n\} \)

- Step 3: For each sentence \( s_i \) in the training set, calculate the probability of each feature \( F_j \), and then use Naïve Bayes classification to compute probabilities of \( s_i \) with both classes (+) and (-).

- Step 4. In the case of the probability of \( s_i \) with class (+) greater than the probability of \( s_i \) with class (-), \( s_i \) will be selected for summary.

- Figure 4 below describes the algorithm of the extraction phase.
EXTRACTION ALGORITHM (TSBN)

Input:
- \( C \): Original text;
- \( K \): Table of topic words and value of its;
- \( n \): Number of sentences that has labeled (+);
- \( n' \): Number of sentences that has labeled (-);

Output:
- \( T \): Text summary

Initialization
\( T = \phi \);
\( S = \text{Split}(C) \); // split sentences from \( C \)
\( F = \phi \); \( F' = \phi \); \( m = 0 \);

1. \textbf{For} each sentence \( s_i \) in \( C \) \textbf{do}
   1.1 \textbf{For} \( j = 1 \) to \( \text{length} \ s_i \) \textbf{do}
   1.1.1 \textbf{If} \( w(j) \in V \) \textbf{then}
   1.1.1.1 \text{match} \_\text{hush}(w(j), K) \) // Matching with \( K \)
   1.1.1.2 \( F(k) \leftarrow n(j) \) // Frequency \( w(j) \) that occur in labeled (+) sentences
   1.1.1.3 \( F'(k) \leftarrow n'(j) \) // Frequency \( w(j) \) that occur in labeled (+) sentences
   1.1.1.4 \( m = m + 1 \); // count topic word
   1.1.1.5 \( W(s_i) = W(s_i) + \log\frac{F(k)}{n} \); // calculate information significant in labeled (+) sentence.
   1.1.1.6 \( W'(s_i) = W'(s_i) + \log\frac{F'(k)}{n'} \); // calculate information significant in labeled (-) sentence.

1.2 \( W(s_i) = W(s_i) + \log\frac{n}{n+n'} + \log\frac{m}{n} + \log\frac{P_{\text{pos}}(s_i)}{n} \); // Probability of \( s_i \) with labeled (+) class
1.3 \( W'(s_i) = W'(s_i) + \log\frac{n}{n+n'} + \log\frac{m}{n} + \log\frac{P_{\text{pos}}(s_i)}{n'} \); // Probability of \( s_i \) with labeled (-) class

1.4 \textbf{If} \( W(s_i) > W'(s_i) \) \textbf{then}
1.4.1 \( T = T \cup W'(s_i) \);
1.5 \( m = 0 \);

Figure 4. Sentence extraction algorithm.
5 EXPERIMENTALS

5.1 Corpus

For experiments, we use Vietnamese text for experimental, with proposed algorithms, we also developed an automatic text summarization system to process text summarization effectively and convenient. During of Vietnamese text study, we have built a Vietnamese text corpus for summary experimental purposes. This corpus is done by selecting documents Vietnamese websites like: http://thongtincongnghe.com, http://echip.com, http://vnexpress.net, http://vietnamnet.vn, http://tin247.com, .... Currently, Corpus includes more than 300 Vietnamese text with 11,670 sentences.

Figure 5 below shows the documents stored .txt file in the database.

![Figure 5. Stored training set in database system.](image)

With each document in training set, human classify into two classes. There are 1,200 labeled (+) sentences. Figure 6. following described labeling process of the training phase. Sentences is assigned a value of 1 is equivalent to the label (+), the sentence is assigned a value of 0 is equivalent to the label (-).
5.2 Building topic words dictionary

When built topic words dictionary, we use Vietnamese text pos and tagging tool. Figure 13 illustrate a text after used pos tagging tool.

Next, we extracted from the above files (that has used pos tagging tool) all nouns (labeled "N"). All nouns will be stored in database. Figure 8 illustrated topic word dictionary and it updated score corresponding to each class (+) and class (-).
With the proposed method, we built a test system. The input text (original text) after preprocessing, saved as a .txt file in system. Figure 9. Is a original text.

5.3 Results and evaluation

With the proposed method, we built a test system. The input text (original text) after preprocessing, saved as a .txt file in system. Figure 9. Is a original text.

Trong những năm qua, botnet đã trở thành nguồn thu nhập ổn định của các nhóm tội phạm mạng bởi chi phí bỏ ra thấp trong khi việc kiếm soạt chứng ngang càng trở nên dễ dàng hơn. Maria Garnayera và Alexey Kadiaev, hai chuyên gia của Kaspersky Lab, đã đề cập những phát hiện của mình về mô hình hoạt động và kinh doanh của các botnet phổ biến trong bài viết “Botnet quảng cáo”, đăng trên website uy tín chuyên về bảo mật securelist.com.

Việc tạo ra lượng truy cập giả mang đến nguồn lợi nhuận thường xuyên cho botnet. Các nhà quảng cáo muốn giả tăng lượng người truy cập vào website của mình bằng cách đặt nhiều liên kết đến trang web hiện thị trên những website thuộc mạng lưới quảng cáo.

Hợp pháp trả tiền cho chủ sở hữu mạng lưới quảng cáo một mức phí cho mỗi lần có người dùng kết nối trực tiếp đến website của mình hoặc gián tiếp từ website của hệ thống quảng cáo. Tuy nhiên, mô hình kinh doanh này sẽ trở nên bất hợp pháp nếu người tham gia trong mạng lưới quảng cáo sử dụng một botnet để tạo lượng truy cập giả vi lợi nhuận.


Theo số liệu từ những tập tin đăng nhập, mã tính tại 235 quốc gia trên toàn cầu đã bị lấy nhiệm trojan này. Các bot Artro liên tục chuyển hướng để người dùng nhấp chuột vào liên kết dẫn đến website đang được quảng cáo, tạo ra lượng truy cập đáng kể.

Tội phạm mạng cũng sử dụng malware download từ bên thứ ba lấy nhiệm cho mã tính cá nhân dễ kiếm tiền. Số lượng mã tính bị lấy nhiệm và lượng người download khởi kiện soạt khién botnet CodecPack hay Artro trở thành công cụ hết sức nguy hiểm trong tay các nhóm tội phạm mạng.

Qua thống kê số lần nhập nhấp chất trúng bính một ngày mà mỗi module thực hiện trên các liên kết, mức thanh toán trung bình từ dịch vụ quảng cáo cho mỗi cú nhấp chuột, số lượng bot tội thiếu trên mạng ma, lợi ích khi bán lượng truy cập, hai chuyên gia Kaspersky Lab đã ước tính thu nhập của tội phạm mạng sở hữu botnet có thể đạt được từ 1.000 đến 2.000 USD một ngày.
Figure. 10 below is the result of the system. Original text has 13 sentences and text summary (target text) 6 sentences.

In the present time, there is no automatic Vietnamese text summarization evaluation system. Therefore, we use the precision measure and compared with some other methods to assess the effectiveness of the proposed method. The two main methods we use the baseline method, an automatic Vietnamese text summarization system that has address at http://labs.baomoi.com/demoTS.aspx (called VTSonline) and an other automatic text summarization at http://www.textcompactor.com/ (called Textcompactor).

Due an other methods often used summary rate, so that to perform this comparison, we implement a summary rate. Summary rate \( r \) is defined by the following formula:

\[
    r = \frac{\text{length of target text}}{\text{length of original text}} \times 100
\]

Table 1 below is a comparison of these methods.
Table 1. Results of experimental.

<table>
<thead>
<tr>
<th>Method</th>
<th>80%</th>
<th>60%</th>
<th>40%</th>
<th>20%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>0.88</td>
<td>0.86</td>
<td>0.82</td>
<td>0.6</td>
</tr>
<tr>
<td>HLT</td>
<td>0.82</td>
<td>0.75</td>
<td>0.69</td>
<td>0.54</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.81</td>
<td>0.8</td>
<td>0.84</td>
<td>0.63</td>
</tr>
<tr>
<td>Textcompactor</td>
<td>0.85</td>
<td>0.82</td>
<td>0.65</td>
<td>0.57</td>
</tr>
<tr>
<td>VTSonline</td>
<td>0.72</td>
<td>0.68</td>
<td>0.51</td>
<td>0.48</td>
</tr>
</tbody>
</table>

The results are expressed as the chart below (Figure 11)

![Figure 11. Comparison with some methods.](image)

6 CONCLUSION

The fundamental problem when processing single syllable language is word processing. When building automatic text summarization system, text classification, text retrieval, ... In general case, word segmentation tool is often embedded in the systems, so processing time for word processing is long special when work with more texts. Addition, segmentation tools have not high accuracy, and haven’t mentioned the noise processing problem in texts.

In this paper, we present a method using Naïve Bayesian and topic words set to solution syllable language text summarization problem. The experimental results show that the proposed method can solution some problems exist in single syllable language text, reduce processing time, computational complexity and quality of summaries is higher.

With this approach, we expect can be extended to cross language summary (cross language) for some single syllable languages in the future and build automatic cross language text effectively.
REFERENCES


Received : June 15, 2013