A Detection and Annotation System for Internet New Words in Taiwan

Chiung-Wei Huang
Department of Electronic Engineering, Ching Yun University
No. 229, Chien-Hsin Rd., Jhong-Li City 320, Tao-Yuan County, Taiwan

Abstract: - In this paper, a system for detecting and annotating Internet New Words, especially used by young men in Taiwan is proposed. At first, an Internet server for dealing with detection and annotation of New Words is constructed. Some of the information retrieval, database, and dynamic Web programming techniques, especially a detecting strategy for New Words are adopted and developed in the server for achieving the goal. At last, not only the overall performance of the system is good, but also the system does provide a nice tool for teachers or parents who want to know what their students or children are talking about and thinking about during the New Words are used in their speaking or writing.

Key-words:- Information Retrieval, Text Annotation, Dynaming Web programming, Internet New Words in Taiwan, Traditional Chinese characters, Phonetic code

1 Introduction

The rapid growth of Internet affects the life style of human beings and the degree of civilization. Some special methods, habits, and styles of communication spread very quickly. Besides, young people these days surfing on the Internet for a very long period of time. Many of them like to chat on the forum or BBS sites searching for the opportunities to make friends or exchange their new information or interesting stuffs. However, New Words are also generated and propagated among their fish-gathering activity [3]. In most cases, those terms came from TV commercials, well-known TV hosts, Pop songs, popular novels, movies or some news events. But there is a special type of New Words caused by the chatting process on the Internet. That is, for quick answering the problem to their friends during chatting, they just use some special abbreviations standing for the original characters or phrases (for example, using a single phonetic code to stand for a Chinese character in Taiwan). It is very informal and may cause misunderstandings. Besides, those New Words are spreading from young men’s talking or posting articles on the Internet to their daily speaking or writing. No matter you like that or not, they are the new ways of young men to express and communicate themselves. Fig. 1 lists the possible categories of New Words nowadays used by young men in Taiwan, i.e., there are Phonetic, English, Numerical, Semantic, and Homophonic types of New Words. The Phonetic type of New Words is in the case that using phonetic codes for short to represent formal Chinese characters in the talking or writing. The New Words of English type are some special abbreviations. The Numerical type of New Words are used from those days we use pagers. The Semantic type New Words are...
generated by the use of some TV or Radio show hosts, advertisements, novel, news, or politicians, etc.

At last, the Homophonic New Words are based on the same or similar pronunciation, e.g., “3Q” stands for “thank you” because the pronunciation of “3Q” in Chinese is quite similar to “thank you.” According to survey, many teachers and parents are not easy to follow the meaning of those New Words [6][7]. Thus, we are motivated to design a system to detect and annotate those articles to help people who want to understand those special New Words used by young men nowadays.

Accordingly, dealing with these kinds of New Words can be treated as the information retrieval of unknown terms. Previously researches have some related experiences on the processing of unknown term [4]. However, the characteristics of Internet New Words have different properties. For example, there are different types of characters need to be dealt with, e.g., phonetic codes, numeric codes, English abbreviations, and Chinese characters. Besides, there might be a hybrid usage of different New Words. Thus, we propose a new approach and establish a system by considering the characteristics of New Words to tackle the detection and annotation on these kinds of special information. Also, we announce the collected New Words on our Web system [2] and provide a service to annotate New Words of young men’s articles, such that viewers can easily understand what young men are talking or writing about.

2 System Architecture

Fig. 2 shows the functional scenario of the proposed system. Due to our goal is to establish a system for detecting and annotating young men’s articles, we have to enable the system with the capability on extracting Web pages of popular forums or articles from hot BBS sites. Then, some preprocessing steps, i.e., the remove of HTML tags and the segmentation of Chinese text, need to be taken for the major processing. Next, the most important part of the finding of New Words and annotation is conducted. Also, the New Words database are refered at the same time. Finally, the article will be annotated and output to the viewer.
In Fig. 3, the working flow of the proposed system is presented and in Fig. 4, the system architecture of the proposed approach is depicted. It has 6 major functional agents, i.e., Web Document Retrieval, HTML tag Remover, Text Segmentation, New Words Detection, New Words Annotation, and User Interface. The Web Document Retrieval agent extracts target Web pages from the Internet. The HTML tag Remover agent helps to remove the html tags for the extracted Web page. And Text Segmentation agent segment the target text into phrases or single character because there is no space among the writing of Chinese characters in sentences. The segmentation helps the processing of interested Chinese characters or the New Words. Besides, the agent of New Words Detection and New Words annotation try to find out those New Words and then annotate them to their original meaning for users.

![Diagram](image)

Fig. 4. System architecture of the proposed approach.

They are the major contributions of our approach. At last, the User Interface is the implementation of a user interface which aims at providing appropriate input methods for users to send their requirments and data to the system and then get their desired annotation outputs. In what following, we further discuss the major functional agents except the User Interface.

- **Web Document Retrieval**
  For retrieving young men’s articles from the Internet, an agent named, Web Document Retrieval, conducts the extraction of Web documents. It is implemented by the Perl module.

- **HTML Tag Remover**
  For removing the HTML tags of Web pages, the HTML tags remover is implemented by calling regular expression function of Perl to fulfill the goal.

- **Text Segmentation**
  Due to it doesn’t have spaces among characters in Chinese text, we have to segment the text into words or characters for later detection and annotation processing. The Academia Sinica of Taiwan has announced a well-known service on text segmentation called CKIP [5]. Thus, we invoke the text segmentation service of CKIP in our Text Segmentation agent to accomplish mission of the text segmentation.

- **New Words Detection**
  Before the discussion of detection, let’s take a view on each type of New Words. In what following, we use the BNF to describe the elements and rules of each type of New Words.

**Basic token:**

- `<digit>` := 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9
- `<phonetic code>` := `ㄆ` `ㄈ` `ㄇ` `ㄈ` `ㄎ` `ㄆ` `ㄗ` `ㄠ` `ㄠ` `ㄢ` `ㄤ` `ㄦ` `ㄦ` `ㄠ` `ㄠ` `ㄠ` `ㄠ`
- `<Chinese character>` := 甲 | 乙 | 丙 | 丁 | 戊 | 己 | 庚 | 辛 | 壬 | 癸 | 甲 | 乙 | 丙 | 丁 | 戊 | 己 | 庚 | 辛 | 壬 | 癸

**Compound token:**

- `<Numeral>` := `{<digit>}`
- `<Phonetic>` := `{<phonetic code>}`
- `<Semantic>` := `{<Chinese character>}`
- `<English>` := `{<alphabet>}`
- `<Homophonic>` := `{<Chinese character>}`

At the first stage of New Words detection, there is a word-type recognizer which distinguishes the characters into numbers (Numeral type), phonetic codes (Phonetic type), English characters (English type), or Chinese characters (Semantic or Homophonic type). After that, the agent invokes dedicate pattern matching routine for each type of New Words. Those agents that deal with the categories of New Words of Numeral, Phonetic, and English are quite simple. They follow the recognition on the rules of compound tokens. But the agent of coping
with Chinese characters has to refer to the New Words database to further distinguish those consecutive characters are of Semantic or Homophonic type.

**New Words Annotation**

After the New Words have been recognized, the annotation agent queries the database of New Words and retrieves the real meaning. Then, annotate and insert the possible meaning after the New Words in the position of the original text.

After the description of our approach, the experimental design will be detailed in Section 3.

### 3 Experiments

In order to verify the performance of our proposed method on the extraction and annotation of New Words, we established an experimental Server and implemented all the components described previously. The hardware platform of the Server is a PC with the Intel Celeron 1.1GHz CPU, 256MB RAM, Ethernet Network Interface Card, and 40GB HD. The software platform is established with Linux FC3 OS, MySQL (database), PHP (dynamic Web programming environment), Perl (programming language), and Apache HTTP Server. Fig. 5 shows the hardware and software environment of the System. During test, some popular BBS or Forums are chosen as the article source. For example, the OpenFind Forum [8], the BBS of the National Tao-Yuan Agricultural and Industrial Vocational High School [9], and the BBS of the Ying-Ge Vocational High School [10]. For evaluate the performance of our approach, we check by using the well-known measure, the recall rate [1], which is defined as equation (1).

\[
\text{Recall} = \frac{\text{Number of Retrieval and Relevant New Words}}{\text{Number of Total Relevant New Words}} \tag{1}
\]

The higher recall rate indicates the approach retrieves more relevant New Words. Table 1 shows the occurrence rate of New Words used in the articles posted on the BBS of Ying-Ge Vocational High School (Jun. 2005). From the result, it seems the phonetic type of New Words occurs most frequently in our survey. That means young men prefer to use phonetic codes in their chatting on the internet.

![Fig. 5. The hardware and software environment of the proposed System.](image)

Table 1. The occurrence rate of New Words used in the articles posted on the BBS of Ying-Ge Vocational High School (Jun. 2005).

<table>
<thead>
<tr>
<th>Term type</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>5.00%</td>
</tr>
<tr>
<td>Homophonic</td>
<td>16.67%</td>
</tr>
<tr>
<td>Semantic</td>
<td>11.67%</td>
</tr>
<tr>
<td>Numeral</td>
<td>3.33%</td>
</tr>
<tr>
<td>Phonetic</td>
<td>63.33%</td>
</tr>
<tr>
<td>Total</td>
<td>100.00%</td>
</tr>
</tbody>
</table>

![Fig. 6. The annotation output. The top text area shows the original text. The bottom text area shows the annotation output. The colored-text in blue are the candidates of New Words, and the text in red color represent the possible meaning on New Words.](image)

Due to our system is implemented to deal with the New Words used in Taiwan, the snapshot of result Web pages are showing in Chinese with Traditional Chinese characters. Fig. 6 demonstrates the annotation output. The colored-text in blue are the candidates of New Words, and the text in red color represent the possible meaning on New Words. Table 2 shows the analysis results of New Words in some articles on the OpenFind forums [Jun. 2005]. Table 3 indicates the analysis results of New Words in some articles on the BBS of the National Tao-Yuan Agricultural and Industrial Vocational High School [TYAV for short, Jun. 2005]. In the results, most of the New Words can be identified. But some articles got lower recall

![Table 2 and Table 3 showing the analysis results of New Words](image)
It is because there still some New Words used by young men, and are not collected and reported by media. So we didn’t collect them in our New Words database in advance. However, after the testing proceeds, we may accumulate more New Words in our database.

Table 2. The analysis results of New Words in some articles on OpenFind forums [Jun. 2005]

<table>
<thead>
<tr>
<th>Article title</th>
<th>No. of articles</th>
<th>Recall Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Do you think that using left feet to brake a car is so strange?</td>
<td>76</td>
<td>81.3%</td>
</tr>
<tr>
<td>Why Toyota’s car is the best selling in Taiwan?</td>
<td>54</td>
<td>81.8%</td>
</tr>
<tr>
<td>Will you go to see the movie “頭文字 d” (a movie title)?</td>
<td>154</td>
<td>50.0%</td>
</tr>
<tr>
<td>Do you like to drive fast?</td>
<td>188</td>
<td>68.9%</td>
</tr>
</tbody>
</table>

Table 3. The analysis results of New Words in some articles on BBS of the National Tao-Yuan Agricultural and Industrial Vocational High School (TYAV) [Jun. 2005]

<table>
<thead>
<tr>
<th>Board title</th>
<th>No. of articles</th>
<th>Recall Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computer game</td>
<td>154</td>
<td>54.5%</td>
</tr>
<tr>
<td>Online game</td>
<td>132</td>
<td>42.9%</td>
</tr>
<tr>
<td>Misc of TYAV</td>
<td>169</td>
<td>72.7%</td>
</tr>
<tr>
<td>Talking about love</td>
<td>187</td>
<td>58.8%</td>
</tr>
<tr>
<td>Shoot the air</td>
<td>146</td>
<td>61.9%</td>
</tr>
</tbody>
</table>

4 Discussion and Conclusion

At last, we may conclude that we have proposed a system for detecting and annotating Internet New Words used by new generation young men in Taiwan. The New Words for detection can be special numbers (Numeral type), English abbreviations (English type), phonetic codes (Phonetic type), or Chinese characters (Semantic or Homophonic type). Also, the performance can be improved when the system tested more Internet articles because more New Words can be accumulated in the database. In what following, we list the contributions of this paper:

1. Propose a heuristic approach to detect and annotate several types of Internet New Words in Taiwan.
2. Release the New Words collected and accumulated by our system.
3. Provide the annotation service to the public, especially for teachers or parents who want to realize what their students or their children are talking or writing about.
4. Provide a translation service for advertising companies to translate their slogans into young men’s favorite style for better promotion.
5. Enable the cell phone portals to provide related services which translate formal text into young men’s favorite style before sending their SMS (Short Message Services).

Acknowledgements

This work is financially supported by the National Science Council of Taiwan under the grant: NSC93-2218-E-231-006. Also, the author would like to thank Mr. C. H. Lee for his effort on the implementation of the proposed approach. Besides, special thanks are sent to Prof. T. Y. Tang who gave the author many helpful comments on the application of the system.

References