Computer-Added Extraction of Chinese New Vocabulary for Language Teaching

Yuen-Hsien Tseng and Zih-Ping Ho
Information Technology Center, National Taiwan Normal University
No.162, Sec. 1, Heping East Road, Taipei City, Republic of China

Abstract - In teaching Chinese, there is a need to collect emerging vocabularies popular in the local Chinese society for updating the teaching materials; this need is particularly a true issue when the teachers are non domestic persons of the society where the Chinese language is in daily use. Non-domestic teachers have less contact with the everyday language used in the domestic society and they would need such information about new terms. In this paper, we propose an approach to extract the candidate new terms from a large and long-lasting stream of news articles. Based on some text mining techniques such as key term extraction, term association analysis, and hot topic detection, we are able to extract candidate vocabularies in support of Chinese language teaching.

Keywords: Computer-Added Education, Chinese New Vocabulary Detection, Information Extraction, Language Teaching, Terms Mining.

1 Introduction

In teaching Chinese, there is a need to collect emerging vocabularies popular in the local Chinese society for updating the teaching materials or for timely use in classroom. This need seems not to be an issue at first sight since Chinese teachers are assumed to know which terms are fashionable in Chinese society’s daily life. Nevertheless, even the knowledge of the Chinese teachers needs to be updated from time to time, just as elder citizens may occasionally find new terms or new usage of terms from their youngsters, such as “火星文” (“Mars language” translated literally). This new term does not really mean a signed language or a set of unusual sequences of characters such that a normal terrestrial would not understand. It is used as metaphor to denote a set of odd vocabulary commonly seen in the Internet chat rooms, discussion lists, messengers, or blogs used by youngsters. Terms like this have become more and more common in the local society and those who do not know its metaphor may not well integrate into the discourse of the local speakers. The need for obtaining new terms for Chinese teaching or learning is particularly a true issue when the teachers are outsiders of the society where the Chinese language is in daily use. Non-domestic teachers have less contact with the everyday language used in the domestic society and they would need such information about new terms. Even if a Chinese teacher has knowledge of them, an evidence-based statistics would help to include those new terms in the updated teaching materials.

To collect those terms to narrow the gap between the language been currently taught and the language been daily used, a technology-enabled mechanism to digest everyday language usage and to report on possible new terms would be of great help. One possible source for the everyday language usage is from news. Daily news reports on various topics concerning the life of a society. Therefore news articles represent a feasible source for finding the candidate terms. However, simple key term extraction is not satisfiable as technology buzzwords or name entities of particular events may often occur. Therefore, it is not trivial to extract those new terms that are suitable for Chinese teaching as a second language. The computer-added preparation of Chinese new vocabulary has thus become a topic worth of study.

In this paper, we propose an approach to sieve out the candidate new terms from a large and long-lasting stream of news. This approach is based on the previously developed techniques for keyword extraction, term association analysis, and hot topic detection [1, 2]. We apply them in a novel way to hopefully meet the need as mentioned above.

2 Literature review

There were few literatures directly related to computer-added selection of Chinese new vocabulary for textbook compilation. We briefly review some of related studies. The previous literature would be discussed in (1) IT for Chinese vocabulary learning; (2) techniques for detecting new Chinese vocabulary.

2.1 IT for Chinese vocabulary learning

Although computer added education were discussed for several decades, there were still few papers about Chinese vocabulary compilation. Blake [3] described a way of distance learning for second language (SL) education and showed that it was effective because SL students could involve in multi-culture environments. Levy [4] stated SL learning should emphasize on grammar, vocabulary, reading, writing, pronunciation, listening, speaking, and culture and IT could play more and more important roles on these topics. In Levy’s study, vocabulary learning software made hyperlinking keywords simple and that would help students recall what the words really meanings.
2.2 Techniques for detecting new Chinese vocabulary

New vocabulary detecting is a pretty new topic in natural language processing (NLP). Until now, new word detecting in computational linguistics pay more attentions to new word collection in an updated dictionary, e.g. Ha et al. [5], Li et al. [6] and Jiang et al. [7].

Kuba et al. [8] stated that they selected and re-evaluated on a manually annotated corpus containing 1.2 million words, and six topics: fiction, school compositions, newspaper articles, computer-related texts, law and short business news. The methods were most common POS tagging approaches, building a tagger based on Hidden Markov Models (HMM). After they testified several different POS tagging methods, they found large corpora would improve the tagging quality; and also found combined methods outperformed using a single POS tagger. Ha et al. [5] presented a SVM (support vector machine) based method that predicted the unknown words, established a two-phase unknown word prediction procedures. The experiment results showed that the experiments were very promising by showing high precision and high recall while high speed. Li et al. [6] defined NWI as a binary classification problem. A SVM trained data were to identify the new words. Pakhomov et al. [9] trained corpus of clinical notes to perform manual annotation. The corpora were used for training POS taggers. If doctors had these POS taggers, the new terms and correct POS taggers of clinic notes would improve the clinic writing.

Brown et al. [10] gave a new idea for automatic measurement the density of POS tagging. In their papers, The Computerized Propositional Idea Density Rater (CPIDER, pronounced “spider”) was proposed. It was a computer program that determined the propositional idea density (P-density) of an English text automatically on the basis of POS tags. The main idea was propositions corresponding roughly to verbs, adjectives, adverbs, prepositions, and conjunctions. The CPIDER counted correlated very closely with the human results. The potential application maybe applied this method into readability and reading comprehension. Hong et al. [11] proposed statistics-based scheme for extraction of new words based on the categorized corpora of Google News retrieved automatically from the Google News site.

Sun et al. [12] stated that the procedure of new words identification and POS tagging were usually separated and the features of lexical information cannot be fully used. They used Latent Dynamic Conditional Random Field and Semi-CRF model to detect new words. The results showed that the proposed method was capable of detecting even low frequency new words together with POS tags. Jiang et al. [7] proposed The Potential Unknown word Detection (PUD) Algorithm was used to detect longest potential unknown words for the single-character model and potential unknown words for the affix model. Their algorithm was statistical based words frequency prediction approaches.

Although numerous literatures were related to automatic detecting new vocabulary, there was only a limited amount of these studies available for Chinese New vocabulary language teaching.

3 Fundamental of Term Networks

Our proposed approach relies on the concept of term co-occurrence or term network. Traditional term co-occurrence (or co-term) network is built by representing terms in vector form which denotes the occurrence of the terms in all documents. To allow more in-depth exploration, we believe that a sentence-level co-term network would be more suitable because terms co-occur in the same sentences tends to exhibit clearer relationship than those in the same documents only.

The seed terms for the network can be obtained from a given list, output of an NLP parser, or any key term extraction algorithms [13]. Term similarities are then calculated based on these vectors to form the similarity matrix of terms [14], which is the input to the co-term network. If there are n terms from m documents, the time complexity can be on the order of $O(n^2m)$, where m steps are required to calculate similarity between any of n term pairs.

The major difference of ours from the above is to limit the terms to be associated to those that co-occur in the same logical segments of a smaller text size, such as a sentence. Association weights are computed in this way for each document and then accumulated over all documents. This changes it into a roughly $O(mnk^3s)$ algorithm, where k is the average number of selected key terms per document and s is the average number of sentences in a document. As can be seen, the larger the n and m, the bigger the difference between $O(mnk^3s)$ and $O(n^2m)$, because k can be kept less than a constant and so can s by breaking large documents into smaller ones.

The technique for term co-occurrence analysis to be used in our work is described as follows: key terms identified from each document based on Tseng’s algorithm [2] are first sorted in decreasing order of their term frequencies (TF), or $TF \times Term\_Length$, or other criterion such as $TF \times IDF$ (Inverse Document Frequency) if the entire collection statistics are known in advance. Then the first k terms are selected for association analysis. A modified Dice coefficient was chosen to measure the association weights as in (1).

$$\text{wgt}(T_i, T_k) = \frac{2 \times S(T_i \cap T_k)}{S(T_i) + S(T_k)} \times \ln(1.72 + S_i)$$ (1)

where $S_i$ denotes the number of sentences in document i and $S(T_i)$ denotes in document i the number of sentences in which term $T_i$ occurs. Thus the first term in Equation (1) is simply the Dice coefficient similarity. The second term $ln(1.72 + S_i)$, where ln is the natural logarithm, is used to compensate for the weights of those terms in longer documents so that weights in documents of different length have similar range of values. This is because longer documents tend to yield weaker Dice coefficients than those generated from the shorter ones. Association weights larger than a threshold (1.0 in our
implementaiton) are then accumulated over all the documents in the following manner as in (2).

\[
sim(T_j, T_k) = \frac{\log(w_k \times n/df_k)}{\log(n)} \times \sum_{i=1}^{n} \text{wgt}(T_{ij}, T_{ik})
\]

where \(df_k\) is the document frequency of term \(k\) and \(w_k\) is the width of \(k\) (i.e., number of constituent words). Variations of the method for dealing with other document types can be found in [15, 16].

4 Research Method

The above term co-occurrence analysis has been applied to Tseng et al [2] to extract junior high school textbook terms from news reports for measuring civic scientific literacy. This work extended their approaches by replacing the textbook terms with those basic contextual terms from a teaching material developed by Hsin et al. [17]. The flow chart of finding the candidate terms is shown in Figure 1. From about 2,900,000 articles of 10 years daily news, we applied the key terms extraction and association analysis techniques to yield 670,000 key terms, each with some related terms. A key term is a frequent topical term in a news article and its related terms are other key terms that frequently co-occurred with the key term in the same sentences. For example, “霸凌” (bullying) is a keyword in a certain article and “學校” (school), and “老師” (teacher) may be its related terms, because bullying often occurs in elementary schools and junior high schools and teachers are often involved in dealing with bullying events. To detect those key terms that are suitable for Chinese teaching, we then match the related terms of the key terms with the basic contextual terms from the teaching material developed by Hsin et al. [17].

The basic contextual terms were compiled to help foreigners to learn basic Chinese terms, such as “school”, “teacher” categorized in its education category. There are twelve categories in this teaching material including personal information, living, occupation, leisure, travels, social relationship, medical cures, educations, shopping, food and bakery, banking, and safety categories. In total, there are about 1100 basic contextual terms and the number of key terms that have matched related terms are about 94,000. As an example, “bullying” is one the 94,000 key terms because two of its related terms “school” and “teacher” are in the set of about 1100 basic contextual terms. To know the trend of each key term over the 10 years of daily news articles, we computed the number of articles the key term occurs (i.e., document frequency) for each year, which constitutes a time series of 10 document frequencies for the key term. We then applied the following trend slope formula, verified by Tseng et al (2009), to know their trend. The key terms were finally sorted based on their trend slope for ease in new term browsing and selection. Tseng et al (2009) proposed to determine whether a topic is a hot topic or not based on the formulas as follows:

\[
\begin{align*}
  x_i &= \frac{1}{n} \sum_{i=1}^{n} i \\
  y_i &= \frac{1}{n} \sum_{i=1}^{n} d_i \\
  slp &= \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}
\end{align*}
\]

In the above formula (3), the more the documents in which a key term occurs in recent years, the higher the score of slp. It is noted in Tseng et al (2009) that it is the rank of the slope that matters in determining the novelty of the terms, rather than the absolute slope values of the terms.

5 Research Results

We not only implemented the above approach, but also built a search system for ease of verification. Figure 2 shows an example of a candidate term. A user inputs a query item, such as “霸凌” (bullying), then the system responds with (1) short snippets of 3 of 195 search results in news database (2) the related terms of the query term “bullying” to know its category context (3) a time series (trend) of the query item. It can be seen that bullying started in 2004 and grew into a salient term year after year. Ten more new terms similar to the above for possible Chinese language teaching were listed.
6 Conclusions

New vocabularies extraction for compilation of vocabulary textbooks are useful to close the gap between the language been taught and the language been used. Thus a software tool which helps collect new daily use Chinese vocabularies automatically is desirable. However, such a technical tool is not trivial to develop. Based on some text mining techniques such as key term extraction, term association analysis, and hot topic detection, we are able to extract candidate vocabularies in support for Chinese language teaching; life style vocabularies could be included into the teaching materials, and students could more easily join to Chinese culture society. The results were welcomed by Chinese teaching professionals, showing the potential usefulness of this work. Future work will collect more corpora and examples for further applications.

7 Acknowledgment

The work is supported in part by the “Aim for the Top University Plan” of National Taiwan Normal University, sponsored by Ministry of Education, Taiwan, R.O.C.

8 References

Figure 2. Example of a search term for verification.

Table 1. The demonstration of trends of selected vocabularies from 2000 to 2009.

<table>
<thead>
<tr>
<th>Term</th>
<th>2000</th>
<th>2001</th>
<th>2002</th>
<th>2003</th>
<th>2004</th>
<th>2005</th>
<th>2006</th>
<th>2007</th>
<th>2008</th>
<th>2009</th>
<th>slope</th>
</tr>
</thead>
<tbody>
<tr>
<td>A: 金融海嘯</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1810</td>
<td>6321</td>
<td>421</td>
</tr>
<tr>
<td>B: 黑心商品</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>21</td>
<td>44</td>
<td>24</td>
<td>28</td>
<td>102</td>
<td>2313</td>
<td>1775</td>
<td>190</td>
</tr>
<tr>
<td>C: 減碳</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>106</td>
<td>2131</td>
<td>3671</td>
<td>1051</td>
<td>8</td>
</tr>
<tr>
<td>D: 山寨</td>
<td>9</td>
<td>8</td>
<td>14</td>
<td>11</td>
<td>6</td>
<td>9</td>
<td>8</td>
<td>4</td>
<td>43</td>
<td>704</td>
<td>40</td>
</tr>
<tr>
<td>E: 產銷履歷</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>9</td>
<td>20</td>
<td>26</td>
<td>10</td>
<td>9</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>F: 部落格</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>15</td>
<td>312</td>
<td>1188</td>
<td>1370</td>
<td>1476</td>
<td>1541</td>
<td>212</td>
</tr>
<tr>
<td>G: 火星文</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>22</td>
<td>193</td>
<td>44</td>
<td>28</td>
<td>34</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>H: 粉絲</td>
<td>55</td>
<td>69</td>
<td>62</td>
<td>96</td>
<td>721</td>
<td>1811</td>
<td>1866</td>
<td>1262</td>
<td>1418</td>
<td>1172</td>
<td>193</td>
</tr>
<tr>
<td>I: 台客</td>
<td>47</td>
<td>70</td>
<td>67</td>
<td>73</td>
<td>67</td>
<td>204</td>
<td>152</td>
<td>174</td>
<td>143</td>
<td>254</td>
<td>20</td>
</tr>
<tr>
<td>J: 詐騙集團</td>
<td>356</td>
<td>389</td>
<td>424</td>
<td>706</td>
<td>954</td>
<td>1020</td>
<td>946</td>
<td>807</td>
<td>702</td>
<td>490</td>
<td>37</td>
</tr>
</tbody>
</table>

Notes: Term A means global financial crisis (belongs to the banking category); B means black heart of goods (shopping); C means carbon reduction (living); D means cottage (shopping); E means traceability resume (food and bakery); F means blogs (social relationship); G means odd wording (social relationship); H means fans (social relationship); I means Taiwanese in non-native countries (social relationship); J means fraud groups (safety).