Applying A Lightweight Chinese Lexical Chain Processing In Web Image Annotation

Chuen-Min Huang
Department of Information Management
National Yunlin University of Science & Technology, Taiwan, R.O.C.
huangcm@yuntech.edu.tw
Tel: +886-9-963119624

Abstract - Traditional CBIR method relies on visual features to identify objects in an image and uses predefined terms to annotate images, thus it fails to depict the implicit meanings. Recent textual-based analysis methods only focus on single term processing thereafter they suffered the disadvantages of fragmented description of the annotation. In this research, we propose a corpus-free, relatively light computation of term segmentation method, namely ‘Chinese Lexical Chain Processing (CLCP),’ to identify compounds from a single web page to obtain anecdotes as a semantic enrichment of the target image. It requires a minimum computation need that allows sharing characters/words and facilitating their use at fine granularities without prohibitive cost. In the experiment, this method achieves a precision rate of 90.04%, and gains acceptance from expert rating and user rating of 86% and 73.7%, respectively. In performance testing, it only takes 0.007 second to process each image in a collection of 1,728 testing data set.

Keywords: Automatic Image Annotation, Chinese Lexical Chain Processing, N-Gram, Lexical Chain

1 Introduction

With the rapid development of image technology and digital devices, more and more web images are stored and displayed on the Internet. As a result, the way to process these images to be retrieved efficiently has become a crucial issue. Content-Based Image Retrieval (CBIR) with its focus on rapid application of voluminous low-level visual features such as color, texture, shape, etc. gains popularity to support efficient searching and browsing images. Due to the visual features explain less semantics, and the annotation relies on limited predefined terms, thereby the results usually are not satisfactory.

A glimpse of related studies would reveal that a couple of supervised learning approach and clustering techniques have been applied to image annotation including Support Vector Machine (SVM) [1, 2], Bayesian [3] and Self-Organizing Feature Map (SOM) [4, 5]. In addition, various text processing techniques that support content identification to analyze textual content based on word co-occurrence, location, and lexical-chained concepts have been elaborated in [6-8]. However, the aforementioned techniques suffer the same disadvantages of heavy computation for multi-document processing, which will consume lots of memory and may incur run-time overhead. In this study, we propose a corpus-free, relatively light computation of term segmentation for single document processing, namely ‘Chinese Lexical Chain Processing (CLCP) method.’ The CLCP method is to identify representative strings from a string in a single document with minimum computation needs that allows sharing characters/words and facilitating their use at fine granularities without prohibitive cost. The results demonstrated that applying our method in a single document is enough to generate content descriptor for image annotation. The precision rate achieves 90.04%, and the acceptance from expert and user rating reach 86% and 73.7%, respectively. The performance testing is also very promising. The remainder of the paper is organized as follows. Section 2 presents the related work including automatic image annotation, keyword extraction, and lexical chain. Then, we address the intent of our experiment in Section 3. The experimental result and evaluation method are described in Section 4. Finally, in Section 5, we draw conclusions and suggest future work.

2 Literature Background

2.1 Automatic Image Annotation

There have been a number of models applied for image annotation. In general, image annotation can be categorized into three types: retrieval-based, classification-based, and probabilistic-based [6]. The basic notion behind retrieval-based annotation is that semantic-relevant images are composed of similar visual features. CBIR have been proposed in 1992 [9]. Since then, more and more studies annotated the images based on this method [10]. CBIR is applied by the use of images features, including shape, color and texture. However, this method is limited by the training data set and the hidden semantic or abstract concepts can’t be extracted because the keywords are confined to pre-defined terms. Consequently, the results of CBIR are usually not satisfactory. The second type, also known as the supervised learning approach, treats annotation as classification using multiple classifiers. The images are classified based on the
extracted features. This method processes each semantic concept as an independent class, and assigns each concept as one classifier. Bayesian [3] and SVM [11] are the most often used approaches. The third type is constructed by estimating the correlations between images and concepts with a particular emphasis on the term-term relationship and intends to solve the problem of ‘synonym’ and ‘homograph.’ Frequent used approaches include co-occurrence model [12], LSA [5], PLSA [13] and HMM [14]. Notwithstanding the efforts made on the enhancement of annotation quality, the aforementioned approaches only focused on single term processing thereafter they suffered the disadvantages of fragmented description of annotation.

2.2 Keyword Extraction

In the field of Information Retrieval, keyword extraction plays a key role in summarization, text clustering/classification, and so on. It aims at extracting keywords that represents the text theme. One of the most prominent problems in processing Chinese texts is the identification of valid words in a sentence, since there are no delimiters to separate words from characters in a sentence. Therefore, identifying words/phrases is difficult because of segmentation ambiguities and the occurrences of newly formed words. In general, Chinese texts can be parsed using dictionary lookup, statistical or hybrid approaches [15].

The dictionary lookup approach identifies keywords of a string by mapping well-established corpus. For example, the Chinese string ‘蔡英文北監探扁會面一小時’ (Ying-wen Tsai went to Taipei prison to visit Shui-bian Chen and talk for an hour) will be parsed as: ‘[蔡英文],’ [北監], ’ [探], ’ [扁], ’ [會面], and ’ [一小時]’ by a well-known dictionary-based CKIP segmentation system in Taiwan. This method is very efficient while it fails to identify newly formed or out-of-the-vocabulary words and it is also blamed for the triviality of the list of the extracted words.

The statistical technique extracts elements by using n-gram (bi-gram, tri-gram, … etc.) computation from the input string. This method relies on the frequency of each n-gram and a threshold to determine the validity of each word. The above string through n-gram segmentation will produce: ‘[蔡英文], [英文], [北監], [監探], [探扁], [扁會面], [會面], [面一], [一小], [小時]; [蔡英文], [英文北], … , [一小時]’ and so on. The application of this method has the benefit of corpus-free and the capability of extracting newly formed or out-of-the-vocabulary words while at the expense of huge computations and the follow-up filtering efforts.

Recently, a number of studies proposed substring [9], significant estimation [16], and relational normalization [17, 18] to identify words based on statistical calculations. The hybrid method conducts dictionary mapping to process the major task of word extraction and handle the leftovers through n-gram computation, which significantly reduces the amount of terms under processing and takes care both the quality of term segmentation and the identification of unknown words. It has gained popularity and adopted by many researchers [19, 20]. Since the most important task of annotation is to identify the most informative parts of a text comparatively with the rest. Consequently, a good text segmentation shall help in this identification.

In the IR theory, the representation of documents is based on the Vector Space Model [21]: a document is a vector of weighted words belonging to a vocabulary V: 

\[ d = \{ w_1, \ldots, w_n \} \]

Each weight \( w_i \) is such that \( 0 \leq w_i \leq 1 \) and represents how much the term \( t_i \) contributes to the semantics of the document \( d \). In the term frequency-inverse document frequency (tf-idf) model, the weight is typically proportional to the term frequency and inversely proportional to the frequency and length of the documents containing the term. The term discrimination value can be used to compute a weight for each word in each document of a collection by combining the term frequency factor with the discrimination value. Some studies [22] [23] proposed methods to assign different weights to words by location.

2.3 Lexical Chain

A lexical chain (LC) is a sequence of words, which is independent of the grammatical structure of the text. Lexical cohesion can be interpreted as the state of cohering for making the sentences of a text, indicated by the use of semantically related vocabulary. Lexical chains (LCs) are sequences of words which are in lexical cohesion relations with each other and they tend to indicate portions of the context that form semantic units; they could serve further as a basis for a segmentation [24]. This method is usually applied in a summarization generation [25]. For instance, the string ‘向量空間模型’ (Vector space model) may be parsed as ‘ [向量], ’ ‘ [空間],’ and ‘ [模型]’ if there is no further merging process undergoing. Thereby the most informative compound ‘向量空間模型’ will be left out. Usually LCs are constructed in a bottom-up manner by taking each candidate word of a text, and finding an appropriate semantic relation offered by a thesaurus. Instead, the paper [26] proposes a top-down approach of linear text segmentation based on lexical cohesion of a text. Some scholars suggested to use machine learning approaches to create a set of rules to calculate the rate of forming a new word by characters for entity recognition including the maximum entropy model (MEM) and hidden Markov model (HMM) and claimed this method was able to achieve reasonable performance with minimal training data [27, 28].

3 Research Design

3.1 Research Model Overview

This study covers three tasks: text processing, image annotation, and image evaluation. The framework of our research is depicted as Figure 1. We used the news title and the text as input data; the image captions as ground truth labels. Then, we conducted CLCP and term weighting for the input data. After that, we generated a list of three
representative words for each image. From the list, we assigned the word with the highest weight as the primary annotation and the rest as secondary annotations. Finally, we evaluated the primary and secondary annotations by using the image caption and human judgment, respectively. In the following section, we will introduce the process of CLCP and the way of term weighting and the word annotation.

\section*{3.2 CLCP}

The fundamental idea of building CLCP was a bottom-up concatenating process based on the significance degree of distribution rate to extract the most meaningful LCs in a string. We treated a news document as a string composed of a series of characters and punctuations. Since there is no delimiter to separate words from characters except for the usage of quotation marks in special occasion, this concatenating process is a challenge without the aid of dictionary.

Most of the traditional studies identify words from the whole context and store all the keywords for further processing as Figure 2. In this way, even a moderate-sized document may require hundreds of thousands of characters, which will consume lots of memory and may incur unacceptable run-time overhead. Due to the number of distinct characters processed is less than that in the document, we adopted a sharing concept to allow reuse of the identical characters. We considered each character as a basic unit from which to build compounds as a composite, which in turn can be grouped to form larger compounds. Since the character and compound will be treated uniformly, it makes the application simple.

By doing so, we adopted the concept of flyweight and composite design patterns proposed by the GOF [29] to implement this design. Figure 3 shows the flyweight as a shared object that can be used in the whole context simultaneously. Figure 4 represents the part-whole hierarchy of texts and the way to use recursive composition. By applying flyweight design pattern, it supports the use of large numbers of fine-grained objects efficiently. By applying composite design pattern, it makes the application easier to add new components. The CLCP steps are depicted as

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure1.png}
\caption{Research model}
\end{figure}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure2.png}
\caption{Traditional document processing}
\end{figure}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure3.png}
\caption{Flyweight design concept}
\end{figure}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure4.png}
\caption{Composite text structure}
\end{figure}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure5.png}
\caption{The detailed description with an example is addressed next.}
\end{figure}

\subsection*{3.2.1 Step 1: Build a directed graph}

A directed graph (or digraph) is a set of nodes connected by edges, where the edges have a direction associated with them. For example, an arc (x, y) is considered to be directed from x to y, and the arc (y, x) is the inverted digraph. Y is the head and x is the tail of the link; y is a direct successor of x, and x is a direct predecessor of y.

We use the string: ‘蔡反擊：宇昌案不是弊案，怎麼一直問，宇昌案國發基金為何一再放棄權利⋯’ (Tsai fired back and stated that Yu Chang case is not a scandal, why did you keep asking for this and why the National Development Fund abandoned its rights repeatedly?) as an example to explain the construction process. After removing duplicate characters and replacing punctuations with new lines, a digraph is built. Figure 6 presents a fraction of the graph in which a solid line indicates the directed link and a dash line means the inverted link.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure6.png}
\caption{A fraction of the graph}
\end{figure}
3.2.2 Step 2: Calculate ADR and concatenate vertices

This step is to calculate the intensity and the degree of a digraph. The intensity is the average distribution rate (ADR); the degree means the number of incident edges. To determine whether two vertices can be concatenated, we applied the criteria listed in (1), (2).

\[ D(i, j) \geq T \cap \text{Digraph}(\text{Node}_i) > 1 \]  \hspace{1cm} (1)

\[ D(p, q) \geq T \cap \text{InvD}(\text{Node}_p) > 1 \]  \hspace{1cm} (2)

Where \( D(i, j) \) and \( D(p, q) \) represent the ADR of the arcs \( i, j \) and \( p, q \); \( \text{Digraph}(\text{Node}_i) \) and \( \text{InvD}(\text{Node}_p) \) indicate the number of directed links for vertex \( i \) and inverted links for vertex \( p \), respectively. \( T \) is the threshold value determined by 10 runs of experiments with the value of 0.1–0.9 assigned to 100 documents and the result was verified by 5 subject specialists with respect to word quality. The result showed that 0.4 outperforms the others as Figure 7, thus we used it as the threshold value. Based on the term frequency theory, it infers that the significance of the concatenation will be proportional to its frequency. Since this study was to generate content descriptor from a single document, the frequency was set as reasonable as possible.

Figure 8 shows the arc [字, 宇] with the expression [1,2] indicating that the intensity is 1 and the degree is 2, therefore it will be concatenated as [字, 宇] because it meets the criteria. In our previous [30], we focused mainly on directed links to extract LCs and failed in identifying some significant concatenations when a vertex has many links which diluted
the intensity of the distribution rate. This study was intended to remedy this problem by processing the directed link and the inverted link in parallel. In Figure 9, the LC \[ \text{[毕业]} \] with the value \([0.33, 2]\) of directed link \([毕业, 毕业]\) will be left out if we didn’t consider its inverted link \([毕业, 毕业]\) with the value \([1, 2]\).

3.2.3 Step 3: Run Iteration
The above steps will be iterated until no concatenation can be found, and this iteration process will generate a series of short and long LCs from the string. A long LC is believed to be more content representative than a short LC could possibly be. In this example, it is obvious that \('毕业案'\) is a better content indicator than either \('毕业'\) or \('毕业\)案\). To reduce the possibility of extracting less representative LCs from the concatenation, we will take a post-processing as the last step to finalize the CLCP.

3.2.4 Step 4: Execute Post-processing
In the final step, significant words are determined by observing the information mutually shared by two-overlapped LCs using the following significance estimation (SE) function as (3).

\[
SE_i = \frac{f_i}{f_a f_b - f_i} \quad (3)
\]

Where \(i\) denotes the LC, to be estimated, i.e., \(i = i_1, i_2, \ldots, i_n\); \(a\) and \(b\) represent the two longest compound substrings of \(LC_i\) with the length \(n-1\), i.e., \(a = i_1 i_2 \ldots i_{n-1}\) and \(b = i_2 i_3 \ldots i_n\). As for \(f_a\), \(f_b\), and \(f_i\) are the frequencies of \(a\), \(b\), and \(i\), respectively. For example, the term \('毕业案'\), we will gain the SE value of 0.83 with its frequency 5 and the frequency 6 of its substring \('毕业'\), as well as the value 5 of the other substring \('案'\). In this case, we will retain term \('毕业案'\) and its substring \('毕业'\) because the frequency of \('毕业案'\) is less than \('毕业'\) indicating \('毕业\) implies useful meanings. Likewise, we will discard the substring \('案'\) because both terms have the same frequency indicating the long term \('毕业案'\) can replace its substring \('案'\). As stated above, since \(f_i < f_a\), we retain both terms, and discard \('案'\) because \(f_i = f_b\).

3.3 Term Weighting
It is suggested that the obvious place where appropriate content identifiers might be found in news is the title and the first paragraph. In addition, we also considered frequency and the length of a word as the indicators of word significance in a document. Given a word \(LC_i\), the term weighting algorithm may be defined as (4).

\[
\text{Weight}_{i} = tfi \times (val_1 + val_2 + length_i) \\
val_1 = \begin{cases} 2, \text{word} \in \text{title} \\ 0, \text{otherwise} \end{cases} \quad val_2 = \begin{cases} 2, \text{word} \in \text{FP} \\ 0, \text{otherwise} \end{cases} \quad (4)
\]

4. Evaluation
In this experiment, we collected 1,738 images-resided web pages from Taiwan news website udn.com as the data sets. To verify our proposed CLCP method can successfully identify the content representation for image annotation, we used image captions as ground truth labels to see whether the primary annotation is included in the list. Subject experts and users were invited to assess the appropriateness of the secondary annotations (rank the second and the third). In the end, we also measured the performance of the CLCP method in a real-time mode.

4.1 Evaluation of Primary Annotation
Since the image captions are written by journalists, it is assumed that a man-made caption would be faithful to an image scenario. Therefore, we considered image captions as the ground truth labels, which will be used to evaluate the accuracy rate of the produced image annotations from the title and text. If the primary annotation matches a substring of the image caption, we will be confident to assure that the CLCP method works satisfactory.

Due to the problem of semantic ambiguity in part of Chinese words, where the interpretation of image annotations may vary from users to users, therefore the exact number of correct annotations of an image will not be clearly identified. For example, the string ‘總統馬英九’ (President Ma Ying-jiu) is meant to be regarded as a single LC, therefore it may not be appropriate to segment it into [總統] and [馬英九] even though these two words are valid. Thus, the recall measurement did not apply to this study. A precision measurement was used to understand the proportion of primary annotations actually matched the image captions as (5).

\[
p = \frac{\text{number of matched PAs in captions}}{\text{total number of PAs}} \quad (5)
\]

Where PAs represent the generated primary annotations from 1,738 documents. After the CLCP processing, the total number of matched PAs in captions is 1,565. We obtained a precision rate of 90.04%.

4.2 Evaluation of Secondary Annotation by Expert
To evaluate the validity of the secondary annotations, we invited five subject experts to participate in the assessment. Thirty pieces of news were randomly selected from which we extracted second and third place in scores of LCs and produced 60 annotations. To reduce ambiguous judgements, each annotation was evaluated based on a method of dichotomic classification to which the annotation represents
the image content. Each expert could only select either ‘agree’ or ‘disagree’ for each annotation. The result showed that the number of check marks of consent is 310 out of 360. It implies that the agreement of the appropriateness of the secondary annotations to the images achieves 86%.

### 4.3 Evaluation of Secondary Annotation by User

To evaluate the appropriateness of the secondary annotation, we conducted another survey to understand the differences between the image annotation and the users' expectation. Sixty graduate and undergraduate students were recruited from National Yunlin university of Science & Technology, Taiwan to participate in the assessment. Sixty pieces of news were randomly selected from which we extracted second and third place in scores of LCs and produced 120 annotations. To assist the assessment, we provided news title, texts and caption for references. Each annotation was evaluated by these participants to understand the degree to which the annotations appropriately address the image content. The result in Table 1 shows that the number of check marks of agreement is much higher than that of disagreement. The agreement rate of user evaluation reaches 73.7%.

**Table 1. Statistics of results of user satisfaction**

<table>
<thead>
<tr>
<th>Highly Agree</th>
<th>Agree</th>
<th>Average</th>
<th>Disagree</th>
<th>Highly Disagree</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1620</td>
<td>1238</td>
<td>684</td>
<td>218</td>
<td>116</td>
<td>3876</td>
</tr>
</tbody>
</table>

### 4.4 Performance Testing

After the validity and acceptance evaluation, we conducted a performance testing with respect to the time spent of processing from an event trigger to system response. Often real-time response times are understood to be in milliseconds and sometimes microseconds. Our testing data sets consist of 1,738 pieces of news; the processing time is 12.52 seconds in total with 0.007 seconds on average for each piece of news.

### 5. Conclusion

In this paper, we propose a corpus-free, relatively light computation of term segmentation for single document processing, namely “Chinese Lexical Chain Processing (CLCP) method” to identify representative terms for image annotation. The CLCP method is based on a hybrid of n-gram and lexical chain processing for image annotation. Unlike recent textual-based analysis methods only focused on single term processing thereafter they suffered the disadvantages of fragmented description of the annotation. We considered each character as a basic unit from which to build compounds as a composite, which in turn can be grouped to form larger compounds. Since the character and compound will be treated uniformly, it makes the application simple.

Our method allows sharing characters/words and facilitating their use at fine granularities without prohibitive cost. Results showed that this method achieves a precision rate of 90.04%, and gains acceptance from expert and user rating of 86% and 73.7%, respectively. In performance testing, it only takes 0.007 second to process each image in a collection of 1,728 testing data set.

Even though the CLCP method is only applied to single-document processing in the current study, with the virtue of corpus-free and lightweight features, it shall gain more benefits from applying in multi-document processing. Other researchers may verify our study using a larger data set or compare with the state-of-the-art algorithms. It is hoped that our research will invite more perspectives on automatic image annotation.

### 6. Acknowledgements

This study was supported by the National Science Council, Taiwan, Republic of China, under the project of special research projects of free software (NSC101-2221-E-224-056).

### References


[29] E. Gamma, R. Helm, R. Johnson, and J. Vlissides, *Design Patterns: Elements of Reusable Object-Oriented Software*. Indianapolis, IN: Pearson Education Co., 1995.