Cognitive-based Multi-Document Summarization Approach

Jingqiang Chen1, Wei Li2

Key Lab of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences
Beijing 100080, China

Nanjing University of Posts and Telecommunications, Nanjing, 210003, China
1cjq@kg.ict.ac.cn, 2weili@kg.ict.ac.cn

Abstract—Automatic text summarization is an important and useful research area in natural language processing and information retrieval. Most of current approaches for text summarization do not make full use of human reading process. This paper proposes a multi-document scanning mechanism by simulating human reading process. The mechanism simulates human memory of words, association between words and three cognitive processes invoked when reading. Changes of human memory of topic words in reading process are used to denote sentences’ significance, based on which sentences are then ordered and extracted to form a summary. Experiments on DUC2007 test data show that our proposing method is efficient and outperforms two baseline methods.

I. INTRODUCTION

Document summarization is of great value to many real world applications, such as snippets generation for search results and news headlines generation. With the explosive growth of the Internet, people are overwhelmed by a large number of accessible texts. Automatic summarization can represent the document with a short piece of text covering the main topics, and help users catch the most relevant information. So document summarization has become one of the most important research topics in the natural language processing and information retrieval communities.

Since Luhn's first paper on summarization in 1958 [1], overwhelming number of studies have been done on automatic summarization, including centroid-based summarization [2], topic-signature-based summarization [3], summarization based on information fusion and sentence fusion [4], [5], summarization based on Rhetorical Parsing [6], summarization based on sentence simplification [7], [8], graph-based summarization [9], [10], supervised summarization [11], [12], summarization based on Maximal Marginal Relevance [13]. These are different approaches to do summarization and are useful in different applications. However, seldom of them take into consideration human reading process. Since text is for human to read, human cognitive factors in reading process cannot be ignored when creating summaries.

To make use of human cognitive factors in reading process is feasible for machine to automatically generate summaries. A preliminary text scanning mechanism has been proposed to deal with text processing problems [14]. It simulates human reading process when reading a single document by bringing in human memory of words, relevancy between words and three cognitive processes, i.e. forget process, recall process and association process, in reading procedure. When scanning text sentence by sentence, human memory of words changes and relevancy between words emerges. Memory variation of a word can be used to identify its significance.

According to Zhuge [15]-[21], reading is a process of constructing semantic link networks of concepts in reader's mind by discovering and browsing the semantic link networks of words weaved by the author. A semantic link network (SLN) is a directed graph consisting of semantic nodes and semantic links. In recent years SLN is extended to semantically linking objects in various spaces to study the fundamental structure of cyber physical society and to create cyber-physical-social intelligence.

Human cognitive process has been considered as an effective aspect to process information in recent years. An associative memory is implemented through self-organizing maps for an intelligent system to dynamically learn about new high-level domains over time [22]. Cognitive modelling and eye tracking are combined to solve visual design problems [23]. A reading recommender is introduced to enhance digital books by utilizing spreading activation over text [24]. A prototype system of machine reading is proposed by chronicling the teaming of NLP and KR&R research groups [20]. And An interactive foraging process is taken into consideration for analysing the interaction between human memory and a network of peers in a multi-agent simulation system [25].

In this paper, we will extend the basic text scanning mechanism to define a multi-document scanning mechanism by simulating human reading process and use it to do multi-document summarization. To model the real cases when human read multiple document, we propose inter-document recall process and forget process in the scanning mechanism. We use changes of topic-related words' impression when scanning a sentence to denote significance of the sentence and extract the most significant sentences to form a summary. Experiments on DUC2007 data set show that our method outperforms two baseline methods MEAD and LEAD.

II. MULTI-DOCUMENT SCANNING MECHANISM

Our multi-document scanning mechanism simulates human reading process by combining word impression, word's
relevancy and three cognitive processes invoked in scanning process. We only counts nouns in a sentence for that noun directly reflects objects while other types of word only render objects.

It is reasonable to assume that documents are scanned in sequence. So the mechanism scans documents one by one, invoking forget process, recall process and association process.

A window is set to represent human local reading area. Let $S_{d,k}$ represent $k^{th}$ sentence in $d^{th}$ document. The window is concentrated on $S_{d,k}$, with $D$ sentences before it and $D$ sentences after it. The range never exceeds the document scope. The global range is from $S_{1,1}$ to $S_{d,k}$ and $S_{d,k}$ is the currently scanning sentence.

A. Word Impression

Word Impression reflects memory of a word in human mind. It changes in reading process. Two kinds of word impressions are considered: Global Word Impression (GWI, i.e. the word impression in the global range) and Local Word Impression (LWI, i.e. the word impression in the local range).

The initial $GWI$ value of the first document is set to 0. And the initial $GWI$ of $d^{th}$ document is calculated based on $GWI$ value when the scanning process of the $(d-1)^{th}$ document is finished. $GWI$ will fade out according to interval time between two sequential documents. The value of interval time is determined by human reading behaviors. One may read documents continuously or discontinuously.

$LWI$ is computed within the local range. During scanning, $GWI$ and $LWI$ of all the nouns in the text are updated. $GWI$ is updated based on $LWI$. Both $GWI$ and $LWI$ range from 0 to $+\infty$.

B. Relevancy between words

Co-occurrence semantic is a basic, simple and operable semantic. Words co-occurring in a paragraph or a sentence have high relevancy. David Hume thinks that people believe things are relevant when the things frequently appear in succession at logic, space, or time. According to this idea, our text scanning mechanism uses number of sentences that two words co-occur to represent relevancy between words.

We define three kinds of relevancy between words in our mechanism: Single Document Relevancy (SR), Local Relevancy (LR) and Global Relevancy (GR). $SR_{d,i}(w,v)$ is defined as number of sentences that $w$ and $v$ co-occur from $1^{st}$ sentence in $d^{th}$ document to $k^{th}$ sentence in $d^{th}$ document. And $SR_{d,i}(w,v)$ is number of sentences the two words co-occur from $1^{st}$ sentence in $d^{th}$ document to the last sentence in $d^{th}$ document. $LR_{d,i}(w,v)$ is the relevancy between $w$ and $v$ in the local range. It is computed based on $GR_{d,i}(w,v)$ and the frequency of two words co-occurring in the local range. $GR_{d,i}(w,v)$ is defined as number of sentences in global range that $w$ and $v$ co-occur when scanning $S_{d,k}$. It is computed from the first sentence in the first document to $k^{th}$ sentence in $d^{th}$ document. Relevancy between words in the global reading area is discounted by their co-occurring times in the local reading scope. $LR$ is used in association process and $SR$ is used in forget process defined in the following. The values of $GR$, $SR$ and $LR$ all range from 0 to $+\infty$.

A two-layer word network is defined in the scanning mechanism. One is Global Word Network (GWN), which is defined in global range and can be formally described as $GWN_{d,k} (GV_{d,k}, GE_{d,k}, GWN_d[]). GWN_{d,k}$ is the global word network generated when $S_{d,k}$ is scanned. $GV_{d,k}$ is the word set and $GE_{d,k}$ is the link set. The other is local word network (LWN). It is defined in local range and can be formally described as $LWN_{d,k} (LV_{d,k}, LE_{d,k}, LWN_d[]). LR_{d,k}[1, D]$.

C. Forget, recall and association process

Forget process, recall process and association process are three cognitive processes invoked in scanning mechanism. GWN and LWN are updated based on relevancy between words in the three processes.

Forget process is to simulate the phenomenon that human forgets slowly at the first, and the forgetting speed then becomes faster and finally slows down again.

Forget time, repeating sentences and link diversity are three factors mainly influencing forgetting speed in the single-document scanning mechanism [14]. Forget time of a word is determined by the number of sentences from the one it last occurs to the currently scanning sentence. The number of repeating sentences is defined as the number of sentences the word occurs within global range. The link diversity is defined as number of links between the word and its neighbors. For multi-document scanning mechanism, repeating documents is another important factor influencing forgetting speed. Like repeating sentences, a word will not be forgotten quickly if it has occurred in many documents.

Combining these four factors, we get Equation 1 and Equation 2:

$$GWI_{d,k}(w) = C_{\text{forget}} \times GWI_{d,k-1}(w)$$

$$= (1 - \frac{1}{\lambda \times (1 + (ft - \alpha)^{2})}) \times GWI_{d,k-1}(w) \quad (1)$$

and

$$\alpha = \text{Max}(\text{MIN}_{a}, \log_{2} \frac{d \prod_{j=1}^{d} (1 + SR(w,j))}{\sum_{i=1}^{d} \text{NL}(w) + \sum_{i=1}^{d} \text{SR}(w,j)}) \quad (2)$$

In Equation 1, $ft$ denotes forget time, $C_{\text{forget}}$ is forgetting speed, and $\lambda$ determines lowest value of $C_{\text{forget}}$ and is usually set to 2. In Equation 2, $\text{MIN}_{a}$ is the minimal value of $\alpha$, which is usually set to 5. $SR(w,j)$ denotes the SR of word $w$ and word $j$ in the $i^{th}$ document. For $d^{th}$ document which is the scanning document, $SR$ is computed from $S_{d,1}$ to $S_{d,k}$. $\text{NL}(w)$ denotes the number of words which are relevant to word $w$ in $i^{th}$ document computed from the first sentence to the last sentence in $i^{th}$ document. For $d^{th}$ document, it is computed from $S_{d,1}$ to $S_{d,k}$.

Big forget process can be defined between two sequential documents. Inter-document forget process is much more complex than intra-document forget process mainly because human behaviors have great uncertainty during the interval time of two documents reading process. Our model uses
interval time in inter-document forget process to model it. Here we let interval time between \( d^\text{th} \) document and \((d+1)^\text{th}\) document be \( I_d \) times of sentence reading time. For simplicity, \( I_d \) is set to \( 1/2 \times D \) in our experiments where \( D \) is the size of reading window. This is to assume that readers will take a rest of a fixed time long after finishing reading each document.

Opposite to forget process, a word’s previous \( GWI \) will be reminded of in recall process. For each word \( w \) in \( S_{d,k} \) being scanned, recall process first try to find the max \( GWI \) of \( w \) within document \( d \). If it fails, recall process will try to search \( d^{-1}\text{th} \) document down to \( 1^\text{st} \) document. We assume inter-document recall will lose \( L \) percent of \( GWI \) of \( w \). \( L \) can be set to 20. The following shows the algorithm for recall process.

**The recall process of \( S_{d,k} \)**

Input: \( GWI_{d,1} \) to \( GWI_{d,k} \)

Output: \( GWI_{d,1} \)

Steps:
- for each word \( w \) in \( S_{d,k} \)
  - if \( w \) has occurred in \( S_{d,1} \) to \( S_{d,k-1} \)
    - \( GWI_{d,k}[w] = \max(GWI_{d,1}[w],...,GWI_{d,k-1}[w]) \)
  - else
    - \( GWI_{d,k}[w] = L \times \max(GWI_{d,1}[w],...,GWI_{d,1,\text{last}(d,k)}[w]) \)
- return \( GWI_{d,1} \)

Forget process is called for words not in \( S_{d,k} \) while recall process is called for words in \( S_{d,k} \). Neither of them increase impression of words in essence. Association process increases \( GWI \) of word \( w \) by adding \( LWI \) of \( w \) to it. When \( S_{d,k} \) is being read, \( LWIs \) of all words are set to 0 initially. \( LWI \) of words in \( S_{d,k} \) will increase by \( \beta \) first, and then the weight propagates to other words in the local range via \( LR \). The usual value of \( \beta \) is set to the noun count of \( S_{d,k} \). After propagation, \( LWI \) is added to \( GWI \) of the words. Here we adopt the propagation algorithm described in [14].

In multi-document scanning mechanism, \( GWI \) values of all the nouns for each scanning sentence can be figured out. Time complexity of the single document scanning mechanism has been proved to be polynomial [14]. Multi-document scanning mechanism is not more complex than single-document scanning mechanism.

**III. Summarization**

For topic-focused multi-document summarization, most of the current methods are to extract the most important sentences from documents about the given topics to form a summary. The difference of these methods lies in how to compute score for the sentence. Our method is to make use of the multi-document scanning mechanism to assign scores to sentences. The main idea is as follows.

Since we are making topic-focused summarization and a topic \( T \) has several keywords related with it, we can sum up their \( GWI \) values when scanning a sentence \( S_{d,k} \). So \( GWI_{d,k}(T) \) can be defined in Equation 3.

\[
GWİ_{d,k}(T) = \sum_{w \in T} GWI_{d,k}(w) \tag{3}
\]

In the following experiments, \( D \) is set to 10 and \( b \) is set to 100. We use \( N(\mu, \sigma) \), to model \( I_d \). \( \mu \) determines expectation and \( \sigma \) determines variation of \( I_d \). Here we set \( \mu=D \). And we mainly consider the case that \( I_d \) varies not too greatly. Thus we set \( \sigma=80 \).

Topic’s \( GWI \) varies when scanning different sentence. Fig. 1 shows an example of Topic \( GWI \) Curve for a document set from DUC2007. The topic name of the document set is "Southern Poverty Law Center" and the topic id is "D0701A". The nouns in topic title and topic narrative are extracted as the topic’s keywords. We order the documents by reporting time and select the first 10 documents for experiment. The numbers in the horizontal axis represent sentence numbers and the numbers in the vertical axis represent \( GWI \) values. There are 442 sentences in all and we show the first 172 sentences for clarity.

![Fig. 1 Example of Topic GWI Curve for a document set whose topic is "Southern Poverty Law Center" from DUC2007](image)

From Fig. 1 we can see that some sentences, e.g. sentence 7, sentence 58, etc., contribute great to topic’s \( GWI \) for that there are big convexities in their position in the curve. Thus we can compute \( GWI \) increment of a sentence comparing to its previous sentence. A better way is to compute its relative \( GWI \) increment. Equation 4 computes increment ratio for \( S_{d,k} \).

\[
IRatio_{d,k}(T) = \frac{GWİ_{d,k}(T) - GWİ_{d,k-1}(T)}{GWİ_{d,k-1}(T)} \tag{4}
\]

We can use \( IRatio \) to denote significance of a sentence. So our summarization algorithm named \( IRatioSumm \) can be given as follows:

- Step 1: Run the multi-document scanning mechanism and compute \( IRatio \) for each sentence;
- Step 2: Order the sentences by \( IRatio \) values in descending order;
- Step 3: Return the first \( SumLen \) sentences as a summary. Here \( SumLen \) is the length of the summary.

DUC2007 data set is used to evaluate our method. We use StanfordCoreNLP [26], a set of natural language analysis tools written in Java, to do sentence splitting and noun extracting. And we compare the results of our summarization method against two baseline summarization methods, MEAD [27] and LEAD, by ROUGE [28]. MEAD is a state-of-the-art summarization system. Here we use the default configuration where length, position and centroid are considered to rank sentences. LEAD is a simple but useful method which just...
takes out first several sentences of document to form a summary. \textit{ROUGE} is a recall-oriented method and has been widely used to evaluate summaries. \textit{ROUGE} has been traditionally used to compute the performance based on the N-gram overlap (\textit{ROUGE-N}) between the summaries generated by the system and the target gold standard summaries. Gold standard summaries have been given in DUC2007 data. Here we use recall version of \textit{ROUGE-1}, \textit{ROUGE-2} and \textit{ROUGE-L} for our evaluation. Table I shows the results.

<table>
<thead>
<tr>
<th>Metric</th>
<th>\textit{IRatioSumm}</th>
<th>\textit{LEAD}</th>
<th>\textit{MEAD}</th>
</tr>
</thead>
<tbody>
<tr>
<td>\textit{ROUGE-1}</td>
<td>0.48705</td>
<td>0.40930</td>
<td>0.45965</td>
</tr>
<tr>
<td>\textit{Recall}</td>
<td>0.12288</td>
<td>0.09612</td>
<td>0.12249</td>
</tr>
<tr>
<td>\textit{ROUGE-L}</td>
<td>0.45330</td>
<td>0.37461</td>
<td>0.41974</td>
</tr>
</tbody>
</table>

From Table I, we can see that our method \textit{IRatioSumm} performs better than both \textit{LEAD} and \textit{MEAD} by \textit{ROUGE-1 Recall} and \textit{ROUGE-L Recall}. And for \textit{ROUGE-2 Recall} metric, the performance of the three methods differs not greatly. These results tell us that human cognitive factors do favor summarization. The reason is that texts are in essence written for human to reading and summaries are generated by human after they read through the texts.

IV. CONCLUSIONS

This paper proposes a multi-document scanning mechanism by simulating human impression of words and relevancy between words emerging and changing in human reading process. Forget process, recall process and association process are brought in to manage impression of words. A sentence scoring method named \textit{IRatio} is defined based on the multi-document scanning mechanism. \textit{IRatio} scores a sentence by \textit{GWI} increments comparing to its previous sentence. A summarization method based on \textit{IRatio} is proposed. Experiments show our summarization method outperforms \textit{MEAD} and \textit{LEAD}.

ACKNOWLEDGEMENT

This work was supported by National Science Foundation of China (61075074) and Research Funding of Nanjing University of Posts and Telecommunications.

REFERENCES