A NEWS CONTENT SUMMARIZER FOR CHINESE CELL PHONES

Yuen-Hsien Tseng

Dept. of Library & Information Science, Fu Jen Catholic University
Taipei, Taiwan, R.O.C. 242
tseng@lins.fju.edu.tw

ABSTRACT

This paper proposes a Chinese news summarization method for the message services of news brief over cell phones. In this method, important sentences were first identified based on the news content. They were matched against the news title to determine a suitable position for combining with the title to become candidates. These candidates were then ranked by their length and fitness for manual selection. In our evaluation, among 40 news stories, over 62.5% of them have their top-ranked candidates be judged good in quality. If all candidates were considered, over 75% of them yield acceptable summaries without manual editing. Since these summaries were concatenated snippets from the news texts, they can be easily integrated and synchronized with other streamed media such as speech or video from the same story.

1. INTRODUCTION

The popularity of cell phones in Taiwan area has reached the highest rate in the world during the last few years. Over 23 million cell phone numbers were used by June 2002, slightly more than the populations of Taiwan [1]. To better utilize this ubiquitous communication device, a number of content providers have provided Chinese news brief services over the cell phone, such as United Daily News [2], Central News Agency [3], and PC Home in Taiwan and New Media in Singapore. The Asahi Shimbum in Japan (the second largest news agency in the world) has provided such news message services with an inexpensive rate since 1999, in the hope that the increase in the number of the readers of their content can finally increase the subscription of their news paper [4]. As multimedia technologies continue to improve, future news service over the cell phone may not only include texts, but also speech, images, or video, integrated and synchronized. However, to reach this vision, the operation cost should be low enough to sustain such services. Therefore, any automated ways to keep the cost low would be of great help.

The news brief shown on a cell phone is different from those on a Web browser. Due to the limited screen size, a length limit is defined for each news message. This is usually 45 Chinese characters in PHS system or 69 characters in other systems, including punctuation marks [2]. Summaries of this kind are longer than a news title but shorter than a long Chinese sentence. For the benefit of the subscribers, the summaries should contain as much content as possible. Also the readability and coherence of the summaries are important factors that should be taken into account.

Automatic summarization techniques for cell phones have been studies in recent years [5-6]. Most existing approaches seek to hide and/or scroll text across the phone's screen to show only necessary information, but our work seeks to provide the most effective snippet of text. Specifically, this paper is to propose a Chinese news summarization method for the message services of news brief over cell phones, with the aim to meet the considerations just described. The proposed method has the potential not only to reduce the cost of current manual operation but also to possibly integrate and synchronize with other media in such services in the future.

2. THE PROPOSED SUMMARIZER

Previous researches have studied the summarization techniques for Chinese news (e.g.,[7]), but none has done for the problem discussed here. To develop an automated Chinese news summarizer subject to the cell phone limitations, an understanding of the style of the news stories and how humans summarize them would be helpful. Table 1 lists three news examples and their English translations [8]. As can be seen, these examples are short, with their bodies having only 1, 2, and 3 sentences, respectively. This is not uncommon for the stories to be transmitted to users' cell phones, although longer stories may be selected as well. Given such short stories, a human summarizer has very few clues to rewrite the story thoroughly to fit the length limit. The best way he or she can do may be to cut and paste the snippets from the news text with minimum editing to avoid garbling the original meaning.

0-7803-9332-5/05/$20.00 ©2005 IEEE

Table 1: Three news examples for summarization. The number in the parenthesis is the number of characters in the preceding sentence.

1	**太空探測器在遙遠的恒星周圍發現水的痕跡** (19) 美國航空航太總署的科學家星期三稱，新近在一顆遙遠的恒星周圍發現了水存在的痕跡，這可以成爲第一個支援除我們自己存在地外生命的證據。 (64) #2001/07/13#
	Space Probe Sees Signs of Water Around Distant Star Newly detected signs of water around a distant star are the first evidence that planetary systems outside our own might be able to support life, NASA scientists said on Wednesday.
2	**專家：世界人口接近頂點 90 億** (13) 科學家星期三預測説，在 2070 年左右，世界人口可能會達到頂峰約 90 億，然後開始下降。(38) #2001/08/03# 澳大利亞人口統計學家在考慮很多因素後計算出到本世紀結束時，地球上的人口會下降至 84 億人。(43)
	Experts: World Population Set to Peak at 9 Billion The world's population will probably peak at about 9 billion around 2070 before it starts to decline, scientists predicted Wednesday. Demographers at a think tank in Austria calculate that by the turn of the century the number of people on the planet will have dropped down to 8.4 billion people.
3	**海洋生物普查行動發現數百種新生物** (16) 有史以來第一次的「海洋生物普查計畫」進行以來，來自 53 個國家的科學家們，平均每星期便發現 3 種新的海洋魚種。(52) #2003/10/23# 這項爲期 10 年、耗資 10 億美元的計畫，動員了來自世界各地的科學家共同合作，目的是爲了將存在海洋中所有種類的生物發掘出來並予以分類。(63) 計畫至今已實施了 3 年，科學家在海裡發現了 15300 多種生物，他們估計還有 5 千多種生物尚未被科學界發掘。(49)
	Marine Life Census Finds Hundreds of New Species New marine fish species are being logged at an average rate of three per week by scientists from 53 countries engaged in the first Census of Marine Life. The 10 year, $1 billion global scientific collaboration aims to identify and catalog all life in the oceans. After their first three years of work, census scientists report over 15,300 species of fish in the sea and estimate 5,000 more are still unknown to science.

Table 2: The summary candidates for the third story in Table 1. They are created by combining the last clause of each body sentence with the title.

1	海洋生物普查行動發現數百種新生物，平均每星期便發現 3 種新的海洋魚種。(34) Marine life census finds hundreds of new species, at an average rate of three per week.
2	海洋生物普查行動發現數百種新生物，目的是爲了將存在海洋中所有種類的生物發掘出來並予以分類。(45) Marine life census finds hundreds of new species, aims to identify and catalog all life in the oceans.
3	海洋生物普查行動發現數百種新生物，他們估計還有 5 千多種生物尚未被科學界發掘。(38) Marine life census finds hundreds of new species; they estimate 5,000 more are still unknown to science.

The snippets to be cut and pasted can be enumerated and then suggested by a computer for manual selection. However the possibilities of such enumeration would be huge if all substrings of the news text are blindly considered. As can be seen from the examples in Table 1, a Chinese sentence is often composed of several comma-separated clauses, which convey the meaning of the sentence in successive sequence. The Chinese clause is quite independent in some circumstances and is thus a useful unit to be combined with others to make a new sentence. Although most of such combined sentences would be invalid, several of them are still meaningful and sometimes more complete in its content, especially for those combined from the beginning and ending clauses.

Take the third story from Table 1 as an example. The title has only 16 characters, falls short of the required length, 45 or 69. The other 3 sentences have 52, 63, and 49 characters, respectively. None of them alone is an ideal summary for the required length. But by concatenating the last clause of each body sentence with the title, as shown in Table 2, each becomes a better choice for summaries of length 45.

This observation gives us clues to effectively enumerate the summary candidates. But there are other problems that need to be considered in order to further reduce the burden of human selection. (1) The number of suggested candidates should be fairly equal for each story. Long stories should not yield far more candidates than short ones. (2) The candidates should be ranked in some sense when they were suggested for selection.

To tackle these problems, we propose the following steps:

Step 1: Sort all the sentences of a news story by their weights and select the best 5 sentences for use in the next step.

Step 2: Create summary candidates by matching each selected sentence with the news title. Calculate the match scores and summary lengths.

Step 3: Sort the candidates by their lengths and scores.

In Step 1, the weight of a sentence in a story of any length is determined by the accumulated weights of the keywords which occur in that sentence, as shown below:

$$weight(S) = \sum_{w \in Keywords \in S} (0.5 + 0.5 \times tf_w / \max\_tf)$$

where tf_w is the term frequency of keyword w and max_tf is the term frequency of the keyword which occurs most in the news story. Here the keywords of a story are those title words that are not non-content-bearing and those maximally repeated patterns that are extracted by Tseng's algorithm [9]. Tseng has shown that Chinese news stories can contain many new keywords, almost 1/3 of repeated words are unknown to a lexicon of 123,226 terms. His algorithm ensures that unknown words can be extracted as well, as long as they occur at least twice in a document.

In Step 2, since titles are guides to a news story, they should better be included in the beginning of the candidates. The ending clauses to be concatenated should better supplement the content of the title. This means that the beginning clauses of a body sentence should be as similar to the title as possible. To spot the position for concatenation and to know the similarity, a dynamic programming (DP) technique is used.

Given two strings A[1..n] and B[1..m], where n<=m, the edit distance between A[1..i] and B[1..j] based on DP is:

$$d[i, j] = min(d[i-1, j] , d[i-1, j-1], d[i, j-1])+c(A[i], B[j])$$

where min is a function that returns the minimum of its 3 arguments, and c(A[i], B[j]) = 0 if A[i]=B[i], and 1 otherwise. The initial values for the distance are: d[0, 0]=0, d[0, j]=0 for j=1..m and d[i, 0] = d[i-1,0]+1 for i=1..n. These initial values follow those proposed in [10], where a similarity measure is defined from the edit distance: exp(d / (d-n)), where exp is the exponent function and d=d[n, j], for some j. This similarity ranges from 0 to 1. But we found that its range does not distribute well for later comparison. So it was changed into:

$$sim = \exp(\frac{d}{d-m-n})$$

The new measure ranges from exp(-n/m) to 1.

The position to determine the ending clauses is at the comma which most closes the position with highest similarity. But since we favor length more than similarity (here length is a direct measure while similarity is just an estimate), the matched comma is changed to its preceding or succeeding comma whenever such changes fit the length limit better.

News stories are often written in a so-called pyramid style where the later the paragraph occurs, the more the details it carries. Thus better summaries often come from the first few sentences. We thus decrease the similarity of the summary candidates composed from the sentences other than the first two by a factor of 0.85, if the number of sentences in the story exceeds 3.

In Step 3, the length of the summary candidate is divided by the required length limit (45 or 69) to yield the length ratio ranging from 0 to 1. Now we now come to a problem of determining the rank of these candidates based on their length ratios and similarities. Ideally this problem can be solved by machine learning methods. But they require manually prepared data to train a classifier to determine the best or to rank the candidates. The effectiveness of a machine classifier heavily depends on the number of training data. Since sufficient training data are hard to prepare, a set of hand-crafted rules are devised instead:

(1) From the candidate list, find the candidate with highest similarity, called it X, and the candidate with largest length ratio, called it Y. If X was Y, then output X and delete X from the candidate list.

(2) If sim(X) > 1.25 x sim(Y) and ratio(X) > 0.75 x ratio(Y), then output X, otherwise output Y. Remove the candidate just output from the list.

(3) Repeat (1) and (2) until there is no candidate.

3. PERFORMANCE EVALUATION

Based on the above steps, the summary candidates of length limit 45 for the third story in Table 1 are exactly those three in Table 2. The (ratio, similarity) values for the candidates are (0.7556, 0.7351), (1.0, 0.7757), and (0.8444, 0.6718), respectively. Step 3 sorts the candidates 1, 2 and 3 into 2, 3, and 1 in decreasing order of rank. As to the quality of the candidates, candidate 2 with rank 1 is correct and coherent in meaning and is perfect in length. Candidate 3 is fair in Chinese expressions. It would become better if the word: “他們” (they) in the beginning of the second clause is deleted. Candidate 1 is also correct and coherent. It carries more interesting content than candidate 2. But it is shorter in length.

To further evaluate the above approach, forty Chinese news stories were tested. They were real-time news (short stories updated per 30 minutes) from China Times [11] during August and September in 2003. Table 3 shows some statistics about these stories.

Table 3. Statistics of the forty news stories.

Average number of sentences per story	2.95
Average number of clauses per sentence	4.08
Average number of characters per sentence	64.54
Average number of characters per clause	15.83
Average number of characters per title	16.63

For each story, summary candidates were generated and ranked. They were then evaluated by a human summarizer who labeled the quality of each candidate G (good), F (fair), or B (bad) if it is correct and coherent, correct with some readability, or unacceptable, respectively. For each story, we only recorded the rank and the quality label of the best candidate. These data were finally accumulated in Table 4. As can be seen, among 40 stories, 62.5% or 65.0% of the first candidates suggested by the system were judged good. If users were able to choose from all the suggested candidates, 80% or 75% summaries can be obtained from a machine without manual editing. About 12.5% or 10% stories yield summaries that are totally unacceptable.

Those best candidates that are unacceptable (9 cases in total) contain undesired conjunctions that break the coherence and/or readability (4 cases), title strings that duplicate the clauses (2 cases), or nothing but the title itself which means that no candidates can be generated under the required length limit (3 cases).

Table 4. (a) Quality statistics for the summary candidates of length limit 45.

Quality Rank	Good	Fair	Bad
1	26 (65.0%)	2 (5.0%)	5 (12.5%)
2	5 (12.5%)	1 (2.5%)	0
3	1 (2.5%)	0	0
4	0	0	0
total	32 (80.0%)	3 (7.5%)	5 (12.5%)

(b) Quality statistics for the summary candidates of length limit 69.

Quality Rank	Good	Fair	Bad
1	25 (62.5%)	6 (15%)	4 (10%)
2	3 (7.5%)	0	0
3	1 (2.5%)	0	0
4	1 (2.5%)	0	0
total	30 (75%)	6 (15%)	4 (10%)

4. POTENTIAL APPLICATIONS

The proposed method recombines the snippets of the text without modifying them. Although it is simple, it is effective for some types of stories. Another advantage of this method is that the other synchronized media such as images, speech, or video of the same story can maintain their synchronization with ease when they are summarized as well since the positions of where to cut and paste are known during the generation of the summary candidates. Thus not only the text summary can be used, but also the other summarized media. Put another way, to achieve speech or video segmentation and summarization for any services, one can use their synchronized texts based on this method.

5. CONCLUSIONS

The fact that the proposed method works for some stories is due to the characteristics of Chinese news. They tell stories in successive details. The clauses at the front and the rear of a sentence are sometimes quite independent. These make clause recombination a choice for summary generation. The remaining work is to evaluate their fitness as summaries and rank them in a correct sense. For the news stories we tested, our proposed method applies to most of them with success.

ACKNOWLEDGEMENT

This work is partly supported by WebGenie Information Ltd. and National Science Council under the grant numbered: NSC 93-2213-E-030-007-.

6. REFERENCES

[1] I-Chin Wang, ”Number One in the World: Cell Phone Numbers used More Than the Populations of Taiwan,” (in Chinese) ETtoday.com, 2002/08/09, accessed on 2005/01/02 at http://www.ettoday.com/2002/08/09/339-1337800.htm.

[2] United Daily News, “egolife: Brief Message Delivery for Cell Phone,” (in Chinese) accessed on 2005/01/02 at http://udn.com/NASApp/LogFriend/UDNSMS/introduction_news.html.

[3] “How to Order Stock News Brief Message from Central News Agency?” (in Chinese), accessed on 2003/12/3 at http://www.suio.com.tw/top/can/can_order_txt.asp.

[4] “Media Challenges: Multi-modal communications,” (in Chinese) accessed on 2003/12/03 at http://marketing.chinatimes.com/item_detail_page/professional_columnist/professional_columnist_content_by_author.asp?MMContentNoID=4369.

[5] O. Buyukkokten, H. Garcia -Molina, A. Paepcke, “Seeing the Whole in Parts: Text Summarization for Web Browsing on Handheld Devices,” The 10th International WWW Conference (WWW10). Hong Kong, China, 2001

[6] Christopher C. Yang, Fu Lee Wang, “Adapting content to mobile devices: Fractal summarization for mobile devices to access large documents on the web,” Proceedings of the 12th intern. conf. on World Wide Web, May 2003, pp.215-224.

[7] Hsin-Hsi Chen, June-Jei Kuo, Sheng-Jie Huang, Chuan-Jie Lin, Hung-Chia Wung, “A summarization system for Chinese news from multiple sources,” Journal of the American Society for Information Science and Technology, Vol. 54, No. 13, Nov. 2003, pp. 1224-1236.

[8] In Table 1, the first two stories were accessed from http://www.1999.com.tw/english/, the third story was access at http://news2.ngo.org.tw/php/ens.php?id=03102302 on 2005/1/5.

[9] Yuen-Hsien Tseng, "Automatic Thesaurus Generation for Chinese Documents", Journal of the American Society for Information Science and Technology, Vol. 53, No. 13, 2002, pp. 1130-1138.

[10] Daniel Lopresti and Jiangying Zhou, “Retrieval Strategies for Noisy Text,” Proceedings of the Fifth Annual Symposium on Document Analysis and Information Retrieval, April 15-17, 1996, pp. 255-269.

[11] “China Times” http://www.chinatimes.com.tw/