Popularity-Based Summarization of Chinese Text: Implicit Weight-Based Features for Newspaper Articles

Lukas Pichl and Nozomi Mikami

Department of Information Science
International Christian University, Osawa 3-10-2, Mitaka, Tokyo, 181-8585 Japan

Abstract

The fully idiographic nature of Chinese language facilitates the algorithms in natural language processing by a tight correspondence between characters and text meaning, and also allows for substantially less preprocessing as compared to the case of other languages. We have recently developed and tested a popularity-based concept to summarization of Chinese text. This paper argues that the popularity based concept, if applied to news articles, naturally combines with the weight-based summarization techniques, and yields an excellent efficiency in Chinese text summarization. The summarization program is available online as java applet (http://cpu.icu.ac.jp/cn/).

Keywords: abstract extraction, summarization of Chinese text, popularity-based algorithm, weight-based summarization

1. Introduction

Automatic text summarization allows for presenting the main information content of text in a shorter form, by means of either sentence extraction or rephrasing abstraction. The latter approach requires understanding of the language grammar and context; thus we adopt the former one, i.e. abstract extraction. Summaries constructed by sentence extraction can always be further paraphrased, in order to present them in a more natural form. Text summaries are a powerful tool for saving time in decision-making; summarization can also be used as a preprocessing tool in more complex information retrieval [1-2]. Much of the recent research developments in the algorithms [3-4], which extract or abstract substances from a range of text data (including the online information) have been motivated by the increase of E-commerce and use of online search engines.

The major language in information retrieval of internet users is English, although the size of the non-English-speaking population keeps growing (820 millions in 2005). In particular, the Chinese speaking internet users already account for the second largest online population at present, and the share of Chinese language increases faster than the growth rate of the internet users [5].

Following the popularity-based concept in webpage ranking on the internet [6], an improved popularity-based approach for English text summarization was developed [7]. Here we further adapt the popularity-based algorithm for Chinese text summarization. It is the purpose of this paper to present the algorithm along with a thorough evaluation by a control group of native Chinese speakers. In particular, we focus on the online news articles, which are important in applications, and also include an implicit ranking system based on the location of sentences in the text. The extracted summaries consist entirely of selected text material copied from within the input document (an upper limit on the size of the summary can be imposed).

The paper is organized as follows. The popularity-based summarization is briefly outlined in Section 2. In particular, we focus on the clustering algorithm for selecting the most popular text sentences, and its combination with the implicit weighing structure of sentence location in news articles. Section 3 evaluates the algorithm on a group of news articles selected from various fields and a diversity of online publishers. Concluding remarks in Section 4 summarize the major advantages of the present approach.

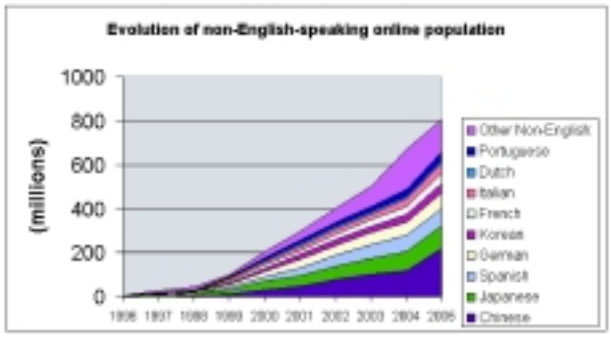

Fig. 1: Relative importance of languages for online-available text over the last 10 years [5].

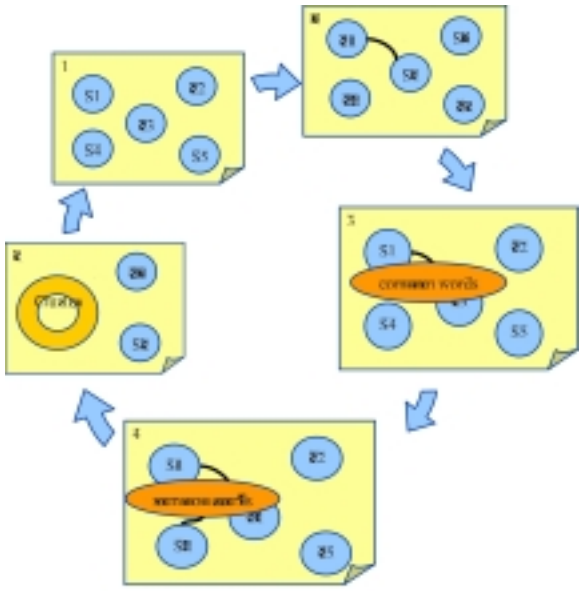

Fig. 2: The outline of the text summarization: (1) Graph Initialization, (2) Cluster Selection, (3) Cluster Annotation, (4) Cluster Growth, and (5) Repeated Initialization.

2. Summarization Algorithm

The summarization algorithm first initializes all sentences as graph vertices, which are then grown by merger operations. Figure 2 shows the main steps in vertex clustering when building the resulting graph. First, two vertices (sentences) exhibiting the largest global similarity are merged. While the two old vertices are deleted, a new vertex is added. This new vertex is annotated with common words of the two parental sentences (a new sentence). Finally, sentences not assigned to the cluster are scanned for a local similarity above a certain threshold value, and possibly merged with the cluster seed as described before. The clustering procedure then restarts again, until all sentences (number N) over a certain similarity threshold τ are exhausted. The detailed flowchart in Fig. 3 shows how the similarity measures and τ is used in the construction of the similarity graph.

There are two key functions quantifying the similarity in constructing the similarity graph. Each takes a pair of sentences on input: the global similarity coefficient (GSC), which is used to create cluster seeds, and the local similarity coefficient (LSC), which is used in growing the clusters. The LSC and GSC are defined as follows.

$$\mathrm{GSC}(\mathrm{S}_i, \mathrm{S}_j) = \frac{2 \times n(\text{common words of } \mathrm{S}_i \ \& \ \mathrm{S}_j)}{n(\mathrm{S}_i) + n(\mathrm{S}_j)}$$

$\forall i \neq j \leq N$
$n(\mathrm{X})$: number of words in vertex X, $0 \leq GSC \leq 1$.

$$\mathrm{LSC}(\mathrm{S}_k, \mathrm{C}) = \frac{2 \times n(\text{common words of } \mathrm{S}_k \ \& \ \mathrm{C})}{n(\mathrm{S}_k) + n(\mathrm{C})}$$

$\forall k \neq i_0, j_0$
C: cluster, $0 \leq \mathrm{LSC} \leq 1$.

A “word” is defined as one Chinese character in what follows. Since the GSC decreases as the similarity among shortened sentences of common words, the iterative algorithm in Figs. 2 and 3 decreases the value of GSC, and stops when the threshold value τ can no longer be reached.

The process of abstract extraction generally consists of the following four parts: Preprocessing, Building Text Graph, Clustering into Themes and Selecting Sentences. In dealing with the Chinese news text, the preprocessing can be minimal. No preprocessing was needed for the news articles dealt with in this work. The other three steps are illustrated in Fig. 3. The weight-based approach for news articles is implicitly applied in this work via the selection of the representative sentence for all cluster groups. Among the original sentences merged into each cluster, the one with the up most position in the news article is selected into the summary. This particular choice corresponds to the implicit sentence location ranking in the news, which can therefore be considered as an application of the weight-based approach. The weight-based approach also naturally comes into play when the number of sentences included in two or more clusters is the same. The cluster with a representative sentence located in the upper part of the article has a higher priority to enter the summary, which is how the news article weighting system again enters into the popularity-based clustering algorithm.

In dealing with Chinese, Java is one of the appropriate programming language tools because it supports Unicode. CJK Unified Unicode Ideograph has the range from U+4E00 to U+9FBB (19968 to 40891) in the Unicode Standard, Version 4.1, and it supports both the simplified Chinese and traditional Chinese. All characters in the input article are converted into the integer data type and stored as a numerical array. By subtracting 19968 from the Unicode character table, the values in the integer data array range from 0 to 20923. In order to separate article sentences, only the delimiters, “。” or “・”, “?” and “!” have been applied.

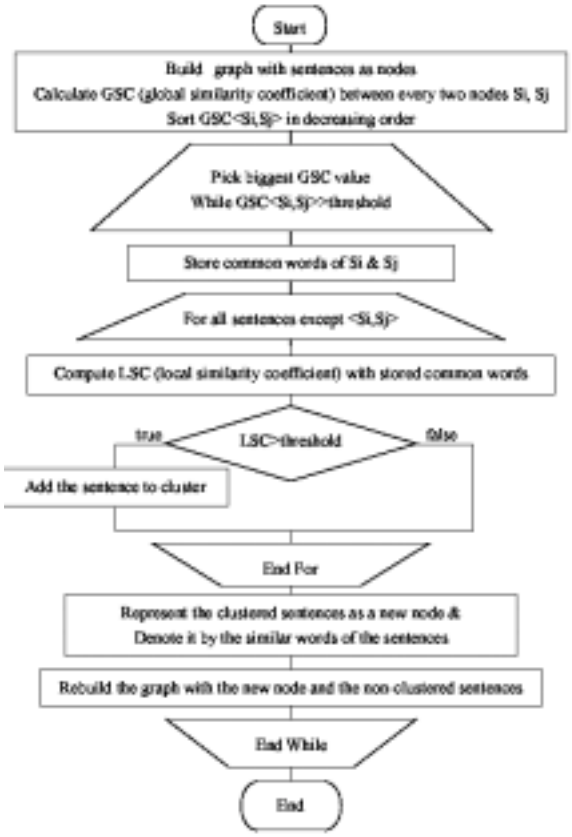

Fig. 3: Flowchart of the summarization algorithm. Note that a "sentence" representing the cluster is defined as a collection of common words, and a separate data structure must be kept in order to construct the final summary.

3. Evaluation and Discussions

In order to evaluate the efficiency of machine text summarization rigorously, we have collected a representative set of newspaper articles, and gathered a 10-member volunteer group. Human-extracted abstracts are compared to the abstracts extracted by the computer program. To compare between the two, we calculate the relevance score (RS) between all pairs of extractors (human vs. human, or human vs. computer). The source text was distributed to human extractors as follows: there were 10 native Chinese-speaking people, each of whom extracted summaries from 10 articles. Among the 10 articles, 5 articles were common for all people, while the other 5 articles were unique. The article length is 20 to 25 sentences, all of which are labeled by Latin alphabet letters. The extracted abstracts for the common articles were used to calculate the spread of summarization among humans. Results were collected by using questionnaire forms shown in Fig. 4. Table 1 shows the results by human selectors A,...,J and the computer program for various

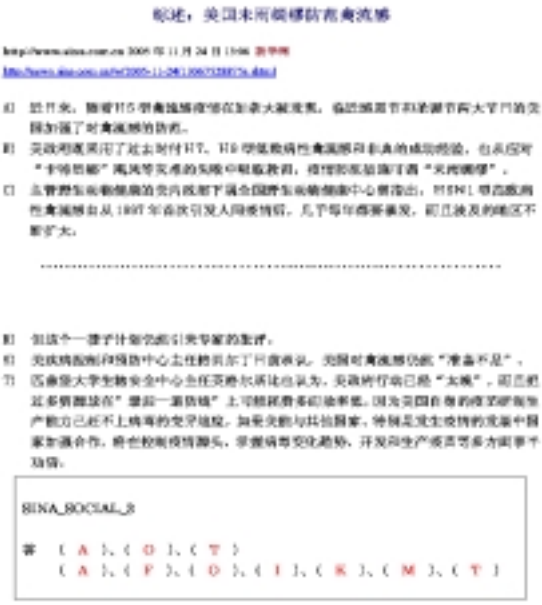

Fig. 4: The form distributed to human extractors of text summaries. The sentences were labeled by Roman alphabet numbers. The extracted summary is shown in the bottom of the form in the parentheses.

threshold values. The program was implemented as a java applet in Fig. 5 [8].

Fig. 5: The summarization program as a java applet.

The five columns in Table 1 contain the sentences extracted as an abstract for one article in each control group (international news, domestic news etc.). The sentences are labeled as a, b, ... Brackets indicate a tie when selecting the representative cluster (upper positions in the text were preferred). Results by a control group of native Chinese speakers (A, B, ..) are immediately followed by machine-extracted abstracts for various thresholds τ of the global and local similarity coefficient. The extraction algorithm is

quadratic in the number of sentences in the worst case, although the complexity is typically close to linear.

4. Conclusion

The relevance score of computer-extracted summaries for threshold value of τ =0.4 computed from Table 1 is 35%, which is the same as the average relevance score in comparing the control group of human-made summaries. We therefore conclude that the automated summary extraction proved as good as the human one. In addition, human analysis of semantic differences among the sentences in extracted summaries mostly indicated rather similar meaning.

Table 1: Human- and machine-extracted summaries

SELECTOR	International	Domestic	Finance	Education	Social
	SINA1	RENMIN1	XINHUA1	CCTV1	TOM1
A	a, s, t	a, i, k	a, c, k	p, u, v	a, c, d
threshold τ =0.40	(a), (b), (c)	(a), (b), l	b, e, c	(b), (c), n	(a), (b), (c)
threshold τ =0.35	(a), (b), l	a, k, n	b, c, d	e, n, p	(a), (b), (c)
threshold τ =0.30	(a), m, l	c, k, l	b, c, d	e, n, o	(a), (b), (c)
threshold τ =0.25	b, f, m	k, l, s	b, c, g	c, e, n	a, (b), d
	CCTV1	TOM1	SINA1	RENMIN1	XINHUA1
B	a, b, d	a, c, u	a, b, w	a, d, v	d, f, l
threshold τ =0.40	a, c, n	(a), (b), q	(a), b, (c)	(a), (b), f	(a), q, m
threshold τ =0.35	a, c, n	b, o, q	(a), b, c	(a), c, f	k, m, q
threshold τ =0.30	a, c, n	b, o, q	b, c, m	c, f, m	k, m, q
threshold τ =0.25	a, f, n	b, o, q	b, f, m	c, f, p	b, m, q
	RENMIN1	XINHUA1	CCTV1	TOM1	SINA1
C	a, l, p	a, c, k	c, e, o	b, i, q	b, n, t
threshold τ =0.40	(a), d, j	(a), f, j	(a), j, p	d, e, m	b, c, p
threshold τ =0.35	d, i, j	i, j, l	h, j, p	b, e, l	a, b, m
threshold τ =0.30	a, d, j	f, i, j	b, h, p	b, e, m	a, b, j
threshold τ =0.25	d, k, n	b, f, j	b, f, p	b, m, q	a, m, p
	TOM1	SINA1	PEOPLE1	XINHUA1	CCTV1
D	a, c, e	a, l, x	a, d, x	a, b, t	a, l, x
threshold τ =0.40	(a), l, o	b, e, h	a, (b), i	(a), f, b	(a), (b), g
threshold τ =0.35	(a), l, o	d, e, h	a, i, k	a, b, f	b, g, n
threshold τ =0.30	d, l, o	d, e, h	a, e, i	a, b, q	e, g, n
threshold τ =0.25	e, l, u	a, e, o	a, b, q	b, f, q	g, n, r
	XINHUA1	CCTV1	TOM1	SINA1	RENMIN1
E	e, i, w	a, g, t	b, r, v	b, c, v	b, o w
threshold τ =0.40	(a), (b), (c)	(a), (b), (c)	a, i, j	b, c, l	(a), (b), d
threshold τ =0.35	(a), e, l	b, h, k	a, i, (c)	a, c, l	(a), b, n
threshold τ =0.30	(a), e, l	b, h, k	b, i, j	a, b, c	d, e, n
threshold τ =0.25	a, e, l	b, h, k	b, (c), i	b, c, m	d, e, n
	SINA2	RENMIN2	XINHUA2	CCTV1	TOM1
F	a, d, m	d, f, p	d, f, k	b, c, e	a, b, c
threshold τ =0.40	(a), h, o	(a), b, d	(a), (b), k	b, g, i	a, c, d
threshold τ =0.35	a, h, o	b, d, q	c, g, k	b, c, i	a, c, d
threshold τ =0.30	a, h, m	a, b, q	d, k, p	c, (d), i	a, b, d
threshold τ =0.25	a, h, o	b, d, q	d, k, p	c, i *	d, g, m
	CCTV2	TOM2	SINA2	RENMIN2	XINHUA2
G	b, j, t	a, i, t	b, i, w	g, p, w	a, s, v
threshold τ =0.40	(a), (b), c	(a), (b), j	(a), h, m	(a), b, (c)	a, g, n
threshold τ =0.35	c, n, t	(a), (b), j	(a), f, m	b, (c), f	a, b, n
threshold τ =0.30	c, f, h	(a), b, c	a, f, m	b, f, q	a, b, n
threshold τ =0.25	c, h, n	a, b, c	a, c, e	b, f, g	f, i, n
	RENMIN2	XINHUA2	CCTV2	TOM2	SINA2
H	b, d, e	a, b, c	a, b, c	f, i, l	h, j, l
threshold τ =0.40	(a), (b), (c)	(a), b, (c)	b, c, e	a, e, s	(a), h, q
threshold τ =0.35	(a), (b), (c)	(a), b, (c)	b, c, i	(b), e, s	(a), h, q
threshold τ =0.30	(a), b, e	b, c, e	c, h, q	e, g, s	h, i, q
threshold τ =0.25	b, g, e	b, c, e	c, f, h	b, e, j	h, i, q
	TOM2	SINA2	RENMIN2	XINHUA2	CCTV2
I	c, p, r	d, q, t	a, h, u	a, n, u	b, k, s
threshold τ =0.40	(a), (b), (c)	(a), (b), (c)	(a), (b), r	a, d, q	d, i, p
threshold τ =0.35	k, n, o	(a), (b), (c)	(a), b, r	a, d, j	a, i, p
threshold τ =0.30	f, n, o	(a), g, h	c, f, r	a, d, m	a, d, h
threshold τ =0.25	g, k, n	d, g, h	c, f, r	a, d, m	a, d, h
	XINHUA2	CCTV2	TOM2	SINA2	RENMIN2
J	a, b, c	a, b, i	a, c, g	a, b, c	a, b, n
threshold τ =0.40	a, b, c	(a), c, g	(a), (b), f	(a), (b), (c)	(a), (b), (c)
threshold τ =0.35	b, c, k	(b), c, g	a, c, e	(a), (b), c	b, d, k
threshold τ =0.30	b, c, k	(b), c, r	a, c, f	a, c, e	b, k, o
threshold τ =0.25	b, d, i	c, d, r	a, f, m	c, g, e	a, b, o

5. References

[1] T. Strzalkowski (Ed.), Natural Language Inf. Retrieval, Kluwer Academic Publishers, 1999.

[2] H. Myaeng, M. Zhou, K.-F. Wong, H.-J. Zhang (Eds.), Inf. Ret. Technol., Asia Information Retrieval Symposium, AIRS 2004, Lecture Notes in Computer Science 3411 (2005).

[3] U. Hahn, I. Mani, "The challenges of automatic summarization," IEEE Comp. **33** (11) 29-36 (2000).

[4] D. R. Radev and K. R McKeown, "Generating the natural language summaries from multiple on - line sources," Computational Linguistics **24** (3) 469-500, (1998).

[5] Global Reach, "Evolut. of the Online Linguistic Populations," 11. Dec. 2005, http://globalreach.biz/globstats/evol.html.

[6] J. Kleinberg, “Authorit. sources in a hyperlinked environment," Journal of the ACM, **46** (5) 604--632 (1999).

[7] P. A. Kumar, K. P. Kumar, T. S. Rao, P. K. Reddy, "An Improved Approach how to Extract Document Summaries Based on the Popularity," Lecture Notes on Computer Science **3433**, 2005, 310-318.

[8] Summarization applet: http://cpu.icu.ac.jp/cn/