ISSN No:-2456-2165

Extraction of Chinese Health News Using Computation of Noun Numbers

Chou-Cheng Chen
Department of Business Management
CTBC Business School
Tainan, Taiwan (R.O.C.)

Abstract:- Significant amounts of health information can be obtained from Chinese newspapers and magazines, but the reader must spend much time to study this. Common methods of extracting information from articles include machine learning, text mining, word cloud sampling or use of algorithms. A high-quality model of machine learning for extracting information must be trained using a large amount of good data. Before high precision and recall of extracting information is obtained from text mining, many keywords should be collected to identify token sentences. This means that both extracting information from machine learning and text mining take up significant amounts of time. Although word cloud systems can quickly identify which words are widely used in the article, the extracted information is often fragmented. Accordingly, the author has created an elegant algorithm to extract health information from Chinese news using computation of noun numbers. Firstly, the title or subtitle of context from Chinese health news of websites were labeled. Secondly, each sentence was separated via identification of commas, periods, and question marks. Thirdly, word segments of context were tagged as parts of speech via natural language processing. Fourthly, the score of each sentence was identified via computation of the number of nouns where the nouns were identified as 3 points and 2 points as nouns detected in the title and subtitle respectively, while other nouns were identified as 1 point. Finally, high scoring sentences were selected via the query of the user. The result of an example can be downloaded from https://drive.google.com/drive/folders/1gYnOXNNz-gp8t2kejq_-qo7ActLpgywu?usp=sharing, showing its efficiency in extracting information from Chinese health news.

Keywords:- *Extraction of Chinese Health News, Computation of Noun Numbers.*

I. INTRODUCTION

There is a substantial increase in health information from Chinese newspapers and magazines in Taiwan every year. The reader can easily obtain health information by using a mobile phone, tablet, laptop, or desktop computer to browse newspapers and magazines including HEHO, The Liberty Times, China Times, United Daily News, Apple Daily, and SET News Channel, et al. [1-6]. Although the reader must spend much time to study and extract information from context, luckily, there are tools that can help save time.

The tools from two famous companies are Blue Planet and Eland Information in Taiwan, which extract important paragraphs or sentences using Machine Learning [7, 8]. Supervised machine learning for extraction of Chinese health news requires much labeled corpora, but consumes much time for the process of labeling data [9, 10]. The other method of extracting information from context is text mining, but this still takes many steps to run the procedure.

The common methods of text mining for extracting information are tokenization of sentences and words [11]. This requires the collection of many terms from user query and execution of many steps to extract token sentences and words, as a computer cannot educe information instantly [12, 13]. The other method of quick extraction from context is word cloud, but this causes fragmentation of information.

Word cloud of English is a direct and visualized method for extracting information from context, but it is unable to obtain contextual information [14]. Although the process of word cloud in Chinese must firstly recognize word segmentation, it is also a straightforward method for extracting information from context, although the method of word cloud in Chinese is also unable to elicit contextual information. This study thus provides an algorithm that can extract information from Chinese health news instantly using computation of nouns in each sentence.

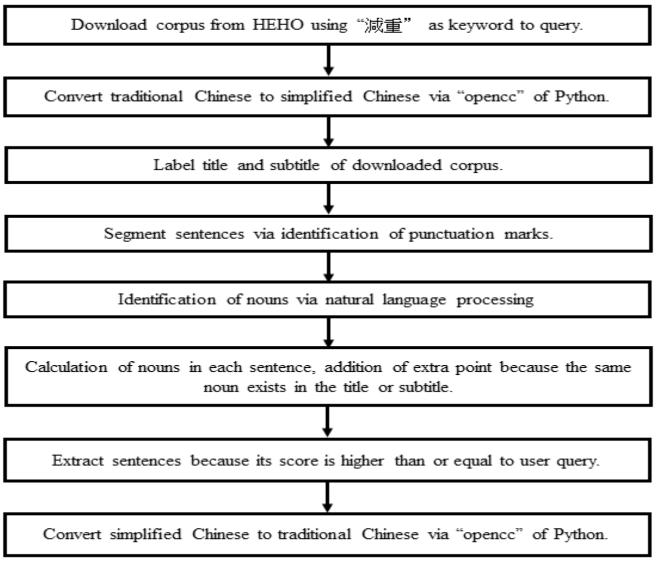
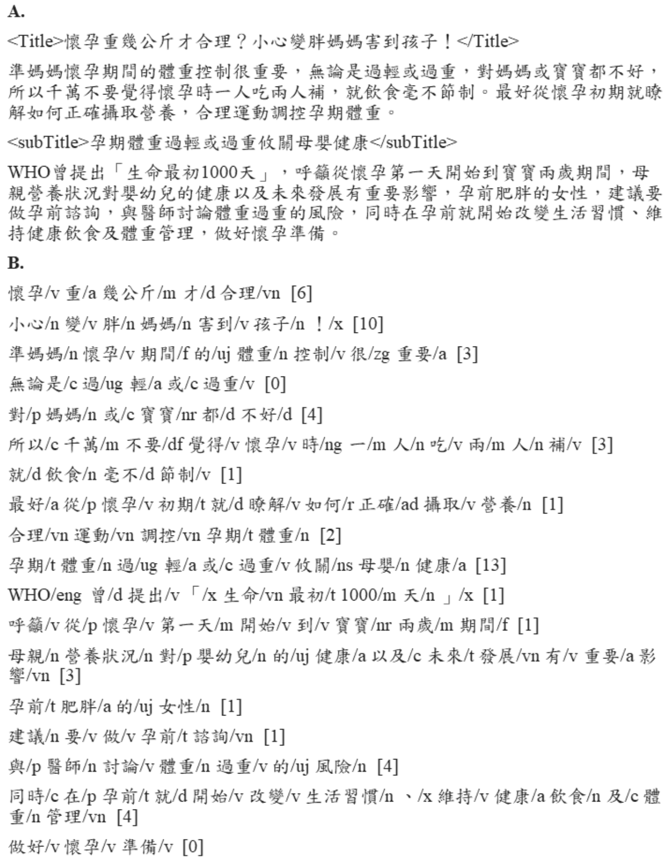

Download corpus from HEHO using “减重” as keyword to query.
Convert traditional Chinese to simplified Chinese via “opencc” of Python.
Label title and subtitle of downloaded corpus.
Segment sentences via identification of punctuation marks.
Identification of nouns via natural language processing
Calculation of nouns in each sentence, addition of extra point because the same noun exists in the title or subtitle.
Extract sentences because its score is higher than or equal to user query.
Convert simplified Chinese to traditional Chinese via “opencc” of Python.

Fig. 1. A flowchart of this study.

The noun as a part of speech is a key piece of information whether in title or context [15]. Any sentence therefore including more nouns is an important sentence. Consequently, this study identifies important sentences using computation of nouns and runs five steps. Figure 1 shows the flowchart of this study. Firstly, the title or subtitle of context from Chinese health news of websites were labeled because the nouns in the title or subtitle are more important than others in context. Secondly, each sentence was separated via identification of punctuation. Thirdly, word segments of context were tagged as parts of speech using "jieba" of python. Fourthly, the score of each sentence was identified via computation of the number of nouns. Finally, sentences were selected via the query of the user identification including 80, 60, 40, or 20 percentiles of scores. To the best of our knowledge, this algorithm can instantly extract key information from Chinese health news without any machine training.



A.

<Title>懷孕重幾公斤才合理？小心變胖媽媽害到孩子！</Title>

準媽媽懷孕期間的體重控制很重要，無論是過輕或過重，對媽媽或寶寶都不好，所以千萬不要覺得懷孕時一人吃兩人補，就飲食毫不節制。最好從懷孕初期就瞭解如何正確攝取營養，合理運動調控孕期體重。

<subTitle>孕期體重過輕或過重攸關母嬰健康</subTitle>

WHO曾提出「生命最初1000天」，呼籲從懷孕第一天開始到寶寶兩歲期間，母親營養狀況對嬰幼兒的健康以及未來發展有重要影響，孕前肥胖的女性，建議要做孕前諮詢，與醫師討論體重過重的風險，同時在孕前就開始改變生活習慣、維持健康飲食及體重管理，做好懷孕準備。

B.

懷孕/v 重/a 幾公斤/m 才/d 合理/vn [6]

小心/n 變/v 胖/n 媽媽/n 害到/v 孩子/n ！/x [10]

準媽媽/n 懷孕/v 期間/f 的/uj 體重/n 控制/v 很/zg 重要/a [3]

無論是/c 過/ug 輕/a 或/c 過重/v [0]

對/p 媽媽/n 或/c 寶寶/nr 都/d 不好/d [4]

所以/c 千萬/m 不要/df 覺得/v 懷孕/v 時/ng 一/m 人/n 吃/v 兩/m 人/n 補/v [3]

就/d 飲食/n 毫不/d 節制/v [1]

最好/a 從/p 懷孕/v 初期/t 就/d 瞭解/v 如何/r 正確/ad 攝取/v 營養/n [1]

合理/vn 運動/vn 調控/vn 孕期/t 體重/n [2]

孕期/t 體重/n 過/ug 輕/a 或/c 過重/v 攸關/ns 母嬰/n 健康/a [13]

WHO/eng 曾/d 提出/v 「/x 生命/vn 最初/t 1000/m 天/n 」/x [1]

呼籲/v 從/p 懷孕/v 第一天/m 開始/v 到/v 寶寶/nr 兩歲/m 期間/f [1]

母親/n 營養狀況/n 對/p 嬰幼兒/n 的/uj 健康/a 以及/c 未來/t 發展/vn 有/v 重要/a 影響/vn [3]

孕前/t 肥胖/a 的/uj 女性/n [1]

建議/n 要/v 做/v 孕前/t 諮詢/vn [1]

與/p 醫師/n 討論/v 體重/n 過重/v 的/uj 風險/n [4]

同時/c 在/p 孕前/t 就/d 開始/v 改變/v 生活習慣/n 、/x 維持/v 健康/a 飲食/n 及/c 體重/n 管理/vn [4]

做好/v 懷孕/v 準備/v [0]

Fig. 2. An example of calculation process. A. Shows some original sentences from the article. B. Shows calculation of noun numbers after the part of speech. The default score of titles is six points, and of subtitles is four points. If "小心", "胖", "媽媽" and "孩子" occurred in other sentences, each identification of word was given three points because of nouns from the title. If "體重", "攸關" and "母嬰" occurred in other sentences, each identification of word was given two points because of nouns from the subtitle. Other nouns in the sentence were given one point. The total score of each sentence is shown in square brackets.

II. METHOD

A. *Collection of Chinese health news*

The 319 contexts of Chinese health news from HEHO were collected using keyword "減重" of query in 2022/3/14 [1]. Each context only kept words and was converted to simplified Chinese via "opencc" of python. "<title>" and "<subtitle>" were inserted before title and subtitle of context respectively, and "</title>" and "</subtitle>" were inserted behind title and subtitle of context respectively.

B. *Sentence segmentation*

"， ", "？", "。\n", "。" and "\n" were replaced with "<|，>", "<|？>", "<|o>", "<|。>" and "<|n>". Each sentence was separated via identification of "<|". All symbols were replaced with punctuation marks after sentence segmentation was performed.

C. *Natural language processing*

All sentences were labeled as identified parts of speech via "jieba", and the keywords of nouns from title and subtitle context marks were used to identify whether they occurred in other sentences.

D. *Computation of noun numbers*

All nouns including "n", "ng", "nr", "nrfg", "nrt", "ns", "nt" and "nz" label were identified and counted for score calculation. The nouns were identified as three points and two points when detected in the title and subtitle respectively, while other nouns were identified as one point. The title and subtitle had six and four points added after computation of noun numbers respectively.

E. *Selection of extracting information*

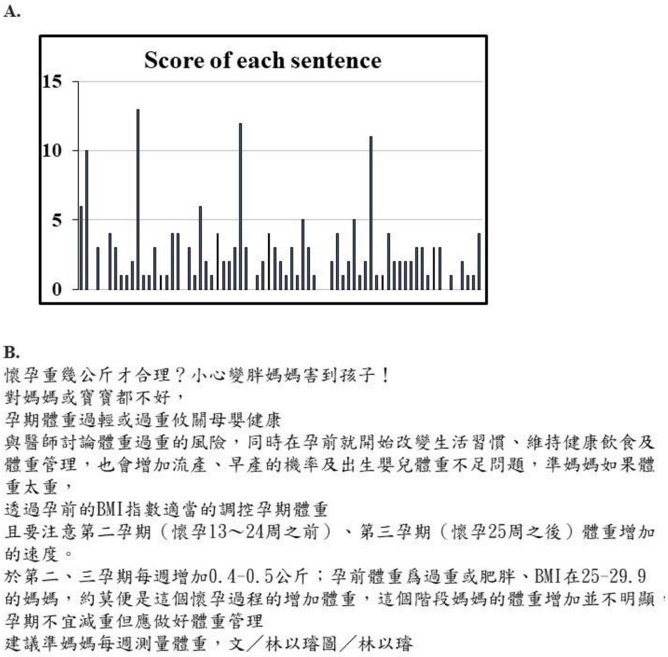
Score of each sentence were identified as percentiles via "numpy" of python, and the user could select sentences via querying the score percentile. Figure 2 shows the scores of calculations from some sentences of an article, and this context can be downloaded from https://heho.com.tw/archives/37007. The sentences from user selection were composed and converted to traditional Chinese, and all punctuations were reverted to original marks.

III. RESULT

The result of context extraction can be downloaded from https://drive.google.com/drive/folders/1gYnOXNNz-gp8t2kejq_-qo7ActLpgywu?usp=sharing. It shows important information can be selected via this algorithm. Figure 3A shows scores of each sentence from an article of https://heho.com.tw/archives/37007, and 3B shows results of 20% concentration, while 80%, 60%, 40% and 20% of context extraction is identified via 20%, 40%, 60% and 80% scores respectively. This study shows computation of nouns in sentences that can efficiently extract information from a body of work.

IV. CONCUSION

To our best knowledge, this study provides another efficient method to extract key information from Chinese health news.



懷孕重幾公斤才合理？小心變胖媽媽害到孩子！
對媽媽或寶寶都不好，
孕期體重過輕或過重攸關母嬰健康
與醫師討論體重過重的風險，同時在孕前就開始改變生活習慣、維持健康飲食及體重管理，也會增加流產、早產的機率及出生嬰兒體重不足問題，準媽媽如果體重太重，
透過孕前的BMI指數適當的調控孕期體重
且要注意第二孕期（懷孕13～24周之前）、第三孕期（懷孕25周之後）體重增加的速度。
於第二、三孕期每週增加0.4-0.5公斤；孕前體重爲過重或肥胖、BMI在25-29.9的媽媽，約莫便是這個懷孕過程的增加體重，這個階段媽媽的體重增加並不明顯，
孕期不宜減重但應做好體重管理
建議準媽媽每週測量體重，文／林以璿圖／林以璿

Fig. 3. An example of result. A. Shows each score of sentences via algorithm. Each bar shows a point of each sentence. B. Shows 20 percent of sentence extraction. The 80th percentile of score is four, and 20% of extraction sentences were identified because scores were higher than or equal to four points.

REFERENCES

[1]. HEHO. https://heho.com.tw/ (accessed)
[2]. T. L. Times. https://health.ltn.com.tw/ (accessed)
[3]. C. T. Group. https://www.chinatimes.com/health/?chdtv (accessed)
[4]. U. D. N. Group. https://health.udn.com/health/index (accessed)
[5]. A. ONLINE. https://tw.feature.appledaily.com/health/ (accessed)
[6]. S. E-Television. https://www.setn.com/Catalog.aspx?PageGroupID=65 (accessed)
[7]. B. P. Inc. "Digital Archives." https://www.blueplanet.com.tw/solutions/digitalarchives (accessed)
[8]. L. Eland Information Co. https://www.eland.com.tw/ (accessed)
[9]. J. Qiu, L. Liao, and P. Li, "Relation Extraction from Chinese News Web Documents Based on Weakly Supervised Learning," presented at the 2009 International Conference on Intelligent Networking and Collaborative Systems, Barcelona Spain, 2009, 219. [Online]. Available: https://ieeexplore.ieee.org/document/5370952?section=abstract
[10].
[11]. Z. Cai, T. Zhang, C. Wang, and X. He, "EMBERT: A Pre-trained Language Model for Chinese Medical Text Mining," Cham, 2021: Springer International Publishing, in Web and Big Data, pp. 242-257.
[12]. D. Johnson. "NLTK Tokenize: Words and Sentences Tokenizer with Example." https://www.guru99.com/tokenize-words-sentences-nltk.html (accessed)
[13]. H. J. Lee, T. C. Dang, H. Lee, and J. C. Park, "OncoSearch: cancer gene search engine with literature evidence," Nucleic Acids Res, vol. 42, no. Web Server issue, pp. W416-21, Jul 2014, doi: 10.1093/nar/gku368.
[14]. H. Hao and K. Zhang, "The Voice of Chinese Health Consumers: A Text Mining Approach to Web-Based Physician Reviews," J Med Internet Res, vol. 18, no. 5, p. e108, May 10 2016, doi: 10.2196/jmir.4430.
[15]. F. Heimerl, S. Lohmann, S. Lange, and T. Ertl, "Word Cloud Explorer: Text Analytics Based on Word Clouds," presented at the 2014 47th Hawaii International Conference on System Sciences, 2014, 1833.
[16]. J. Chen and H. Zhuge, "Automatic generation of related work through summarizing citations," Concurrency and Computation: Practice and Experience, vol. 31, no. 3, p. e4261, 2019.