International Journal of Multimedia and Ubiquitous Engineering
Vol.11, No.6 (2016), pp.379-388
http://dx.doi.org/10.14257/ijmue.2016.11.6.34

Research of Theme Statement Extraction for Chinese Literature Based on Lexical Chain

WanLi Feng

*School of Computer and Software Engineer, Huaiyin Institute of Technology,
HuaiAn, JiangSu, 223003, P. R. China
fengwanli@hyit.edu.cn*

Abstract

The paper proposes an extraction approach for short Chinese documents topic sentences based on lexical chains and context technology, with which narrative clues of documents are found out first. Multiple lexical chains are constructed subsequently before the extraction of a topic-information-rich one as the keyword sequence of topic sentences. Experimental results demonstrate that this approach covers information of documents to the greater extent, and that it avoids possible repetitive expression that is caused by different keyword sequences in expressing the same topic sentences. Therefore, it achieves an obviously better effect than the statistic approach.

Keywords: *Chinese Literature; lexical chain; theme statement extraction*

1. Introductions

Chinese documents topic sentences extraction is widely applied to automatic categorization, automatic summarization, and automatic indexing. Not only is it an indispensable underpinning and prerequisite to the above operations, but it also plays a significant role in constructing internet database. The statistic approach is commonly used in topic sentences extraction. It determines weights of candidate keywords, and sifts through them for the phrase with the largest weight as the final topic keyword. Therefore, the core of document keywords determination is to calculate the weight of candidate keywords, which is determined by the topic it reflects. In general, the better a phrase reflects topics of documents, the larger weight it gains. According to previous studies, word frequency and word location take important effect in documents topic determination. They are also referred to in massive schemes to calculate candidate phrases weight, resulting being unsatisfactory though. Currently, most efforts to extract keywords are done in determining candidate keywords weight based on phrase statistics and in selecting the keyword of documents from candidate keywords beyond certain threshold. The documents topic keywords extraction approach proposed by Wang Jun is extraction of keywords for automatic indexing from indexed structural Chinese corpus [1]. According to the position of phrases and their statistic feature, Zhang Yonggang, *et al.* proposed to extract automatically documents keywords from Chinese documents based on statistic technology [2]. As this requires massive statistical work on numerous Chinese documents, the calculation is great in amount. In addition, highly frequent phrases may not necessarily expound topics of documents, which mean that the keywords extracted from them may also lack precision. Hirst initialized an approach relating to lexical chain that is constructed by similar or relevant phrases [3]. Corresponding relation to the documents structure enables the lexical chain to provide important clues of the structure and topic. Various constructing methods of lexical chain are developed subsequently. Given differences between Chinese and English phrases, Suo Hongguang *et al.* proposed such construction algorithm of Chinese lexical chains as primarily choosing the most matchable phrase among current constructed lexical chains as the keywords of the topic.

Since this algorithm takes into account semantic relationships among phrases rather than the documents as a whole, the constructed lexical chain fails to express precisely the semantic structure of documents [4]. As a result, it may easily cause semantic misjudgment. Since each short Chinese document has a single topic at large, the paper determines lexical chains in the foundation of corresponding relationship between documents' structure and lexical chains as well as context technology, and selects keywords of the topic among information-rich phrases that are can represent the topic. The results demonstrate that the lexical chain approach can determine topics of short Chinese documents in a more effective manner.

2. Extraction of Short Chinese Documents Topics

2.1. Construction of Lexical Chains

The paper considers it that the meaning of a polysemous word is mostly determined by its contexts that are determined herein as the words before and behind it, and obtains the word's semantics in its contexts in this way. There are over 1500 sememes defined by HowNet. In the tree structure formed by them, sememes with closer semantics stay closer to each other. They are divided into two categories. One is fundamental sememes, mostly connoting DEF that expresses semantic information of words; the other is relational sememes, which stands for the structual feature among different DEF [5]. For example, the DEF of the word "university" is equal to {InstitutePlace|场所: domain={education|教育}, modifier={HighRank|高等}, {study|学习:location={~}},{teach|教:location={~}}}. It means that the "university" is a place in the educational domain, being of high rank for people to "study and teach", where InstitutePlace|场所, education|教育, HighRank|高等, study|学习 and teach|教 are fundamental sememes, while words including domain, location and modifier are relational sememes. With a collection of multiple DEF sets, a polysemous word has its semantics determined generally in HowNet by being projected to only one DEF in the sets.

Definition 1: In the Chinese documents vector sets $SM|(w_1,w_2,\cdots,w_m)$ it is assumed that φ_{is} and $\varphi_{jt}$are respectively the sememe of polysemous words w_i and w_j. w_i and w_j belongs to the same semantic category if their collections of fundamental sememes are in the same set.

The significance of definition 1 is to construct semantic categories that satisfy equivalence relations based on it that a word's connotation is determined by its fundamental sememes that subsume it in its DEF. The paper obtains semantics of a polysemous word by finding the right semantic category of it from its multiple semantic categories. The approach to constructing Chinese documents lexical chain is to search the contextual information w_i of a wordw_i, which is in Chinese documents, in a way that produces great affinity between one of the semantic categories of w_j and that of w_j.

Definition 2: The affinity between the semantic category φ_{is} and φ_{jt} is 1 if their relational sememes contain their fundamental sememes, or 0 if not.

Since semantic categories have their main information expressed by fundamental sememes and their relational feature by relational sememes, this definition is able to demonstrate whether the projected semantic information betweenφ_{is} and φ_{jt}has affinity or not [6].

Definition 3: If $CR(\varphi_{is})$ represents the total amount of relational sememes ofφ_{is}, then the degree of affinity between relational sememes of $\varphi_{is}$and that of $\varphi_{jt}$is as follows:

$$CD(\varphi_{is},\varphi_{jt}) = \frac{IR(\varphi_{is},\varphi_{jt})+IR(\varphi_{jt},\varphi_{is})}{CR(\varphi_{is})+CR(\varphi_{jt})} \tag{1}$$

The eq (1) expresses the degree of affinity between the semantic information of two different semantic categories, including subordination and dominance relationship.

Definition 4: If $SI(\varphi_{is},\varphi_{jt})$ represents the intersection size of the semantic sememes sets between φ_{is} andφ_{jt}, then the degree of affinity between the two relational sememes is as follows:

$$\mathrm{SR}(\varphi_{is},\varphi_{jt})=\frac{\mathrm{SI}(\varphi_{is},\varphi_{jt})}{\mathrm{CR}(\varphi_{is})+\mathrm{CR}(\varphi_{jt})} \tag{2}$$

In the above eq(2) the $SR(\varphi_{is},\varphi_{jt})$ describes the degree of affinity between the relational sememes of two semantic categories, aiming at measuring the relational structure of semantic categories.

Definition 5: If μ_k represents one of the fundamental sememes ofφ_{is}, and $dep(\mu_k)$ represents the depth of $\mu_k$in the sememes tree, then the difference of levels between φ_{is} and φ_{jt} is as follows:

$$LD(\varphi_{is},\varphi_{jt})=\begin{cases}\sum_{\mu_k\in\varphi_{is},V_l\in\varphi_{jt}}|Dep(\mu_k)-Dep(V_l)|\ , If\ \mu_k\ ,V_l\ are\ in\ the\ same\ tree\\ \infty \qquad otherwise\end{cases} \tag{3}$$

The eq(3) shows the difference of levels of the fundamental sememes sets between two semantic categories, where the closer the semantics between two semantic categories are, the smaller difference of levels their semantic categories have [7].

Definition 6: The degree of similarity of fundamental sememes of two semantic categories is as follows:

$$\mathrm{BS}(\varphi_{is},\varphi_{jt})=\frac{1}{\mathrm{LD}(\varphi_{is},\varphi_{jt})+1} \tag{4}$$

The eq(4) shows that the degree of similarity of fundamental sememes of two semantic categories is measured by their positions in the sememes tree.

Definition 7: the similarity degree of semantics of two semantic categories φ_{is} and $\varphi_{jt}$is as follows:

$$\mathrm{SS}(\varphi_{is},\varphi_{jt})=\lambda\times\mathrm{BS}(\varphi_{is},\varphi_{jt})+(1-\lambda)\times\mathrm{SR}(\varphi_{is},\varphi_{jt}) \tag{5}$$

In the eq(5) λ is a parameter to adjust the similarity degree between fundamental sememes and relational sememes in semantic categories. Given that fundamental sememes can better project the semantic information of words, the value of λ is generally set to be or be above 0.6 in the experiment [7-8].

Definition 8: the degree of affinity between $\varphi_{is}\ and\ \varphi_{jt}$ is as follows:

$$R(\varphi_{is},\varphi_{jt})=\frac{\mathrm{CD}(\varphi_{is},\varphi_{jt})+\mathrm{SS}(\varphi_{is},\varphi_{jt})}{2} \tag{6}$$

The eq(6) expresses the average value of similarity degree and affinity degree between semantic categories. It reflects how close two semantic categories may be.

2.2. Construction Algorithm of Lexical Chains

The steps of construction of short Chinese documents lexical chains are as follows:

Step 1: determine the contexts of the word w_i

Step 1-1: Set each word w_i in the Chinese documents vector $(w_1,w_2,\cdots,w_m)$ as one of the vertexes of a graph. Two vertexes, if their corresponding words have the semantic affinity greater than the threshold θ, are connected by a line with its weight as the similarity degree between two semantics. Then the Chinese documents vector $(w_1,w_2,\cdots,w_m)$ constructs a unidirectional figure G with weight in it as shown in Figure 1.

Step 1-2: For Figure G, if the relationship between one of the semantic categories φ_{is} of the word w_i and one of the semantic categories φ_{jt} of the word w_j satisfies the conditions of $R(\varphi_{is}, \varphi_{jt}) \geq \theta$, then w_j is chosen as the contextual information of w_i

Step 2: construct Chinese documents lexical chains

Step 2-1: Take each of the connected components of graph$G$. Assume that from the first word of the Chinese documents vector w_1 (or the second word w_2 if the first word is an isolated point or is already subsumed in a lexical chain and so on) there has already been certain lexical chain L_p that contains words with a number of p. Next, choose the context w_p of the final word L_p to be linked onwards. According to step 1-2, there are possibly numerous w_p, which is taken based on such method. Let the semantic sequences L_p is $\varphi_{i_1 1}, \varphi_{i_2 2}, \cdots, \varphi_{i_k k}$ and φ_{pl} is the semantic category $w_p$determined by step 1-2. According to equation (8), the similarity degree between w_p and L_p is computed as follows:

$$LSim(L_p, w_p) = \max_{n=1}^{p} \{SS(\varphi_{it}, \varphi_{pl})\} \tag{7}$$

Step 2-2: Repeat step 2-1 until there are no more untapped vertexes in Figure G that meet the above conditions.

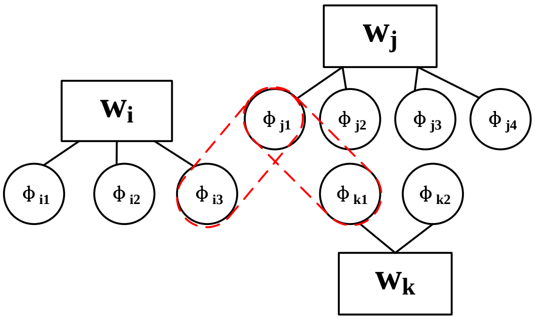

Figure 1. The Diagram of Similarity Degree of a Chinese Document

2.3. Extraction Approach to Topic Sentences

For specific Chinese documents, there are many lexical chains constructed based on Section 1.1 herein, the one that can project more Chinese documents information is then extracted for its words collections to be the keywords collections in the documents. Assume w as words in the Chinese documents lexical chains, and $H(w_i)$as the weight of the word w_i defined in Document [9-10], then the weight of the Chinese documents lexical chain Le is as follows:

$$WL(Le) = \sum_{i=1}^{|Le|} H(w_i) \times log_{2}^{(|Le|)} \tag{8}$$

From eq (8) it is seen that the weight of lexical chain has relations with their word numbers and word weight. The more word numbers the lexical chain contains, the broader distribution the lexical chain has for its information. And the heavier weight the lexical chain contains, the more significant the information is. The algorithm herein uses eq(8) to choose words set with the heaviest weight in the lexical chain as the collection of keywords that construct topic sentences, and then constructs topic sentences under the construction principles of characteristics of words in [11]. The flowchart of Chains for auto getting keywords is depicted as shown Figure 2.

3. Analyses of the Algorithm and Experimental Results

3.1. Analysis of the Algorithm

The topic sentences extraction approach herein is divided into two parts: construction of lexical chains and construction of topic phrases. The time complexity in obtaining lexical meaning: according to the algorithm, in order to obtain lexical meaning, it is necessary to scan word space in the documents in a forward sequence, and to deal with multiple semantic categories of the word successively. Assume the Chinese documents word space given in document [7] is n, and the utmost total semantic category of the word in HowNet is k, then the number of vertexes in diagram G is $k \times n$ [12-13]. As there are at most n lexical chains in the Chinese documents, the time complexity to obtain word meanings is$O(n^2)$. In order to construct topic sentences, the first matter is to scan the n lexical chains, and determines the collection of keywords to construct topic sentences. According to the above analysis, the time complexity is $O(n^2)$. Therefore, the total time complexity is computed as $O(n^2)$ [14].

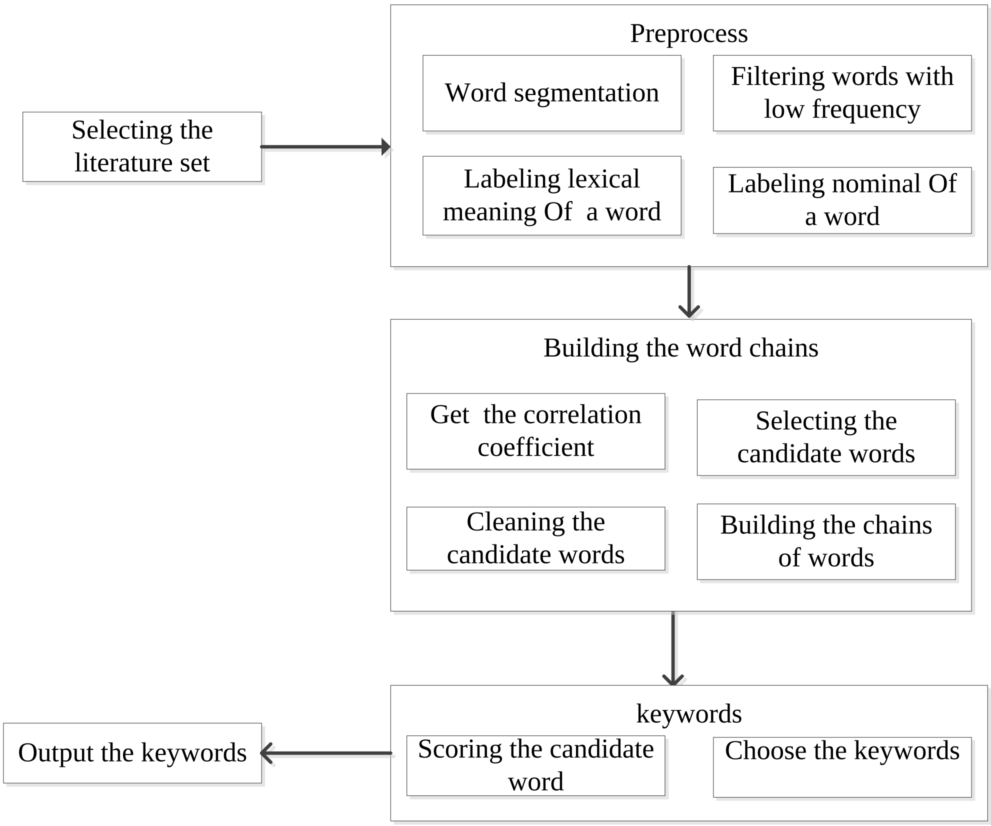
Preprocess
Word segmentation
Filtering words with low frequency
Selecting the literature set
Labeling lexical meaning Of a word
Labeling nominal Of a word
Building the word chains
Get the correlation coefficient
Selecting the candidate words
Cleaning the candidate words
Building the chains of words
keywords
Output the keywords
Scoring the candidate word
Choose the keywords

Figure 2. The Flowchart of Chains for Auto Getting Keywords

3.2. Experiment Results and Analysis

The paper takes a short Chinese document as an example to analyze the construction of lexical chains and the results of keyword extraction, and compares it with the keywords extracted according to the statistic information. This short Chinese document is as follows:

Congratulations! The account number of your Fetion has been drawn as our online lucky user for this National Day in 2008. You will get a surprising 58,000 RMB and a Samsung Q30 laptop from the head office. Please log in FEIXIN18.CN first. For the prize and money award, please contact 089888155951. Your verification code is 6108. This is sent as a system message.

In order to facilitate the description, this document is named as Fetion document, and its keyword set that is extracted from the statistic information is: Fetion, lucky, prize, head office, account, Samsung, surprising, money award, National Day, laptop, verification code. Parts of the lexical chains that top the outputted weight as the lexical chain is constructed include “lucky user” and “laptop”. When the algorithm is over, the final outputted topic sentences are “lucky user”, “prize”, “money award”, “Samsung” and “laptop”.

The Table 1 summarizes the above results, its firstly column presents the algorithm name and the following columns respectively are the keywords found by the corresponding algorithm. As Table 1 shown that that due to lack of analysis of topic sentences, the keyword extraction algorithm based on statistic information produces some highly frequent word that cannot reflect its topic sentence, such as “Fetion” and “lucky”. Meanwhile, it may produce too many similar words that lead to repetitiveness of information. Whereas the topic sentences extracted by the lexical chain approach cover information of documents to the greater extent, and avoid possible repetitive expression as different keyword sequences express different topic sentences.

Table 1. The Result Produced by Statics and Chains for Demo Short Chinese Document

Algorithms	Keywords						
	1	2	3	4	5	6	7
Statics	lucky	Fetion	Prize	Samsung	account	Money	--
Chains	Fetion	User	Laptop	Award	Surprising	Account	Lucky

The key to construction of topic sentences is the extraction of keyword sets. However, there is great subjectivity in judging extraction results, and different people may get different results even from the same Chinese document. Thus it is difficult to find any document as evaluation corpus for practical application. The paper chooses 500 Chinese documents with clearer boundaries as testing corpus, expanding from festival greetings, fraud and false information, illegal advertisements, school lives, to pornography [15-16]. It also indicates the topic for each of the Chinese documents by hand. Comparison is made for Chinese documents topic extraction experiments between the approach herein (marked as Lexical chains) and the statistic approach in [17] (marked as Statistics).

3.2.1. Analysis of Time Complexity

For explaining the time performance of Lexical chains and Statistics, the literature set which is from HIT [18] is chosen and test by the two algorithms, its result are presented as Table 2 and Figure 2. In Table 2, the firstly column shows the name of chosen algorithm which is the same as that of Table 2 and the following columns respectively shows the mean and associated standard variance of the runtimes under the different sample numbers. As Table 2 shown that with the increase of the sample number, the mean runtime gradually increase which fully demonstrate that the algorithm of Chains is far better that of Statics. Figure 3 shows different time spent for Lexical chains (5.08s) and Statistics (4.42s) in the experiment on 500 Chinese documents that have clear boundaries and are classified into five categories.

Both of the time complexities of Lexical chains based on lexical chains and Statistics are $O(n^2)$. The time spent for Statistics is a bit shorter than that for Lexical chains when the number of the sample is over 100. This is because those candidate keywords wait to be determined when topic sentences are extracted under the construction principle of characteristics of words in [19]. Thus the steps of Lexical chains exceed those of Statistics.

Table 2. The Mean and Standard Variance of Runtime(S) about the Statics and Chains in Various Sample Number

Algorithms	Sample(100)	Sample(200)	Sample(300)	Sample(400)	Sample(500)
Statics	1.10±0.02	2.11±0.03	2.56±0.01	4.12±0.02	5.08±0.03
Chains	1.02±0.01	1.91±0.01	2.44±0.02	3.78±0.02	4.42±0.02

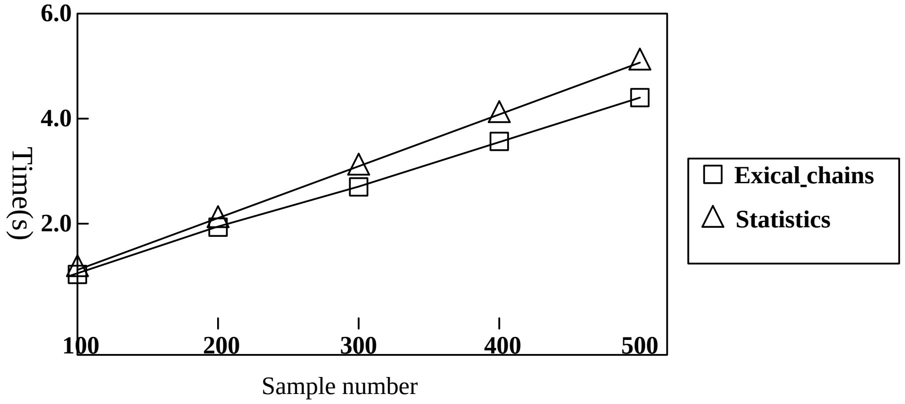

Figure 3. The Comparison of Time Spent Between the Algorithm of Lexical Chains and Statistics

3.2.2. Analysis of Accuracy Rating of Chinese Documents Topic Extraction

The accuracy rating of Chinese documents topic extraction can reflect the property of the algorithm directly if it is defined as the division of the total topic numbers outputted from this algorithm by the topic numbers that conform to the manually marked ones [20]. The accuracy rating between Lexical chains and Statistics is shown in Table 3 and Figure 3. The Table 3 presents the accuracy rate of Statics and Chains in various sample number in which the means of each column is the same as that of Table 2, and the mean accuracy rating of each algorithm is almost increasingly with the increment of the sample number. However, there is exception that the accuracy rate of Statics is decreasing when the sample number is from 200 to 300 while that of Chains is decreasing when the sample number from 300 to 400. Figure 4 presents the content of Table 3 in line chart. As shown Figure 4, it is seen that the line of Statistics oscillates up and down in a great magnitude. The reason for this phenomenon is that it fails to observe the topic distribution of Chinese documents when being extracted, and thus cannot totally cover topic information with its keywords [21]. Whereas the Lexical chains line has a general upward trend, which is because it covers topic information well by avoiding disadvantages of Statistics on one hand, and on the other hand choosing the most descriptive lexical chain among those subsumed in the same topic information so that it escapes from repetitive information.

Table 3. The Mean Accuracy Rating (%) of Statics and Chains in Various Sample Number

Algorithms	Sample(100)	Sample(200)	Sample(300)	Sample(400)	Sample(500)
Statics	81.12%	83.12%	86.23%	85.20%	86.79%
Chains	70.01%	72.36%	68.21%	75.63%	72.31%

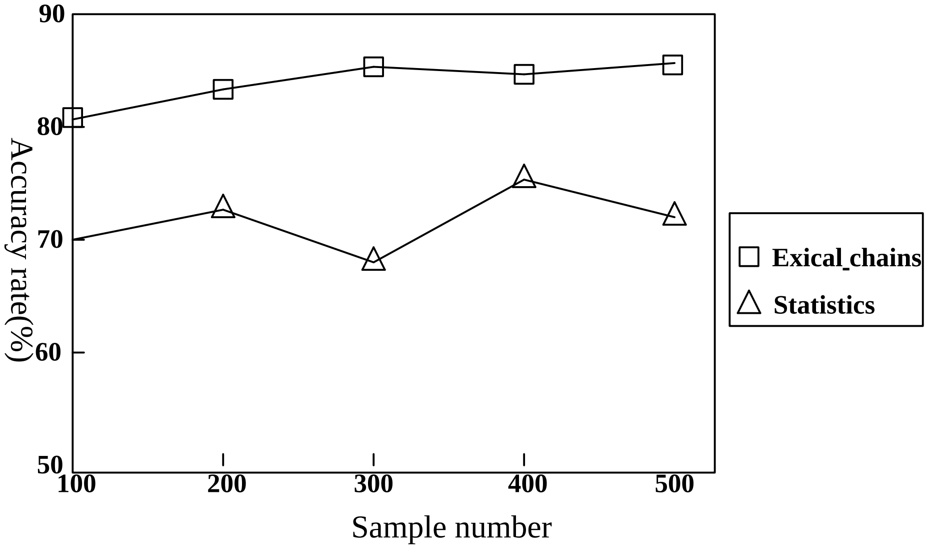

Figure 4. The Comparison of the Accuracy Rating of Lexical Chains and Statistics

For comprehensively comparing the performance of two chosen algorithm, the criteria of similarity degree for Chinese document (LSim), the weight of the Chinese documents lexical chain (WL), Recall, Accuracy and the keyword number are selected in which LSim the mean similarity of word vector, WL is the weight coefficient, Recall is the percentage of all correct information or related information detected which is also called as the true positive rate, Accuracy is the percentage of corrected information that are relevant. Table 4 shows that the method of Chains is basically better than that of Statics in all criterions. However, there is an exception in Accuracy in which the Statics is better than that of Chains. This is because that the criteria of Accuracy and Recall are conflict and the value of Accuracy is low when the value of Recall is high.

Table 4. The Mean Index of the Statics and Chains Algorithms in HIT Set of the Short Literatures

Algorithms	$LSim$(%)	WL(%)	Recall (%)	Accuracy (%)	#Keyword
Statics	69.25	58.32%	92.47%	84.23%	12.25
Chains	75.23	62.98%	94.87%	83.25%	15.37

4. Conclusions

The paper proposes a keyword extraction approach based on lexical chain. It covers information of Chinese documents to the greater extent, and has little repetitive expression. Experimental results show that the information content of Chinese documents topics by this approach far outweighs that by Statistics. Such approach probes into the semantic level of Chinese documents to find its deep topic information, thus achieving a more ideal effect.

The experimental results support the claim that Lexical chain is valuable for learning to extract keywords from short Chinese text. A keywords extraction algorithm incorporating such specialized knowledge (word chains) performed significantly better than an algorithm without such knowledge such as statistics. This extractor method can make a keyword list for an average short text and the speed of extractor makes it possible to be used in applications where it would not be economically feasible to use human-generated keywords. Subjective human evaluation of the keywords generated by this method suggests that about 94.87% of the keyword is acceptable to the readers. This level of performance should be satisfactory for a wide variety of applications.

However, a drawbacks of our approach is that many synonyms are often used in short Chinese text for avoid boring the reader with repetition and increasing the diversity and flexibility of Chinese. Unfortunately, the repetition of synonyms is a major clue for Lexical chain that a candidate word is a keyword. We believe that the results could be significantly improved by adding some methods of synonym detection to the keywords extraction algorithms. The current version of Lexical chain works with documents in the simplified Chinese documents. We are currently working on adding traditional Chinese as the future works.

Acknowledgments

This research was supported in part by the HuaiAn science technology program under grant HAG09061, HAC2015008 and 491712124.

References

[1] W. Jun, “Updating Thesaurus via Extracting Keywords from Metadata”, Journal of Chinese Information Processing, vol. 19, no. 6, **(2005)**, pp. 36-42.

[2] Z. Yonggang, L. Yinghong, Y. Zhenxiang and Y. J. Min, “Research On Statistics- Based Automatic Extraction Of Chinese Keyphrase”, Journal of Jiangnan University (Natural Science Edition), vol. 9, no. 1, **(2010)**, pp. 26-29.

[3] A. Hirst, “Beata B.M.A Study on Automatically Extracted Keywords in Text Categorization. Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics”, Sydney. Australia, **(2006)**, pp. 537-544.

[4] E. Gonenc and C. Ilyas, “Using Lexical Chains for Keyword Extraction. Information Processing and Management”, vol. 43, no. 6, **(2007)**, pp. 1705-1714.

[5] S. Guang, L. Shu and C. Ying, “A Keyword Selection Method Based On Lexical Chains”, Journal of Chinese Information Processing, vol. 20, no. 6, **(2006)**, pp. 25-30.

[6] L. Z. Mao, L. Ting and L. Sheng, “The Research Progress of Statistical Word Sense Disambiguation”, Acta Electronice Sinice, vol. 34, no. 2, **(2006)**, pp. 333-343.

[7] L. J. Ling, “High Quality Algorithm For Chinese Short Messages Text Clustering Based On Semantic”, Computer Engineering, vol. 35, no. 10, **(2009)**, pp. 201-205.

[8] L. Sujian, “Research of Relevancy between Sentences Based On Semantic Computation”, Computer Engineering and Application, vol. 38, no. 7, **(2002)**, pp. 75-76.

[9] E. Gonenc and C. Ilyas, “Using Lexical Chains Foreword Extraction”, Information Processing and Management, vol. 43, no. 6, **(2007)**, pp. 1705-1714

[10] L. J. Ling, “Dimensionality Reduction of Short Message Text Classification and Thematic Extraction of Semantic”, Computer Engineering and Applications, vol. 46, no. 23, **(2010)**, pp. 159-161.

[11] L. Y. Chao, W. X. Long, X. Z. Ming and L. B. Quan, “Mining Construction Rules of Chinese Keyphrase Based on Rough Set Theory”, Acta Electronice Sinice, vol. 35, no. 2, **(2007)**, pp. 371-374.

[12] T. Weidong and Z. Yongliang, “Answer Extraction Scheme Based On Answer Pattern and Semantic Feature Fusion”, Computer Engineering And Applications, vol. 47, no. 13, **(2011)**, pp. 127-130.

[13] S. Gaofeng and G. Shumin, “Chinese Web Page Feature Extraction By Optimizing Comprehensive Heuristics Based on GA”, CAAI Transaction On Intelligent Systems, vol. 9, no. 4, **(2014)**, pp. 474-479.

[14] Z. Ling, R. Han and W. Jing, “Automatic Ontology Construction Based On Clustering Nucleus”, Wuhan University Journal of Natural Sciences, no. 2, **(2015)**, pp. 167-174.

[15] W. Xiaomei, H. Sixing, C. Bo and J. Donghong, “Biotsa: Annotating Token Semantic Association To Support Biomedical Text Mining”, Wuhan University Journal of Natural Sciences, no. 2, **(2015)**, pp. 210-218.

[16] Z. Jing, L. Chaozhen, J. Donghong and L. Xiaohui, “Framework Construction and Application for Global Health Information Platform”, Wuhan University Journal of Natural Sciences, no. 2, **(2015)**, pp. 228-236.

[17] D. M. Blei, A. Y. Ng and M. I. Jordan, “Latent Dirichlet Allocation”, Journal of Machine Learning Research, vol. 3, **(2003)**, pp. 993-1022.

[18] L. Zhiyuan, “Research on Keyword Extraction Using Document Topical Structure”, Beijing: Tsinghua University, **(2011)**.

[19] L. Gang and D. Qiangbin, “Keywords Automatics Indexing Based On Lexical Chains”, Document Information Knowledge, vol. 12, no. 3, **(2011)**, pp. 67-71.

[20] Z. Zheng, W. Lei and L. Huan, “On Similarity Preserving Feature Selection”, IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 3, **(2013)**, pp. 619-632.

[21] W. Xindong, Y. Kui, D. Wei, H. Wang and X. Zhu, “Online Feature Selection with Streaming Features”, IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 35, no. 5, **(2013)**, pp. 1178-1192.

Authors

Wanli Feng, was born in 1973, received his B.S. degree in Software Engineering from Tsinghua University, Beijing, China in 2003, his M.S. degree in computer science and technology from Southeast University, Nanjing, China in 2010. He is an associate professor at Huaiyin institute of technology, Huai’an, China. His current research interests include image process, software design and data mining.