Automatic Keyphrase Extraction System for Establishing Text-Significant Concept Relationship

[*]Cong-Xun XIE [a], Hsueh-Fu LU[a], Cheng-Ting CHEN[b], Sheng-Yung CHENG[a] & Jia-Sheng HEH[a]

[a]*Department of Imformation and Computer Engineering,*
[b]*Department of Applied Linguistics and Language Studies* ,
Chung Yuan Christian University, *Taiwan*
*natogood@gmail.com

Abstract: This paper explores several unsupervised approaches to automatic keyword extraction using online news articles. After utilizing Stanford Parser to extract keywords and key phrases, SingleRank calculation method was employed to assign specific scores for key phrases in a single document. After keyphrase extraction and evaluation, a graphic structure can be established which is named TSCR (Text-Significant Concept Relationship) structure system. In order to investigate the effectiveness of the system for assisting English reading comprehension, a questionnaire was conducted, and the results confirm the positive influence of the system.

Keywords: Machine learning, keyphrase extraction, TextRank

1. Introduction

Keywords in a document provide important information about the content of the document. They can help users search easily and decide whether the information was needed. They can also be used for any text document title or any language processing task such as text categorization and information retrieval. When reading documents or papers, keywords can also be important words marked by users. After reading, whenever the users want to review the document or paper, they just need to scan these marked words which will remind them quickly of what the document or paper is about.

Many efforts have been made toward keyword extraction for text domain, but less works on how to use them for language learning purpose. So far, most frequently usage of keyword extraction is for information retrieval or search engine. Moreover, many existing researches just compare them with precision and recall which algorithms are better. Therefore, in this study, we use keywords extraction to construct graph structure for reading assistance. With the intention of better efficacy, this study chooses to extract “keyphrase” instead of keywords. The “keyphrase” could be composed with more than one word including nouns and adjunctions.

When reading documents or papers in English, people whose first language is not English usually like to mark words they do not know, and then use dictionaries to find out the meanings, which will waste much time. If keywords or keyphrases can be extracted first to construct a meaningful graph structure, the reading comprehension process should be facilitated. Hence, in this study, we do not only extract keyphrases but also calculate the keyphrase scores and rank them. Before extracting keyphrases, we use Stanford Parser to parse documents based on POS (part of speech) and tokenize, so that we can easily find

the words we want. In addition, after extracting keyphrases, we need to determine how important the keyphrases are by calculating the scores. Ranking the keyphrases can help choosing the important ones as top keyphrases and eliminating the keyphrases which are not that important. Using top keyphrases to construct graph structure can help users to read documents or papers.

2. Literature Review of Keyphrase Extraction

Early research which extracted keyphrase from a single document is conducted by Krulwich and Burkey (1996). They use heuristics which are based on syntactic clues, such as the use of italics, the presence of phrases in section headers, and the use of acronyms to find keyphrases. Muñoz (1996) uses an unsupervised learning algorithm to discover two-word keyphrases. The algorithm is based on Adaptive Resonance Theory (ART) neural networks. Barker and Cornacchia (2000) propose a simple system for choosing noun phrases from a document as keyphrases. Mihalcea and Tarau (2004) recommend the TextRank model to rank keywords based on the co-occurrence links between words which make use of "voting" or "recommendations" between words to extract keyphrases. Wan and Xiao (2008), on the other hand, offers to use a small number of nearest neighbor documents to provide more knowledge to improve single document keyphrase extraction. Liu et al. (2009) evaluate the system performance in different ways, including comparison to human annotated keywords using F-measure and a weighted score relative to the oracle system performance, as well as a novel alternative human evaluation.

Supervised machine learning algorithms have been introduced to categorize keyphrases. According to Wan and Xiao (2008), "GenEx (Turney, 2000) and Kea (Frank et al., 1999; Witten et al., 1999) are two typical systems, and the most important features for classifying a candidate phrase are the frequency and location of the phrase in the document." In addition, Medelyan and Witten (2006) put forward the system called KEA++ that enhances automatic keyphrase extraction by using semantic information on terms and phrases gleaned from a domain-specific thesaurus. Nguyen and Kan (2007) focus on keyphrase extraction in scientific publications by using new features that capture salient morphological phenomena found in scientific keyphrases.

Different from the abovementioned studies, this research uses an unsupervised method for keyphrase extraction which involves assigning a saliency score to each candidate phrase by considering various features.

3. The Proposed Text-Significant Concept Relationship (TSCR) Structure

Before constructing the graph structure, three steps should be followed. The first step is to parse the paragraph for constructing the grammar tree, and then to find the words consisting of nouns and adjunctions. Next, noun phrases or keyphrases contain nouns and adjunctions will be calculated. After calculating the keyphrases, a certain score will be acquired for ranking purpose. Third, according to the score, the higher the number on the rating scale, the more important the keyphrases are, thereby the graph structure could be constructed by the important keyphrases in order to assist reading process.

Figure 1 is the graph structure that establishes step processes which is called Text-Significant Concept Relationship (TSCR). In this paper, SingleRank analysis method was used to construct graph structure for reading assistance.

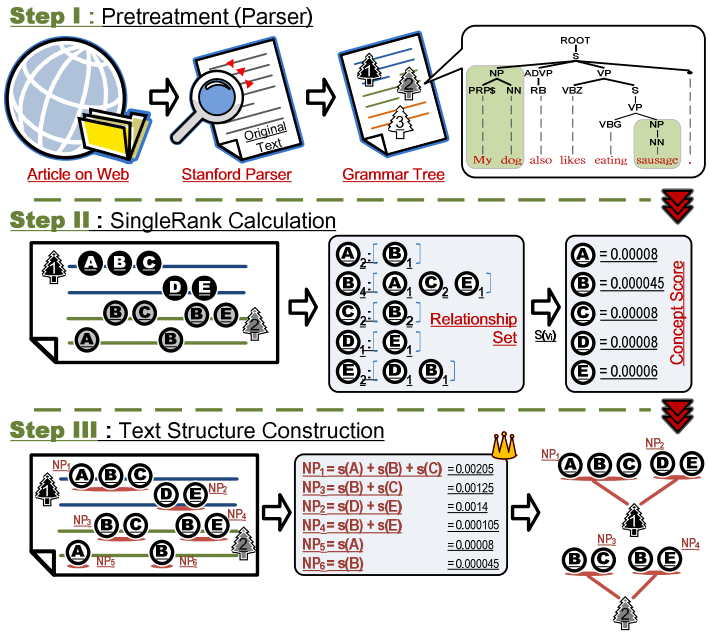

Figure 1: TSCR processes

3.1 Stanford Parser

In step 1, over 100 news documents were recruited and the paragraphs were parsed using the Stanford Parser. Later, noun phrases (NP) and phrases containing nouns and adjunctions were found by applying DFS search in order to generating the grammar tree.

3.2 SingleRank calculation

Step 2 is about how to choose the keyphrases utilizing SingleRank calculation method. Top keyphrases will be selected while low score keyphrase will be discarded. The SingleRank value is:

$$S(v_i) = (1-d) + d * \sum_{V_j \in In(V_i)} \frac{w_{ij}}{\sum_{V_k \in Out(V_j)} w_{kj}} S(v_j)$$

where S(vi) denotes the one of the score of the keyphrase containing word vi . Simultaneously, vj denotes co-occurrence with vi ; wij denotes weight between i and j; and d represents damping factor which is 0.85. Every keyphrase should be evaluated because not all high frequency keyphrases are important. Hence, researchers must evaluate the total scores of the keyphrases and rank them. Keyphrases with top scores were chosen when the ones with low scores were cut off.

3.3 Graph structure

Step3 is to establish TSCR structure after choosing top keyphrases, and then use this structure to assist reading. The system that constructs the TSCR graph structure will be presented in the next section. How to assist reading will also be discussed.

4. An Implementation of TSCR System

The goal of this research is to extract keyphrases and rank them. Then, top keyphrases will be selected based on the score to construct TSCR for helping users to facilitate reading process.

In this phase, the system was implemented according to figure 1 processes. Every step in the system will be showed after parsing and calculating from paragraphs. The system has two sections, server and client. The Server is responsible for parsing the paragraphs, finding the keyphrases, as well as calculating the keyphrases' scores and ranking them. The Client is accountable for showing paragraphs with high score keyphrases and words. Finally, the TSCR structure with the highest score keyphrase will be demonstrated.

Figure 2 illustrates the server interface which consists of two parts showing the news we extracted from The China Post website. Section A denotes three categories we chose: "Art & Leisure," "Health," and "Sports". Section B displays the list for news titles with the date and the difficulty level.

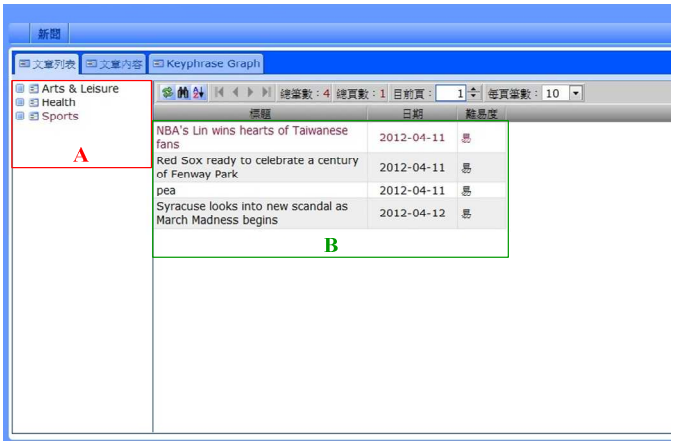

Figure 2: Establish classifies and articles

Figure 3 shows the server interface that high score keyphrases were parsed from the paragraphs, so that the keyphrases were identified and calculated.

Figure 3: The scores with keyphrase

Figure 4 shows the client interface which contains three parts. Part A is the content of the chosen news. Part B is the list of words. Part C lists the top keyphrases. In this interface, we can also see the top sentences in the paragraphs. While the red color indicates top keyphrases, the yellow color reveals top sentences.

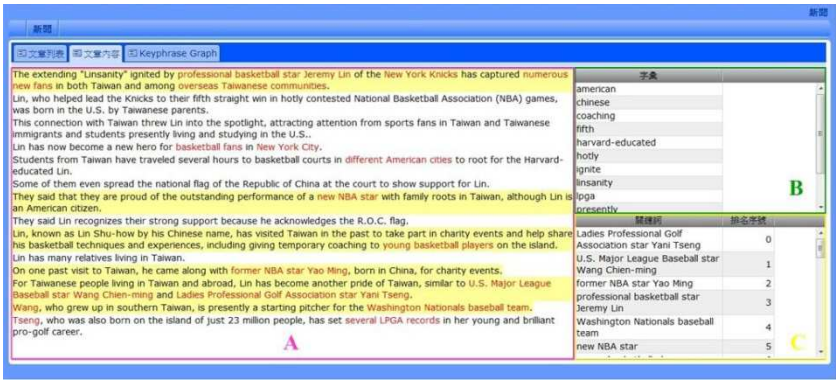

Figure 4: Interface with paragraph content, words and keyphrases

Figure 5 is the client interface showing the TSCR structure composed based on top keyphrases, and the function of the TSCR structure that is able to assist reading paragraphs. In this structure, *s* represents the sentence with one or more keyphrases within. The main purpose of this structure is to help students to improve their English reading ability, and to understand the meanings quickly.

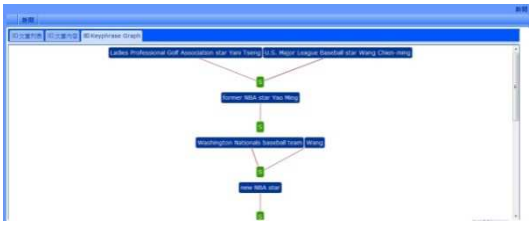

Figure 5: TSCR structure

Figure 6 is the client interface illustrating sentence hints with TSCR structure. When moving cursor to the sentence's icon, the full sentence will appeal including top keyphrase. Hence, users will have more hints to realize what the paragraph means.

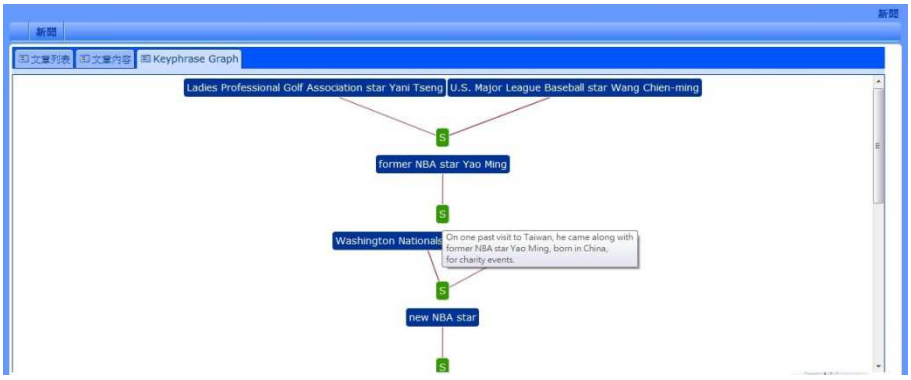

Figure 6: Sentence hint in TSCR structure

5. Experiment with Questionnaire

The system was implemented and the TSCR structure was established. Next, we want to know how effective the system is for helping users in reading English paragraphs. Thirty users were randomly selected from a university in northern Taiwan as our pilot participants. A questionnaire was designed to investigate users' personal background,

computer-using experiences, and effectiveness of using the system. The questionnaire consists of five important questions regarding using the system:

1) After using the system, do you think the system helpful for you?
2) After using the system, do you think the system can improve your English reading ability?
3) After using the system, do you think the system can assist vocabulary learning?
4) After using the system, do you think the system is interesting?
5) After using the system, do you think it can help reading comprehension within the TSCR structure?

Table 1 reports the statistics of the first question: "Do you think the system helpful for you?" The result illustrates that over 60% of the users deem the system helpful for users.

Table 1: statistics for helpfulness of the system

Effective	Number	Percent	Effective Percent	Accumulation Percent
Strongly Disagree	0	0.0	0.0	0.0
Disagree	3	10.0	10.0	10.0
Normal	8	26.7	26.7	36.7
Agree	15	50.0	50.0	86.7
Strongly Agree	4	13.3	13.3	100.0
Total	30	100.0	100.0	

Table 2 shows the statistics regarding the second question: "Can the system improve the reading ability?" The result reveals that over 70% of users agree that using the system can improve their reading ability.

Table 2: statistics for effectiveness of improving reading ability

Effective	Number	Percent	Effective Percent	Accumulation Percent
Strongly Disagree	0	0.0	0.0	0.0
Disagree	2	6.7	6.7	6.7
Normal	6	20.0	20.0	26.7
Agree	20	66.7	66.7	93.3
Strongly Agree	2	6.7	6.7	100.0
Total	30	100.0	100.0	

Table 3 illustrates the statistics about the effectiveness of improving vocabulary learning. Over 80% users agree or strongly agree that using the system can improve vocabulary learning.

Table 3: statistics for effectiveness of improving vocabulary learning

Effective	Number	Percent	Effective Percent	Accumulation Percent
Strongly Disagree	0	0.0	0.0	0.0
Disagree	1	3.3	3.3	3.3
Normal	4	13.3	13.3	16.7
Agree	22	73.3	73.3	90.0
Strongly Agree	3	10.0	10.0	100.0
Total	30	100.0	100.0	

Table 4 demonstrates the statistics about "how interesting the system is." Over 60% users consider the system interesting.

Table 4: statistics for how interesting the system is

Effective	Number	Percent	Effective Percent	Accumulation Percent
Strongly Disagree	0	0.0	0.0	0.0
Disagree	1	3.3	3.3	3.3
Normal	10	33.3	33.3	36.7
Agree	15	50.0	50.0	86.7
Strongly Agree	4	13.3	13.3	100.0
Total	30	100.0	100.0	

Table 5 presents the statistics regarding how effective the system is to assist reading within the TSCR structure". The result exposes that 76,7 % of users agree or strongly agree that using the system can help reading within the TSCR structure.

Table 5: statistics for the effectiveness of assisting reading within the TSCR structure

Effective	Number	Percent	Effective Percent	Accumulation Percent
Strongly Disagree	0	0.0	0.0	0.0
Disagree	1	3.3	3.3	3.3
Normal	6	20.0	20.0	23.3
Agree	20	66.7	66.7	90.0
Strongly Agree	3	10.0	10.0	100.0
Total	30	100.0	100.0	

According to the above statistics, we know that the system with TSCR structure can assist reading for users. The users also consider this system useful because the system can find keyphrases they could not find out. By employing the system, users can figure out the paragraph outline easily and be able to guess the meanings.

6. Conclusion

In this paper, a TSCR structure for assisting the reading of English was established. This study is significantly different from most previous works. Documents from online newspaper are selected, and unsupervised machine learning training was used for analyzing the documents in the system. The top keyphrases chosen by the system were similar to professor-marked keyphrases. The mainly function of using top keyphrases is to help users to read. However, this research did not combine another paper for precision and recall. It merely uses single document for reading assistance. Moreover, a questionnaire was conducted, and the users confirm the effectiveness of the system.

For the future work, we plan to add other teaching methods such as grammar structure and other statement analysis. We also need to expand TSCR structure by turning document keyphrases into cloud type.

Acknowledgements

This research is supported as part of project "NSC 101-2631-S-003 -001 –CC2"by National Science Council, Taiwan, Republic of China.

References

[1] Krulwich, B., and Burkey, C. 1996. Learning user information interests through the extraction of semantically significant phrases. In AAAI 1996 Spring Symposium on Machine Learning in Information Access.

[2] Muñoz, A. 1996. Compound key word generation from document databases using a hierarchical clustering ART model. Intelligent Data Analysis, 1(1).

[3] Frank, E.; Paynter, G. W.; Witten, I. H.; Gutwin, C.; and Nevill-Manning, C. G. 1999. Domain-specific keyphrase extraction. Proceedings of IJCAI-99, pp. 668-673.

[4] Witten, I. H.; Paynter, G. W.; Frank, E.; Gutwin, C.; and Nevill-Manning, C. G. 1999. KEA: Practical automatic keyphrase extraction. Proceedings of Digital Libraries 99 (DL'99), pp. 254-256.

[5] Barker, K., and Cornacchia, N. 2000. Using nounphrase heads to extract document keyphrases. In Canadian Conference on AI.

[6] Turney, P. D. 2000. Learning algorithms for keyphrase extraction. Information Retrieval, 2:303-336.

[7] Mihalcea, R., and Tarau, P. 2004. TextRank: Bringing order into texts. In Proceedings of EMNLP2004.

[8] Medelyan, O., and Witten, I. H. 2006. Thesaurus based automatic keyphrase indexing. In Proceedings of JCDL2006.

[9] Nguyen, T. D., and Kan, M.-Y. 2007. Keyphrase extraction in scientific publications. In Proceedings of ICADL2007.

[10] Wan, Xiaojun and Jianguo Xiao. (2008). “Single document keyphrase extraction using neighborhood knowledge”, Proceedings of the 23rd AAAI Conference on Artifical Intelligence, pp.855-860.

[11] Feifan Liu;Deana Pennell;Fei Liu;Yang Liu, “Unspervised Approaches for Automatic Keyword Extraction Using Meeting Transcripts”, Proceedings of Annual Conference of the North American Chapter of the ACL, June 2009, pages 620–628.