International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:1, No:11, 2007

Using Heuristic Rules from Sentence Decomposition of Experts' Summaries to Detect Students' Summarizing Strategies

Norisma Idris, Sapiyan Baba, and Rukaini Abdullah

***Abstract*—**Summarizing skills have been introduced to English syllabus in secondary school in Malaysia to evaluate student's comprehension for a given text where it requires students to employ several strategies to produce the summary. This paper reports on our effort to develop a computer-based summarization assessment system that detects the strategies used by the students in producing their summaries. Sentence decomposition of expert-written summaries is used to analyze how experts produce their summary sentences. From the analysis, we identified seven summarizing strategies and their rules which are then transformed into a set of heuristic rules on how to determine the summarizing strategies. We developed an algorithm based on the heuristic rules and performed some experiments to evaluate and support the technique proposed.

***Keywords*—**Summarizing strategies, heuristic rules, sentence decomposition.

I. Introduction

SUMMARY writing is included in the English syllabus for secondary schools in Malaysia to evaluate student's comprehension of a given text. It requires students to employ several summarizing strategies to produce the summary. These strategies are important in summary writing as the basic rules to determine what to include and eliminate, how to organize information and how to ensure that the summary retains the meaning of the original text. There are five basic rules that are essential to produce adequate summaries [1]: 1) deletion of trivial information; 2) deletion of redundant information; 3) generalization; 4) sentence topic selection; and 5) invention. Careful instruction in teaching summary writing positively influenced student's use of summarization strategies and the quality of their summaries [2]. Thus, in school, summarizing is taught by using the instructional rules step-by-step that are similar to the basic rules above.

Manuscript received 30 November, 2007. This work was supported by the Ministry of Higher Education Malaysia and University of Malaya under Fundamental Research Grant Scheme FP050/2006A.

Norisma Idris, is with the Department of Artificial Intelligence, University of Malaya, 50603 Kuala Lumpur, Malaysia (phone: 603-79676395; e-mail: norisma@um.edu.my).

Sapiyan Baba is currently a Professor at the Department of Artificial Intelligence, University of Malaya, 50603 Kuala Lumpur, Malaysia (phone: 603-79676343; e-mail: pian@um.edu.my).

Rukaini Abdullah is the head of the Department of Artificial Intelligence, University of Malaya, 50603 Kuala Lumpur, Malaysia (phone: 603-79676378; e-mail: rukaini@um.edu.my).

Summarization is one of the best learning techniques to evaluate students' comprehension [3]. Hence, automated summarization assessment has drawn a lot of interest in recent years. There are a few systems developed for this purpose, e.g. *Summary Street*® [4][5], *Laburpen Ebaluaka Automatikoa* or *LEA* [3] and Summarization Assessment Strategies Model [6]. *Summary Street*® is a computer-based assessment system that provides an environment where students can get feedback about the content of their written summary. The tool employed *Latent Semantic Analysis (LSA)* which used a machine learning method to construct the semantic representations that mirror the way human knowledge is structured. The system compares the similarity in meaning between a student's summary and the original text. It gives immediate visual feedback based on measures such as content knowledge, writing mechanics, length, redundancy, and plagiarism. *Laburpen Ebaluaka Automatikoa* or *LEA* is an automatic summary evaluation environment which takes evaluation decision based on human expertise modeling, to train students in summarization skills and also to assess human summary evaluation. It gives feedback on the coherence, content coverage and cohesion, the use of language and adequacy of the summary. Like Summary Street, LEA also employed LSA as the tool to measure on domain knowledge and summarization skills. Another example is modeling summarization assessment strategies using LSA, which is an effort to model the way teachers assess students' summaries. The model is based on the automatic detection of five macrorules which are copy, paraphrase, generalization, construction and off-the-subject. These macrorules were implemented in the LSA framework where each summary sentence is compared with each sentence of the original text.

Although previous works (e.g. as in [3][4][5]) have presented an invaluable contribution towards the development of the summarization assessment system, their focus is only on the output summary, viz. the completeness of the information presented in the summary and the quality of the summary. Previous study has shown that student's difficulties in summarizing were linked to the students' use of strategic skills [7]. However, automated summarization assessment for detecting students' summarizing strategies is still lacking. In addition, previous analysis which was done to study the performance of students in summary writing and the summarizing strategies employed by students [8] suggested

that students were very weak at summarizing and the marks assigned to the students do not reflect their skills in summarizing. Thus, we proposed a computer-based summarization assessment that can be used to detect the summarizing strategies used by students. Given a student summary and the original text, the system should be able to identify what strategies are used for a summary sentence. The method involves a set of heuristic rules constructed from the analysis of the decomposition of summary sentences written by experts.

The rest of the paper is organized as follows. Section 2 discusses the process of decomposition of expert-written summaries. Section 3 formulates the problem for detecting the summarizing strategies and presents the general heuristic rules based on the results from the process in previous section. Section 4 shows the example of the algorithm and Section 5 discusses the experiment done. Finally in Section 6, we report the conclusion and the progress of the project.

II. Sentence Decomposition of Experts' Summaries

Sentence decomposition of experts' summaries involves analyzing summary sentences of the experts to determine how these sentences are constructed. In summarization assessment process, detecting student's summarizing strategies is not a common practice amongst teachers in schools. However, the teachers' skills in summarization can be used as rules to detect students' summarizing strategies. Thus, we need to consider how these experts produce their summary sentences. Therefore, a study was conducted to uncover the summarizing strategies and how they used the strategies in producing their summary sentences. The result of the study is used to develop a set of heuristic rules for the system. The overview of the system development is depicted in Fig. 1.

The subjects in this study are experienced secondary school English teachers. They were asked to summarize an article of about 2000 words into a summary of less than 200 words. We collected 6 samples from the experts and the document analysis method is used to analyze the summaries. Each summary was decomposed into sentences since human's effort in summarizing process are generally focus at the sentence level rather than the paragraph or passage level. This is because a sentence offers better control over compression and is accepted as a traditional linguistic unit in syntactic analyses [9]. Each summary sentence is compared to the original text so that we can mark exactly *which* phrases are taken from the original text by the experts and *how* the phrases are joined together to produce the summary sentence. The outline of the process is described as follows:

Given an expert summary and the original text;

i. *Decompose the summary into summary sentences*
ii. *For each summary sentence, search for the sentence(s) in the original text which is/are close to the summary sentence*
iii. *Compare the summary sentence and the sentences from the original text and identify the strategies used to construct the summary sentence*

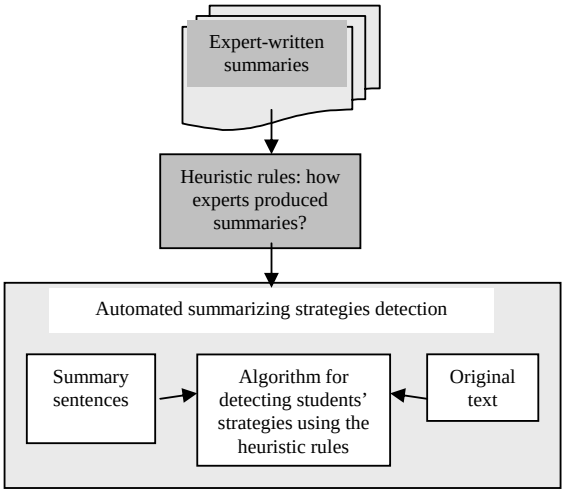

Fig. 1 The overview of the system development

Based on the analysis, we have identified 7 types of strategies that have been used by these experts to produce their summary sentences which are:

- **Deletion**
  In the deletion process, trivial or redundant information are eliminated from the sentence in the original text. From the analysis, this process involves units such as stop phrases, explanations, examples, scenarios and elaborations.
- **Sentence combination**
  In the sentence combination process, phrases from more than one sentence are merged into one sentence. This process is typically applied together with other strategies such as deletion, generalization and paraphrasing so that the summary sentences produced are short but informative. These sentences are usually combined using conjunction words such as *and, but, although, though, by, where, which,* and *because* or even by coma (,) and are most likely to come from the same paragraph.
- **Topic sentence selection**
  In the topic sentence selection, usually only one main sentence is chosen to represent the main idea of the whole paragraph. This strategy involves identifying relevant information from the original text to be included in the summary. From our analysis, the human experts made use of cue phrases (e.g., *"It is concluded that"*, *"She discovered that*", *"She claims that*") and their location (which is either the first sentence or the last sentence of the paragraph) in order to identify the theme of the text.
- **Syntactic transformation**
  In syntactic transformation, the order of the words in a sentence or the syntactic structure of a sentence is changed. These changes may affect the location of the words in the summary text as compared to the original text.

- **Paraphrasing**
  In the paraphrasing process, a phrase or word in the original sentence is replaced with a similar phrase/word in the summary sentence. This refers to the use of similarity concept involving synonym, where different words that have same meaning.
- **Generalization**
  In the generalization process, a list of words or items is replaced by a more general word in the same class. This refers to the use of relatedness concept such as hypernymy *(is-a relationship)*, meronymy *(part-of relationship)* or any kind of frequent association.
- **Invention**
  In the invention process, the meaning of the sentences is constructed by using the expert's own words. In this case, our assumption is: if more than 75% of words in a summary sentence are composed by using their own words, then it is considered that the sentence is produced by using the invention strategy. This assumption is due to cases where some words are unique words and cannot be replaced by other related words. For example, person's name, diseases, place, etc.

The strategies identified in this analysis can be divided into two approaches. In the first approach, they used the *cut and paste technique* to retain its principal meaning by reusing words from the original text. This technique comprises strategies such as deletion, sentence combination, topic sentence selection, paraphrase, generalization and syntactic transformation. Most professional summarizers use this technique to summarize text document [10]. In addition, previous study has also shown that human expert also rely on this technique to summarize to avoid presenting different ideas from the original text [11]. In fact, this technique has been used in automatic summarization system and cross language summarization system [11][12]. In the second approach, the experts read and try to understand the whole text before they produce the summary by using their own words or invention. Most of the summary sentences are not taken from the original text but constructed from scratch based on their understanding of the text. It is also found that a single summary sentence is often constructed by using more than one strategy. For example, sentence combination is often used with the deletion, generalization or paraphrase strategies. The results from this analysis are transformed into general heuristic rules discussed in the next section.

III. Developing the Heuristic Rules

The summarizing strategies found from the decomposition of experts' summaries were transformed into a set of heuristic rules to detect the summarizing strategies. These rules are given in Table I below:

TABLE I
A Set of Heuristic Rules to Detect Summarizing Strategies

Strategy	Heuristic Rules
Deletion	*A summary sentence is produced by deletion if two words or more:* • *are found from the same sentence in the original text* • *are sequence of words in adjacent positions in the original text* • *retain their relative precedent relation as in the original text* • *are less than the number of words of the original sentence*
Sentence Combination	*A summary sentence is produce by sentence combination if the words are found:* • *in different sentences from original text which come from nearby sentences or from the same paragraph* • *applied together with other strategies* • *involve the used of conjunction words to combine phrases from different sentences*
Topic sentence selection	*A summary sentence is produced by topic sentence selection if:* ▪ *cue phrase or* ▪ *location cue (the first or last sentence of the paragraph in original text)* *is found in the sentence.*
Syntactic transformation	*A summary sentence is produced by syntactic transformation if:* ▪ *the words are found from the same sentence in the original text* ▪ *the order of the words are different from the original text*
Paraphrase	*A summary sentence is produced by paraphrase if:* ▪ *the words in the sentence are similar/synonym to the words found in any sentence in original text*
Generalization	*A summary sentence is produced by generalization if:* ▪ *a set of words in the sentence is related to the same class or word found in any sentence in original text*
Invention	*A summary sentence is produced by invention if:* ▪ *most of the words in summary sentence is not found in any sentence in original text but are semantically related by lexical relationships.*

IV. The Algorithm

Our task is to determine which sentence in the original text corresponding to the words in the summary sentence, and identify how the summary sentence is constructed *as* defined below:

Given a student-written summary sentence,

- *Which words in the summary sentence come from the original text?*
- *Where are the words located in the original text?*
- *How is the summary sentence constructed from these words?*

A summary sentence, *S*, consist of sequence of words that can be represented as: (W_1, W_2, W_3, …… , W_m).

The position of a word in the original text can be represented by their sentence position and word position: *(SPOS, WPOS).*

Some common words occur more than once in the text. Thus, multiple occurrences of a word found in the document can be represented as: *{($SPOS_1$, $WPOS_1$), ($SPOS_2$, $WPOS_2$), ($SPOS_3$, $WPOS_3$)……}*

In order to determine which sentence in the original text corresponding to the words in the summary sentence, we need to locate the positions of the words in the original text and find the best sequence. The best sequence would be one which summary sentence matches closely the sequence of any sentence in the original text.

For example, we translated the first rule on how to detect the '*deletion strategy*' into a set of algorithm as shown below:
Given a summary sentence,

- *Locate the words in the original text and list all the positions of possible sequence according to the sentence positions, SPOS.*
  e.g. Pos(W_1) = ($SPOS_1$, $WPOS_1$) and Pos(W_2) = ($SPOS_2$, $WPOS_2$) where $SPOS_1 = SPOS_2$
- *Determine whether the word positions, WPOS, retain their positions in the document.*
  e.g. Pos(W1) = ($SPOS_1$, $WPOS_1$) and Pos(W2) = ($SPOS_2$, $WPOS_2$) where $WPOS_1 < WPOS_2$
- *For each possible sentence in original text that is used to construct the summary sentence, calculate the length, l, and the distance between words to find the best sequence of the position of the words, where ,*

$$d = \sum_{i=1}^{l-1} \left[WPOS_{i+1} - WPOS_i \right] / l \qquad (1)$$

V. The Example of the Experiment

This algorithm is implemented in Lisp. Consider a summary sentence *"I started towards the shore"* extracted from a student's summary. The words in the summary sentence are found in the original text where each of the word is represented by their sentence positions and words positions. Most of the common words occur more than once in the text. For example, the word *'I'* is located in few locations in the original text, ((1, 3), (4, 1), (4, 7), (5, 3), ..), as shown in Table II below.

After all the positions of possible sequence are listed according to the sentence positions, we obtained few possible sequences, as shown in Table III. However, from the calculation of the distance and length of the possible sentences, the closest distance is chosen. Therefore, the best sequence of the word positions should be ((4 1) (4 2) (4 3) (4 4) (4 5)) which indicate that the summary sentence was produced from sentence 4 in the original text, as presented in Table IV. In this example, the summary sentence was proved to be produced by using deletion strategy since it complies with the rules in the algorithm for detecting the strategy.

TABLE II
The Location of Words from Summary Sentence in the Original Text

Word	*I*	*started*	*towards*	*the*	*shore*
location of words	(1 3)	(4 2)	(4 3)	(1 9)	(1 15)
	(4 1)	(22 7)	(6 7)	(1 14)	(3 9)
	(4 7)			(3 8)	(4 5)
	(5 3)			(3 17)	(17 22)
	(6 1)			(4 4)	(22 5)
	(7 1)			(6 10)	(33 10)
	(8 1)			(6 14)	
	(9 5)			(9 19)	
	(10 2)			(10 7)	
	(10 19)			(10 13)	
	:			:	

TABLE III
The Calculations of *L* and *D* for Possible Sentences

SPOS						***l***	***d***
1	(1 3)	(1 9)	(1 15)			3	4
4	(4 1)	(4 2)	(4 3)	(4 4)	(4 5)	5	0.8
4	(4 7)	(4 2)	(4 3)	(4 4)	(4 5)	5	1.6
5	(5 3)					1	-
6	(6 1)	(6 7)	(6 10)			3	3.25
7	(7 1)					1	-
8	(8 1)					1	-
9	(9 5)	(9 19)				2	7
10	(10 2)	(10 7)				2	3.7
10	(10 19)	(10 7)				2	6

TABLE IV
The Best Sequence of the Positions of Words Found in the Original Text

I	*started*	*towards*	*the*	*shore*
(1 3)	(4 2) →	(4 3)	(1 9)	(1 15)
(4 1) ↗	(22 7)	(6 7)	(1 14)	(3 9)
(4 7)			(3 8)	(4 5) ↗
(5 3)			(3 17)	(17 22)
(6 1)			(4 4) ↘	(22 5)
(7 1)			(6 10)	(33 10)
:			(6 14)	
			:	

VI. Conclusion

This paper presents how a set of heuristic rules are constructed from analyzing the expert's summaries using summary sentence decomposition. These rules are used to develop a tool for detecting students' strategies in summary writing. Currently, we are in a process of developing and

testing the algorithm of the heuristic rules discussed in this paper. This algorithm would be embedded in an educational tool that would help teachers to detect the ability of their students in applying the strategies in summarizing. It is also can be used to help students to hone their skills in summary writing.

ACKNOWLEDGMENT

We gratefully acknowledge teachers and lecturers who have been involved and given their full corporation in this project.

REFERENCES

[1] A. L. Brown, and J. D. Day, "Macrorules for summarizing texts: The development of expertise" *Journal of Verbal Learning and Verbal Behavior*, 1983, Vol. 22, pp. 1-14.

[2] Gurdashan Kaur. "The Role of Summarization Instruction in the Comprehension of Expository Texts", Unpublished Master dissertation. Faculty of Education, University of Malaya, 1997.

[3] I. Zipitria, A. Arruarte, J. A. Elorriaga, and A. Díaz de Ilarraza, "Towards a Cognitive Model of Summary Evaluation", in J. Mostow & P. Tedesco (Eds.) *Proceedings of the ITS'2004 on Modeling Human Teaching Tactics and Strategies*, 2004, pp. 25-34. Maceió, Brazil.

[4] D. Wade-Stein, and E. Kintsch, "Summary Street: Interactive Computer Support for Writing", *Cognition and Instruction*, 2004, Vol. 22, pp 333 - 362.

[5] M. Franzke, E. Kintsch, D. Caccamise, N. Johnson, and S. Dooley, "Summary Street®: Computer support for comprehension and writing" *Journal Educational Computing Research*, 2005, Vol. 33(1) 53-80.

[6] B. Lemaire, S. Mandin, P. Dessus, and G. Denhière, "Computational cognitive models of summarization assessment skills", *Proceedings of the 27th Annual Meeting of the Cognitive Science Society (CogSci' 2005)*. Stresa, Italy, July 21-23, 2005, pp.1266-1271.

[7] P. Winograd, "Strategic Difficulties in Summarizing Texts", *Reading Research Quarterly*. 1984, 19(4):404—425.

[8] I. Norisma, Mohd. Sapiyan Baba, and Rukaini Abdullah, "An Analysis on student-written summaries: A Step towards Developing an Automated Summarization Assessment", *International Conference on Electrical Engineering and Informatics (ICEEI2007)*. Institut Teknologi Bandung, Indonesia, June 17 – 19, 2007, pp. 550 – 553.

[9] I. Mani, *Automatic Summarization*. Amsterdam: John Benjamins Publishing Company, 2001, Chapter 3.

[10] B. Endres-Niggemeyer, K. Haseloh, J. Miller, S. Peist, I. Santini de Segel, A. Sigel, J. Wheeler, and B. Wollny, B. *Summarizing Information*. Springer, Berlin, 1998.

[11] H. Jing, and K. McKeown, "Cut and Paste Text Summarization", In *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics*, 2000, pp. 178-185.

[12] M. L. Nguyen, and S. Horiguchi, "Accuracy enhancement for the decomposition of human-written summary", *International Journal of Computer Processing of Oriental Languages, 2005*, Vol. 18, No. 1, pp 53-74.