Document-polishing Support System for Creating Top-Down Structure

Satomi Yamate and Wataru Sunayama

Graduate School of Information Sciences, Hiroshima City University 3-4-1, Ozuka-Higashi, Asa-Minami, Hiroshima 731-3194, Japan

Abstract. In recent years, with the growth of electronic communication such as e-mail, blogs, and on-line reports, we have had many opportunities to write documents. Although a document is used to transfer information or intention, it may be hard to read even if the writer made strong efforts to make it readable. Therefore, there is need for a system that can supply objective values for polishing documents. In this study, a system that can support document writing in a top-down structure is proposed. According to the experimental results, the proposed system could be effectively used to write documents with a top-down structure.

Keywords: document structure-polishing support, document writing support, paragraph relationship evaluation, text mining

1 Introduction

In this study, a system that can support document writing in a top-down structure is proposed. The definition of top-down structure is given as "Paragraphs giving the conclusions and key points are presented before paragraphs that describe the details." Evaluation values for verifying a top-down structure are calculated by the connections between paragraphs. That is, the system supports the user in creating comprehensible documents by supplying evaluation values reflecting the quality of a top-down structure.

2 Related works

Among the research work on evaluating more than two paragraphs, the coherence of a document has been measured by its discourse structure [1] or word connections [2]. Although these works evaluated whether a document is coherently related to a topic, this study's system deals with the branches to sub-topics so that it can evaluate the relationships between paragraphs that include a conclusion or key points as well as the reasons leading to the conclusion. That is, the proposed system visualizes the relationship between paragraphs as a tree structure and thus supports users in improving the document's structure.

In systems for evaluating the content and constitution of a document, automatic rating systems have been suggested [3, 4]. However, those systems are not for use by the writer but by an evaluator, so writers cannot use them easily.



Fig. 1. Example of tree structure construction (each paragraph (node) and its link having maximum conditional probability)

In the research on systems to improve the user's comprehension of document structure, one work displays a concept map of keywords [5], another system displays the distance among paragraphs as a map [6], and a third system displays a document's structure as a tree structure [7]. Although these systems have been used for comprehending documents, they do not output guidance on how to modify a document. In this study, the proposed system supplies concrete advice for constructing a top-down structured document.

Real relationships among paragraphs cannot be identified by the proposed system because the method does not use semantic information but uses word appearance information only. However, the system aims at displaying the most simple and objective structure in order to assure the current status of a document calculated by the words connection.

3 Definition and Evaluation of Top-down Structure

3.1 Construction of Tree Structure

In order to evaluate a top-down structure, a conditional probability for evaluating the connection between paragraphs A and B is defined as Eq.(1), where W_A denotes the set of words (nouns) included in paragraph A and the function $n(W_A)$ denotes the number of elements included in set W_A . This value means the probability of paragraph B inheriting the contents of paragraph A, since the value becomes larger if paragraph B includes many words used in paragraph A.

$$Relation(A,B) = \frac{n(W_A \bigcap W_B)}{n(W_A)}$$
(1)

Figure 1 also shows an example of creating tree structure when the number of paragraphs is six.

- 1. Let each paragraph $P_i(i = 1, 2, ..., N)$ be a node.
- 2. Calculate conditional probabilities $Relation(P_i, P_j)(1 \le i < j \le N)$. Probabilities from smaller paragraphs to larger paragraphs are calculated to evaluate the story line.

- 3. Create links $(P_{ki} \rightarrow P_i)$ between each paragraph $P_i(i = 1, 2, ..., N)$ and P_{ki} whose conditional probability $Relation(P_k, P_i)(k < i)$ becomes the maximum.
- 4. Create tree structure by combining the created (N-1) links as the first paragraph becomes a root node.

3.2 Calculation of Evaluation Value for Top-down Structure

This section describes how to calculate the evaluation value Value(Tr) for the top-down structure of created tree structure Tr.

- 1. Let $PV(P_i)$ be zero as the evaluation value for the sub-tree from each paragraph $P_i(i = 1, 2, ..., N)$.
- 2. For the paragraphs P_i that have two or more links (K is number of links), calculate evaluation value $BV(B_{ij})$ for each branch $B_{ij}(j = 1, 2, ..., K)$, and calculate $PV(P_i)$ as the products of them. The evaluation value for each branch $BV(B_{ij})$ is given by Eq.(2) as the sum of the link evaluation values from each branch to the farthest leaf node (at which the number of links becomes the maximum), where P_a and P_b denote a pair of paragraphs in the route $Route_{ij}$ from paragraph P_i to the leaf node. The link evaluation value located between paragraphs C and D is defined as Eq.(3).
- 3. Calculate the top-down-structure evaluation value Value(Tr) for tree structure Tr as the sum of evaluation values for each paragraph $PV(P_i)$.

$$BV(B_{ij}) = \sum_{P_a, P_b \in Route_{ij}} Link(P_a, P_b) + 0.9$$
⁽²⁾

$$Link(C, D) = \min\{Relation(C, D), 0.5\}$$
(3)

3.3 Calculation of Normalized Top-down-structure Evaluation Value

The basic value is given by the top-down-structure evaluation value of the basic document structure. A basic document structure with N paragraphs is defined as a structure that gives the conclusion in the first paragraph and describes the details using G branches, where G is given as Eq.(4) and INT(X) means the maximum integer that is not more than X. In addition, the conditional probability for each link from the first paragraph is set as 1/G, and the probabilities for other links are set as 0.4. Finally, the top-down-structure evaluation value Value(Tr) is transformed to normalized Evaluation(Tr) as Eq.(5) by using basic value Base(N), where a document has N paragraphs.

$$G = INT(\sqrt{N-2}) + 1 \tag{4}$$

$$Evaluation(Tr) = \frac{Value(Tr)}{Base(N)} \times 80$$
(5)



Fig. 2. Screen of document-polishing support system for document structure

4 Document-polishing Support System for Document Structure

Figure 2 shows the interface display of our document-polishing support system for document structure. The system consists of three parts.

4.1 Paragraph-ordering Modification

The system provides an interface panel that allows users to change the paragraph order by drag & drop operations. In addition, the system displays advice for each paragraph that helps the users to construct top-down structure easily according to the following algorithm.

- 1. Calculate Average of conditional probabilities $Relation(P_i, P_j), (i < j)$.
- 2. Count number of links $Link(P_i)$ whose conditional probability is higher than the Average for each link that starts from paragraph P_i $(P_i \rightarrow P_j, (i < j))$.
- 3. Display upward or downward arrows as the difference between sorted paragraph order by value of $Link(P_i)$ and current order.

4.2 Paragraph-structure Modification

The system provides the tree structure and its top-down-structure evaluation value described in Section 3. Users can move each paragraph from the current position to an ideal position by drag & drop operations and thus check the evaluation value after such movement. That is, the link from a dragged node is cut and the link to the nearest node from the dragged node is created when the user drops the node. However, users must understand that this movement is a temporary simulation.



Fig. 3. Change in top-down-structure evaluation values by document-polishing

4.3 Paragraph-connection Modification

The system provides two panels for editing. In the left-center panel, users can edit moved paragraphs in the left panel when the user pushes the button "moved paragraph." In the right-center panel, users can edit the paragraphs that are currently connected to or will next be connected to the moved paragraph when the user pushes the "connected paragraph" or "connect to paragraph" buttons, respectively.

In these panels, common words are displayed in yellow and different words are displayed in blue. The advice to connect paragraphs more strongly is displayed as either "add yellow words" or "delete blue words." On the other hand, the advice to connect paragraphs more weakly is displayed as either "delete yellow words" or "add blue words." Therefore, users can modify their document by following these suggestions and thus construct the ideal structure of paragraphs.

In addition, these panels can display any two paragraphs. Therefore, users can display paragraphs that should be more strongly connected and thus modify the connection accordingly.

5 Evaluation Experiment for Document-polishing Support System for Document Structure

5.1 Experiment Settings

Test subjects were 12 university and graduate school students majoring in engineering. They were instructed to write a 2000-character document in about eight paragraphs entitled "About the xxx (a noun) that I love." They were also instructed to write paragraphs that include more than 150 characters or five sentences rather than compose short paragraphs.

Next, they used the proposed document-polishing support system. When they used the system, they were instructed to modify their document as a topdown structure: "Please describe the conclusions or key points before describing the details." They were instructed to stop their participation in the experiment when they used the system for more than 30 minutes and the evaluation value became larger than 80 points or in any case when they used the system for more than 60 minutes.

5.2 Experimental Results

Figure 3 shows the changes in top-down-structure evaluation values after using this document-polishing technique. Ten test subjects were able to achieve 80 points. Since the top-down-structure evaluation values were significantly increased in all test subjects, including those whose points were lower than 80, the proposed system effectively assisted the test subjects in polishing their documents.

In our breakdown of points changed, we found that all test subjects confirmed their tree structure and modified their paragraphs to create stronger connections. Therefore, the system could assist users by displaying the document structure and the top-down-structure evaluation value with the goal of reaching 80 points.

6 Conclusions

This paper proposed a document-polishing support system for better document structure. Specifically, it assists users in constructing top-down structured documents by focusing on paragraph connections. According to the experimental results, users could modify the documents in a way that achieves a top-down structure. In particular, users who have above-average writing skills could become conscious of the document's structure and thus modify it both globally and locally.

References

- [1] Barzilay, R. and Lapata, M.: Modeling local coherence: An entity-based approach, *Computational Linguistics*, 34(1), 1-34 (2008)
- [2] Nishihara, Y. and Sunayama, W.: Document Visualization using Light and Shadow based on Topic Relevance, *International Journal of Intelligent Information Pro*cessing, 2(2), 1-8 (2011)
- [3] Burstein, J.C.: The e-rator scoring engine: automated essay scoring with natural language processing, In Sherimis, M.D. and Burstein, J.C. editors, *Automated essay scoring: a cross-disciplinary perspective*, Lawrence Erlbaum, 113-21 (2003)
- [4] Warschauer, M. and Ware, P.: Automated writing evaluation: defining the classroom research agenda, *Language Teaching Research*, 10(2), 1-24 (2006)
- [5] Villalon, J. and Calvo, R.A.: Concept Maps as Cognitive Visualizations of Writing Assignments, *Educational Technology & Society*, 14(3), 16-27 (2011)
- [6] O'Rourke, S.T., Calvo, R.A., and McNamara, D.S.: Visualizing Topic Flow in Students' Essays, *Educational Technology & Society*, 14(3), 4-15 (2011)
- [7] Cesarini, F., Gori, M., Marinai, S., and Soda, G.: Structured document segmentation and representation by the modified X-Y tree, In Proceedings of the Fifth International Conference on Document Analysis and Recognition, 563 - 566 (1999)