Document Composition and Content Selection Evaluation

Shijian Lu
CSIRO ICT Centre
Locked Bag 17,
North Ryde NSW 1670 Australia
Shijian.Lu@csiro.au

Abstract *Our work is concerned with the design of adaptive hypertext systems that produce documents tailored to their intended reader. In our approach, a system composes document on-the-fly, assembling existing text fragments. One of our challenges in this approach is to support the technical writer who configures the system. The task of the technical writer is to specify the structure of the documents to be generated, together with their applicability conditions. To perform their task, authors need to know what information is available. In this paper, we examine the impact of different strategies for presenting the existing text fragments on the task of document composition. We focus in particular on the impact on the quality of the resulting documents. We found that people compose better documents when existing text fragments are presented in a structured way.*

Keywords document composition, information reuse, document quality, evaluation, method.

1. Introduction

Communication is important to a successful organisation. Effective communication is often enabled by coherent documents. Many people are familiar with or have the experience of writing documents for some specific purpose. For example, a communicator in a research organisation may be asked to create project flyers for potential clients or the general public. Producing documents this way is a manual and laborious process. Now, the ever increasing availability of information content and advancement of natural language generation technology have made it possible to augment the task of composing documents from existing content segments. Myriad [8] delivery platform represents one such research initiative towards enabling rapid document composition.

At the core of Myriad is the VDP (virtual document planner) which embodies a plan-based approach to discourse generation based on [5]. It is through discourse operators that one defines the types of text to be produced. However, authoring discourse operators is not an easy task as expertise and knowledge from many areas are required. In order to reduce the requirement threshold, we have introduced the concept of content structure [1] which is inspired by the RST (rhetorical structure theory) [4]. A content structure is composed of content nodes organised vertically as a hierarchy and horizontally as RST units. A design supporting tool Constructor [3] has been developed. Using Constructor, content structures can be visually defined from which discourse operators can be generated automatically.

We sought to evaluate our first prototype of Constructor, to investigate both the feasibility of the approach and the usability of the prototype. We started our study by an expert user evaluation in recreating Scifly [7]. One of the main difficulties uncovered was in finding what data was available to include in a document (i.e., what data could a document designer exploit to compose a document). While Constructor provides a list of retrieval services, these are currently displayed as a flat list. The expert user found it hard to locate relevant retrieval services. We conjectured that providing structure to present existing content fragments would make it easier to locate appropriate content fragments. As a result, (1) document would get composed more quickly and (2) the result would be of better quality. To test our hypotheses, we did a user experiment. It confirmed our first hypothesis regarding the time [2]. Here, we report our findings on hypothesis (2) regarding the quality of the resulting documents.

In the next section, we describe our experiment set-up before presenting our analysis on the composed documents. In Section 4, we discuss flyer content quality evaluation. The paper concludes in Section 5.

2. The experiment

Our objective was to understand how different presentations of the existing content fragments (to be used to compose a new document) would affect the task of document composition. We focus here on the quality of the resulting documents. Our hypothesis is that providing structure to present existing content fragments would result in documents of higher quality than the documents composed when the text fragments were presented to authors without structure. Twelve CSIRO employees were randomly selected to participate in the experiment and were randomly divided into two equal sized groups. One subject could not finish the experiment because of time constraints. In the remaining subjects, there were five scientists, four research engineers and two administrative staff. Among them, there were eight men and three women.

All subjects were asked to perform the same two tasks: (1) compose a flyer about a single project ("Web Service Integration") and (2) compose a flyer two projects (specifically, "Under Water Vehicle" and "Virtual Critical Care Unit"). One group of participants

(Group A) were given the existing text fragments in an unstructured list while the other group (Group B) were given a structured list. In both tasks, subjects were presented with information regarding ten projects, from which they had to pick the appropriate information for the project(s) of the flyer they were writing. The information was made available through the two interfaces.

The domain data for the experiment was a subset of the data used in the Scifly application [7]. The input material was presented in HTML format in a browser interface. Subjects used a web browser to access the data. The interface consisted of two parts: the list part and content part. The actual content in the content part was changed according to what was selected from the list. Initially, the content part was empty. There were 205 items in the list. The list was represented in two different ways: an unstructured list and a semantically structured list.

Essentially, participants copied selected content from the web browser (from the input material), pasted it into an MS Word document, structured and ordered the content appropriately. All experiment sessions were video recorded. Pre-experiment and post-experiment questionnaires were also used. In addition, all sessions were observed, and subjects' actions were noted on paper. The result on time completion has been analysed elsewhere [2]. In the next section, we will analyse the project flyers composed by the subjects.

3. Analysis of composed flyers

At the end of the experiment, two sets of flyers have been composed by subjects: single project flyers and two-project combined flyers. After close examination of the resulting flyers, two interesting characteristics are discovered. One is about the presence of different types of content fragments. The other is about flyers length which shows close correlation with input material organisation.

3.1.Classifying content fragments

By studying the generated flyers, it is found that content fragments included in these flyers fall into four different types: namely, important content fragments, repetitive content fragments, irrelevant content fragments and marginal content fragments. Important content fragments are those which are important to flyers quality which based subjects' response in the post-experiment questionnaire. It is found that on average the number of important content fragments for the two groups are very close: 6.8 for Group A and 7.2 for Group B. Repetitive information refers to content fragments which are repeated multiple times in a flyer. We found that the total number of Repetitive fragments for Group A is 7 which are much larger than Group B's 2. Irrelevant information refers to content fragments which are not related to the interested project, research laboratory, or ICT Centre. Marginal content fragments are relevant to the topics of flyers. But, they are neither important, nor repetitive. There were 13.3 marginal fragments for Group A and 6.2 for Group B.

3.2. Flyer length

We look at the flyer length in terms of the number of pages and the number of content fragments which a flyer contains. Although no explicit length limit is given before the experiment, there is a big difference between the two groups. In terms of number of pages in the resulting flyers, Group A ranges from 1 to 4 pages while the spread for Group B is between 1 and 2. The average number of pages is 2.4 for Group A and 1.4 for Group B. Evidently, the average page number for Group A of 2.4 is much bigger than 1.3 for Group B. Indeed, the t-test shows that the difference is statistically significant.

In terms of number of content fragments, Group A ranges from 7 to 21, while the spread for Group B is between 7 and 12. The average number of content fragments is about 15 for Group A and about 11 for Group B. The difference, however, is not statistically significant.

4. Flyer quality evaluation

In order to assess the effect of organisation of input material on flyer composition quality, we need a reasonable method. As mentioned earlier, flyer quality is used to refer to content selection rather than flyer structure or flyer layout. To our knowledge, content selection evaluation has not been studied in the context of document composition. However, it has been studied in the context of summarisation [6] [9]. The Pyramid method [6] is an empirically motivated method for evaluating the quality of summarisation.

Our analysis of flyer content described in Section 3 concludes that there are four different types of content fragments. These different types of content fragments play different roles in flyer quality. Specifically, repetitive and irrelevant content fragments will contribute negatively towards flyer quality while important and marginal content fragments will contribute positively and neutrally towards flyer quality. Furthermore, irrelevant content fragments will do more damage to flyer quality than repetitive ones. Based on the observation that, in document composition, there are not only positive contribution content but also negative contribution content, the X-method is developed which is an extension of the Pyramid method for dealing with the impact pollutant content fragments. The X-method is also a X of content fragments which consists of two opposing pyramids (e.g., *Figure 1*): one positive pyramid and one negative pyramid. Like the Pyramid method [6], based on a pool of human input, a pyramid of important content fragments (ICC) will be computed. Each ICC has a weight corresponding to the number of nomination it gets as described in Section 3.1. For example, the content fragment "*contact*" is nominated by 8 subjects while "*background*" is nominated by 6 subjects. This forms the

positive pyramid of IICC arranged by weight in descending order. Unlike the Pyramid method, a pyramid of pollutant content fragments is also formed. The default weights for irrelevant and repetitive content fragments are -*n* and -$\frac{n}{2}$, here, n is the number of tiers in the positive pyramid.

With X-method, the score of a flyer D_{pyr} is the ratio of the sum of the weights of its content fragments D to the sum of weights of the optimal flyer with the same number of content fragments $D_{\max}$.

$$D_{pyr} = \frac{D}{D_{\max}} \qquad (1)$$

Where, D *-- the sum of the weights of its content fragments;*

$D_{\max}$ *-- the sum of weights of the optimal flyer with the same number of content fragments.*

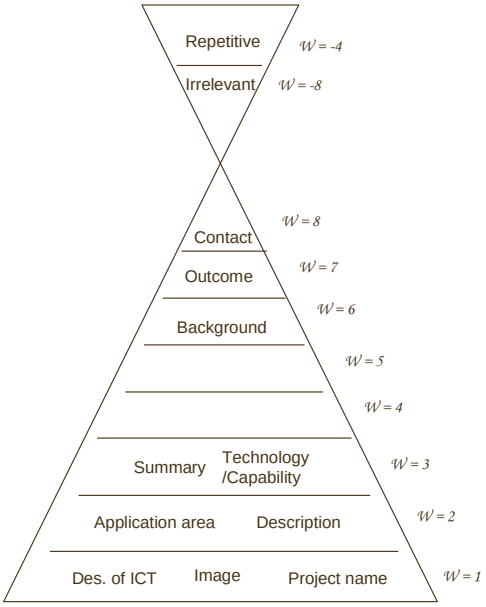

Figure 1. Content fragments weights for the X-method.

Suppose the positive pyramid has *n* tiers, T_i, with tier T_n on top and T_1 on the bottom. The weight of content fragments in tier T_i will be i. Let $|T_i|$ denote the number of content fragments in tier T_i. Let D_i be the number of content fragments in the flyer that appears in T_i. Content fragments in a flyer that do not appear in the pyramid are assigned weight zero. The total content fragment weight *D* is:

$$D = \sum_{i=1}^{n} iD_i - \frac{n}{2}\sum_{j=1}^{2} jR_j \qquad (2)$$

Where,

R_1 *-- the number of repetitive content fragments in a flyer;*

R_2 *-- the number of irrelevant content fragments in a flyer;*

The maximum content score for a flyer with x content fragments is:

$$D_{\max} = \sum_{i=j+1}^{n} iT_i + j\left(x - \sum_{i=j+1}^{n} T_i\right) \qquad (3)$$

Applying equations (1), (2), and (3) to the experiment data for Task 1, the X-method scores can be calculated for different flyers produced by the two groups of subjects (Figure 2). As you can see, the average score for the Group A is 0.77 compared to 0.91 for Group B. That means that the average flyer quality for the structured group is better than those for the unstructured group. However, the t-test (Table 1) shows that the difference is not statistically significant.

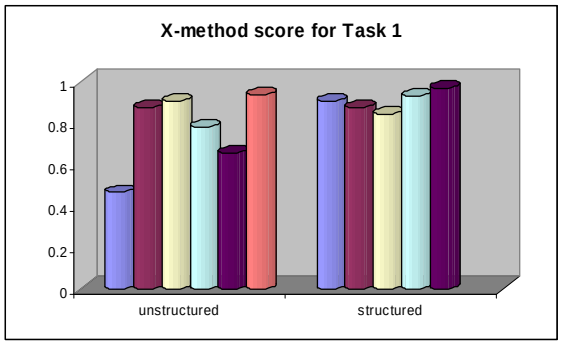

Figure 2. Quality score for Task 1

Table 1. T-test results in terms of the X-method scores for task 1

Task	$Mean_A$-$Mean_B$	t	df	P one-tailed
1	-0.1662	-1.67	9	0.064629

Similarly, the X-method scores for the different flyers composed by the two groups of subjects are calculated which is listed (Figure 3). Interestingly, this time, Group A on average scored 0.78 which is considerably higher than 0.61 scored by Group B. That means that the structured group composed better two-project combined flyers than the structured group. However, the difference is not statistically significant as demonstrated by t-test (Table 2).

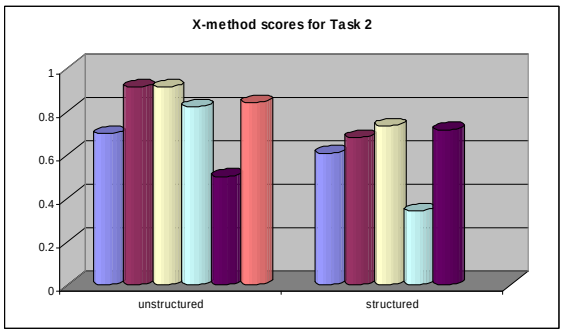

Figure 3. The quality score for Task 2

Table 2. T-test results in terms of the quality scores for task 2

Task	$Mean_A$- $Mean_B$	t	df	P one-tailed
2	0.1652	+1.71	9	0.06072

5. Discussion and conclusion

In this paper, we have presented an empirical study on document composition. It is found that the unstructured group produced significantly longer flyers than the structured group for both tasks. We surmise that this length discrepancy is caused by the difference in the organisation of content fragments. Since it is not easy to find information a subject needs from the unstructured list, subjects were inclined to get hold on all information they may come across related to the target project. In contrast, finding needed content fragments is not an issue for the structured group. Consequently, they were able to focus on strategic issues and more conscientious about the proper length of produced flyer. It is also found that there were considerably more pollutant content fragments in the resulting flyers produced by the unstructured group than those by the structured group.

The study of the effect on document quality is limited to content selection and no consideration is paid towards document structure and layout. It is found that, on average, subjects who used the structured input composed considerably better single project flyers than those who used unstructured input. But, when it comes to two-project combined flyers, subjects who used unstructured input produced considerably better flyers that those who used structured input. However, the difference in both cases is not statistically significant.

In conclusion, organisation of input information has a strong impact on document quality in document composition. Providing the topic of a document is clearly defined, semantically structured input would have positive impact on document composition quality. Another contribution of the paper lies in the development of the X-method for evaluating content selection in document composition. In practical terms, we will incorporate structure in future Constructor development for presenting retrieval services.

Acknowledgements: We are grateful to Cécile Paris and John Colton for their comments to early drafts of the paper. Our appreciation also goes to all the people who participated in our experiment.

References

[1] Lu, S., Paris, C. and Wu, M. (2005) 'Document modelling for customised information delivery', Proceeding of The Tenth Australasian Document Computing Symposium (ADCS 2005), Sydney, pp.11-18.

[2] Lu, S. and Paris, C. (2007): Specifying adaptive documents: an authoring tool prototype and user studies. To appear in the Special issue on Authoring of Adaptive and Adaptable Hypermedia of the International Journal of Learning Technology, edited by Alexandra Cristea and Rosa Carro.

[3] Lu, S. and Paris, C. (2006): Authoring Content Structure for Adaptive Documents. In the Proceedings of the International Workshop on Authoring of Adaptive and Adaptable Hypermedia at the 4th International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems, Dublin, Ireland, June 21-23, 2006.

[4] Mann, W.C. and Thompson, S.A. (1988) 'Rhetorical Structure Theory: Toward a functional theory of text organisation', Text, vol.8 (3), pp.243-281.

[5] Moore, J. and Paris, C. (1993) 'Planning Text for Advisory Dialogues: Capturing Intentional and Rhetorical Information', Journal of Computational Linguistics, vol.19 (4), pp.651 – 694.

[6] Nenkova, Ani and Passonneau, Rebecca J. (2004). Evaluating content selection in summarization: The pyramid method. In Proceedings of the Joint Annual Meeting of Human Language Technology (HLT) and the North American chapter of the Association for Computational Linguistics (NACL), Boston, MA.

[7] Paris, C. and Colineau, N. (2006) 'Scifly: tailored corporate brochures on demand', CSIRO Tech Report, No. 06/268, www.csiro.au/scifly.

[8] Paris, C., Wu, M., Vander Linden, K., Post, M. and Lu, S. (2004). Myriad: An Architecture for Contextualized Information Retrieval and Delivery, in AH2004: International Conference on Adaptive Hypermedia and Adaptive Web-based Systems. August 23-26, The Netherlands. pp. 205-214.

[9] Radev, D.R., Teufel, S., Saggion, H., Lam, W., Blitzer, J., Qi, H., elebi, A., Liu, D., Drabek, E.: Evaluation challenges in large-scale document summarization. In Proc. of the 41st Annual Meeting of the Association for Computational Linguistics. (2003) pp. 375—382.