# Variable Length On-Line Document Generation

Mick O'Donnell
Department of Artificial Intelligence,
University of Edinburgh,
80 South Bridge, Edinburgh. EH1 1HN, UK.
email: micko@aisb.ed.ac.uk

## 1 Introduction

Adaptive hypertext involves variation in either the content of, or linking between, hypertext documents in response to some variation in the context of browsing. This paper is concerned with one form of content adaption: the adaption of the length of a document to suit the needs of the user. We call these documents *variable length documents* (VLDs). In such documents, the user designates how long the document should be, and it is presented at that length.[1]

The value of such a technique is obvious – some users want more detail and explanation, while others want or need less. A static hypertext document can only offer one level of detail. A variable-length document allows the user to choose their level of verbosity.

However, VLDs have not been practical given the current level of technology. Some approaches (e.g., Rino & Scott 1996) have discussed document summarisation on the basis of full natural language generation (NLG). However, the cost of authoring knowledge to support fully-generated documents has prohibited this approach, even if we allow that NLG has reached the required degree of robustness. Ono *et al* (1994) proposes building VLDs on the basis of *automatic document structure recognition*. However, I am yet to be convinced that such recognition is reliable on free text as yet.

This paper proposes an alternative technique for establishing VLDs, which substantially reduces the effort needed to get such documents on-line. Our technique involves the marking up of an existing natural language document using a document mark-up tool which we have developed, called the *RST-Tool* (see O'Donnell 1997b). The mark-up of the document is used to determine optimal locations for pruning of the text. Documents so marked-up can then be used for variable-length presentation, on the web or in some other hypertext environment.

Document mark-up involves indicating the *rhetorical structure* of the text, in terms of *Rhetorical Structure Theory* (RST – Mann & Thompson 1987). RST structures a text in terms of a dependency structure, showing the rhetorical dependence between units of text. For instance, the first sentence of this paper is dependent on the second sentence, and stands in the relationship of BACKGROUND. The head of the dependency relation is called the nucleus, while the dependent text is called the *satellite*.

A common hypothesis about RST is that satellites are less essential to the text's goals than the nuclei. Thus, to produce a document of a particular length, we need only prune off branches of the RST tree until the required word limit is reached. The method of pruning is described in section 2. (Ono *et al* (1994) also prunes RST structures to achieve text summarisation. See O'Donnell (1997a) for a comparison of the two approaches).

Strictly speaking, hypertext involves text which can be clicked upon to reach some other body of related text. The technique described here does not so much apply to hyperlinking, but

[1] See http://toros.ces.cwru.edu/veli/papers.html for work by Veli J. Hakkoymaz, applying this idea to multimedia presentations.

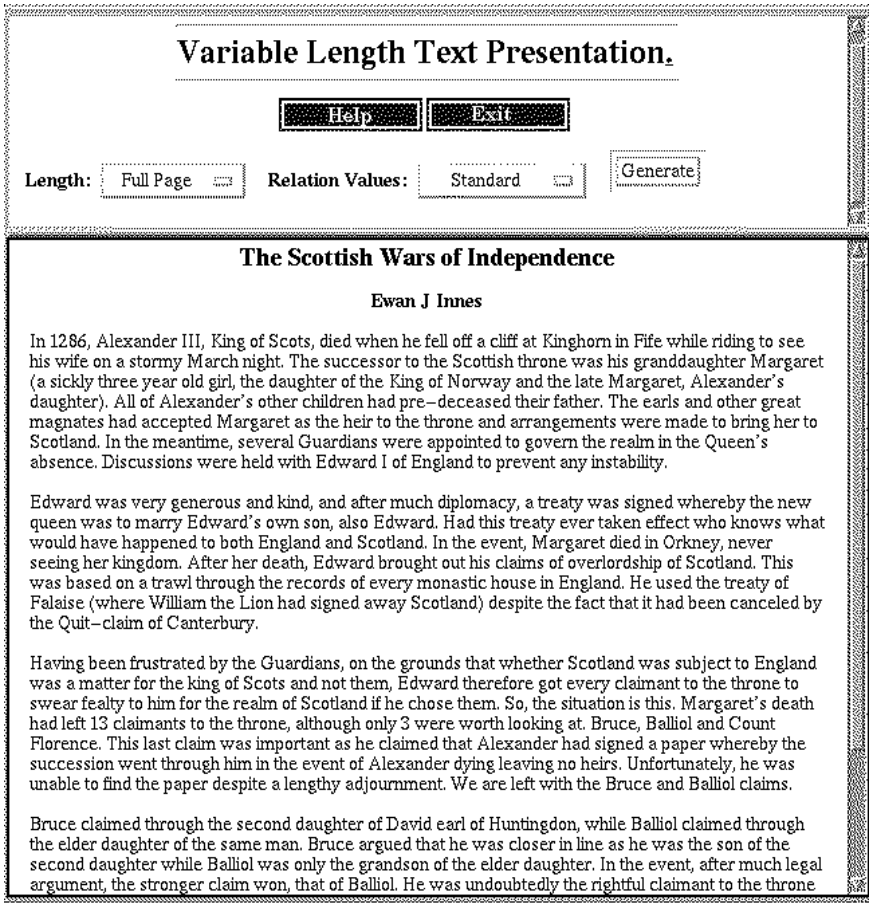

Figure 1: The VLDP interface

to the content of the nodes which are linked. Applying the technique to a hypertext document would change the contents (length) of each node in the document. A future development of the system will allow the user to zoom in on text by clicking on it: as a result of clicking on sentence punctuation, the full text dependent on that sentence will be presented.

The work reported here has been carried out as part of the *ILEX* project, whose goal is to produce dynamically generated descriptions of objects in museums. See Knott *et al* (1996) for details.

## 2 Variable-Length Document Presentation

After a text has been marked up with the RST-Tool, it can be be registered with a cgi-script which knows how to present the document at variable lengths. The document is initially presented full-length (see figure 1), but the user can select from a set of reduced lengths. Figure 2 shows two versions of the same document, although with a 200 word limit set (see section 2.3 for more detail on user-model variation). The system can be seen working at http://cirrus.dai.ed.ac.uk:8000/cgi-bin/jewel-start?start/summariser.

The marked-up document is structured as a dependency tree, with each node of the tree being a segment of text. Each branch of the tree represents a dependency relationship between two text nodes. The process of pruning is then as follows. Figure 3 show the dependency analysis of a single sentence of the text. Note that while RST usually does not deal with dependency within the clause, for this application I provided a set of intra-clausal relations. Pruning of clausal adjuncts is an important source of summarisation without meaning-loss.

The markup tool also allows the inclusion of *multinuclear* structures (a node whose children are text nodes of equal status, e.g., *Sequence*, *Joint*), and *schemas*, what are sometimes called

"story grammars", allowing a sequence of named elements of structure, e.g., INTRODUCTION, BODY, CONCLUSIONS, BIBLIOGRAPHY, etc. Both of these structures are handled similarly to RST structures, so will not be discussed further.

## 2.1 Assigning relevance scores to text nodes

Each RST-relation type is assigned a relevance rating. For instance, ELABORATION may have a score of 0.40 (low relevance), while PURPOSE might be scored more highly. The first step, before pruning, is to assign each segment of the text a relevance score, between 0.0 and 1.0. The root of the tree is assigned a relevance value of 1.0. Each of its satellites is then assigned a relevance based on this value times the relevance value of the relationship linking it. Through a process of recursive descent, we assign each node in the tree the relevance level of its parent, multiplied by the relevance score of the relation which connects it. For instance, if the top-node in figure 3 had a relevance of 0.70, and the COOCURRENCE relation was valued at 0.6, then the text *when he fell off a cliff* would have relevance 0.42. Nodes lower in the RST-tree (less nuclear) will thus have lower relevance than higher nodes (more nuclear), and will thus be the first to be pruned.

This is a simple mechanism, but it has shown good results in producing reasonable texts at whatever degree of verbosity. There are however some cases where this method breaks down – nuclearity does not always reflect centrality of information. Sometimes an author introduces information in a rhetorically unimportant place, yet that information may be needed later to understand the argument. One example of this in the summary shown earlier is where the original text had said: *he was faced with constant pressure from Edward to sign. He refused to do so.* In the summary, "to sign" was pruned, but it was actually a central concept, and the anaphoric "so" failed because of its pruning.

The text-nodes are then placed in a queue, position based on their relevance score.

## 2.2 Pruning the RST-tree

When a request is received to display the text at a particular length, the system needs to determine which text-nodes to display. Taking each node in turn from the relevance queue (starting with the most relevant), the program checks to see if including this text node will push the word-count over the limit. If not, it adds the node to the nodes-to-be-expressed list, and increments the words-so-far count. When the word-limit is exceeded, the procedure then turns to expressing the selected nodes. The nodes are expressed in the order in which they appeared in the original full text.

---

*How&Why Summary*: Alexander III, King of Scots, died. The successor to the Scottish throne was his granddaughter Margaret. The earls and other great magnates had accepted Margaret as the heir to the throne and arrangements were made to bring her to Scotland. Several Guardians were appointed **to govern the realm**. Discussions were held with Edward I **to prevent any instability**. A treaty was signed **whereby the new queen was to marry Edward's own son**. Margaret died. Edward brought out his claims of overlordship. **He used the treaty of Falaise**. ...

---

*Where&When Summary*: **In 1286,** Alexander III, King of Scots, died **at Kinghorn in Fife**. The successor to the Scottish throne was his granddaughter Margaret. The earls and other great magnates had accepted Margaret as the heir **to the throne** and arrangements were made to bring her to Scotland. **In the meantime,** several Guardians were appointed. Discussions were held with Edward I. A treaty was signed. Margaret died **in Orkney**. **After her death,** Edward brought out his claims of overlordship **of Scotland.** ...

Figure 2: Summaries with different weighting sets

---

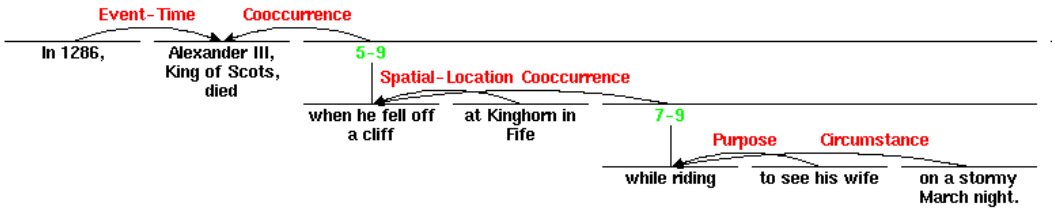


Figure 3: A Typical RST Analysis

Note that the satellites of a node will always have lower or equal relevance than the node itself, so we never include a satellite in the nodes-to-be-expressed list if its nucleus is not, which can produce incoherent text.

### 2.3 User-Variation of relation weightings

The actual values associated with each relation are not fixed, but can be varied by the user. The user can select values which reflect their interests, highlighting some types of rhetorical relations, and ignoring others. For instance, figure 2 demonstrated the slight difference of information (bold font) included in the text when switching between a relation set preferring spatial and temporal location/extent vs. a system which emphasises causes, purposes, reasons, etc.

## 3 Preserving Coherence in Dynamic Document Presentation

When summarising a document, we do damage to various aspects of the document's coherency. Four areas which are at risk are:

- **Paragraphing**: Deleting sentences without changing paragraph boundaries would produce a text of many short paragraphs, reducing readability. Rather than attempt to repair document paragraphing, we have found it easier to throw away the original paragraphing, and re-determine paragraph boundaries. Our algorithm, which will be described elsewhere, optimises within the trade-off between two factors: *Paragraph Rhythm*: paragraphs should be roughly the same size; and *Rhetorical Clustering*: paragraphs should represent material closely related in terms of rhetorical structure.
- **Punctuation**: When deleting an intra-sentence nucleus, we may also delete the punctuation it carries. For instance, in (N: *Edward surrendered,*)(S: *in 1245.*), deletion of the satellite leaves us with a sentence terminated by a comma. Our system ensures all sentences start with a capital, and recovers the sentence-terminating punctuation from any pruned segments where necessary.
- **Referring Expressions**: When deleting sections of a text, we may destroy the referential cohesion of a text in two ways. Firstly, we might delete the introduction of an entity, which provided the entities name, or other characteristics which allow the reader to identify the entity correctly. A second problem involves changing the referential environment of entities. References which are contextually unambiguous in the full text may be brought into close proximity to other entities which are potential confusers. In the system as implemented so far, there has been no attempt to correct these problems. A future version will allow a user to mark up the co-reference of NPs in the text, allowing some degree of repair to reference problems after pruning.
- **Discourse Markers**: Markers of rhetorical relations are usually attached to satellites, and so there is no problem when the satellite is pruned. However, in some peoples analyses, some relations are marked on the nucleus, not the satellite. In others, both the nucleus and satellite are marked (e.g., if/then). When we delete the satellite, we should ensure

that the discourse marker is also removed from the nucleus. However, due to the rarity of nucleus marking in our corpus, this problem has not been a problem so far.

# 4 Summary

This paper has described a system for presenting variable-length on-line documentation, which allows the user to select the degree of verbosity of the text presented. The results so far on a small-scale have shown that reasonable-quality texts can be produced dynamically. The cost of document mark-up stops this approach being used on texts of short display-life, but makes it economical for documents of longer duration where length-variability has value.

Apart from text-length, VLDs allow the user a small degree of content-control, in that the user can determine the relevance of each RST relation (or of elements of a schema).

The major problem for the system involves restoring coherence after text-pruning, particularly in areas of reference, discourse markers, paragraphing and punctuation. The problems of paragraphing and punctuation have been solved, and solutions are suggested for the other two areas.

Regardless of the problems of this approach, the system is up and running on-line. New documents are being added as time allows, to test the generalisability of the approach.

# 5 Bibliography


Knott, Alistair, Chris Mellish, Jon Oberlander & Mick O'Donnell. 1996. "Sources of Flexibility in Dynamic Hypertext Generation". Proceedings of the 8th International Workshop on Natural Language Generation, Herstmonceux Castle, UK, 13-15 June.

Mann, William C. & Sandra Thompson, 1987. "Rhetorical Structure Theory: A Theory of Text Organization". Technical Report ISI/RS-87-190.

O'Donnell, Michael. 1997a. "Variable Length On-line Document Presentation. Proceedings of the 6th European Workshop on Natural Language Generation. March 24 - 26. Gerhard-Mercator University, Duisburg, Germany.

O'Donnell, Michael. 1997b. "RST-Tool: An RST Analysis Tool". Proceedings of the 6th European Workshop on Natural Language Generation. March 24 - 26. Gerhard-Mercator University, Duisburg, Germany.

Ono, Kenji, Kazuo Sumita, & Seiji Miike. 1994. "Abstract generation based on rhetorical structure extraction". Proceedings of the 15th International Conference on Computational Linguistics (COLING-94), Vol. 1. August 5-9, Kyoto, Japan.

Rino, L.H.M. & Scott, D.R. 1996. "A Discourse Model for Gist Preservation". In Dibio L. Borges and Celso A.A. Kaestner (eds.), Advances in Artificial Intelligence (Proceedings of the 13th Brazilian Symposium on Artificial Intelligence), pp. 131-140. Springer-Verlag, Germany.

Sparck Jones, Karen. 1993. "What might be in a summary?", Information Retrieval 93: Von der Modellierung zur Anwendung (Ed. Knorz, Krause and Womser-Hacker), Konstanz: Universitatsverlag Konstanz), 9-26.