Experimenting with the Interaction between Aggregation and Text Structuring

Hua Cheng
Division of Informatics
University of Edinburgh
80 South Bridge, Edinburgh EH1 1HN, UK
Email: huac@dai.ed.ac.uk

Abstract

In natural language generation, different generation tasks often interact with each other in a complex way, which is hard to capture in the pipeline architecture described by Reiter (Reiter, 1994). This paper focuses on the interaction between a specific type of aggregation and text planning, in particular, maintaining local coherence, and tries to explore what preferences exist among the factors related to the two tasks. The evaluation result shows that it is these preferences that decide the quality of the generated text and capturing them properly in a generation system could lead to coherent text.

1 Introduction

In automatic natural language generation (NLG), various versions of the pipeline architecture specified by Reiter and Dale ((Reiter, 1994) and (Reiter and Dale, 1997)) are usually adopted. They successfully modularise the generation problem, but fail to capture the complex interactions between different modules. Take aggregation as an example. It combines simple representations to form a complex one, which in the mean time leads to a shorter text as a whole. There is no consensus as to where aggregation should happen and how it is related to other generation processes ((Wilkinson, 1995) and (Reape and Mellish, 1999)).

We think that the effect of aggregation spreads from text planning to sentence realisation. The task of text planning is to select the relevant information to be expressed in the text and organise it into a hierarchical structure which captures certain discourse preferences such as preferences for global coherence (e.g. the use of RST relations (Mann and Thompson, 1987)) and local coherence (e.g. center transitions as defined in Centering Theory (Grosz et al., 1995)). Aggregation affects text planning by taking away facts from a sequence featuring preferred center movements for subordination. As a result, the preferred center transitions in the sequence are cut off. For example, comparing the two descriptions of a necklace in Figure 1, *2* is less coherent than *1* because of the shifting from the description of the necklace to that of the designer. To avoid this side effect, aggregation should be considered in text planning, which might produce a different planning sequence.

Aggregation is also closely related to the task of referring expression generation. A referring expression is used not only for identifying a referent, but also for providing additional information about the referent and expressing the speaker's emotional attitude toward the referent (Appelt, 1985). The syntactic form of a referring expression affects how much additional information can be expressed, but it can only be determined after sentence planning, when the ordering between sentences and sentence components has been decided. This demands that the factors relevant to referring expression generation and aggregation be considered at the same time rather than sequentially to generate referring expressions capable of serving multiple goals.

In this paper, we are concerned with a specific type of aggregation called embedding, which shifts one clause to become a component within the structure of an NP in another clause. We focus on the interaction between maintaining local coherence and embedding, and describe how to capture this interaction as preferences among related factors. We believe that if these preferences are used properly, we would be able to generate more flexible texts without sacrificing quality. We implemented the preferences

1. This necklace is in the Arts and Crafts style. Arts and Crafts style jewels usually have an elaborate design. They tend to have floral motifs. For instance, this necklace has floral motifs. It was designed by Jessie King. King once lived in Scotland.

2. This necklace, which was designed by Jessie King, is in the Arts and Crafts style. Arts and Crafts style jewels usually have an elaborate design. They tend to have floral motifs. For instance, this necklace has floral motifs. King once lived in Scotland.

Figure 1: An aggregation example

in an experimental generation system based on a Genetic Algorithm to produce museum descriptions, which describe museum objects on display. The result shows that the system can generate a number texts of similar qualities to human written texts.

2 Embedding in a GA Text Planner

To experiment with the interaction between maintaining local coherence and embedding, we adopt the text planner based on a Genetic Algorithm (GA) as described in (Mellish et al., 1998). The task is, given a set of facts and a set of relations between facts, to produce a legal rhetorical structure tree using all the facts and some relations. A fragment of the possible input is given in Figure 2.

A genetic algorithm is suitable for such a problem because the number of possible combinations is huge, the search space is not perfectly smooth and unimodal, and the generation task does not require a global optimum to be found. The algorithm of (Mellish et al., 1998) is basically a repeated two step process - first sequences of facts are generated by applying GA operators (crossover and mutation) and then the RS trees built from these sequences are evaluated. This provides a mechanism to integrate various planning factors in the evaluation function and search for the best combinations of them.

To explore the whole space of embedding, we did not perform embedding on structured facts or on adjacent facts in a linear sequence because these might restrict the possibilities and even miss out good candidates. Instead, we defined an operator called *embedding mutation*. It randomly selects two units (say U_i and U_k) mentioning a common entity from a sequence $[U_1, U_2, ..., U_i, ..., U_k, ..., U_n]$ to form a list $[U_i, U_k]$ representing an embedding. The list substitutes the original unit U_i to produce a new sequence $[U_1, U_2, ..., [U_i, U_k], ..., U_n]$, which is then evaluated and ordered in the population.

3 Capturing the Interactions as Preferences

A key requirement of the GA approach is the ability to evaluate the quality of a possible solution. We claim that it is the relative preferences among factors rather than each individual factor that play the crucial role in deciding the quality. Therefore, if we can capture these preferences in a generation system properly, we would be able to produce coherent text. In this section, we first discuss the preferences among factors related to text planning, based on which those for embedding can be introduced.

3.1 Preferences for global coherence

Following the assumption of RST, a text is globally coherent if a hierarchical structure like an RST tree can be constructed from the text. In addition to the semantic relations and the *Joint* relation[1] used in (Mellish et al., 1998), we assume a *Conjunct* or *Disjunct* relation between two facts with at least two identical components, so that semantic parataxis can be treated as a combining operation on two subtrees connected by the relation.

Embedding a *Conjunct* relation inside another semantic relation is not preferred because this could convey wrong information, for example, in Figure 3, *2* cannot be used to substitute *1*. Also a semantic relation is preferred to be used whenever possible. Here is the preferences concerning the use of relations, where "A>B" means that A is preferred over B:

[1] In (Mellish et al., 1998), a *Joint* relation is used to connect every two text spans that do not have a normal semantic relation in between.

```
fact(choker,is,broad,fact_node-1).
fact('Queen Alexandra',wore,choker,fact_node-2).
fact(choker,'can cover',scar,fact_node-3).
fact(band,'might be made of',plaques,fact_node-4).
fact(band,'might be made of',panels,fact_node-5).
fact(scar,is,'on her neck',fact_node-6).
...
rel(in_that_reln,fact_node-2,fact_node-3,[]).
rel(conjunct,fact_node-4,fact_node-5,[]).
```

Figure 2: A fragment of the input to the GA text planner

1. The necklace is set with jewels in that it features cabuchon stones. Indeed, an Arts and Crafts style jewel usually uses cabuchon stones. An Arts and Crafts style jewel usually uses oval stones.

2. The necklace is set with jewels in that it features cabuchon stones. Indeed, an Arts and Crafts style jewel usually uses cabuchon stones and oval stones.

Figure 3: Conjunct and semantic relations

Heuristic 1 *Preferences among features for global coherence:*
a semantic relation > Conjunct > Joint > parataxis in a semantic relation

3.2 Preferences for local coherence

In Centering Theory, Rule 2 specifies preferences among center transitions in a locally coherent discourse segment: sequences of *continuation* are preferred over sequences of *retaining*, which are then preferred over sequences of *shifting*. Instead of claiming that this is the best model, we use it simply as an example of a linguistic model being used for evaluating factors for text planning.

Another type of center transition that appears frequently in museum descriptions is *associate shifting*, where the description starts with an object and then moves to a closely associated object or perspectives of that object. Our observation from museum descriptions shows that *associate shifting* is preferred by human writers to all other types of movements except for *center continuation*.

Oberlander et al. (1999) define yet another type of transition called *resuming*, where an utterance mentions an entity not in the immediate previous utterance, but in the previous discourse. The following is the preferences among features for local coherence:

Heuristic 2 *Preferences among center transitions and semantic relations:*
Continuation > Associate shifting > Retaining > Shifting > Resuming
a semantic relation > Joint + Continuation

3.3 Preferences for embedding

For a randomly produced embedding, we must be able to judge its quality. We distinguish between a *good, normal* and *bad embedding* based on the features it bears[2]. A *good embedding* is one satisfying all following conditions:

1. The referring expression is an indefinite, a demonstrative or a bridging description (as defined in (Poesio et al., 1997)).
2. The embedded part can be realised as an adjective or a prepositional phrase (Scott and de Souza, 1990)[3].
3. The embedded part does not lie between text spans connected by semantic parataxis or hypotaxis (Cheng, 1998).
4. There is an available syntactic slot to hold the embedded part.

[2]We do not claim that the set of features is complete. In a different context, more criteria might have to be considered.

[3]We assume that syntactic constraints have been inserted before in text planning, using Meteer's Text Structure (Meteer, 1992) for example.

A *good embedding* is highly preferred and should be performed whenever possible. A *normal embedding* is one satisfying condition 1, 3 and 4 and the embedded part is a relative clause. A *bad embedding* consists of all those left.

To decide the preferences among embeddings and center transitions, let's look at the paragraphs in Figure 1 again. The only difference between them is the position of the sentence "*This necklace was designed by Jessie King*", which can be represented in terms of features of local coherence and embedding as follows:

> *the last three sentences in 1: Joint + Continuation + Joint + Shifting*
> *the last two sentences plus embedding in 2: Joint + Resuming + Normal embedding*

Since *1* is preferred over *2*, we have the following heuristics:

Heuristic 3 *Preferences among features for embedding and center transition:*
Continuation + Shifting + Joint > Resuming + Normal embedding
Good embedding > Normal embedding > Joint > Bad embedding
Good embedding > Continuation + Joint

4 Justifying the Evaluation Function

We have illustrated the linguistic theories that can be used to evaluate a text. However, they only give evidence in qualitative terms. For a GA-based planner to work, we have to come up with actual numbers that can be used to evaluate an RS tree.

We extended the existing scoring scheme of (Mellish et al., 1998) to account for features for local coherence, embedding and semantic parataxis. This resulted in the rater 1 in Table 1[4], which satisfied all the heuristics introduced in Section 3.

We manually broke down four human written museum descriptions into individual facts and relations and reconstructed sequences of facts with the same orderings and aggregations as in the original texts. We then used the evaluation function of the GA planner to score the RS trees built from these sequences. In the meantime, we ran the GA algorithm for 5000 iterations on the facts and relations for 10 times. All human texts were scored among the highest and machine generated texts can get scores very close to human ones sometimes (see Table 2 for the actual scores of the four texts). Since the four human texts were written and revised by museum experts, they can be treated as "nearly best texts". The result shows that the evaluation function based on our heuristics can find good combinations.

To justify our claim that it is the preferences among generation factors that decide the coherence of a text, we fed the heuristics into a constraint-based program, which produced a lot of raters satisfying the heuristics. One of them is given in Table 1 as the rater 2. We then generated all possible combinations, including embedding, of seven facts from a human text and used the two raters to score each of them. The two distributions are shown in Figure 4.

The qualities of the generated texts are normally distributed according to both raters. The two raters assign different scores to a text as the means of the two distributions are quite different. There is also slight difference in standard deviations, where the deviation of Rater 2 is bigger and therefore it has more distinguishing power. Despite these differences, the behaviours of the two raters are indeed very similar as the two histograms are of roughly the same shape, including the two right halves which tell us how many good texts there are and if they can be distinguished from the rest. The difference in standard deviations is not significant at all. So the distributions of the scores from the two raters show that they behave very similarly in distinguishing the qualities of texts from the same population.

As to what extent the two raters agree with each other, we drew the scatterplot of the scores, which showed a strong positive linear correlation between the variables representing the two scores. That is, the higher the score from rater 1 for a given text of the population, the higher the score from rater 2 tends to be. We also calculated the Pearson correlation coefficient between the two raters and the corre-

[4]The table only shows the features we are concerned with in this paper.

Features/Factors	Values 1	Values 2
Semantic relations		
a Joint	-20	-46
a Conjunct or Disjunct	10	11
a relation other than Joint, Conjunct or Disjunct	21	69
a Conjunct inside another semantic relation	-50	-63
a precondition not satisfied	-30	-61
Center transitions		
a Continuation	20	7
an Associate shifting	16	1
a Shifting	14	-3
resuming a previous center	6	-43
Embedding		
a Good embedding	6	3
a Normal embedding	3	0
a Bad embedding	-30	-64
Others		
topic not mentioned in the first sentence	-10	-12

Table 1: Two different raters satisfying the same constraints

	text 1	text 2	text 3	text 4
scores of the human texts	170	22	33	24
highest scores of the generated texts	167	24	31	25
average scores of the generated texts	125.7	18.9	26.1	9.3

Table 2: The scores of four human written texts

lation was .9567. So we can claim that for this data, the scores from rater 1 and rater 2 correlate, and we have fairly good chance to believe our hypothesis that the two raters, randomly produced in a sense, agree with each other on evaluating the text and they measure basically the same thing.

Since the two raters are derived from the heuristics in Section 3, the above result partially validates our claim that it is the relevant preferences among factors that decide the quality of the generated text.

5 Summary and Future work

This paper focuses on the complex interactions between embedding and planning local coherence, and tries to capture the interactions as preferences among related features. These interactions cannot be easily modelled in a pipeline architecture, but the GA-based architecture offers a mechanism to coordinate them in the planning of a coherent text. The result shows to some extent that capturing the interactions properly in an NLG system is important to the generation of coherence text.

Our experiment could be extended in many aspects, for example, validating the evaluation function through empirical analysis of human assessments of the generated texts, and experimenting with the interaction between aggregation and referring expression generation. The architecture based on the Genetic Algorithm can also be used for testing interactions between or within other text generation modules. To generalise our claim, a larger scale experiment is needed.

Acknowledgement This research is supported by a University of Edinburgh Studentship.

References

Douglas Appelt. 1985. Planning english referring expressions. *Artificial Intelligence*, 26:1–33.

Hua Cheng. 1998. Embedding new information into referring expressions. In *Proceedings*

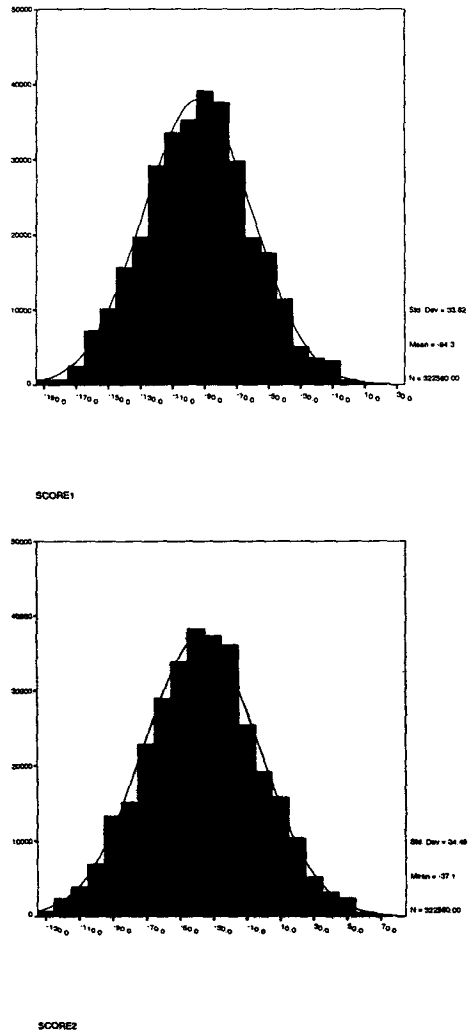

Figure 4: Histogram of the scores from rater 1 (top) and rater 2 (bottom)

of COLING-ACL'98, pages 1478–1480, Montreal, Canada.

Barbara Grosz, Aravind Joshi, and Scott Weinstein. 1995. Centering: A framework for modelling the local coherence of discourse. *Computational Linguistics*, 21(2):203–226.

William Mann and Sandra Thompson. 1987. Rhetorical Structure Theory: A Theory of Text Organization. Technical Report ISI/RR-87-190, Information Sciences Institute, University of Southern California.

Chris Mellish, Alistair Knott, Jon Oberlander, and Mick O'Donnell. 1998. Experiments using stochastic search for text planning. In *Proceedings of the 9th International Workshop on Natural Language Generation*, Ontario, Canada.

Marie Meteer. 1992. *Expressibility and The Problem of Efficient Text Planning.* Communication in Artificial Intelligence. Pinter Publishers Limited, London.

Jon Oberlander, Alistair Knott, Mick O'Donnell, and Chris Mellish. 1999. Beyond elaboration: Generating descriptive texts containing it-clefts. In T Sanders, J Schilperoord, and W Spooren, editors, *Text Representation: Linguistic and Psycholinguistic Aspects.* Benjamins, Amsterdam.

Massimo Poesio, Renata Vieira, and Simone Teufel. 1997. Resolving bridging references in unrestricted text. Research paper hcrc-rp87, Centre for Cognitive Science, University of Edinburgh.

Michael Reape and Chris Mellish. 1999. Just what is aggregation anyway? In *Proceedings of the 7th European Workshop on Natural Language Generation*, pages 20–29, Toulouse, France.

Ehud Reiter and Robert Dale. 1997. Building applied natural language generation systems. *Natural Language Engineering*, 3(1):57–87.

Ehud Reiter. 1994. Has a consensus nl generation architecture appeared, and is it psycholinguistically plausible? In *Proceedings of the 7th International Workshop on Natural Language Generation.*

Donia Scott and Clarisse Sieckenius de Souza. 1990. Getting the Message Across in RST-based Text Generation. In R. Dale, C. Mellish, and M. Zock, editors, *Current Research in Natural Language Generation*, pages 47–73. Academic Press.

John Wilkinson. 1995. Aggregation in Natural Language Generation: Another Look. Technical report, Computer Science Department, University of Waterloo.