Generating Discourse across Several User Models: Maximizing Belief while Avoiding Boredom and Overload

Ingrid Zukerman
Department of Computer Science
Monash University
Clayton, VICTORIA 3168
AUSTRALIA

Richard McConachy
Department of Computer Science
Monash University
Clayton, VICTORIA 3168
AUSTRALIA

Abstract

In this paper we present a content planning system which takes into consideration a user's boredom and cognitive overload. Our system applies a constraint-based optimization mechanism which maximizes a probabilistic function of a user's beliefs, and uses a representation of boredom and overload as constraints that affect the possible values of this function. Further, we discuss two orthogonal policies for relaxing the parameters of the communication process when these constraints are violated: conveying less information or breaking up the material into smaller chunks.

1 Introduction

It is generally accepted that to generate competent discourse, a speaker must take into consideration the beliefs and inferences of the addressee. In fact, several discourse planning systems rely on some sort of user model to generate appropriate descriptions, e.g., [Paris, 1988; Cawsey, 1990; Zukerman and McConachy, 1994]. However, little attention has been paid to the possibility that the user model being targeted may not fit the addressee. If the addressee is more competent than expected by the speaker, the addressee may be bored by the discourse. In contrast, if the addressee is less competent than expected, s/he may be overloaded with too much new information.

Our content planning mechanism addresses these problems by taking into consideration the speaker's uncertainty regarding which user model an addressee belongs to. Given a communicative goal and probabilities that the user belongs to a range of user models, our mechanism uses a constraint based optimization procedure to generate a set of *Rhetorical Devices (RDs)* which maximizes the material believed correctly by the user across the different user models, subject to constraints that restrict the user's boredom and cognitive overload. We have considered two policies for dealing with the violation of these constraints: (1) conveying less information, and (2) breaking up the information to be conveyed into smaller chunks. The generated RDs are then organized by means of the discourse structuring procedure described in [Zukerman and McConachy, 1993a], and rendered in English by means of the Functional Unification Grammar described in [Elhadad, 1992].

Without constraints
Substances containing ionic bonds are not flexible, e.g., salt crystals are brittle. Substances containing network-covalent bonds are also not flexible, e.g., diamonds are brittle. Substances containing ionic bonds have high melting points, and substances containing network-covalent bonds do not melt. [However, substances containing non-network-covalent bonds, e.g., ice, have low melting points.]
With constraints (boredom and overload)
Salt crystals are brittle, and diamonds are brittle. Salt crystals have high melting points, and diamonds do not melt. [However, ice has a low melting point.]

Table 1: Sample texts

Table 1 illustrates the discourse generated by our mechanism to convey information about different types of chemical bonding to a weak student. The text on the top was unconstrained, while the text on the bottom was generated using policy (1) above. The text on the top includes several examples to ensure that all the intended information is acquired, while in the text on the bottom, the abstract information is entirely replaced by concrete examples. When the constraints regarding boredom and overload are less stringent, it is sufficient to remove the last sentence in the discourse on the top [in square brackets]. This is because this sentence is generated to contradict a possible erroneous inference, which is of lower significance than the propositions in the input (Table 2).

2 Input and Output of the Content Planner

Our system receives two types of input: propositional and user model related. The propositional input contains (1) a set of propositions to be conveyed; (2) the degree of belief the user is expected to achieve with respect to each proposition; and (3) the significance of each proposition, i.e., how important it is that the user believes it. Propositions are grouped into aspects such as operation, domain and structure based on their main predicate. Table 2 shows the input from which the texts

Propositions to be Conveyed	Intended Belief	Sig.
¬[ionic-bond has-prop flexible]	*BELIEVED*	HIGH
¬[net-cov.-bond has-prop flexible]	*BELIEVED*	HIGH
[ionic-bond has-MP high]	*BELIEVED*	HIGH
¬[net-cov.-bond has-MP]	*BELIEVED*	HIGH

Table 2: Input that yields the sample texts

in Table 1 are generated.

The user model related input is a space of user models accompanied by the probability that the user belongs to each model in this space. In the current implementation we maintain five stereotypical user models: *excellent*, *good*, *average*, *mediocre* and *weak*. Each user model represents the *beliefs* of a particular type of user, his/her *inferences*, and his/her *profile* (a detailed description of these models appears in [Zukerman and McConachy, 1993b]).

Since it is easier to make broad assessments rather than pinpoint numerical assessments with respect to a user's beliefs, we represent a user's conjectured beliefs by means of the following *qualitative belief states* [Bonarini *et al.*, 1990]: {*BELIEVED, RATHER BELIEVED, CONTRADICTORY, UNKNOWN, RATHER DISBELIEVED, DISBELIEVED*}.

Our mechanism uses inference rules to model two main types of inferences: (1) *Direct inferences*, which reproduce directly the content of the discourse; and (2) *Indirect inferences*, which produce inferences that add information to what was said. The indirect inference rules considered in our model are based on those described in [Zukerman and McConachy, 1993b], e.g., *generalization*, *specialization* and *similarity*.

The profile attributed to a particular type of user determines the correctness and strength of the initial beliefs in the model of the user, and the degree of belief in the inference rules. For example, the profile of a mediocre student is characterized by weak convictions with respect to facts and inference rules, and lack of discrimination between correct and incorrect beliefs, and between sound and unsound inferences [Sleeman, 1984].

The output of the content planner is a set of RDs, such as those in Table 3 (which yield the sample texts in Table 1). The RDs are related to each other by means of discourse relations such as prerequisite, cause and elaboration [Mann and Thompson, 1987] (not shown in Table 3).

3 Definitions

Boredom takes place when too much known or easily inferable information is being presented, or when the discourse is too long. Overload occurs when a user stops paying attention due to the difficulties associated with digesting the information being presented.

3.1 Acquiring correct beliefs

In the process of generating discourse, speakers make sure that important propositions are conveyed, while placing less emphasis on less important propositions. Our function of belief reflects this behaviour by taking

Without constraints
Negate+Instantiate [ionic-bond has-prop flexible]
Negate+Instantiate [net-cov.-bond has-prop flexible]
Assert [ionic-bond has-MP high]
Negate [net-cov.-bond has-MP]
Assert [non-net-cov.-bond has-MP low]
Instantiate [non-net-cov.-bond]
With constraints (boredom and overload)
Instantiate ¬[ionic-bond has-prop flexible]
Instantiate ¬[net-cov.-bond has-prop flexible]
Instantiate [ionic-bond has-MP high]
Instantiate ¬[net-cov.-bond has-MP]
Instantiate [non-net-cov.-bond has-MP low]

Table 3: Sets of RDs that yield the sample texts

into consideration both the significance of the intended propositions and the degree to which these propositions are believed upon completion of the discourse.

$$BEL = \sum_{p} f_{BEL}(p) Sig(p) \qquad (1)$$

where $Sig(p)$ is the significance or importance of a proposition, and

$$f_{BEL}(p) = \begin{cases} 1 & \text{if } |bel_{act}(p)| \geq |bel_{int}(p)| \text{ and} \\ & sign(bel_{act}(p)) = sign(bel_{int}(p)) \\ \frac{bel_{act}(p)}{bel_{int}(p)} & \text{otherwise.} \end{cases}$$

$bel_{act}(p)$ is the 'actual' degree of belief in p after presenting a piece of discourse[1], and $bel_{int}(p)$ is the intended degree of belief in p. If the actual belief in p exceeds or equals the intended belief in p (and both have the same sign) then $f_{BEL}(p) = 1$. Otherwise, $f_{BEL}(p)$ is the ratio between the actual and the intended belief, and is less than 1. This ratio is positive when the actual belief has the same orientation as the intended belief, and it is negative when the orientation of the actual belief is the opposite of that of the intended belief.

When generating discourse, speakers will often anticipate and correct erroneous inferences a hearer is likely to make from the discourse. Such inferences are opportunistically addressed by the discourse planning mechanism described in [Zukerman and McConachy, 1994], which is used by our content planner. However, they are not really part of the original communicative goal, and hence, are not included in our function of belief acquisition.

3.2 Boredom

Our function of boredom reflects two different aspects of discourse: (1) the length of the text; and (2) the type of information in the text, i.e., the amount of inferable and previously known information that is mentioned, and the amount of understood information that is (unnecessarily) presented. The first factor usually causes boredom in weaker students, while the second factor usually affects

[1] The 'actual' degree of belief is conjectured by means of a function which simulates a user's change in belief as a result of a piece of discourse. This function depends on the user's ability and on the complexity and abstractness of the information [Zukerman and McConachy, 1993b].

stronger students. Thus, we have two separate formulas for boredom.

Boredom due to the length of the discourse
People normally get bored with instructional material that is "too long". Therefore, we must ensure that the presented material is shorter than a certain threshold. A good approximation to the length of a piece of discourse is simply the number of RDs in the discourse $|\{RD\}_{said}|$[2]. Thus, the first of the above factors is expressed by the following formula:

$$|\{RD\}_{said}|/TRD_{max}(M) \quad (2)$$

where $TRD_{max}(M)$ is the maximum length of instructional text a user who belongs to model M can tolerate[3].

Boredom due to unnecessary RDs
A set of RDs which is optimal for one user model is likely to be sub-optimal with respect to other user models. For example, a piece of discourse that is perfect for an average student may be too verbose for an excellent student. Thus, the second of the above factors is expressed by the following formula:

$$|\{RD\}_{said} - \{RD\}_{reduced}(M)| \quad (3)$$

where $\{RD\}_{said}$ is the set of RDs generated, and $\{RD\}_{reduced}(M)$ is the most reduced version of $\{RD\}_{said}$ that can still convey the intended propositions to a user who belongs to a particular model M.

$\{RD\}_{reduced}(M)$ is obtained by removing RDs from $\{RD\}_{said}$ so long as the communicative goal is still achieved with respect to model M. Owing to the interactions between the RDs in a set of RDs, the reduction of a set of RDs requires exhaustive enumeration. During this process, when removing an RD, we also remove the RDs that depend only on this RD, i.e., the sets of RDs that convey prerequisite and referring information for this RD only, and the RDs that contradict erroneous inferences from this RD.

3.3 Overload

Overload occurs when an addressee is unable to integrate the information being presented, thereby causing working memory to fill up with individual pieces of information, and the knowledge acquisition process to eventually shut down [Just and Carpenter, 1987]. An important factor that affects overload is the total intended shift in belief as a result of the presentation of a piece of discourse. The higher the total intended shift in belief, the more will generally need to be said, and the harder it will be for the addressee to integrate the presented information.

We distinguish between three types of propositions for the purpose of predicting cognitive overload: P – propositions that were previously unknown or correctly believed by the user; P' – propositions that were wrongly believed by the user and must now be contradicted; and $\hat{P}$ – propositions that were wrongly inferred by the user as a result of discourse planned to convey P and/or $\neg P'$ and must now be contradicted. The difficulty associated with the different types of shifts in belief is represented by F factors as follows. $F_P(M)$ reflects the amount of effort required to acquire additional information, $F_{P'}(M)$ reflects the amount of effort required to reverse a previous belief, and $F_{\hat{P}}(M)$ reflects the amount of effort required to reverse a new inference ($F_P(M) < F_{\hat{P}}(M) < F_{P'}(M)$). These factors depend on the type of the user, e.g., a strong student usually has stronger convictions than a weak student, and hence will have more difficulty reversing a belief.

The following formula expresses the total weighted shift in belief experienced by a user who belongs to a particular model M when attempting to achieve an intended degree of belief with respect to a set of propositions.

$$\begin{aligned} T_{SHIFT}(M) \quad = \quad & \sum_{p \in P} f_{SHIFT}(p) \cdot F_P(M) + \\ & \sum_{p \in P'} f_{SHIFT}(p) \cdot F_{P'}(M) + \\ & \sum_{p \in \hat{P}} f_{SHIFT}(p) \cdot F_{\hat{P}}(M) \end{aligned} \quad (4)$$

where $f_{SHIFT}(p)$ represents the contribution of proposition p to T_{SHIFT}. This contribution is the absolute value of the difference between the actual and the previous belief in p.

$$f_{SHIFT}(p) = |bel_{act}(p) - bel_{old}(p)|$$

Thus, the requirement to avoid a total shift in belief which results in cognitive overload is expressed by the following formula:

$$T_{SHIFT}(M)/belshift_{max}(M) \quad (5)$$

where $belshift_{max}(M)$ is the maximum shift in belief a user who belongs to model M can tolerate.

4 Belief without Boredom or Overload

The objective of the optimization process is to plan discourse that achieves a required degree of belief with respect to a list of intended propositions without violating the boredom and overload constraints. The belief objective must be achieved probabilistically with respect to all user models.

We assume a distribution of the user models where the highest probability mass is allocated to a *target* model,

[2] Clearly, this approximation can sometimes be wrong, but it is not productive to realize a piece of discourse currently being planned just in order to measure its exact length.

[3] When a system such as ours is in actual use, $TRD_{max}(M)$ and other such thresholds introduced later in this paper should be empirically obtained. However, in the current research their values are determined to test the effect of different values on the discourse, while ensuring that they make sense relative to each other, e.g., the explanations that weak students can tolerate without getting bored are typically shorter than those tolerated by strong students.

and most of the probability mass is allocated in the vicinity of the target model. This assumption is justified by the observation that if a teacher believes that a student is likely to be, say, an average student, then the probability that the student is good or mediocre is higher than the probability that the student is excellent or weak.

A result of our assumption regarding the target model is that when posting the overload and boredom constraints we do not need to consider all the possible user models. Rather, for boredom caused by the presentation of unnecessary RDs, it is sufficient to consider the target model and higher models, viz models of more competent students; and for cognitive overload and boredom caused by discourse that is too long, it is sufficient to consider the target model and lower models, viz models of less competent students. This is explained as follows. If a student who belongs to the target model is not experiencing boredom due to the presentation of unnecessary RDs, weaker students are unlikely to think that unnecessary RDs are being presented, while stronger students may find that the presented discourse contains unnecessary RDs. In contrast, if a student who belongs to the target model is not experiencing overload and s/he is not getting bored by discourse that is too long, these constraints will certainly be satisfied for stronger students, but may still be violated with respect to weaker students.

Further, due to our assumption regarding the manner in which the probability mass is allocated and the fact that we only have five user models, it is sufficient to post constraints with respect to the two lower and two higher models adjacent to the target model (in addition to the target model itself).

These observations lead to the following formulation of our objective function:

$$\max\{\sum_{M}\{\sum_{p} f_{BEL}(p)Sig(p)\}Prob(M)\} \qquad (6)$$

subject to the following constraints:

Overload

$$T_{SHIFT}(M_T)/belshift_{max}(M_T) \leq 1$$
$$T_{SHIFT}(M_{T-1})/belshift_{max}(M_{T-1}) \leq O(M_{T-1})$$
$$T_{SHIFT}(M_{T-2})/belshift_{max}(M_{T-2}) \leq O(M_{T-2})$$

Boredom due to length

$$|\{RD\}_{said}|/TRD_{max}(M_T) \leq 1$$
$$|\{RD\}_{said}|/TRD_{max}(M_{T-1}) \leq L(M_{T-1})$$
$$|\{RD\}_{said}|/TRD_{max}(M_{T-2}) \leq L(M_{T-2})$$

Boredom due to unnecessary RDs

$$|\{RD\}_{said} - \{RD\}_{reduced}(M_T)| \leq B(M_T)$$
$$|\{RD\}_{said} - \{RD\}_{reduced}(M_{T+1})| \leq B(M_{T+1})$$
$$|\{RD\}_{said} - \{RD\}_{reduced}(M_{T+2})| \leq B(M_{T+2})$$

The thresholds $O(M_{T-i})$, $L(M_{T-i})$ and $B(M_{T+i})$ are greater than or equal to their counterparts for M_T, and they depend on the relative probabilities of their user models. For example, if the probability of model M_{T-1} is close to the probability of M_T, then the overload and length thresholds will be just a little over 1. That is, $T_{SHIFT}(M_{T-1})$ can be only a little higher than $belshift_{max}(M_{T-1})$, and similarly $|\{RD\}_{said}|$ can be only a little higher than $TRD_{max}(M_{T-1})$. In contrast, if the probability of M_{T-1} was low compared to that of M_T, then the thresholds would be higher, meaning that $T_{SHIFT}(M_{T-1})$ and $|\{RD\}_{said}|$ could be much higher than $belshift_{max}(M_{T-1})$ and $TRD_{max}(M_{T-1})$ respectively.

This is a non-linear integer optimization problem even without the constraints. Thus, none of the gradient-based optimization techniques is suitable, and a weak search method must be applied.

4.1 The optimization process

Typically, a weak search algorithm contains an expansion step and a selection step. During expansion it generates a set of alternatives, and during selection it determines which alternative is to be the basis for the next expansion. This process iterates until the algorithm finds a solution that achieves the optimal value for the objective function while satisfying all the constraints.

Below we describe the expansion and selection steps (these steps can be slotted into any weak search algorithm, e.g., Graphsearch). The input to our algorithm is a set of propositions to be conveyed accompanied by their significance and intended degree of belief.

Expansion

1. For each proposition to be conveyed, determine the RDs that increase the user's belief in this proposition.
2. For each user model, determine the *minimally sufficient* sets of RDs which convey all the propositions, where a set of RDs is minimally sufficient if the removal of any RD causes the set to stop conveying the intended information. If necessary, a minimally sufficient set of RDs also includes RDs that contradict a user's possible erroneous inferences. Compute the extent to which these sets of RDs satisfy the constraints.

Selection

1. Select the minimally sufficient set of RDs that satisfies all the constraints and for which the objective function has the highest positive value. If there are several such sets of RDs, select the set which satisfies the constraints to the largest extent, i.e., the set which is farthest from boredom and from overload.

   If there are no minimally sufficient sets of RDs that yield an objective function with a positive value while satisfying all the constraints, then relax the requirements of the problem (Sections 4.2 and 4.3), and select for expansion the 'best' set of RDs that satisfies the constraints.
2. Determine which prerequisite propositions must be known to the user in order to understand the set of RDs selected for expansion, and which referring expressions are required to identify the concepts in this set of RDs.

The determination of RDs that increase a user's belief in a proposition (Step 1, Expansion) is performed as described in [Zukerman and McConachy, 1994].

Our algorithm computes sets of RDs (Step 2, Expansion), rather than simply collating together the RDs that convey individual propositions, because these RDs typically interact with each other in two possible ways: (1) an RD planned for one proposition may convey other propositions, thereby making their RDs obsolete; and (2) an RD may yield erroneous inferences which require the generation of additional RDs to correct them.

The algorithm for computing minimally sufficient sets of RDs may keep sets of RDs that subsume each other, so long as they are generated for different user models. For example, this happens if $\{RD_1, RD_2\}$ is minimally sufficient for model $M_{average}$, but it is not sufficient for M_{weak}, where $\{RD_1, RD_2, RD_3\}$ is required.

If there are no minimally sufficient sets of RDs that satisfy all the constraints, then it is impossible to generate a single piece of discourse that conveys the intended information without incurring boredom and/or overload. In this case, we consider two orthogonal approaches for generating discourse which satisfies the constraints (Step 1, Selection). These approaches relax the following requirements of the problem: (1) the communicative goal, i.e., less information is conveyed (Section 4.2); or (2) the single-discourse requirement, i.e., the material is broken up into smaller chunks to be presented sequentially in a session (Section 4.3). The first approach yields new minimally sufficient sets of RDs that satisfy all the constraints, thus Step 1 of Selection is performed successfully this time. The second approach yields a (possibly partial) sequence of sets of RDs, each of which is optimal for conveying a smaller chunk of propositions, thereby satisfying all the constraints as well.

Once a set of RDs has been selected for expansion, our algorithm determines prerequisite propositions and referring expressions required for understanding this set of RDs (Step 2, Selection). This is performed as described in [Zukerman and McConachy, 1994]. The expansion-selection process is then repeated to compute minimally sufficient sets of RDs that convey these prerequisite propositions.

During the next iteration, the constraints are checked for the set of RDs generated so far plus its referring expressions and the sets of RDs which convey its prerequisite propositions. This is necessary because it is possible that when a referring expression or a set of RDs that conveys prerequisite propositions is added to a main set of RDs, the overload and/or boredom constraints are violated (even though individually neither set of RDs could violate these constraints). In this case, the 'convey less information' policy for dealing with constraint violation is applied with respect to the original intended propositions, rather than their prerequisite propositions. This is because if we decide to satisfy a constraint by not conveying a particular prerequisite proposition, we affect the understanding of all the RDs that rely on this proposition, and hence, the understanding of possibly several intended propositions conveyed by these RDs.

To illustrate the optimization process, consider a situation where the intended propositions are as shown in Table 4, and the probabilities that the student belongs to the different student models are as follows: *excellent* (E) – 0.2, *good* (G) – 0.5, *average* (A) – 0.2, *mediocre* (M) – 0.1, *weak* (W) – 0.0. These probabilities may be assigned subjectively or may be obtained by testing the performance of the students in a class. Table 5 contains the initial minimally sufficient sets of RDs generated in Step 2 of the Expansion procedure, accompanied by the student model for which these sets of RDs are minimally sufficient, the value of the objective function, and the constraints that are violated by each set of RDs. Note that $\{RD_1, RD_2, RD_3\}$ and $\{RD_1, RD_2, RD_4, RD_5, RD_6\}$ are both minimally sufficient with respect to the good and the average student models. The constraints relevant to this problem are:

Propositions	Degree of Belief	Sig.
p_1	*BELIEVED*	HIGH
p_2	*BELIEVED*	HIGH
p_3	*RATHER BELIEVED*	MEDIUM
p_4	*RATHER DISBELIEVED*	LOW

Table 4: Sample intended propositions

Min. Sufficient Sets of RDs	Student Model	Obj. Func.	Violated Constraints
$\{RD_0, RD_1\}$	E	-0.3	—
$\{RD_1, RD_2, RD_3\}$	G,A	2.75	—
$\{RD_1, RD_2, RD_4, RD_5, RD_6\}$	G,A	2.9	$\{c_8\}$
$\{RD_1, RD_2, RD_3, RD_4, RD_5, RD_6\}$	M	3.0	$\{c_6, c_8\}$

Table 5: Initial minimally sufficient sets of RDs

$c_1 : T_{SHIFT}(M_{good})/belshift_{max}(M_{good}) \leq 1$
$c_2 : T_{SHIFT}(M_{ave})/belshift_{max}(M_{ave}) \leq O(M_{ave})$
$c_3 : T_{SHIFT}(M_{med})/belshift_{max}(M_{med}) \leq O(M_{med})$
$c_4 : |\{RD\}_{said}|/TRD_{max}(M_{good}) \leq 1$
$c_5 : |\{RD\}_{said}|/TRD_{max}(M_{ave}) \leq L(M_{ave})$
$c_6 : |\{RD\}_{said}|/TRD_{max}(M_{med}) \leq L(M_{med})$
$c_7 : |\{RD\}_{said} - \{RD\}_{reduced}(M_{good})| \leq B(M_{good})$
$c_8 : |\{RD\}_{said} - \{RD\}_{reduced}(M_{excel})| \leq B(M_{excel})$

The set of RDs selected for expansion is $\{RD_1, RD_2, RD_3\}$ since it violates no constraints and its objective function has the highest value. In the next iteration, our procedure generates minimally sufficient sets of RDs that convey the prerequisite propositions of $\{RD_1, RD_2, RD_3\}$ for the different student models (no RDs are generated for some models, if a student who belongs to them is presumed to believe these propositions to an adequate extent). The sets of RDs that convey prerequisite information have their own objective function, whose value must be positive. If all these sets violate one or more constraints, the discourse planning process must proceed either as described in Section 4.2 or 4.3.

4.2 Relaxing the communicative goal

In this approach, the system abandons part of the communicative goal, i.e., it decides to convey some propositions to a lesser extent than originally specified and/or to give up conveying some propositions altogether. To perform this task, we remove RDs from each of the candidate minimally sufficient sets of RDs, and determine the

Sets of RDs after Goal Relaxation	Obj. Func.	Previous Set of RDs
$\{RD_1, RD_2, RD_5, RD_6\}$	2.7	$\{RD_1, RD_2, RD_4, RD_5, RD_6\}$
$\{RD_1, RD_2, RD_3, RD_4\}$	2.8	$\{RD_1, RD_2, RD_3, RD_4, RD_5, RD_6\}$
$\{RD_1, RD_2, RD_3, RD_5\}$	2.9	

Table 6: Sample sets of RDs after goal relaxation

effect of these removals on our objective function. During this process, as when satisfying boredom constraints (Section 3.2), when removing an RD, we also remove the RDs that depend only on this RD.

Owing to the relationships between RDs, the process of removing RDs (and their dependents) until no constraints are violated essentially requires exhaustive enumeration. It generates candidate sets of RDs by removing in turn each possible RD (and its dependents) from each minimally sufficient set of RDs, and repeating this process until each resulting set of RDs satisfies all the constraints. If a minimally sufficient set of RDs has n RDs, initially n different alternative sets of RDs are spawned. The alternatives which satisfy all the constraints are then stored, and the alternatives which still violate constraints are passed on to the next iteration. RDs whose removal led to successful alternatives are not considered in later iterations because the removal of these RDs from the currently unsuccessful alternatives will lead only to sets of RDs that are subsumed by the currently successful alternatives. These subsumed sets of RDs are superfluous since they yield a lower objective function than the currently successful sets of RDs.

Table 6 contains some of the sets of RDs which are generated when this policy is applied to the last two sets of RDs in Table 5. These reduced sets of RDs satisfy all the constraints while yielding objective functions whose values are lower than before.

4.3 Relaxing the single-discourse requirement

When a set of RDs can be presented in several stages, the system must decide which RDs can be conveniently presented together in one chunk, and also in which order the different chunks of RDs should be presented. The objective of this procedure is to separate a given set of propositions into chunks, such that when the sets of RDs which convey these chunks are presented in sequence, they convey all the intended propositions. Two factors that affect the coherence of a sequence of sets of RDs are (1) *intra-connectivity*, which measures the type and number of discourse relations that link the RDs within each set; and (2) *inter-connectivity*, which measures the relations between RDs that are mentioned in different sets of RDs in this sequence. The higher the intra-connectivity and the lower the inter-connectivity of a set of RDs, the more suitable this set of RDs is for being presented separately.

The procedure outlined below splits the propositions to be conveyed into chunks, and generates a partial ordering of the sets of RDs that convey these chunks. It receives as input a list of propositions to be conveyed $\{p\}$, and the aspect of each proposition $\{a\}$.

Procedure *Generate-RD-chunks*($\{p\}$,$\{a\}$)

1. Separate the propositions to be conveyed along their aspects. Given n aspects $\{a_1, \ldots, a_n\}$, this step yields n sets of propositions[4].
2. Build all the i-tuples of the aspects, where $i = 1, \ldots, n-1$: $\{a_1, \ldots, a_n, (a_1, a_2), \ldots, (a_{n-1}, a_n), \ldots, (a_2, a_3, \ldots, a_n)\}$. Generate a chunk of propositions that corresponds to each i-tuple of aspects, and then generate all the possible supersets of chunks of propositions, so that each superset contains all the intended propositions, and each proposition appears only once in each superset.
3. Apply the procedure described in Section 4.1 (without relaxing the problem requirements) to generate an optimal set of RDs that conveys each chunk of propositions (including referring expressions and sets of RDs that convey prerequisite propositions). Remove the sets of RDs which violate any constraints or whose objective function has a negative value.
4. Generate supersets composed of these sets of RDs, so that each superset of sets of RDs conveys one superset of chunks of propositions.
5. Compute the combined score for the connectivity of each superset S of sets of RDs as follows:

$$C(S) = \sum_{\{RD^i\} \in S} \begin{array}{l} Intra\_connectivity(\{RD^i\}) - \\ Inter\_connectivity(\{RD^i\}) \end{array}$$

   If the inter-connectivity of a set of RDs is greater than zero, then note its relations to the other sets of RDs in its superset. These relations are used to order the presentation of the different sets.
6. Select the superset S_{max} of sets of RDs with the highest value for $C(S)$.

For example, for three aspects, the i-tuples are: $\{a_1, a_2, a_3, (a_1, a_2), (a_1, a_3), (a_2, a_3)\}$, and the chunks of propositions are: $\{[p_{a_1}], [p_{a_2}], [p_{a_3}], [p_{a_1}, p_{a_2}], [p_{a_1}, p_{a_3}], [p_{a_2}, p_{a_3}]\}$. The supersets of the chunks of propositions are: $\{[p_{a_1}], [p_{a_2}, p_{a_3}]\}$, $\{[p_{a_2}], [p_{a_1}, p_{a_3}]\}$ and $\{[p_{a_3}], [p_{a_1}, p_{a_2}]\}$, where each chunk is to be conveyed separately from the other chunk in its superset. For instance, the optimal set of RDs that conveys the propositions in $[p_{a_1}, p_{a_2}]$ and the optimal set of RDs that conveys the propositions in $[p_{a_3}]$ are generated and put in a superset. Finally, in order to compute the connectivity of the superset that conveys $\{[p_{a_3}], [p_{a_1}, p_{a_2}]\}$, the score of the optimal set of RDs that conveys $[p_{a_1}, p_{a_2}]$ is added to the score of the optimal set of RDs that conveys $[p_{a_3}]$. If there is an inference from an RD that conveys a proposition in $[p_{a_1}, p_{a_2}]$ to a proposition in $[p_{a_3}]$ or vice versa,

[4] We assume that the propositions related to a single aspect are always conveyed together. The violation of this assumption would lead to undesirable discourse where the steps of a procedure or the parts of an object are conveyed in separate sets of RDs. In this case, the organization of the underlying knowledge base must be modified.

1. [The atomic mass of a nucleus is the number of neutrons plus protons in it.]
2. A nucleus with a low binding energy is unstable.
3. [A nucleus with a huge atomic mass is also unstable, e.g., U235.]
4. An unstable nucleus is easily split. An easily split nucleus is fissionable fuel,
5. which is used in fission reactors.

Table 7: Sample discourse for an average student with/without boredom due to length

this inference is noted, so that the discourse relation between the RDs that convey these propositions can be expressed in the discourse, and the two sets of RDs can be ordered.

5 Results

The system was run on several inputs with both policies for dealing with constraint violations. It generates introductory discourse in technical areas such as nuclear fission, chemistry and biology. The following observations were made based on the system's output for the 'convey less information' relaxation policy.

When the boredom constraints are turned off there is no penalty for excessive length or unnecessary RDs, hence the generated texts contain several examples to ensure that the material is conveyed (top of Table 1). Overload affects the system's output only if a shift in belief that is too large for the more probable user models is required.

When boredom due to length is activated, RDs that convey propositions of lower significance tend to be omitted first. For instance, in Table 7, Sentence 3 is removed since its significance is low, and Sentence 1 is then removed since atomic mass is defined only because of its use in Sentence 3. If a proposition with a higher significance requires many RDs, then these RDs become good candidates for omission.

When both types of boredom constraints are activated, if the probabilities of the user models are evenly distributed around the target model, the only way to satisfy both sets of constraints is to convey very little information, yielding an objective function with a low value. In this case, following accepted teaching practices, the system relaxes the unnecessary-RDs constraints, giving a higher priority to the requirements of the weaker user models.

When the 'break-up the material' relaxation policy is used, the discourse becomes longer since some information is repeated in order to link the different sets of RDs.

6 Conclusion

We have offered a content planning system which takes into account a speaker's uncertainty regarding the user model to which an addressee belongs, and considers two possible outcomes of generating discourse aimed at a user model that does not fit the addressee: boredom and overload. Our system applies a constraint-based optimization mechanism which uses a probabilistic function of a user's beliefs as its objective function, and a representation of boredom and overload as constraints that affect the possible values of this function. Further, we have discussed two orthogonal policies for relaxing the parameters of the communication process when constraints are violated, viz relaxing the communicative goal and relaxing the single-discourse requirement.

Since the purpose of this research is to investigate the effect of the above mentioned factors on discourse, optimal algorithms which perform exhaustive enumeration were devised. Using these algorithms, our system takes 30-60 seconds of CPU time on a SPARCstation2 to generate English text with 15-20 RDs. Since these times are not acceptable for an interactive system, next we intend to compare the performance of these algorithms with that of sub-optimal but more time-efficient algorithms.

References

[Bonarini *et al.*, 1990] A. Bonarini, E. Cappelletti, and A. Corrao. Network-based Management of Subjective Judgments: a Proposal Accepting Cyclic Dependencies. Technical Report 90-067, Department of Electronics, Politecnico di Milano, 1990.

[Cawsey, 1990] A. Cawsey. Generating Explanatory Discourse. In R. Dale, C. Mellish and M. Zock (eds.), *Current Research in Natural Language Generation*, pages 75-102, Academic Press, 1990.

[Elhadad, 1992] M. Elhadad. FUG: The Universal Unifier User Manual Version 5.0. Technical Report, Columbia University, New York, New York, 1992.

[Just and Carpenter, 1987] M.A. Just and P.A. Carpenter. *The Psychology of Reading and Language Comprehension*. Allyn and Bacon, Inc., 1987.

[Mann and Thompson, 1987] W.C. Mann and S.A. Thompson. Rhetorical Structure Theory: A Theory of Text Organization. Technical Report ISI/RS-87-190, Information Sciences Institute, Los Angeles, California, 1987.

[Paris, 1988] C.L. Paris. Tailoring Object Descriptions to a User's Level of Expertise. *Computational Linguistics* 14(3):64-78, 1988.

[Sleeman, 1984] D. Sleeman. Mis-Generalization: An Explanation of Observed Mal-rules. In *Proceedings of the Sixth Annual Conference of the Cognitive Science Society*, pages 51-56, Boulder, Colorado, 1984.

[Zukerman and McConachy, 1993a] I. Zukerman and R.S. McConachy. An Optimizing Method for Structuring Inferentially Linked Discourse. In *AAAI-93 Proceedings – the National Conference on Artificial Intelligence*, pages 202-207, Washington, D.C., 1993.

[Zukerman and McConachy, 1993b] I. Zukerman and R.S. McConachy. Consulting a User Model to Address a User's Inferences during Content Planning. *User Modeling and User Adapted Interaction* 3(2):155-185, 1993.

[Zukerman and McConachy, 1994] I. Zukerman and R.S. McConachy. Being Concise versus Being Shallow: Two Competing Discourse Planning Paradigms. In *ECAI94 Proceedings – The European Conference on Artificial Intelligence*, pages 515-519, Amsterdam, The Netherlands, 1994.