DMSumm: Um Gerador Automático de Sumários*

Thiago Alexandre Salgueiro Pardo e Lucia Helena Machado Rino
{taspardo,lucia}@dc.ufscar.br
Departamento de Computação, Centro de Ciências Exatas e de Tecnologia.
Universidade Federal de São Carlos
Rodovia Washington Luiz, km 235 – Monjolinho. Caixa Postal 676
13565-905 São Carlos - SP

Resumo. Este artigo apresenta uma abordagem particular para a sumarização automática, mais especificamente, a geração de sumários dirigida por objetivos comunicativos. O sumarizador implementado, chamado DMSumm (*Discourse Modeling SUMMarizer*), com base em um modelo de discurso que faz um mapeamento entre relações discursivas (relações semânticas, intencionais e retóricas), visa a produção de sumários coerentes a partir de uma mensagem que corresponde à interpretação do texto que se quer sumarizar. O DMSumm segue a arquitetura clássica da geração automática de textos de três passos, isto é, a seleção de conteúdo, o planejamento textual e a realização lingüística, sendo que o planejamento textual é, de fato, a implementação do modelo de discurso utilizado.

Palavras-chave: Sumarização Automática, Planejamento Textual, Avaliação

1. Caracterização do Trabalho

A sumarização automática é o processo de seleção das informações mais importantes de um texto, aqui chamado de texto-fonte, para produzir uma versão mais curta do mesmo, visando determinado usuário e tarefa. Nos dias atuais, com a crescente quantidade de informação disponível (principalmente *on-line*) e o tempo cada vez mais reduzido que o leitor tem para apreender o máximo dessa informação, a sumarização se tornou uma área promissora de pesquisa, movida por interesses acadêmicos, comerciais e governamentais.

O trabalho relatado neste artigo explora a abordagem fundamental para a sumarização, propondo a implementação de um modelo discursivo (Rino, 1996), buscando validar tal modelo. O sumarizador implementado, chamado DMSumm (*Discourse Modeling SUMMarizer*), possui os três processos clássicos da geração automática de textos, isto é, a seleção de conteúdo, o planejamento textual e a realização lingüística, conforme mostra a Figura 1. Neste ponto, deve-se esclarecer que o DMSumm não é um sumarizador automático de textos, propriamente ditos, pois não parte de um texto-fonte para gerar seu sumário, mas sim da mensagem do texto-fonte, chamada aqui de mensagem-fonte, a qual é uma representação interna resultante da interpretação do texto que se quer sumarizar. Sob essa ótica, o DMSumm sumariza a mensagem-fonte, sendo classificado, portanto, como um gerador automático de sumários, já que não possui o processo de interpretação automática do texto-fonte na mensagem-fonte.

A mensagem-fonte é constituída de três componentes (vide Figura 1): o objetivo comunicativo (OC), a proposição central (PC) e a base de conhecimento (BC), todos considerados correspondentes ao texto-fonte. O objetivo comunicativo é a motivação primeira para a existência de qualquer discurso, consistindo no objetivo que o escritor pretende alcançar com seu texto. A proposição central é a informação principal que o escritor deseja transmitir com a veiculação do texto e é sobre ela que deverá incidir a construção do

* This work has been funded by FAPESP – Fundação de Amparo à Pesquisa do Estado de São Paulo.

discurso. A base de conhecimento corresponde ao conteúdo informativo do texto-fonte, isto é, ela contém todas as unidades de informação recuperáveis do mesmo, que servirão para produzir o conteúdo do sumário. Neste trabalho, convencionou-se que a proposição central é uma dessas unidades de informação e, logo, estará sempre presente na base de conhecimento.

O DMSumm trabalha da seguinte forma:

- Na seleção de conteúdo, ele recebe a mensagem-fonte como entrada e tem a função de diminuir o conteúdo informativo para a produção dos sumários, podando a base de conhecimento. São usadas heurísticas de poda para selecionar o conteúdo relevante;
- No planejamento textual, considera-se a base de conhecimento já podada como dado de entrada. Juntamente com o objetivo comunicativo e a proposição central originais, essa mensagem-fonte reduzida é usada pelo planejador para produzir as estruturas retóricas dos possíveis sumários, as quais são chamadas de planos (retóricos) do texto. Tal tarefa é realizada com base no modelo de discurso de Rino, fazendo um mapeamento de relações semânticas e intencionais nas relações retóricas dos planos de texto, utilizando, para isso, operadores de plano, os quais são um artifício computacional para a montagem da estutura e do conteúdo textual;
- A realização lingüística produz os sumários, propriamente ditos, a partir dos planos de texto, utilizando, para isso, *templates* definidos em função de uma gramática e de um léxico de uma determinada língua natural.

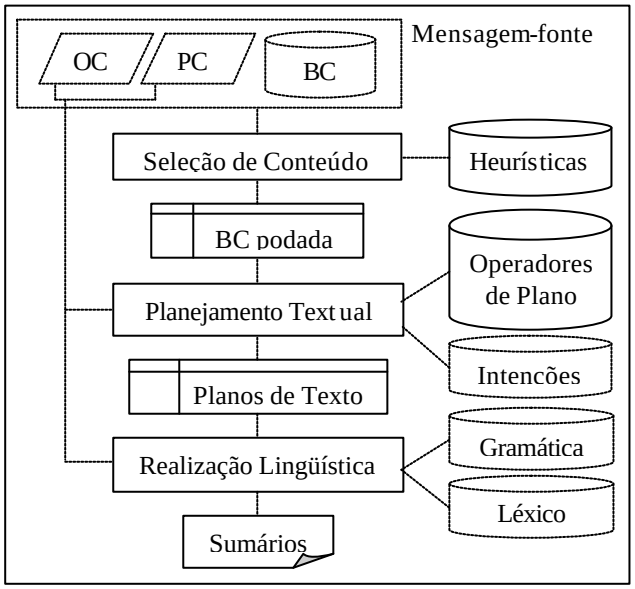

Figura 1 – Arquitetura do DMSumm

Nesse ponto, é importante dizer que o modelo de discurso de Rino é independente de língua, conforme discutido por Rino. No DMSumm, portanto, somente o processo de realização lingüística, mais especificamente, os *templates*, é dependente da língua determinada. Para o DMSumm, foram desenvolvidos *templates* tanto para o português quanto para o inglês.

As premissas deste trabalho são de que os sumários gerados automaticamente devem (1) satisfazer o objetivo comunicativo e (2) preservar a proposição central do texto-fonte. O objetivo comunicativo é o responsável por: (a) garantir a coerência dos sumários gerados automaticamente e (b) selecionar as proposições do texto-fonte que se relacionarão à proposição central nos sumários, garantindo, portanto, sua

preservação. A idéia do DMSumm é, então, buscar a satisfação do objetivo comunicativo pela determinação de sub-objetivos que, em última instância, indicam ao sistema quais os segmentos de discurso relevantes para compor o sumário e como organizá-los para tal fim.

A próxima seção apresenta o modelo de discurso fundamental do DMSumm, sendo que seus principais processos, assim como uma ilustração de seu funcionamento, são apresentados na Seção 3. A Seção 4 discute o desempenho do DMSumm, segundo alguns critérios de avaliação que conduzem a algumas considerações finais (Seção 5).

2. Modelo de Discurso do DMSumm

O modelo de discurso de Rino é um modelo tripartido, fazendo um mapeamento de relações semânticas e intencionais em relações retóricas. A seguir, os modelos correspondentes a essas relações são descritos. Por fim, serão mostrados o modelo de Rino e como tais relações interagem.

2.1. O Modelo Semântico

O modelo semântico é utilizado para organizar tanto a macro quanto a microestrutura da base de conhecimento da mensagem-fonte.

A macroestrutura segue o modelo Problema-Solução (Winter, 1976; 1977; Jordan, 1980; 1984). Esse modelo foi definido com base em análise lingüística de corpora de textos científicos, os quais apresentam uma seqüência lógica de componentes, constituída, de um modo geral, por segmentos que indicam o *problema*, a *solução*, os *resultados* e sua *avaliação*. Variações desta estrutura podem incluir, ainda, segmentos que indiquem a *situação* e a *conclusão* do discurso.

No nível microestrutural, o texto é segmentado em unidades mínimas de significado, ou seja, proposições elementares. Essas proposições, por sua vez, são associadas semanticamente por relações semelhantes às relações clausais de Jordan (1992). Na representação da base de conhecimento, tais proposições são etiquetadas por sua função na macroestrutura do discurso. Por exemplo, o texto *Using Computers in Manufacturing* a seguir (Jordan, 1980, p.225)[1] é segmentado conforme suas proposições elementares, numeradas de 1 a 6 (Em geral, uma proposição corresponde a uma sentença. Entretanto, 3a e 3b indicam duas proposições distintas em uma mesma sentença). A base de conhecimento da Figura 2 ilustra a estrutura informativa desse texto. As informações grifadas se referem aos macrocomponentes do Problema-Solução; os nós intermediários, em itálico, se referem às relações semânticas da microestrutura da base de conhecimento; as folhas se referem às proposições etiquetadas (*sit=situação*, *probl=problema*, *sol=solução*, *met=método* e *prop=proposição genérica*). Este texto será usado no decorrer deste trabalho para ilustrar o processo de sumarização.

Using Computers in Manufacturing

1. Whether you regard computers as a blessing or a curse, the fact is that we are all becoming more and more affected by them.

2. The general level of understanding of the power and weaknesses of computers among manufacturing managers is dangerously low.

3a. In order to counteract the lack of knowledge, the Manufacturing Management Activity Group of the IprodE is organizing a two-day seminar on "Computers and manufacturing management"

3b. to be held at the Birmingham Metropole Hotel at the National Exhibition Centre from 21-22 March 1979.

4. The seminar has been specially designed by the IprodE for managers concerned with manufacturing processes and not for computer experts.

5. The idea is that delegates will be able to share the experiences of other computer users and learn of their successes and failures.

---

[1] É utilizado um texto em inglês por ser ele reconhecidamente estruturado segundo o modelo Problema-Solução.

6. The seminar will consist of plenary sessions followed by syndicates where delegates will be arranged into small discussion groups.

2.2. O Modelo Intencional

No nível intencional, o modelo de discurso de Rino se baseia na *Grosz and Sidner Discourse Theory*, ou GSDT (Grosz and Sidner, 1986). Esse modelo assume que, quando um escritor escreve seu texto, ele estrutura mentalmente o conteúdo do texto em função de suas intenções. Estas, por sua vez, remetem a relações intencionais entre segmentos de discurso, individualizando o discurso e tornando-o coerente. Assim, segmentos de discurso são inter-relacionados em função de suas contribuições para a satisfação das intenções dos outros segmentos e para a intenção do discurso como um todo. A contribuição de qualquer segmento de discurso é expressa, portanto, pelas intenções subjacentes a cada um de seus segmentos. São quatro as relações intencionais da GSDT:

- X *dominates* Y (X *DOM* Y): a satisfação da intenção subjacente a Y contribui para a satisfação da intenção subjacente a X;
- X *satisfaction-precedes* Y (X *SP* Y): a satisfação da intenção subjacente a X deve preceder a satisfação da intenção subjacente a Y;
- X *supports* Y (X *SUP* Y): a satisfação da intenção subjacente a X fornece subsídios para que se consiga inferir a intenção subjacente a Y;
- X *generates* Y (X *GEN* Y): a satisfação da intenção subjacente a X implica a intenção subjacente a Y.

Rino adicionou às relações intencionais originais a relação de simetria *symmetry* (SYMM), segundo a qual dois segmentos discursivos podem contribuir um com o outro indistintamente. Essa relação é definida pela seguinte implicação lógica: $DOM(X,Y) \wedge DOM(Y,X) \Rightarrow SYMM(X,Y)$.

No modelo de Rino, as relações intencionais são definidas em função das etiquetas das proposições da base de conhecimento. Para o texto *Using Computers in Manufacturing*, por exemplo, algumas das possíveis relações intencionais podem ser dadas por:

- DOM(*sol*,*probl*): a satisfação da intenção subjacente ao problema contribui para a satisfação da intenção subjacente à solução;
- SP(*probl*,*sol*): a satisfação da intenção subjacente ao problema deve preceder a satisfação da intenção subjacente à solução.

2.3. O Modelo Retórico

A base retórica do DMSumm é responsável por estruturar coerentemente as unidades informativas do texto-fonte que foram consideradas relevantes para compor o sumário. No modelo de Rino, a organização retórica se baseia, principalmente, na teoria *Rhetorical Structure Theory*, ou RST (Mann and Thompson, 1987). Pelo menos teoricamente, essa teoria define um conjunto de relações capaz de representar todas as possíveis relações existentes entre quaisquer segmentos discursivos. Nos casos padrões, essas relações são binárias e indicam uma informação principal, *nuclear*, e uma informação complementar, *satélite*. Em geral, o satélite (S) influencia de alguma forma a interpretação que o leitor faz da informação apresentada no núcleo (N). Quando se tratam de unidades proposicionais igualmente relevantes, têm-se as relações multinucleares, isto é, com vários núcleos, mas nenhum satélite.

A definição de cada relação retórica pressupõe a existência de quatro tipos de informação: *restrições sobre N*, *restrições sobre o S*, *restrições sobre a combinação de N e S*, além do *efeito* que se espera que a relação em foco cause no leitor. É importante notar que, embora a RST tenha sido desenvolvida para a interpretação textual, ela tem sido explorada para a geração textual e é assim que ela é vista neste trabalho: como uma teoria para organizar um sumário. Neste sentido, faz-se uso de suas restrições de nuclearidade também para preservar a proposição central de um texto-fonte: esta será mantida como a informação mais nuclear em um plano de texto. A Figura 3 mostra uma das possíveis estruturas retóricas referente à base de conhecimento da Figura 2 (arcos em negrito indicando os núcleos das relações). Nesta estrutura, a propo-

sição central é indicada pela proposição 2 do texto, já que esta é a proposição mais nuclear da estrutura ilustrada.

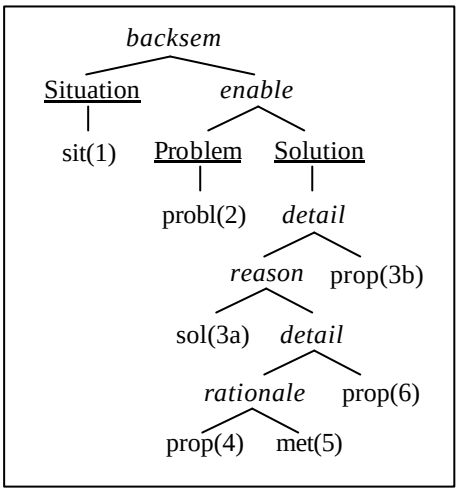

Figura 2 – Base de conhecimento do texto *Using Computers in Manufacturing*

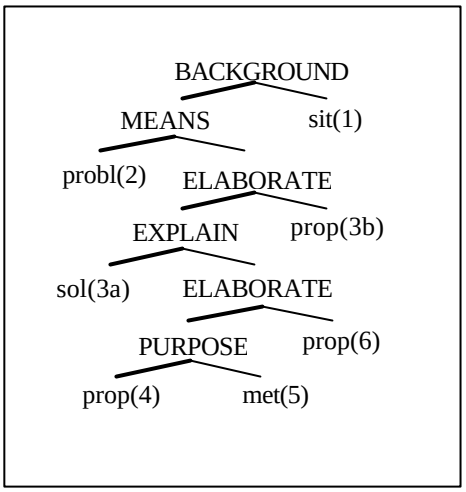

Figura 3 – Estrutura retórica para o texto *Using Computers in Manufacturing*

2.4. O Modelo de Discurso de Rino

2.4.1. Objetivos Comunicativos

Em seu modelo, Rino utiliza somente os objetivos comunicativos *descrever*, *relatar* e *discutir*. No cenário discursivo, tais objetivos são refinados a fim de observar as premissas da GSDT. Isto é feito por meio de uma rede sistêmica de objetivos (Rino, 1996, p. 194). Percorrendo-se essa rede, os objetivos comunicativos são expandidos em sub-objetivos que, por sua vez, também são expandidos, até se chegar aos atos primitivos, isto é, aos objetivos finais que introduzem no plano de texto a informação de conteúdo que fará parte do sumário. O objetivo comunicativo será satisfeito se seus sub-objetivos forem satisfeitos.

As diferentes estratégias para a derivação dos objetivos comunicativos permitem gerar diferentes planos de texto para uma mesma mensagem-fonte. Ainda, observando-se que diferentes interpretações são possíveis para um mesmo texto-fonte frente às expectativas e ao conhecimento de mundo do leitor, são permitidas variações da proposição central e do objetivo comunicativo na mensagem-fonte, dando margem para a produção de uma variedade ainda maior de planos. A proposição central, a despeito das variações possíveis, é sempre mantida como a informação mais nuclear nos planos, observando, assim, a restrição de preservação da proposição central e sua centralidade no discurso.

É durante a expansão dos objetivos comunicativos em outros objetivos que as relações retóricas são inseridas nos planos de texto. Para que uma relação seja inserida no plano de texto, entretanto, ela deve ser abonada pelo contexto informativo e intencional, conforme é discutido na subseção seguinte.

2.4.2. Regras de Organização do Discurso

A produção dos planos de texto durante o planejamento é feita pelo mapeamento de relações semânticas e intencionais em relações retóricas. Os mapeamentos possíveis foram determinados por análise de corpora e podem ser encontrados em (Rino, 1996; Rino and Scott, 1996). Esses mapeamentos indicam quais restrições nos níveis semântico e intencional devem ser verificadas para gerar a estrutura retórica. Como exemplo de mapeamento, considere o seguinte caso: se uma relação *enable* ocorre entre duas proposições X e Y (X *enables* Y) na base de conhecimento e as restrições Y *DOM* X e X *SP* Y são verificadas no nível intencional, então a relação *means* (Y é o meio para X) ou a relação *purpose1* (X é o objetivo que se

alcança através de Y) pode ocorrer no nível retórico. Apenas uma delas será inserida no plano de texto. Um exemplo real consiste na aplicação deste mapeamento para a base de conhecimento da Figura 2: (1) existe uma relação *enable* entre *probl*(*2*) e *sol*(*3a*); (2) e supondo que o repositório intencional contenha as relações intencionais DOM(*sol*,*probl*) e SP(*probl*,*sol*); (3) então é possível gerar a relação retórica MEANS(*probl*(*2*),*sol*(*3a*)) ou a relação PURPOSE1(*sol*(*3a*),*probl*(*2*)). É importante dizer que as relações semânticas são relações informativas (por abordarem o conhecimento do domínio), enquanto as relações retóricas são discursivas e, logo, são delineadas a partir das intenções, as quais, por sua vez, determinam as informações a serem incluídas no plano de texto.

Computacionalmente, no DMSumm, o planejamento dirigido por objetivos comunicativos, o qual faz o mapeamento entre as relações discursivas, é feito por meio de *operadores de plano* (Moore and Paris, 1993). Os operadores consistem nos seguintes campos:

- *Name*: identificação do operador, sem efeito operacional;
- *Header*: objetivo que o operador deve atingir se aplicado com sucesso;
- *Effect*: caracterização do efeito atingido no leitor se o objetivo é satisfeito;
- *Constraints*: lista de condições a serem observadas para a aplicação do operador;
- *Nucleus*: informação derivada a partir da aplicação do operador. Toda aplicação de operador bem sucedida deve resultar na expansão de outros sub-objetivos ou na introdução da informação especificada neste campo no plano de texto;
- *Satellite*: informação adicional que pode ser derivada a partir da aplicação do operador. Um operador pode ter zero ou mais satélites, podendo ser opcionais.

O planejamento dirigido por objetivos é feito incrementalmente, de forma *top-down*, isto é, parte-se do objetivo comunicativo original, o qual vai sendo expandido em outros sub-objetivos, aplicando-se outros operadores, até se chegar a atos primitivos, quando, então, delineiam-se as proposições elementares.

A Figura 4 mostra um operador de plano que faz o mapeamento exemplificado acima, considerando que se quer *relatar* um conceito/evento *X* qualquer. As restrições (*constraints*) são referentes às seguintes assunções: a) *know* acessa o conhecimento de mundo do leitor (conforme Moore e Paris, 1993); b) *isa* acessa dependências semânticas, expressas na base de conhecimento; c) as demais remetem às restrições intencionais. Todas essas restrições verificam se o operador é aplicável ao discurso, para sua expansão e conseqüente organização retórica.

Name	relatar-por-means
Header	relatar(X)
Effect	know(reader,X)
Constraints	not know(reader,X), isa(X,Y,enable), sp(X,Y), dom(Y,X)
Nucleus	relatar(X)
Satellite	know(reader,means(X,Y))

Figura 4 – Operador de plano para relatar X por meio da relação *means*

3. Descrição do Gerador Automático de Sumários

Esta seção apresenta brevemente os processos do DMSumm, isto é, a seleção de conteúdo, o planejamento textual e a realização lingüística, assim como um exemplo completo de sumarização de uma mensagem-fonte passo a passo. Mais detalhes podem ser encontrados em Pardo e Rino (2001) e Pardo (2002).

3.1. Seleção de Conteúdo

O processo de seleção de conteúdo parte da mensagem-fonte e é baseado na aplicação de heurísticas. Estas são definidas em (Rino and Scott, 1994) e incluem 12 heurísticas que contemplam a macro e a micro-estrutura da base de conhecimento. No nível macroestrutural, por exemplo, é possível excluir todos os

macrocomponentes do modelo Problema-Solução, com exceção de *Problema* e *Solução*. No nível micro-estrutural, as heurísticas são definidas em função das relações semânticas. Por exemplo, pode-se excluir detalhes (relação semântica *detail*) ou informações de contexto (relação *backsem*). A Figura 5 mostra a base de conhecimento da Figura 2 podada pela heurística que exclui detalhes.

3.2. Planejamento Textual

O planejamento textual, a processo principal do DMSumm, tem como entrada a mensagem-fonte do sumário, composta pela base de conhecimento podada e pelo objetivo comunicativo e proposição central originais. Neste processo, é possível construir mais de um plano de texto, pois há várias estratégias comunicativas. Neste caso, diferentes seqüências de operadores de plano são delineadas.

Para a base de conhecimento podada da Figura 5, supondo-se que se deseja relatar o problema, tem-se como objetivo comunicativo *relatar* e como proposição central *probl(2)*. A Figura 6 mostra um dos planos de texto produzidos automaticamente dentre os 10 possíveis a partir das estratégias discursivas do DMSumm.

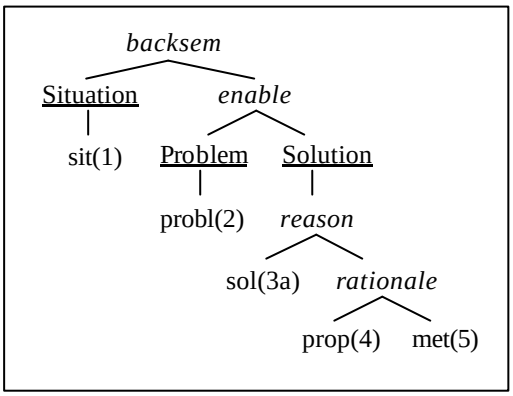

Figura 5 – Base de conhecimento podada

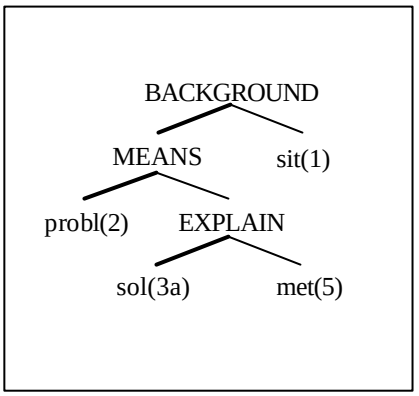

Figura 6 – Plano de texto

3.3. Realização Lingüística

Nesta etapa, *templates* são utilizados para gerar o sumário, propriamente dito, a partir do plano de texto. Decisões de ordem e agrupamento são tomadas, delimitando a forma do sumário, e marcadores discursivos são inseridos, quando necessário. As decisões de ordem dizem respeito à ordem entre núcleos e satélites das relações retóricas, enquanto as decisões de agrupamento determinam como núcleos e satélites serão apresentados, isto é, se aparecerão em *uma mesma sentença*, em *sentenças adjacentes* ou em *parágrafos distintos*. As decisões de ordem e agrupamento, assim como a lista de marcadores discursivos para cada relação retórica, seguem a proposta de Marcu (1997) para o inglês. Para o português, algumas decisões de ordem e agrupamento foram adaptadas. O texto abaixo mostra um sumário gerado automaticamente para o plano de texto da Figura 6. Os marcadores inseridos são apresentados em negrito.

> *Whether you regard computers as a blessing or a curse, the fact is that we are all becoming more and more affected by them.*
>
> *The general level of understanding of the power and weaknesses of computers among manufacturing managers is dangerously low.* ***Considering this****, in order to counteract the lack of knowledge, the Manufacturing Management Activity Group of the IProdE is organizing a two-day seminar on computers and manufacturing management,* ***and, explaining****, the idea is that*

delegates will be able to share the experiences of other computer users and learn of their successes and failures.

4. Avaliação

Vários experimentos foram feitos para avaliar o DMSumm, enfocando, principalmente, as premissas básicas do sistema, isto é, a satisfação do objetivo comunicativo e a preservação da proposição central, conforme relatado em Pardo e Rino (2002). Outros critérios, como textualidade, também foram considerados. Para os experimentos, foi utilizado o *Theses Corpus* (Pardo, 2002), contendo 10 introduções de teses e dissertações da área de computação, tendo, em média, 530 palavras cada introdução. Esse corpus foi escolhido pelo fato dos textos apresentarem a estutura Problema-Solução e serem acompanhados por sumários autênticos (isto é, aqueles produzidos pelos próprios autores dos textos).

Foram considerados dois pontos de decisão (White et al., 2000) principais para a avaliação dos sumários automáticos, a saber: textualidade e preservação da idéia principal. Considera-se que a idéia principal de um texto é mantida no sumário se este satisfaz o objetivo comunicativo original e preserva a proposição central do texto-fonte, ou seja, observa as premissas básicas deste trabalho. Com base nisso, os sumários automáticos e autênticos foram julgados por 10 juízes lingüistas computacionais e falantes nativos do português do Brasil. A Tabela 1 indica as seis notas que os juízes podiam atribuir a tais dados. O Gráfico 1 sintetiza o julgamento dos sumários automáticos (linha sólida) e autênticos (linha pontilhada). Deve-se ressaltar que, para este experimento, a medida Kappa (Carletta et al., 1997) foi calculada em 0.78, indicando uma concordância significativa entre os julgamentos dos diversos juízes.

Tabela 1 - Pontos de decisão e notas correspondentes para avaliação de sumários

Textualidade	Idéia Principal	Nota
Mantida	Preservada	6
Prejudicada	Preservada	5
Mantida	Parcialmente preservada	4
Prejudicada	Parcialmente preservada	3
Mantida	Não preservada	2
Prejudicada	Não preservada	1

De acordo com os julgamentos mostrados no Gráfico 1, 62% dos sumários foram considerados satisfatórios, 22% foram considerados aceitáveis e 16% foram considerados ruins. Outros dados percebidos foram: (a) 67% dos sumários automáticos mantiveram a textualidade (enquanto 90% dos sumários autênticos mantiveram a textualidade); (b) 61% dos sumários automáticos preservaram somente parcialmente a idéia principal e 31% a preservaram totalmente (enquanto todos os sumários autênticos preservaram totalmente a idéia principal).

Foi medida também a informatividade semântica dos sumários automáticos em relação aos autênticos, isto é, o quanto de informação foi reproduzida no sumário automático em relação a todo o conteúdo do texto-fonte (Mani, 2001). A medida é calculada de tal forma que, quanto mais curto um sumário e mais conteúdo informativo ele possuir, maior sua classificação. Portanto, quanto maior a classificação (o máximo é 1), mais informativo é o sumário. O Gráfico 2 mostra a informatividade semântica dos sumários. A média da informatividade semântica para os sumários automáticos foi calculada em 0.41, enquanto a média para os sumários autênticos foi calculada em 0.55.

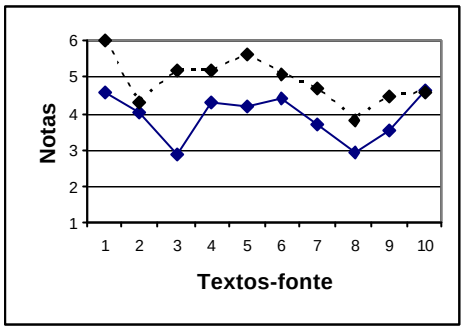

Gráfico 1 – Julgamento dos sumários

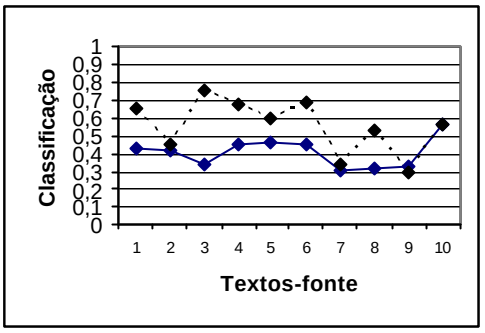

Gráfico 2 – Informatividade semântica dos sumários

Além da informatividade semântica, calculou-se a legibilidade dos sumários (automáticos e autênticos) e dos textos-fonte, conforme sugerido por Mani. Mudanças drásticas na legibilidade indicam, por exemplo, que a situação comunicativa dos sumários pode ser diferente da situação comunicativa dos textos-fonte, o que é indesejável para a sumarização automática. Entretanto, neste experimento, tanto os sumários quanto os textos-fonte apresentaram graus de legibilidade similares.

As medidas de precisão (*precision*), cobertura (*recall*) e *f-measure* foram calculadas. No contexto da sumarização automática, elas são definidas da seguinte forma:

- a precisão indica o quanto de informação do sumário autêntico os sumários automáticos apresentaram em relação a tudo o que apresentaram, ou seja, indica o quão próximos os sumários automáticos estão dos autênticos;
- a cobertura indica o quanto de informação do sumário autêntico os sumários automáticos apresentaram, ou seja, indica o grau de informatividade dos sumários automáticos em relação aos autênticos;
- a *f-measure* é uma distribuição da combinação da precisão e da cobertura, sendo, portanto, uma medida única de eficiência de um sistema. No caso da sumarização automática, ela indica a performance de um sistema em produzir sumários próximos dos ideais.

O DMSumm produziu sumários com 44% de precisão e 54% de cobertura, com uma *f-measure* de 0,48.

5. Considerações Finais

O trabalho apresentado consiste em uma abordagem profunda para a sumarização automática. Foi implementado um gerador automático de sumários, chamado DMSumm – *Discourse Modeling SUMMarizer* – utilizando, como base, um modelo de discurso que faz um mapeamento entre relações discursivas, isto é, relações semânticas, intencionais e retóricas. O objetivo principal consistiu na validação do modelo de discurso utilizado.

As premissas básicas do gerador são que um sumário deve satisfazer o objetivo comunicativo do texto-fonte e preservar sua proposição central. A satisfação do objetivo comunicativo é garantida pelo próprio processo de geração do sumário, o qual é guiado por objetivos comunicativos. A proposição central, por sua vez, é sempre preservada, pois é mantida como a informação mais nuclear durante a estruturação discursiva do sumário.

Por contemplar somente a geração automática de sumários, o sistema não incorpora ainda a interpretação dos textos-fonte, que tem sido realizada manualmente. Além dessa limitação, algumas outras podem ser percebidas: a) o realizador lingüístico é rudimentar, dado o uso de *templates*; b) as relações retóricas do modelo de Rino não são exaustivas, pois foram delineadas a partir de estudos de corpora. Apesar dessas limitações, a avaliação do DMSumm mostrou que a produção de sumários automáticos coerentes é pos-

sível e que o modelo de discurso de Rino é promissor, mas também mostrou que o sistema ainda pode ser aprimorado.

Os próximos passos consistem na realização de uma avaliação mais robusta, utilizando um corpus maior e composto de textos completos (em vez de introduções somente). Além da nova avaliação, foi iniciada uma pesquisa sobre como automatizar o processo de interpretação de umtexto para a geração da sua mensagem-fonte.

Referências

Carletta, J.; Isard, A.; Isard, S.; Kowtko, J. C.; Soherty-Sneddon, G.; Anderson, A. H. (1997). The reliability of a dialogue structure coding scheme. *Computational Linguistics*, Vol. 23, No. 1, pp. 13-32.

Grosz, B. and Sidner, C. (1986). Attention, Intentions, and the Structure of Discourse. *Computational Linguistics*, Vol. 12, No. 3.

Jordan, M.P. (1980). Short Texts to Explain Problem-Solution Structures – and Vice Versa. *Instructional Science*, Vol. 9, pp. 221-252.

Jordan, M.P. (1984). Structure, style and word choice in everyday English texts. *TESL* 15, Vols. 1 & 2.

Jordan, M. P. (1992). An Integrated Three-Pronged Analysis of a Fund-Raising Letter. In W. C. Mann and S. A. Thompson (eds), *Discourse Description: Diverse Linguistic Analyses of a Fund-Raising Text*, pp. 171-226.

Mani, I. (2001). *Automatic Summarization*. John Benjamins Publishing Co., Amsterdam.

Mann, W.C. and Thompson, S.A. (1987). *Rhetorical Structure Theory: A Theory of Text Organization*. Technical Report ISI/RS-87-190.

Marcu, D. (1997). *The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts*. PhD Thesis, Department of Computer Science, University of Toronto.

Moore, J.D. and Paris, C. (1993). Plannig Text for Advisory Dialogues: Capturing Intentional and Rhetorical Information. *Computational Linguistics*, Vol. 19, No. 4, pp. 651-694.

Pardo, T.A.S. (2002). *DMSumm: Um Gerador Automático de Sumários*. Dissertação de Mestrado. Departamento de Computação. Universidade Federal de São Carlos. São Carlos – SP.

Pardo, T.A.S. and Rino, L.H.M. (2002). DMSumm: Review and Assessment. In E. Ranchhod and N. J. Mamede (eds.), *Advances in Natural Language Processing*, pp. 263-273 (Lecture Notes in Artificial Intelligence 2389). Springer-Verlag, Germany.

Pardo, T.A.S. and Rino, L.H.M. (2001). A Summary Planner Based on a Three-Level Discourse Model. In the *Proceedings of the Natural Language Processing Pacific Rym Symposium* – NLPRS. Tokyo, Japan.

Rino, L.H.M. (1996). *Modelagem de Discurso para o Tratamento da Concisão e Preservação da Idéia Central na Geração de Textos*. Tese de Doutorado. IFSC-USP. São Carlos - SP.

Rino, L.H.M. and Scott, D. (1996). A Discourse Model for Gist Preservation. In D. L. Borges and C. A. A. Kaestner (eds.), *Advances in Artificial Intelligence*, pp. 131-140 (Lecture Notes in Artificial Intelligence 1159). Springer-Verlag, October.

Rino, L.H.M. and Scott, D. (1994). *Automatic generation of draft summaries: heuristics for content selection*. ITRI Techn. Report ITRI-94-8. University of Brighton, England.

White, J. S.; Doyon, J. B.; Talbott, S. W. (2000). Task Tolerance of MT Output in Integrated Text Processes. In *ANLP/NAACL 2000: Embedded Machine Translation Systems*, pp. 9-16. Seattle, WA.

Winter, E.O. (1976). *Fundamentals of Information Structure*. Hatfield Polytechnic, Hertfodshire, England.

Winter, E.O. (1977). A Clause-Relational Approach to English Texts. A Study of Some Predictive Lexical Items in Written Discourse. *Structional Science*, Vol. 6, No. 1, pp. 1-92.