UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

DMSumm: Um Gerador Automático de Sumários

Thiago Alexandre Salgueiro Pardo

SÃO CARLOS

ABRIL, 2002

Pardo, Thiago Alexandre Salgueiro.
P226dm DMSumm: um gerador automático de sumários / Thiago Alexandre Salgueiro Pardo. -- São Carlos : UFSCar, 2002.
103 p.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2002.

1. Programas de computador. 2. Sumarização automática. 3. Modelagem de discurso. 4. Planejamento textual. I. Título.

CDD: 005.1 (20ª)

“Se não houver frutos, valeu a beleza das flores.
Se não houver flores, valeu a sombra das folhas.
Se não houver folhas, valeu a intenção da semente.”

AGRADECIMENTOS

A meus pais, aos quais nem palavras tenho para expressar o quanto os amo, e a meu irmão, por ser muito mais que um irmão.

A meus avós, que sempre me apoiaram e se orgulharam até mesmo das pequenas coisas que fiz.

A minha querida orientadora, amiga, conselheira... quase mãe, que é presente e um presente em minha vida desde o início de minha graduação, por tudo que passamos juntos e ainda passaremos. Toda e qualquer palavra que eu diga é insignificante diante de tudo o que ela significa para mim.

A meus primos e aos amigos Fábio, Adriana e Lola, que aliviaram a maratona do mestrado.

Aos amigos Xico, Calebe, Denis, Soninha, Bart e Alan, meus companheiros dentro e fora da Federal.

Aos mais que amigos Andréia, Valéria, Gawa, Raquel, Ariani, Juliana, Helena, Leandro, Jorge, Camilla, Lucia S., Marcelo, Alice, Graça e Bruno, pela amizade, pelos cafezinhos, pelas contribuições para este trabalho. Em especial, à Valéria, pelos bate-papos que nada contribuíram para este trabalho.

Ao NILC – Núcleo Interinstitucional de Lingüística Computacional, pelo apoio e uso de suas instalações.

À FAPESP – Fundação de Amparo à Pesquisa do Estado de São Paulo, pelo suporte financeiro a este trabalho.

ÍNDICE GERAL

1. INTRODUÇÃO 1

1.1. CONTEXTUALIZAÇÃO 1
1.2. ABORDAGENS PARA A SUMARIZAÇÃO AUTOMÁTICA 2
1.3. ARQUITETURA GENÉRICA PARA A SUMARIZAÇÃO AUTOMÁTICA 4
1.4. DELIMITAÇÃO DA PROPOSTA DE SUMARIZAÇÃO AUTOMÁTICA 5

2. CARACTERIZAÇÃO DO TRABALHO 8

2.1. PREMISSAS E HIPÓTESE 8
2.2. PROCESSOS DO DMSUMM 9

3. MODELO DE DISCURSO DO DMSUMM 12

3.1. O MODELO SEMÂNTICO 12
3.2. O MODELO INTENCIONAL 15
3.3. O MODELO RETÓRICO 17

4. TRABALHOS QUE CORROBORAM O MODELO DE RINO 20

5. PROCESSOS DO DMSUMM 26

5.1. SELEÇÃO DE CONTEÚDO 27
5.2. PLANEJAMENTO TEXTUAL 29
5.2.1. Modelo de Discurso 30
5.2.1.1. Objetivos Comunicativos 30
5.2.1.2. Regras de Organização do Discurso 32
5.2.2. Do Modelo à Implementação 33
5.2.2.1. Metodologia de Desenvolvimento do DMSumm 33
5.2.2.2.Implementação 34
5.3. REALIZAÇÃO LINGÜÍSTICA 39

6. O SISTEMA DMSUMM 42

6.1. ESCOPO DO SISTEMA 42
6.2. IMPLEMENTAÇÃO 42

7. AVALIAÇÃO DO DMSUMM 50

7.1. EXPERIMENTO 1: IDENTIFICAÇÃO DA IDÉIA PRINCIPAL DOS TEXTOS-FONTE 50
7.2. EXPERIMENTO 2: AVALIAÇÃO DOS SUMÁRIOS AUTOMÁTICOS 51
7.2.1. Identificação do Sumário Autêntico 51
7.2.2. Avaliação da Qualidade dos Sumários Automáticos com Base em Dois Pontos de Decisão .. 53
7.2.3. Avaliação da Textualidade e da Preservação da Idéia Principal de Forma Isolada 54
7.2.4. Avaliação da Proximidade entre Sumários e Textos-Fonte 60
7.2.5. Avaliação da Informatividade dos Sumários 62
7.3. PRECISÃO E COBERTURA DOS SUMÁRIOS AUTOMÁTICOS 64
7.4. AVALIAÇÃO DO MODELO DE DISCURSO DO DMSUMM 65
7.4.1. Nível Semântico 65
7.4.2. Nível Retórico 66
7.4.3. Nível Intencional 67
7.5. RESULTADOS DA AVALIAÇÃO 68

8. CONSIDERAÇÕES FINAIS 69

8.1. LIMITAÇÕES DESTE TRABALHO 69
8.2. CONTRIBUIÇÕES DESTE TRABALHO 71

8.3. TRABALHOS FUTUROS 72

APÊNDICE A – MODELO PROBLEMA-SOLUÇÃO 75

A.1. ETIQUETAS DO MODELO PROBLEMA-SOLUÇÃO 75
A.2. EXTRAÇÃO DA ESTRUTURA INFORMATIVA DE TEXTOS-FONTE 76

APÊNDICE B – RELAÇÕES SEMÂNTICAS 80

APÊNDICE C – RELAÇÕES RETÓRICAS 84

APÊNDICE D – DEFINIÇÃO DOS OPERADORES DE PLANO 86

REFERÊNCIAS BIBLIOGRÁFICAS 88

ÍNDICE DE FIGURAS

Figura 1 – Níveis de representação e grau de abstração 3
Figura 2 – Arquitetura genérica de um sistema de sumarização automática 5
Figura 3 – Arquitetura do DMSumm 10
Figura 4 – Base de conhecimento do texto *Using Computers in Manufacturing* 15
Figura 5 – Plano RST para o texto *Using Computers in Manufacturing* 19
Figura 6 – Classificação de Marcu 22
Figura 7 – Classificação de O'Donnel 24
Figura 8 – Arquitetura do DMSumm 26
Figura 9 – Seleção de conteúdo 29
Figura 10 – Operador de plano para relatar X por meio da relação *means* 35
Figura 11 – *Template* para a relação *means* 40
Figura 12 – Interface da seleção de conteúdo 43
Figura 13 – Base de conhecimento podada 43
Figura 14 – Interface do planejamento textual 45
Figura 15 – Planos de texto 46
Figura 16 – Interface da realização lingüística 47
Figura 17 – Sumários 49
Figura 18 – Base de conhecimento com causa e efeito 67
Figura 19 – Planos de texto com causa e efeito 68
Figura 20 – Base de conhecimento do Texto 1 77
Figura 21 – Base de conhecimento do Texto 2 79

ÍNDICE DE TABELAS

Tabela 1 – Rede sistêmica 31
Tabela 2 – Mapeamento de relações semânticas e intencionais em relações retóricas 33
Tabela 3 – Ordem e agrupamento das relações retóricas 40
Tabela 4 – Pontos de decisão e notas correspondentes par avaliação de sumários..... 51

ÍNDICE DE GRÁFICOS

Gráfico 1 – Média das notas dos sumários por texto 54
Gráfico 2 – Distribuição dos julgamentos dos sumários automáticos por nota..... 54
Gráfico 3 – Distribuição dos sumários automáticos de acordo com a textualidade 56
Gráfico 4 – Distribuição dos sumários automáticos de acordo com a preservação da idéia principal..... 57
Gráfico 5 – Índice de legibilidade médio dos sumários para cada texto-fonte..... 61
Gráfico 6 – Índice de legibilidade dos textos-fonte 62
Gráfico 7 – Informatividade semântica média dos sumários por texto-fonte..... 63

RESUMO

O trabalho relatado nesta dissertação é sobre a Sumarização Automática de Textos, um ramo da Lingüística Computacional que, com o advento da internet, tem recebido grande atenção nos últimos tempos, pois a quantidade de informação *on-line* é enorme e os leitores têm cada vez menos tempo para apreender o máximo desta informação.

Esta dissertação apresenta a implementação de uma abordagem profunda para a sumarização automática, mais especificamente, a geração de sumários dirigida por objetivos comunicativos. Com base em um modelo de discurso que faz um mapeamento entre relações profundas do discurso (relações semânticas, intencionais e retóricas), a geração visa a produção de sumários coerentes a partir de uma mensagem que corresponde à interpretação do texto que se quer sumarizar. O gerador de sumários apresentado, chamado DMSumm – *Discourse Modeling Summarizer* – 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. O DMSumm visa, sobretudo, validar o modelo de discurso utilizado, com especial enfoque no planejamento textual, o qual, de fato, é o principal foco do modelo de discurso adotado.

ABSTRACT

The reported work in this dissertation is about Automatic Text Summarization, a Computational Linguistics research field, which, with the internet advent, has received great attention nowadays, because the amount of on-line information is huge and the readers have had less and less time for apprehending the maximum of such information.

This dissertation presents the implementation of a deep approach for automatic summarization, more specifically, the summary generation driven by communicative goals. Based on a discourse model that defines a mapping between deep discourse relations (semantic, intentional and rhetorical relations), generation aims at producing coherent summaries from a message that corresponds to the interpretation of the text to be summarized. The summary generator presented, called DMSumm – Discourse Modeling Summarizer – follows the classical automatic text generation architecture of three steps, i.e., content selection, text planning and linguistic realization. DMSumm aims at validating the discourse model used, with special attention given to text planning, which, indeed, is the main focus of the adopted discourse model.

1. Introdução

1.1. Contextualização

A sumarização 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/ou tarefa (Mani and Maybury, 1999). 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.

No campo da sumarização humana encontramos vários tipos de sumários: resenhas de notícias jornalísticas, sinopses do movimento da bolsa de valores, sumários de textos novelísticos, *abstracts* de livros e teses, resumos de previsões meteorológicas, etc. Cada um desses tipos envolve pressuposições e características diversas, assim como conteúdos, estruturas e objetivos variados, dependendo das intenções do autor. Esses sumários podem ser classificados como informativos, indicativos ou críticos. O sumário informativo, ou autocontido, abrange todo o conteúdo do texto-fonte, dispensando a leitura deste; o indicativo, ou indexador, por sua vez, serve apenas para indicar o conteúdo do texto a que se refere, utilizado, por exemplo, quando se quer fazer a busca por documentos que se refiram a algum determinado tópico; o sumário crítico avalia o conteúdo de um texto. A utilidade de cada um desses tipos depende diretamente da tarefa a que se destina.

Variações dessa natureza são comuns na produção humana de sumários, possibilitando a produção de mais de um sumário para um texto-fonte. Apesar dessas variações serem bastante naturais e até mesmo intuitivas para os humanos, elas constituem um grande problema para a sumarização automática. De fato, Sparck Jones (1999) distingue quatro classes de fatores que afetam a sumarização automática: (a) a natureza do texto-fonte; (b) o objetivo do sumário; (c) o formato de saída do sumário; (d) a audiência a que se destina o sumário. A natureza do texto-fonte refere-se, por exemplo, ao tipo/gênero do texto, se é científico ou não, se é ou não um hipertexto, etc, sendo essencial para a delimitação do escopo e desenvolvimento de um sumarizador automático. O objetivo do sumário depende diretamente da informação que se quer comunicar ao leitor e da tarefa a que se destina o sumário, sendo a motivação primeira

para a própria existência do sumário e, de forma geral, de qualquer expressão escrita. Em outras palavras, sem um objetivo comunicativo, um sumário, ou mesmo seu texto-fonte correspondente, não teria porque existir. O formato de saída do sumário é a forma como este será apresentado ao leitor, por exemplo, em um arquivo do tipo texto, em hipertexto, por meio de multimídia, estruturado ou não em parágrafos, etc. Por fim, a audiência a que se destina o sumário interfere, sobremaneira, no conteúdo e na forma que o sumário apresenta: para leitores leigos, é necessário introduzir mais detalhes ou mais claramente o assunto; para especialistas, somente um breve relato da idéia principal do texto-fonte pode bastar. A idéia principal, neste caso, consiste na informação principal que o autor do texto quer transmitir ao leitor, a qual pode estar diluída (esparsa) ou expressa em uma única proposição central na mensagem do texto.

Várias outras características, delineadas a partir da análise do processo humano de sumarização, também irão interferir no projeto e desenvolvimento de sumarizadores automáticos. Rino (1996a) ressalta, por exemplo, que sumários remetem, necessariamente, a eventos ou textos originários dos mesmos, devendo transmitir, sempre, a idéia principal destes. Além disso, sumários devem ser construídos tendo em mente que não pode haver perda do significado original essencial, muito embora contenham menos informação e possam apresentar diferentes estruturas em relação a suas fontes.

Com vista nos fatores que afetam a sumarização, a seguir são discutidas algumas abordagens para a sumarização automática.

1.2. Abordagens para a Sumarização Automática

As abordagens mais comuns para a sumarização automática são a profunda e a superficial. A primeira tem uma certa influência da psicologia cognitiva. Como tal, afirma que o processamento humano da linguagem deve envolver o entendimento da língua, argumentando que um bom sistema automático de processamento de língua natural deve fazer uso intensivo de regras gramaticais complexas, habilidades de inferência lógica e conhecimento de mundo (Sampson, 1987). A sumarização automática, nessa abordagem, é baseada em teorias lingüísticas e modelos formais para simular a inteligência humana em um eficiente processamento da língua. Como essa é uma tarefa muito complexa, alguns pesquisadores utilizam técnicas superficiais (estatísticas ou empíricas), de mais simples aplicação. Tais métodos são ditos “cegos”,

pois não levam em consideração critérios de compreensão ou conhecimento lingüístico profundo para a produção de sumários.

Entre os recursos contemplados na abordagem profunda, encontram-se, por exemplo, relações discursivas e outros tipos de conhecimento, os quais são normalmente utilizados para estruturar o discurso e condensar seu conteúdo proposicional para a produção dos sumários. As relações profundas do discurso, que constituem a base deste trabalho, são relações que remetem ao nível conceitual do discurso, tratando, por exemplo, de seu conteúdo proposicional e de sua estruturação coerente, de modo que ao texto correspondente possa ser relacionado um significado. Essas relações encontram-se nos níveis mais abstratos do discurso, considerando-se que sua produção ou compreensão exige tanto o tratamento dos componentes lingüísticos (por exemplo, fonética-fonologia, morfologia e sintaxe), quanto dos componentes extralingüísticos (por exemplo, semântica e pragmático-discursivo, os quais remetem ao significado e ao conhecimento de mundo, respectivamente). Esses componentes se relacionam em termos do grau de abstração, como ilustra a Figura 1: quanto mais abstratos, mais complexo é o tratamento computacional, devido à dificuldade de representação em um modelo formal. As relações profundas sob enfoque na geração de sumários deste trabalho referem-se aos níveis semântico e pragmático-discursivo.

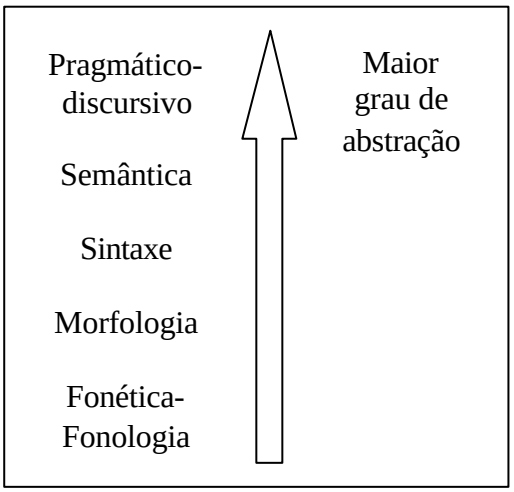


Figura 1 – Níveis de representação e grau de abstração

Os métodos superficiais utilizam outros recursos para formar um sumário, por exemplo, a identificação de palavras-chave do texto (Luhn, 1958; Black and Johnson, 1988) ou o próprio título do texto (Edmundson, 1969), a identificação de frases indicativas (Edmundson, 1969; Larocca et. al., 2000) ou, ainda, a localização das

informações principais no texto (Baxendale, 1958), para a identificação das sentenças mais relevantes do texto-fonte. Quando comparados aos recursos da abordagem profunda, esses métodos são de mais simples aplicação computacional. Entretanto, podem levar a sumários mal construídos, desconexos e que contenham referências anafóricas mal resolvidas, decorrentes da extração de sentenças isoladas do texto-fonte para formar o sumário.

Há, ainda, uma abordagem intermediária entre a superficial e a profunda chamada de abordagem baseada em entidades (Mani and Maybury, 1999), que procura modelar as entidades do texto-fonte e suas relações com a finalidade de extrair padrões do texto para ajudar na determinação de suas informações mais relevantes. Dentre as relações, destacam-se, por exemplo, as relações de similaridade entre entidades ou de proximidade, assim como relações ontológicas, lógicas e sintáticas.

Mais detalhes sobre as abordagens profunda e superficial, assim como experimentos realizados, podem ser encontrados em (Martins et. al., 2001).

Este trabalho, especificamente, adota somente a abordagem profunda para a sumarização automática, fazendo uso de relações semânticas, retóricas e intencionais, conforme o modelo de discurso de Rino (1996a), que será detalhado na Seção 2. Para subsidiar o processamento profundo na geração de sumários, apresentaremos, antes, uma arquitetura genérica para a sumarização automática, seguida de considerações específicas sobre a proposta de sumarização deste trabalho.

1.3. Arquitetura Genérica para a Sumarização Automática

Sparck Jones (1999) define uma arquitetura genérica para a sumarização automática composta de três processos principais, conforme mostra a Figura 2.

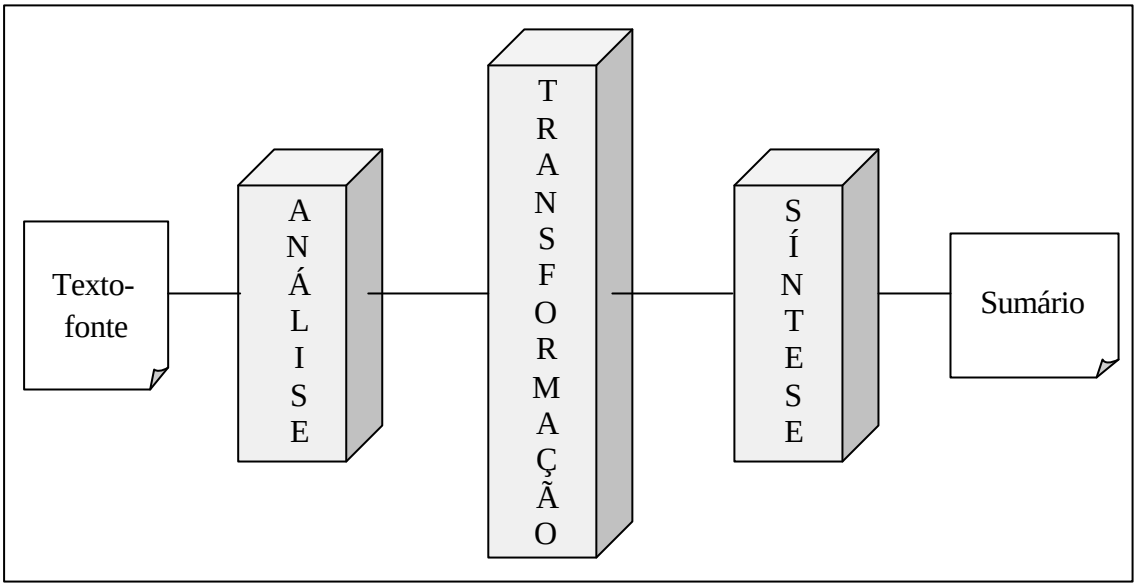


Figura 2 – Arquitetura genérica de um sistema de sumarização automática

O processo de análise corresponde, basicamente, à interpretação do texto-fonte, resultando em uma representação interna conceitual, correspondente à sua mensagem, abstraída da forma lingüística original. Essa mensagem remete aos níveis do discurso e ao grau de abstração (Figura 1): a representação interna deve ser formal o suficiente para permitir a manipulação computacional.

A transformação é o processo principal dessa arquitetura, devendo gerar uma representação interna do sumário a partir da representação interna do texto-fonte. Nesta etapa, processos de sumarização são aplicados de forma a condensar e estruturar o conteúdo do texto-fonte, produzindo, assim, uma mensagem condensada que corresponderá ao conteúdo do sumário.

O processo de síntese é responsável por expressar em língua natural a representação interna condensada, produzindo o texto condensado propriamente dito.

A delimitação deste trabalho e a forma como o sumarizador automático desenvolvido se encaixa na arquitetura genérica acima são mostradas na subseção seguinte.

1.4. Delimitação da Proposta de Sumarização Automática

O trabalho desta dissertação é a implementação de uma abordagem profunda para a sumarização automática baseada em um modelo discursivo, buscando validar tal modelo. O sumarizador apresentado aqui, 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.

Neste ponto, deve-se esclarecer que o DMSumm não é um sumarizador automático de textos propriamente dito, pois não parte de um texto-fonte para gerar seu sumário, mas sim da mensagem do texto-fonte. 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.

Com relação aos processos do DMSumm, a seleção de conteúdo e o planejamento textual consistem na fase de *transformação* da arquitetura genérica de sumarização sugerida por Sparck Jones. A aplicação desses processos à representação interna do texto-fonte, a mensagem-fonte, produz uma representação interna do sumário, chamada mensagem-fonte do sumário. Especificamente, o processo de seleção de conteúdo selecionada a informação da mensagem-fonte que fará parte do sumário, enquanto o processo de planejamento textual estrutura coerentemente essa informação, produzindo os chamados planos de texto. Por fim, o processo de realização lingüística, correspondente à fase de *síntese* da arquitetura genérica, expressa em língua natural os planos de texto.

O processo mais importante do DMSumm é o planejamento textual, por ser este o módulo principal do modelo de discurso de Rino (1996a). Nesse modelo, a estruturação coerente do sumário é determinada em função da combinação de conhecimento semântico e intencional de forma a satisfazer o objetivo comunicativo e preservar a proposição central do texto-fonte. O foco deste trabalho está justamente nessa estruturação discursiva que, diferentemente de outros trabalhos baseados em métodos superficiais (por exemplo, Larocca Neto et al., 2000; Knight and Marcu, 2000a; Black and Johnson, 1988; Baxendale, 1958), métodos profundos (por exemplo, Marcu, 1998a, 1998b; O'Donnel, 1997; Hovy, 1988; McKeown, 1985) ou até mesmo em cenários comunicativos limitados (por exemplo, Cawsey, 1993; Maybury, 1992; Moore and Paris, 1993), constitui um modelo independente de língua.

A caracterização detalhada do DMSumm é descrita na Seção 2, mostrando as premissas e hipótese deste trabalho.

A Seção 3 apresenta as bases do modelo de discurso de Rino (1996a), ilustrando e mostrando suas funções no DMSumm.

A Seção 4 faz uma breve revisão de trabalhos que corroboram, em algum aspecto, as premissas e hipótese do DMSumm.

A Seção 5 descreve os processos do DMSumm, detalhando-os e apresentando o raciocínio subjacente a eles.

A Seção 6 apresenta o sistema computacional propriamente dito, identificando e ilustrando cada um de seus módulos.

A Seção 7 descreve as várias propostas de avaliação do DMSumm, tanto em relação à *performance*, quanto em relação ao modelo e às decisões de implementação.

As considerações finais são apresentadas na Seção 8, apontando as principais limitações, contribuições e as possíveis melhorias e extensões deste trabalho.

2. Caracterização do Trabalho

2.1. Premissas e Hipótese

Este trabalho segue as premissas de Rino (1996a) para a sumarização automática. Com vista em seu modelo de discurso, Rino assume que, para a geração de um sumário coerente a partir de um texto-fonte qualquer, deve-se considerar as seguintes premissas:

1. Todo texto-fonte tem um objetivo comunicativo e uma proposição central associados a ele;
2. É possível extrair de um texto-fonte sua estrutura informativa, usada para elaborar o discurso;
3. O sumário gerado deve satisfazer o objetivo comunicativo do texto-fonte;
4. O sumário gerado deve preservar a proposição central do texto-fonte.

Como esse trabalho privilegia somente a geração do sumário, correspondente às fases de *transformação* e *síntese* da arquitetura genérica de sumarização da Figura 2, as premissas 1 e 2 assumem que o processo de análise é viável, de forma a produzir a mensagem-fonte, a qual é composta pelo objetivo comunicativo e pela proposição central do texto-fonte e por uma estrutura informativa referente ao conteúdo desse texto.

O objetivo comunicativo representa a motivação para a existência de qualquer discurso. É por meio do objetivo que o escritor, ao escrever um texto, define que informação fará parte do texto e como ela será relacionada ao restante das informações do texto. A premissa 3 diz que um sumário deve refletir esse objetivo comunicativo, satisfazendo-o durante seu processo de geração. Por essas razões, diz-se que a geração do sumário é dirigida por objetivos comunicativos.

A proposição central se refere à informação principal que o escritor do texto quer que o leitor apreenda durante sua leitura. Dessa forma, como atesta a premissa 4, um sumário deve preservar, no mínimo, essa informação principal.

A hipótese deste trabalho é que é possível estruturar coerentemente um sumário a partir da satisfação do objetivo comunicativo especificado e pela interação entre conhecimentos de diferentes naturezas, conforme será explicado na Seção 5. Com relação à coerência de um sumário, um sumário será coerente se for significativo para o leitor. A coerência pode ser refletida no sumário por meio da coesão. Diz-se que a

coesão é a expressão superficial da coerência, a qual está no nível profundo. Se um sumário é coerente e coeso, o sumário possui textualidade.

A próxima subseção detalha brevemente os processos do DMSumm, discutindo onde e como as premissas básicas deste trabalho são observadas.

2.2. Processos do DMSumm

Conforme já mencionado, o dado de entrada do DMSumm é uma mensagem correspondente à interpretação do texto-fonte. Neste trabalho, ela é obtida pelo processo humano de interpretação do texto-fonte e é composta de três unidades de informação: a proposição central, o objetivo comunicativo e a base de conhecimento (premissas 1 e 2). A base de conhecimento refere-se ao conteúdo do texto-fonte propriamente dito e é todo o conhecimento disponível para a produção dos sumários. No DMSumm, tanto o objetivo comunicativo quanto a proposição central são informações pontuais; a base de conhecimento é um componente hierárquico semanticamente estruturado que também contém a proposição central.

A Figura 3 indica como a mensagem-fonte é manipulada no DMSumm:

1. Primeiramente, a seleção de conteúdo recebe como entrada esse dado, tendo a função de reduzir o conteúdo informativo disponível para a produção dos sumários. Esse processo é composto de duas tarefas: a) podar a base de conhecimento por meio de heurísticas e b) reproduzir tanto o objetivo comunicativo quanto a proposição central na base de conhecimento podada. A saída deste processo, a base de conhecimento podada, e o objetivo comunicativo e a proposição central originais constituem os dados de entrada para o planejamento textual, denominados mensagem-fonte do sumário.
2. A seguir, o planejamento textual, cuja função é estruturar o discurso, recebe como entrada o dado de saída da seleção de conteúdo, isto é, a mensagem-fonte do sumário. A partir desta, são montados os planos de texto (estruturas retóricas, as quais, neste trabalho, são referenciadas também como estruturas discursivas e estruturas de coerências) de possíveis sumários com base no modelo de discurso de Rino (1996a), mapeando relações semânticas (da base de conhecimento) e intencionais (do repositório de intenções) em relações retóricas. Para isso, operadores de plano são utilizados. Eles são um artifício computacional para se montar a estrutura e o conteúdo textual. Os operadores foram definidos de forma a

privilegiar todos os mapeamentos entre as relações profundas do modelo de discurso de Rino. Eles são disparados pelo objetivo comunicativo e pela proposição central da mensagem-fonte do sumário, produzindo hierarquias de contribuição de objetivos, as quais dão origem aos planos de texto.

3. Por fim, o processo de realização lingüística, cuja função é produzir os sumários propriamente ditos, expressa os planos de texto em uma determinada língua natural pela aplicação de *templates* definidos em função de gramáticas e léxicos específicos de cada língua natural.

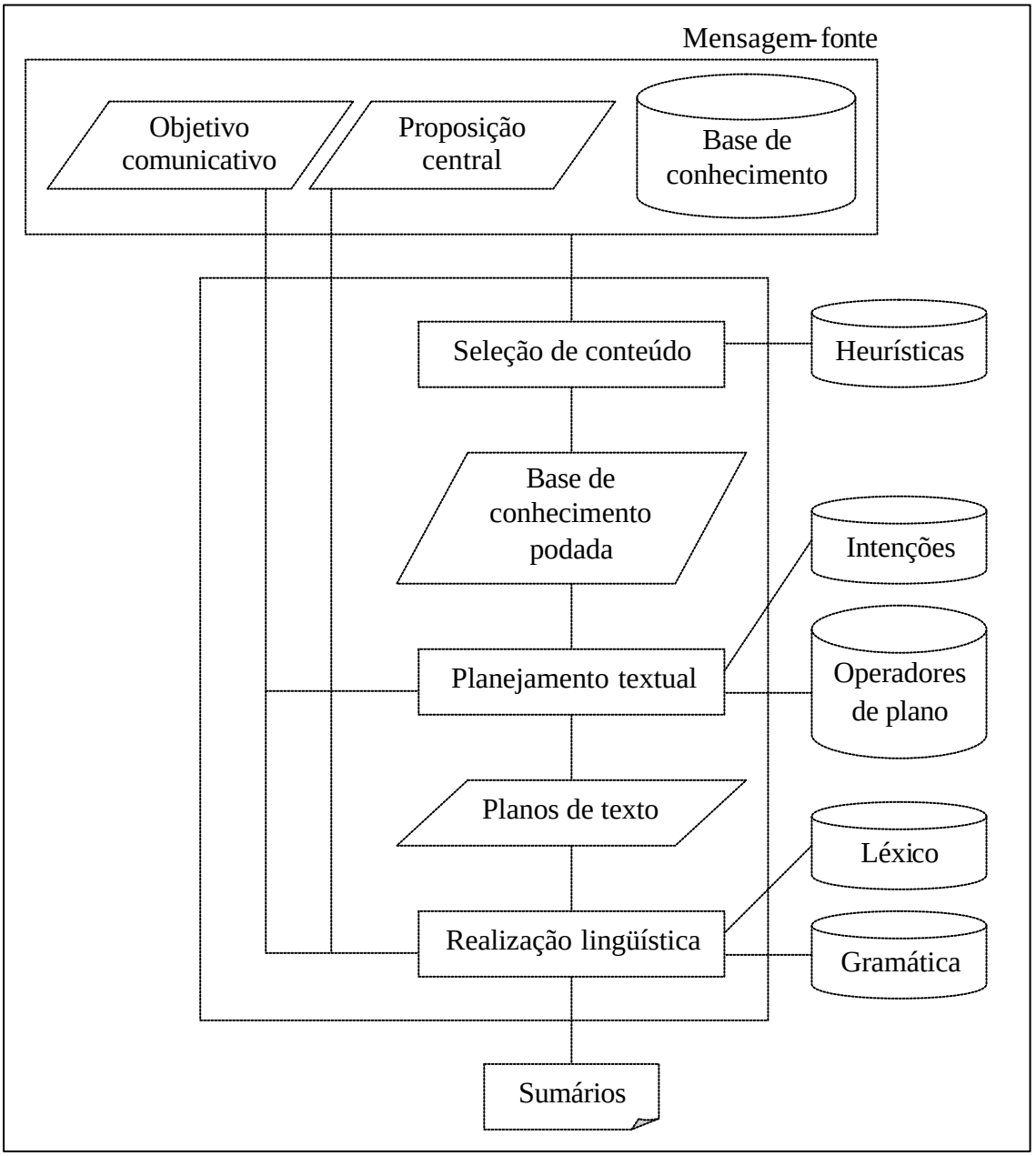


Figura 3 – Arquitetura do DMSumm

Os processos do DMSumm serão melhor explorados na Seção 5 após a descrição do modelo de discurso fundamental para o planejamento textual.

3. Modelo de Discurso do DMSumm

Nesta seção são descritos os níveis de representação do modelo de discurso de Rino (1996a), o qual é a base para a implementação do DMSumm.

Rino diz que organizar discurso coerentemente implica o tratamento de três níveis de representação profunda diferentes, mas complementares: o semântico, o intencional e o retórico, sendo os dois primeiros a base para produzir o último, conforme será mostrado na Seção 5. A seguir são discutidas as principais características de cada um desses níveis.

3.1. O Modelo Semântico

No modelo de discurso de Rino, o modelo semântico organiza tanto a macro quanto a microestrutura textual para formar a base de conhecimento da mensagem-fonte. Essa base contém somente a representação informacional do domínio.

O modelo Problema-Solução (Winter, 1976, 1977, 1979; Jordan, 1980, 1984) é utilizado para organizar a macroestrutura do texto-fonte. A escolha do modelo Problema-Solução justifica-se naturalmente pela própria seqüência lógica dos componentes em um discurso (normalmente científico, domínio este utilizado por Rino em seu trabalho), no qual é possível identificar, por exemplo, os componentes *problema*, *solução*, *resultados* e *avaliação*. Variações desta estrutura também podem ser encontradas, como a presença de segmentos que indiquem a *situação* e a *conclusão* no discurso.

O modelo Problema-Solução pode ser sistematicamente encontrado em textos de gêneros e domínios variados, conforme demonstram Hoey (1983) e Jordan (1980, 1984), sendo também independente de uma língua natural específica, conforme discutido em Rino (1996b). O Apêndice A mostra alguns textos em português cuja análise resulta em estruturas desse modelo.

No nível microestrutural, o texto é segmentado em unidades mínimas de significado, ou seja, proposições elementares. Essas proposições, por sua vez, são relacionadas por meio de relações semânticas, semelhantes às relações clausais de Jordan (1992), e classificadas de acordo com etiquetas que indicam sua função na macroestrutura textual. Os Apêndices A e B listam e definem, respectivamente, as possíveis etiquetas das proposições e as relações semânticas utilizadas neste trabalho.

No que diz respeito ao *reconhecimento* da macro e da microestrutura, Winter (1977), assim como muitos outros autores, afirma que os sinais coesivos na superfície textual ajudam a identificar a interconexão entre as unidades textuais, consistindo, assim, em vocabulários de sinais coesivos que auxiliam a estruturação textual. Esses vocabulários podem ser divididos em três tipos:

- Vocabulário 1: palavras de classe fechada, por exemplo, conjunções (*however*, *in spite of*, etc);
- Vocabulário 2: conectivos (*alternatively*, *also*, etc);
- Vocabulário 3: palavras de classe aberta e termos discursivos convencionais que evidenciam os macrocomponentes do discurso (*problem*, *solution*, *result*, *comparison*, etc).

Como exemplo de estruturação semântica, considere o texto '*Using Computers in Manufacturing*' a seguir (Jordan, 1980, p.225)[1], cujos segmentos, correspondentes às unidades mínimas de significado, são numerados para referência. A base de conhecimento da Figura 4 ilustra a estrutura informativa desse texto, constituindo toda a informação disponível para a produção dos sumários no DMSumm. 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 desta dissertação 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"

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

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.

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

Analisando o texto nos termos de Winter, pode-se notar o uso de sinais coesivos que auxiliam o leitor a identificar a função dos trechos de texto na estrutura global discursiva do texto. Por exemplo: o uso de *the fact* (proposição 1) pode indicar o contexto ou a *situação* em que o problema do texto ocorre; o advérbio *dangerously* (proposição 2), por sua conotação negativa, pode indicar a presença do *problema* abordado no texto; *in order to* (proposição 3) geralmente indica a *solução* para um problema, assim como a presença de gerúndios (*is organizing*), já que este fenômeno lingüístico normalmente é usado para mostrar a realização de uma tarefa que altera o estado das coisas no mundo, neste caso, buscando a *solução* para um problema existente em um contexto específico; palavras como *March*, *Hotel* e *Centre* podem indicar detalhes de um evento. É importante notar que tais palavras sinalizadoras e outros marcadores discursivos são ambíguos, podendo ser empregados em muitos contextos diferentes para expressar muitas relações diferentes. Desta forma, os vocabulários sugeridos por Winter constituem apenas *pistas* para o *reconhecimento* da estrutura textual.

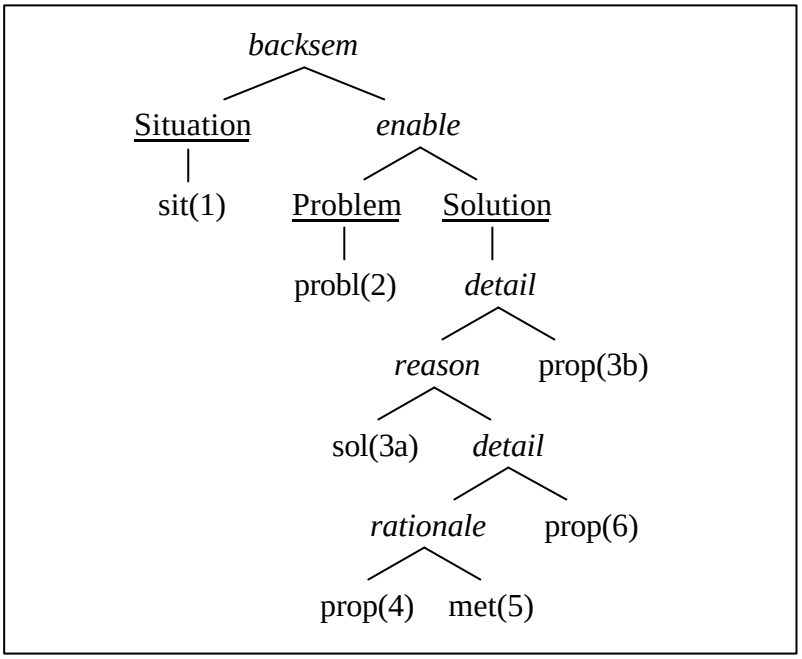


Figura 4 – Base de conhecimento do texto *Using Computers in Manufacturing*

3.2. O Modelo Intencional

No nível intencional, o modelo de discurso de Rino se baseia na GSDT (*Grosz and Sidner Discourse Theory* – 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, as quais remetem a relações intencionais entre segmentos de discurso, individualizando o discurso e tornando-o coerente. As relações intencionais são importantes, portanto, para explicar a estrutura do discurso e sua própria origem. Segundo a GSDT, o reconhecimento das relações intencionais é que permite ao leitor recuperar o que o escritor quer comunicar com o texto, isto é, as intenções do escritor.

Na GSDT, 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. Nesse modelo, qualquer contribuição é expressa pelas intenções subjacentes aos segmentos, definida de acordo com quatro relações intencionais:

- X *dominates* Y (X *DOM* Y): indica que a satisfação da intenção subjacente a X contribui para a satisfação da intenção subjacente a Y;
- X *satisfaction-precedes* Y (X *SP* Y): indica que a satisfação da intenção subjacente a X deve preceder a satisfação da intenção subjacente a Y;

- X *supports* Y (X *SUP* Y): indica que a satisfação da intenção subjacente a X fornece subsídios para que o leitor consiga inferir a intenção subjacente a Y (no sentido de crença);
- X *generates* Y (X *GEN* Y): indica que a satisfação da intenção subjacente a X implica (ou faz parte de um conjunto de intenções que implica) a intenção subjacente a Y (no sentido de "ação").

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: $DOM(X,Y) \wedge DOM(Y,X) \Rightarrow SYMM(X,Y)$.

Rino supõe que a organização retórica de um discurso pode ser delineada a partir das relações intencionais. Mais especificamente, tais relações remetem aos segmentos de informação da base de conhecimento e, portanto, ao seu relacionamento semântico. As relações intencionais são, assim, definidas em função das etiquetas das proposições da base de conhecimento, seguindo o modelo Problema-Solução. 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.

Em termos práticos, tem-se que a intenção de se descrever uma solução para um problema deve ser precedida pela intenção de se descrever o próprio problema (sp(probl,sol)) e que a apresentação de um problema contribui para que o leitor entenda sua solução (dom(sol,probl)).

A associação de intenções ao conteúdo informativo de um texto pode introduzir uma vasta gama de variações discursivas. É importante notar que as intenções podem variar mesmo que a base de conhecimento seja fixa, dadas as possíveis variações das intenções do escritor durante a escrita do texto-fonte.

3.3. O Modelo Retórico

A retórica é essencial para qualquer sistema de processamento profundo, pois ela é responsável por estruturar coerentemente as informações de um texto.

No modelo de Rino, a organização retórica se baseia, principalmente, na teoria RST (*Rhetorical Structure Theory* – Mann and Thompson, 1987), que é atualmente a teoria retórica mais influente nos modelos de geração automática (Reiter and Dale, 2000). Pelo menos teoricamente, essa teoria define um conjunto de relações capaz de representar todas as possíveis relações existentes entre segmentos discursivos. Em casos padrões, essas relações se estabelecem entre dois segmentos discursivos, sendo um nuclear (N) e outro complementar (S – "satélite"), indicando, respectivamente, a informação principal e uma informação adicional, a qual influencia de alguma forma a interpretação que o leitor faz da informação nuclear. Quando ambas as informações relacionadas são igualmente importantes, diz-se que se tem uma relação multinuclear, isto é, com mais de um núcleo e nenhum satélite.

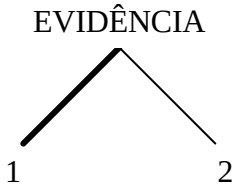
A definição de cada relação pressupõe a existência de quatro tipos de informação necessários, quer para determiná-la durante um processo de interpretação textual, quer para decidir pelo uso de uma relação particular durante a geração textual:

- Restrições sobre o núcleo (N);
- Restrições sobre o satélite (S);
- Restrições sobre a combinação do núcleo e do satélite;
- Efeito: especificação do efeito que a relação em questão causa no leitor, quando este interpreta um texto, ou do efeito pretendido pelo escritor, quando este seleciona tal relação para compor seu texto.

Abaixo é dado um exemplo de definição de uma relação retórica e sua aplicação a um trecho de texto segmentado em unidades mínimas de significado (o arco em negrito indica o núcleo da relação).

Nome da relação: EVIDÊNCIA
Restrições sobre N: o leitor pode não acreditar em N com um grau de satisfação qualquer. *Restrições sobre S*: o leitor acredita em S ou o acha válido. *Restrições sobre N + S*: a compreensão do leitor sobre S aumenta a crença do leitor em N.
Efeito: a crença do leitor em N aumenta.

1. O programa realmente funciona.
2. Em somente alguns minutos, eu consegui acessar todos os meus dados.



No modelo de Rino, são consideradas somente 15 relações retóricas porque ela se baseou em corpora restritos. Além de relações da RST, Rino também considerou as relações de coerência de Hobbs (1985). Entretanto, sua organização e aplicação seguem, fundamentalmente, a Teoria RST. É importante notar que, nesse modelo e, portanto, no DMSumm, a garantia da preservação da proposição central no sumário deve-se ao fato desta ser sempre mantida como a informação mais nuclear nos planos de texto (que são estruturas retóricas) (premissa 4). Ilustrações disso serão apresentadas na Seção 5.

A Figura 5 mostra uma das possíveis estruturas retóricas referente à base de conhecimento da Figura 4. Nesta estrutura, a proposição central é indicada pela proposição 2 do texto, já que esta é a proposição mais nuclear do plano RST ilustrado. A definição das relações retóricas do modelo de Rino pode ser encontrada no Apêndice C. É importante notar que, apesar da semelhança das relações semânticas com as retóricas, as relações semânticas são estritamente dependentes do domínio e remetem ao conteúdo informativo, enquanto as relações retóricas remetem ao conhecimento discursivo organizacional do conteúdo informativo.

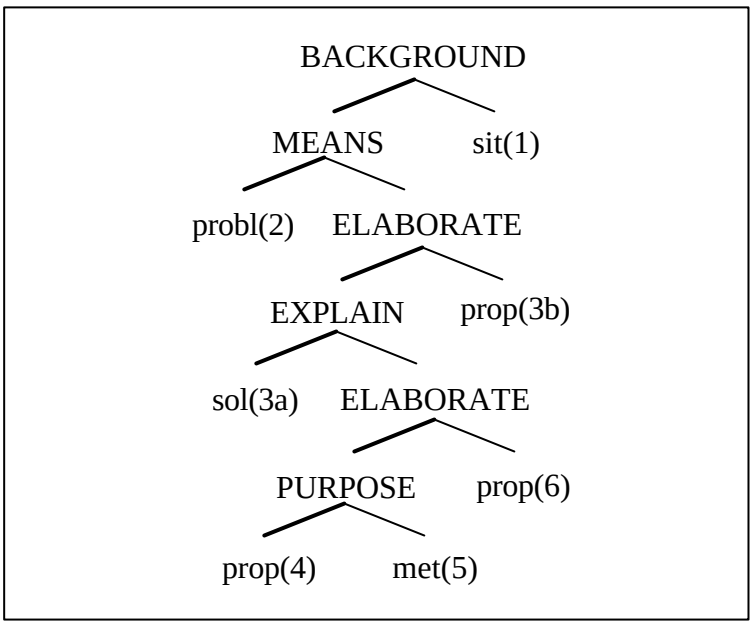


Figura 5 – Plano RST para o texto *Using Computers in Manufacturing*

No contexto da sumarização automática, as premissas dignas de nota são as seguintes: a) é possível construir uma árvore retórica relacionando-se coerentemente segmentos discursivos pela escolha apropriada de relações RST; b) a nuclearidade pode corresponder à relevância, isto é, um núcleo RST pode fornecer informação mais relevante do que seu satélite; c) como conseqüência, escolhas de sumarização podem ser feitas com base na nuclearidade (Mann and Thompson, 1987; Sparck Jones, 1993; Marcu, 1996, 1998a, 1998b).

A próxima seção mostra alguns trabalhos de sumarização que se relacionam, em algum aspecto, com este trabalho, procurando mostrar, quando é o caso, como os outros trabalhos corroboram as premissas e hipótese tomadas como base neste trabalho.

4. Trabalhos que Corroboram o Modelo de Rino

No nível semântico, Hoey (1983) sugere que sumários podem ser construídos com base nos macrocomponentes do modelo Problema-Solução. Por exemplo, um sumário mínimo pode ser obtido mantendo-se somente os segmentos discursivos correspondentes aos componentes *Problema* e *Solução*, que, segundo ele, podem representar as partes principais do discurso científico. Para que isso seja possível, Hoey adota como premissa o fato de textos poderem ser estruturados conforme o modelo Problema-Solução, corroborando a premissa 2 deste trabalho, isto é, uma estrutura informativa pode ser construída a partir de um texto-fonte.

Hutchins (1987) diz que a sumarização deve preservar principalmente a macroestrutura textual, já que evidências comprovam que os detalhes microestruturais não são retidos pelo leitor. Hutchins aponta o uso de quatro macroregras para a sumarização[2]:

- *Delição*: exclusão de partes da estrutura textual;
- *Generalização*: substituição de hipônimos por seus hiperônimos ou de detalhes por uma idéia ou frase mais abstrata que envolva os detalhes, deixando-os implícitos;
- *Construção*: esta regra, semelhante à de generalização, age como seu complemento sintagmático: enquanto a generalização é basicamente pragmática, a construção depende da microestrutura textual. Neste caso, a macroproposição construída é resultante de uma seqüência de microproposições. Por exemplo, a seqüência *"Pedro comprou tijolos, areia, cimento, colocou os alicerces, ergueu paredes..."* permite construir a sentença *"Pedro construiu uma casa"*;
- *Composição delição-construção*: consiste em excluir sentenças que expressam pré-condições ou motivações para a descrição de ações.

No DMSumm, algumas das heurísticas do processo de seleção de conteúdo são baseadas nas idéias de Hutchins. As heurísticas serão abordadas na próxima seção.

Com relação à modelagem discursiva como base para a sumarização automática, Sparck Jones (1993) afirma que todo texto tem um objetivo comunicativo e que este deve ser levado em consideração durante a geração de um sumário, versando também sobre a importância da proposição central do texto na escolha das informações que comporão esse sumário. Essas afirmações dizem respeito às premissas 1, 3 e 4 deste

[2] Essas regras foram originalmente propostas por Van Dijk (1979).

trabalho, as quais falam, respectivamente, sobre o fato de todo texto ter um objetivo comunicativo e uma proposição central associados a ele e sobre a satisfação do objetivo comunicativo e da preservação da proposição central no sumário correspondente.

Mais especificamente, com relação às relações retóricas de um texto, Sparck Jones explorou em seu trabalho o fato do núcleo dessas relações fornecerem informação mais importante que o satélite, sendo úteis, portanto, para a identificação de informação relevante para compor um sumário em um processo de sumarização automática, corroborando as idéias originais de Mann e Thompson (1987). Mais recentemente, trabalhos importantes baseados nessa idéia surgiram, conforme discutido abaixo.

Em seu trabalho, Marcu (2000a) desenvolveu o primeiro analisador retórico de textos de domínio irrestritos para o inglês. Dado um texto, esse analisador extrai sua estrutura retórica mais provável, utilizando, como base, os marcadores discursivos e suas funções no discurso. Por exemplo, o marcador *however* pode indicar a presença de uma relação retórica de contraste. Marcu utiliza, no total, 77 relações retóricas para estruturar os textos, enquanto a RST estabeleceu originalmente cerca de 25 relações.

A partir da estrutura retórica extraída automaticamente de um texto-fonte, Marcu aplica um algoritmo para computar o grau de importância de cada proposição da estrutura retórica, considerando que as proposições mais importantes apresentam informações mais relevantes do texto, sendo adequadas, portanto, para compor o sumário desse texto.

O cálculo da importância das proposições é baseado na saliência (ou nuclearidade) e na profundidade em que cada proposição aparece na estrutura retórica. Quanto mais saliente e mais acima estiver na estrutura, maior a importância de uma proposição. A Figura 6 mostra a estrutura retórica do texto *Using Computers in Manufacturing*, onde, ao lado de cada relação dessa estrutura, é indicada a proposição mais saliente da sub-árvore correspondente. A ordem de importância das proposições para compor o sumário é dada por 2>1>3a>3b>4>6>5, onde 2 é mais importante que 1, que é mais importante que 3a, etc. Analisando-se passo a passo como se chegou a essa ordem de importância entre as proposições, percorrendo a estrutura retórica de baixo para cima, tem-se:

1. entre o núcleo 4 e o satélite 5 da relação *purpose*, o núcleo 4 sobe para a relação *purpose*;
2. entre o núcleo 4 e o satélite 6 da relação *elaborate*, o núcleo 4 sobe para a relação *elaborate*;

3. entre o núcleo 3a e o satélite 4 da relação *explain*, o núcleo 3a sobe para a relação *explain*;

4. entre o núcleo 3a e o satélite 3b da relação *elaborate*, o núcleo 3a sobre para a relação *elaborate*;

5. entre o núcleo 2 e o satélite 3a da relação *means*, o núcleo 2 sobe para a relação *means*;

6. entre o núcleo 2 e o satélite 1 da relação *background*, o núcleo 2 sobe para a relação *background*.

Quantas dessas proposições serão utilizadas para compor o sumário dependerá do tamanho desejado do sumário.

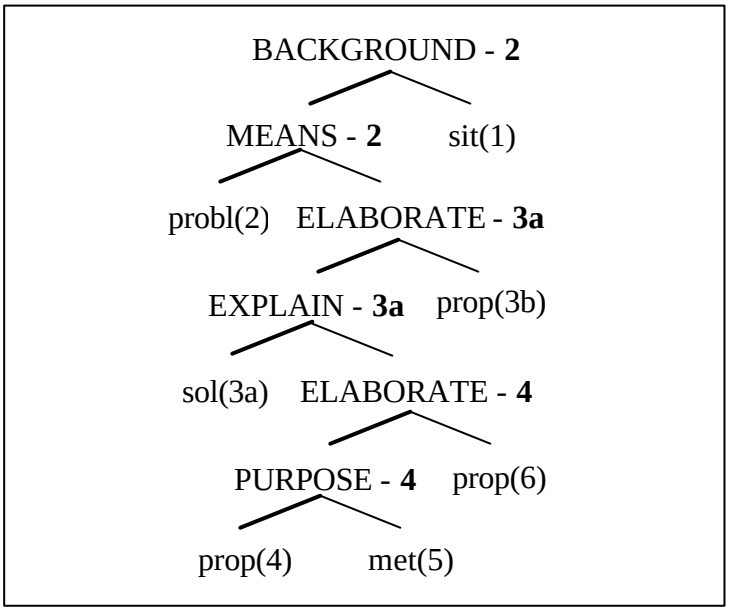


Figura 6 – Classificação de Marcu

Em experimentos realizados, Marcu produziu resultados automáticos com 57% de cobertura e 51% de precisão, produzindo uma *f-measure* de 53,8%.

O'Donnel (1997), em seu trabalho, atribui valores de *importância* para as proposições da estrutura retórica. Entretanto, O'Donnel não possui um analisador retórico automático, fazendo essa análise de forma manual.

O valor de cada proposição é calculado com base no valor da proposição mais nuclear mais próxima à proposição em foco multiplicado pelo valor (peso) da relação retórica pela qual a proposição mais nuclear é imediatamente dominada. Os pesos das relações são pré-definidos de acordo com sua *taxa de relevância* em um processo de

comunicação. Por exemplo, O'Donnel atribui o peso 0.40 para a relação *elaboration*, enquanto atribui o peso 0.70 para a relação *purpose*, já que esta deve apresentar informação mais importante para a satisfação do objetivo do texto do que a relação *elaboration*.

A proposição mais nuclear na estrutura retórica recebe automaticamente o valor máximo (1.0); os demais valores são calculados a partir deste de forma recursiva, conforme se atinge os níveis mais profundos da estrutura. Desta forma, as proposições mais acima na estrutura receberão valores maiores do que as mais abaixo. O'Donnel afirma que essa técnica pode produzir resultados ruins em alguns casos, já que, de acordo com seus experimentos, nem sempre a nuclearidade reflete a centralidade da informação, pois o autor do texto pode apresentar informações importantes em lugares do texto retoricamente irrelevantes.

A Figura 7 mostra a estrutura retórica do texto *Using Computers in Manufacturing* e os valores de cada proposição. Pesos fictícios foram associados às relações retóricas (*purpose*=0,70; *explain* e *means*=0,60; *background*=0,50; *elaborate*=0,40), resultando na seguinte ordem de importância: 2>3a>1>4>5>3b>6. Analisando-se passo a passo, de cima para baixo na estrutura retórica, tem-se:

1. a proposição 2, por ser a mais nuclear, recebe o valor máximo 1;
2. o valor da proposição 1 é igual ao valor da proposição 2 multiplicado pelo valor da relação *background*, ou seja, 1 x 0,50 = 0,50;
3. o valor da proposição 3a é igual ao valor da proposição 2 multiplicado pelo valor da relação *means*, ou seja, 1 x 0,60 = 0,60;
4. o valor da proposição 3b é igual ao valor da proposição 3a multiplicado pelo valor da relação *elaborate*, ou seja, 0,60 x 0,40 = 0,24;
5. o valor da proposição 4 é igual ao valor da proposição 3a multiplicado pelo valor da relação *explain*, ou seja, 0,60 x 0,60 = 0,36;
6. o valor da proposição 6 é igual ao valor da proposição 4 multiplicado pelo valor da relação *elaborate*, ou seja, 0,36 x 0,40 = 0,14;
7. o valor da proposição 5 é igual ao valor da proposição 4 multiplicado pelo valor da relação *purpose*, ou seja, 0,36 x 0,70 = 0,25.

Como no trabalho de Marcu, quantas dessas proposições serão utilizadas para compor o sumário dependerá do tamanho desejado do sumário.

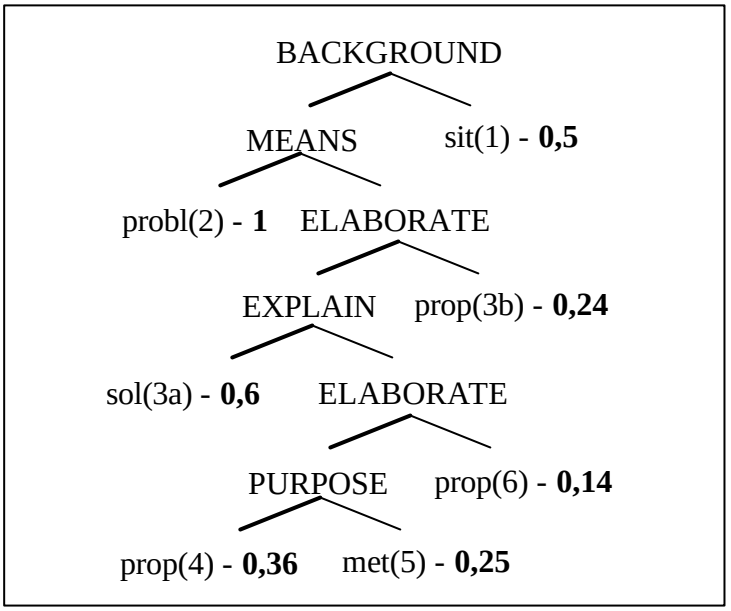


Figura 7 – Classificação de O'Donnel

Tanto o trabalho de Marcu como o de O'Donnel consideram, de uma forma ou de outra, que a informação mais importante a compor um sumário é a informação mais nuclear em uma estrutura retórica. Da mesma forma, neste trabalho, a proposição central de um sumário é sempre a informação mais nuclear no plano de texto. É desse modo que se garante que a proposição central será preservada durante a geração do sumário (premissa 4).

O trabalho de Teufel e Moens (1999), seguindo a abordagem de Kupiec et al. (1995), usa uma aproximação diferente da de Marcu e da de O'Donnel para produzir sumários com base na estrutura retórica do texto, não atribuindo importância às proposições, mas selecionando informações da estrutura retórica para preencher *esquemas* retóricos de sumários.

Teufel e Moens procuraram identificar que organizações retóricas são comumente utilizadas em um sumário, identificando as relações *purpose* (ou *problem*), *scope* (ou *methodology*), *results* e *conclusion* (ou *recommendations*) como as principais. Com base nisso, para formar o sumário, as proposições imediatamente dominadas por essas relações em uma estrutura retórica são utilizadas para preencher os esquemas. Os melhores resultados conseguidos atingiram precisão e cobertura de 66%.

Como será mostrado na próxima seção, de acordo com o modelo de Rino, diferentemente dos trabalhos relacionados nesta seção, não se seleciona de forma alguma as informações do plano de texto que comporão o sumário de um determinado

texto-fonte. A plano de texto produzido durante o planejamento textual já é considerado como sendo a própria estrutura retórica do sumário, isto é, o plano do sumário.

5. Processos do DMSumm

De acordo com a arquitetura do DMSumm (Seção 2), reproduzida abaixo por conveniência (Figura 8), os processos são executados em *pipeline*: a seleção de conteúdo poda a base de conhecimento, fornecendo um conteúdo informativo reduzido para o planejamento textual, cujos planos de texto produzidos são, então, lingüisticamente realizados. Esses processos são explicados e ilustrados a seguir, com especial enfoque no planejamento textual.

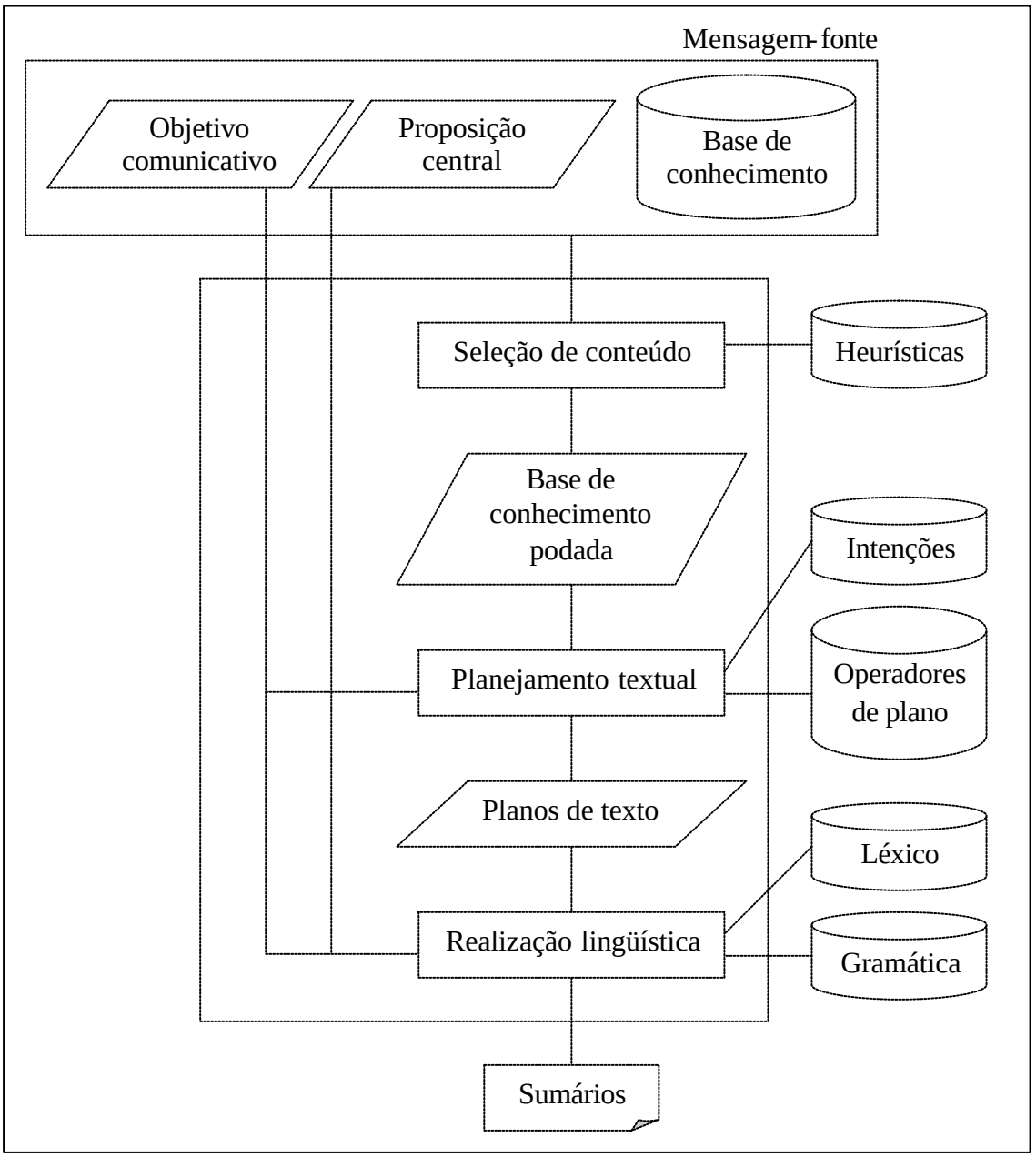


Figura 8 – Arquitetura do DMSumm

5.1. Seleção de Conteúdo

O processo de seleção de conteúdo parte da mensagem-fonte e é baseado na aplicação de heurísticas (Rino and Scott, 1994), as quais podam a base de conhecimento segundo componentes do modelo Problema-Solução e critérios de sumarização baseados no inter-relacionamento das unidades de informação da base de conhecimento, isto é, na organização semântica da base de conhecimento. Abordagens similares têm sido exploradas na literatura, considerando, em muitos casos, o gênero do discurso científico. Conforme discutido na Seção 4, especialmente dirigido pelo paradigma Problema-Solução, Hoey (1983) sugere que um esqueleto de sumário razoavelmente satisfatório pode ser conseguido extraindo-se componentes representativos de cada elemento do padrão Problema-Solução, ou, no mínimo, os componentes *problema* e *solução*. Hutchins (1977) e Rino (1996a) discordam: Hutchins afirma que este método ignora muita informação essencial, enquanto Rino enfatiza que não há garantia de que a proposição central seja expressa no sumário final, principalmente quando se considera que o escritor do texto nem sempre privilegia o *problema* ou a *solução* como proposições centrais de seu discurso.

As heurísticas da seleção de conteúdo procuram evitar que a seleção da informação seja fortemente conectada aos componentes de uma estrutura Problema-Solução. Estendem-se as sugestões acima por meio de uma poda mais informada da base de conhecimento, considerando critérios semânticos, em adição às premissas (3 e 4) relativas ao objetivo comunicativo e à proposição central. Nesse estágio, o papel do objetivo comunicativo é indicar o tipo de informação que deve ser mantida na base de conhecimento durante a poda. Por exemplo, se o objetivo comunicativo for *discutir* e a proposição central for *X* (uma proposição qualquer), qualquer informação que permita a discussão de X será relevante e não poderá ser excluída da base de conhecimento. Desta forma, considerando tais critérios, a proposta de Rino corrobora a de Sparck Jones (1993) e é reforçada por Jordan (2000), autores que argumentam que o objetivo comunicativo deve ser levado em consideração durante a seleção da informação a compor o sumário. As heurísticas e suas justificativas são mostradas abaixo:

- H1: manter apenas o problema e a solução – os componentes *problema* e *solução* são suficientes para o leitor ter uma idéia sobre o conteúdo do texto;
- H2: excluir resultados – resultados, que são informações substanciais, devem ser excluídos de um sumário indicativo;

- H3: excluir pré-condição de um evento – não é crucial para o discurso, pois o leitor pode inferi-la;
- H4: excluir propósitos de instrumentos ou eventos – os propósitos podem ser recuperados pelo leitor;
- H5: excluir argumentação (prova) de uma asserção – o leitor pode inferi-la ou deduzi-la;
- H6: excluir o efeito de relações causais – o efeito pode ser recuperado pelo leitor;
- H7: excluir atributos/detalhes/exemplos de entidades ou eventos – esses elementos são desnecessários em um sumário;
- H8: excluir avaliações – em uma comunicação técnica e objetiva, expressões pessoais devem ser evitadas;
- H9: excluir razões de eventos ou procedimentos – as razões podem ser supérfluas em certos contextos;
- H10: excluir cenário contextual – o cenário pode ser recuperado pelo leitor.

Por exemplo, dada a base de conhecimento do texto *Using Computers in Manufacturing* (Figura 4), a Figura 9 mostra os passos da aplicação da heurística para se excluir detalhes. As partes em negrito indicam as relações a serem excluídas em cada passo. A estrutura informativa resultante depois do passo 2 é a base de conhecimento podada.

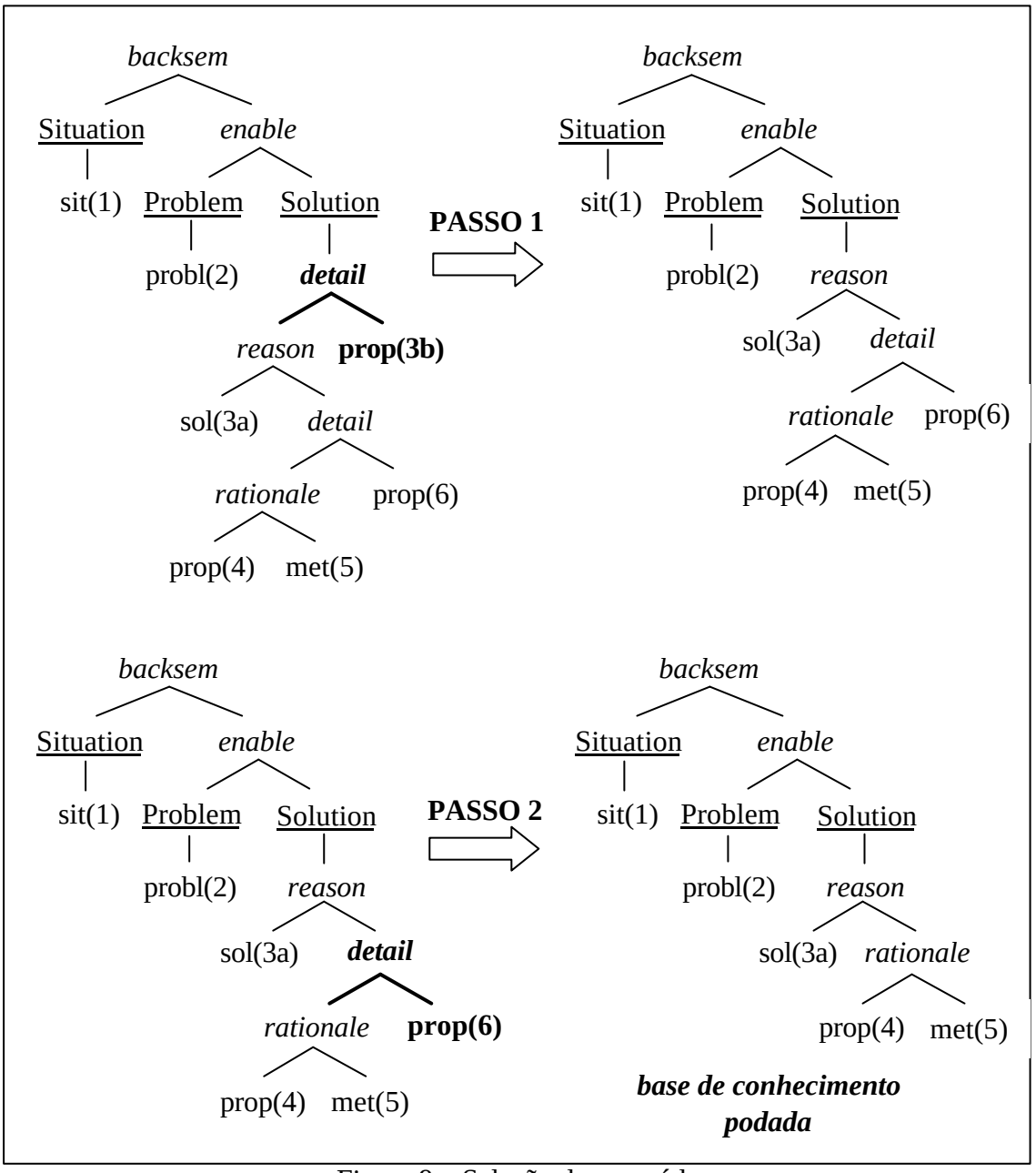


Figura 9 – Seleção de conteúdo

5.2. Planejamento Textual

O planejamento textual, o foco deste trabalho, tem como entrada a mensagem-fonte do sumário, composta pela base de conhecimento podada e pelo objetivo comunicativo e proposição central originais (vide Figura 8). Neste processo, todos os planos de texto possíveis podem ser construídos, os quais, após a realização lingüística, poderão dar origem a sumários de diferentes tamanhos e estruturas comunicativas e lingüísticas variadas (supondo, por exemplo, diferentes audiências). Essa diversidade na geração de sumários depende diretamente do objetivo comunicativo especificado na mensagem-fonte do sumário e da disponibilidade de estratégias discursivas variadas.

As subseções seguintes detalham o modelo de discurso de Rino, que é a base para o planejamento textual, e a forma como ele foi implementado.

5.2.1. Modelo de Discurso

5.2.1.1. Objetivos Comunicativos

Em seu modelo, Rino utiliza somente os objetivos comunicativos *descrever*, *relatar* e *discutir*, cujas definições são:

- *Descrever*: o escritor pretende fazer com que o leitor conheça e seja capaz de identificar entidades ou características de objetos, processos, métodos ou resultados;
- *Relatar*: o escritor apresenta planos, idéias, processos ou mudanças de estado com a intenção de fazer o leitor conhecê-los (ou saber como funcionam), compreendê-los ou se tornar competente para executá-los;
- *Discutir*: o escritor pretende influenciar as crenças e/ou o comportamento do leitor, freqüentemente com a intenção de persuadi-lo a alterar seu comportamento ou conhecimento adquirido.

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. 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, objetivos finais que introduzem no plano de texto a informação que fará parte do sumário. O objetivo comunicativo será satisfeito se seus sub-objetivos forem satisfeitos, isto é, levarem a atos primitivos (premissa 3). Além disso, é a partir do objetivo comunicativo que se escolhem as informações que se relacionarão com a proposição central no sumário por meio da determinação da contribuição entre os segmentos discursivos.

A rede sistêmica de objetivos é sintetizada na Tabela 1 abaixo. Ela foi modificada da proposta original devido à necessidade de uma maior sistematização dos objetivos comunicativos e como eles se dão na organização discursiva do sumário. Para tal sistematização do modelo de Rino, utilizou-se como base outras propostas de contribuições comunicativas, destacando-se as propostas de Moore e Paris (1993), Moore (1995) e Maier e Hovy (1991).

Tabela 1 – Rede sistêmica

Objetivos / RELs	***Descrever***	***Relatar***		***Discutir***	
	know-about	*know*	*competent*	*bel*	*persuade*
Purpose1					
Means					
Purpose2					
Justify1					
Evidence					
Justify2					
Nonvolres					
Nonvolcause					
List					
Contrast					
Elaborate					
Evaluate					
Explain					
Sequence					
Background					

Essa rede representa um complexo relacionamento entre os objetivos comunicativos e como eles se dão nos níveis de estratégias retóricas e de estados mentais considerados pelo autor em relação ao leitor hipotético. Como exemplo desse relacionamento profundo, um escritor pode querer transmitir uma mensagem para o leitor *discutindo* um conceito por meio de uma relação retórica *means*, de modo que o leitor *acredite* na veracidade de intenção do escritor ou que seja *persuadido* a acreditar nela (linha 2 da Tabela 1). *Acreditar* ou *persuadir* são possíveis estados mentais a se atingir no leitor. Por outro lado, para *discutir* um conceito por meio da relação *purpose1*, pode-se somente *persuadir* o leitor (linha 1 da Tabela 1).

São as diferentes estratégias para se percorrer essa rede sistêmica que permitem gerar diferentes planos de texto para uma mesma mensagem-fonte. Cada estratégia é chamada de estratégia *discursiva* ou *comunicativa.* 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 (premissa 4).

Os objetivos comunicativos *descrever*, *relatar* e *discutir* são, ainda, intercambiáveis, isto é, a partir de um desses objetivos, pode-se expandir os outros, visto que um texto dificilmente será puramente uma descrição, um relato ou uma discussão. Por exemplo, caso se tenha o objetivo *descrever*, durante a estruturação discursiva, pode-se expandir o objetivo *relatar* ou *discutir* a partir do primeiro.

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.

5.2.1.2. 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 são mostrados na Tabela 2, a qual foi proposta por Rino (1996a) para indicar os casos possíveis de estruturação textual a partir de uma mensagem-fonte. Rino definiu os possíveis mapeamentos com base em estudo de corpus. Logo, esses mapeamentos não são exaustivos, podendo ser expandidos no futuro.

A tabela indica quais restrições nos níveis informativo (semântico) e intencional (que se dá em função das etiquetas do modelo Problema-Solução) devem ser verificadas para gerar a estrutura retórica (conforme indicado pela direção das flechas abaixo da tabela). Como exemplo de mapeamento, considere o primeiro caso da Tabela 2, o qual indica que, 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 condensada da Figura 9:

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*))

Tabela 2 – Mapeamento de relações semânticas e intencionais em relações retóricas

Caso	Relações semânticas	Relações retóricas	Relações intencionais
1	enable(Y,X)	purpose1(Y,X) means(X,Y)	X sp Y Y dom X
2	rationale(Y,X)	purpose2(X,Y) justify1(X,Y)	Y sup X X dom Y not X sp Y
3	proof(X,Y)	evidence(X,Y) justify2(X,Y)	Y sp X X dom Y Y sup X
4	cause(X,Y)	nonvolres(Y,X) nonvolcause(X,Y)	Y sp X X dom Y Y gen X
5	simSem(X,Y)	list(X,Y)	X symm Y
6	difSem(X,Y)	contrast(X,Y)	X symm Y
7	attribute(X,Y) detail(X,Y) exemplify(X,Y)	elaborate(X,Y)	Y sup X
8	evalSem(X,Y)	evaluate(X,Y)	X dom Y
9	reason(X,Y)	explain(X,Y)	Y gen X
10	sequence(X,Y)	sequence(X,Y)	X sp Y
11	backSem(X,Y)	background(X,Y)	not X sp Y Y sup X

5.2.2. Do Modelo à Implementação

5.2.2.1. Metodologia de Desenvolvimento do DMSumm

O DMSumm foi desenvolvido por meio de prototipação baseada em exemplos. Inicialmente, foram utilizados estudos de caso já realizados por Rino (1996a), cujos textos são científicos e do domínio da Física. Posteriormente, foram escolhidos textos de autores proeminentes (por exemplo, Jordan e Hoey), abrangendo textos dos mais diversos tipos e gêneros, por exemplo, textos científicos de computação, textos jornalísticos de diversos assuntos e partes de relatórios técnicos. Foram feitos mais de 20 estudos de caso, incluindo textos em português e em inglês, corroborando Rino (1996b) sobre o fato de o modelo de discurso utilizado ser independente de língua, podendo ser aplicado a uma grande variedade de textos. Desses textos foram extraídas as mensagens-fonte, os dados de entrada do sistema, a partir dos quais o sistema foi delineado. Ao mesmo tempo, de acordo com a análise dos resultados do sistema para os estudos de caso, foi-se aprimorando o sistema.

Os repositórios de operadores de plano e de intenções do sistema (vide Figura 8) foram desenvolvidos de forma gradual, por demanda, conforme os estudos de caso foram sendo realizados. Ao fim da prototipação do DMSumm, foi possível produzir repositórios genéricos desses elementos que, a princípio, se aplicam a qualquer mensagem-fonte, conforme discutido a seguir.

5.2.2.2.Implementação

Computacionalmente, no DMSumm, o planejamento dirigido por objetivos comunicativos, o qual faz o mapeamento entre as relações profundas do discurso, é feito por meio de *operadores de plano* (Moore and Paris, 1993; Maybury, 1992). Derivados da RST, 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, ou seja, às proposições elementares. É importante notar que diferentes grupos de operadores de plano levam a diferentes estratégias de produção de planos de texto. O planejamento é feito de forma *top-down* porque o objetivo do sistema é encontrar a rede de sub-objetivos que satisfaçam o objetivo comunicativo da mensagem-fonte e, assim, este é o único dado de entrada possível.

A Figura 10 mostra um operador de plano que faz o mapeamento do Caso 1 da Tabela 2, considerando que se quer *relatar* um conceito/evento *X* qualquer. As restrições são referentes às seguintes assunções: a) *know* acessa o conhecimento de mundo do leitor (conforme Moore e Paris, 1993); b) *isa* acessa dependências semânticas que ocorrem na base de conhecimento; c) as outras verificam restrições intencionais,

com o objetivo de tornar o operador aplicável durante o planejamento textual. Outros tipos de restrições são diretamente relacionados com a sumarização, por exemplo, se um segmento é um detalhe, ele pode não ser expandido por um operador de plano (isto tem relação com as heurísticas da seleção de conteúdo).

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

O repositório de operadores de plano do DMSumm contém 89 operadores definidos de forma a cobrir todos os mapeamentos possíveis da Tabela 2 (conforme relatado em Pardo e Rino, 2001), respeitando, sempre, a rede sistêmica de objetivos da Tabela 1. Os padrões de definição de operadores são mostrados no Apêndice D.

O repositório de intenções, por sua vez, desenvolvido com base nos estudos de caso realizados para a prototipação do DMSumm, possui 186 relações intencionais que se mostraram estáveis no decorrer da prototipação. Entretanto, esse repositório deve ser testado para textos de gêneros ainda não considerados neste trabalho, para se verificar sua robustez.

É importante dizer que o plano de texto não é gerado diretamente pelos operadores. É produzida, antes, uma estrutura intencional, da qual será, ao fim do planejamento, extraído o plano de texto. A estrutura intencional guarda o histórico da geração textual, isto é, a expansão dos objetivos comunicativos, a produção das relações retóricas e a inserção das proposições no plano de texto.

Abaixo é ilustrada a montagem de um plano de texto para a base de conhecimento condensada da Figura 9, mostrando os operadores de plano aplicados e suas instanciações, a montagem da estrutura intencional e a extração gradativa do plano de texto desta. Assumiu-se que o objetivo comunicativo é *relatar* e a proposição central é *probl*(*2*). Neste ponto, deve-se ressaltar o seguinte:

- pode-se produzir um mesmo plano de texto através da aplicação de vários grupos de operadores de plano diferentes – o exemplo mostra apenas uma forma de produção;
- sempre que se gera um ato primitivo *inform*, a proposição em foco é inserida no plano de texto;

- a ocorrência de *cohererel* (passo 3) indica a inserção da relação retórica no plano de texto.

PASSO 1:

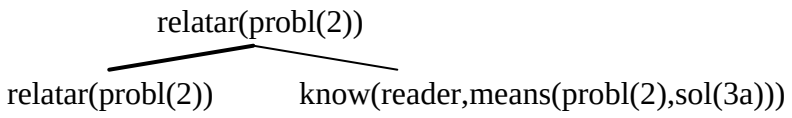
Operador de plano aplicado:

Name	relatar-por-means
Header	relatar(X)
Effect	know(reader,X)
Constraints	isa(X,Y,enable), sp(X,Y), dom(Y,X)
Nucleus	relatar(X)
Satellite	know(reader,means(X,Y))

Operador de plano instanciado:

Name	relatar-por-means
Header	relatar(probl(2))
Effect	know(reader,probl(2))
Constraints	isa(probl(2),sol(3a),enable), sp(probl,sol), dom(sol,probl)
Nucleus	relatar(probl(2))
Satellite	know(reader,means(probl(2),sol(3a)))

Estrutura intencional:



Plano de texto:

Ø

PASSO 2:

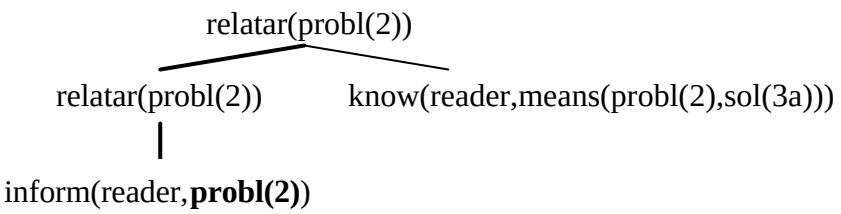
Operador de plano aplicado:

Name	relatar-informando
Header	relatar(X)
Effect	know(reader,X)
Constraints	nil
Nucleus	inform(reader,X)
Satellite	nil

Operador de plano instanciado:

Name	relatar-informando
Header	relatar(probl(2))
Effect	know(reader,probl(2))
Constraints	nil
Nucleus	inform(reader,probl(2))
Satellite	nil

Estrutura intencional:



Plano de texto:

probl(2)

PASSO 3:

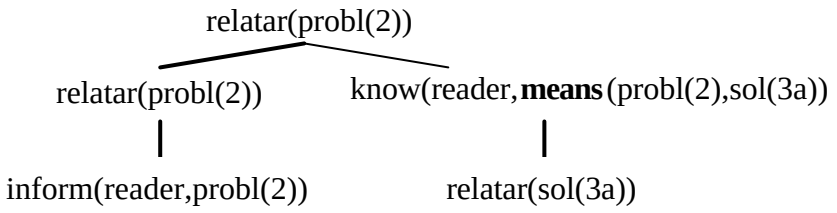
Operador de plano aplicado:

Name	identificar-por-means
Header	know(reader,means(X,Y))
Effect	know(reader,means(X,Y))
Constraints	isa(X,Y,enable), sp(X,Y), dom(Y,X)
Nucleus	relatar(Y), cohererel(means(X,Y))
Satellite	nil

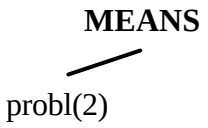
Operador de plano instanciado:

Name	identificar-por-means
Header	know(reader,means(probl(2),sol(3a)))
Effect	know(reader,means(probl(2),sol(3a)))
Constraints	isa(probl(2),sol(3a),enable), sp(probl,sol), dom(sol,probl)
Nucleus	relatar(sol(3a)), cohererel(means(probl(2),sol(3a)))
Satellite	nil

Estrutura intencional:



Plano de texto:



PASSO 4:

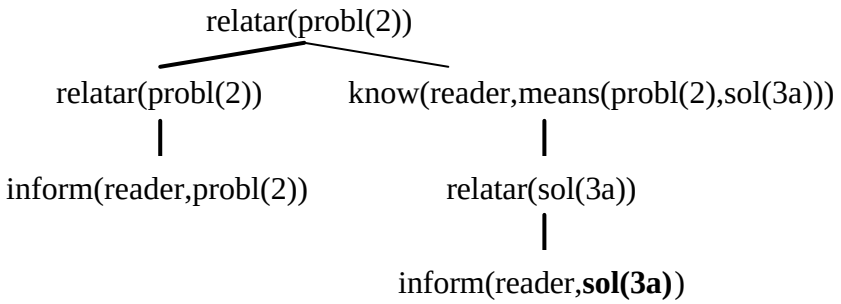
Operador de plano aplicado:

Name	relatar-informando
Header	relatar(X)
Effect	know(reader,X)
Constraints	nil
Nucleus	inform(reader,X)
Satellite	nil

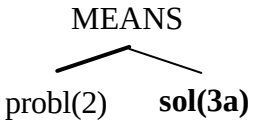
Operador de plano instanciado:

Name	relatar-informando
Header	relatar(sol(3a))
Effect	know(reader,sol(3a))
Constraints	nil
Nucleus	inform(reader,sol(3a))
Satellite	nil

Estrutura intencional:



Plano de texto:



5.3. Realização Lingüística

No DMSumm, a realização lingüística é um processo secundário, dada a ênfase na verificação do planejamento textual, para validação do modelo de Rino. Assim, a realização é bastante simplificada, considerando-se somente *templates* cuja parte variável remete somente aos marcadores discursivos, mas com parte fixa consistindo dos segmentos textuais inteiros, correspondentes a cada unidade informativa da base de conhecimento (*canned texts*). A realização segue, portanto, uma estratégia extremamente simples e restrita que deve ser aprimorada em outra versão do DMSumm. Devido a essa simplificação, optou-se por resolver as dependências contextuais dos segmentos informativos manualmente, para manter a clareza e a independência entre os segmentos.

Os *templates* são disparados pelas relações retóricas do plano de texto a ser realizado e introduzem, quando necessário, marcadores discursivos no texto. O elenco de marcadores discursivos para cada relação retórica foi extraído do trabalho de Marcu para o inglês (1997), o qual se deu por meio de estudo de corpus. Alguns exemplos são dados a seguir:

- Purpose: *in the face of it*;
- Means: *considering this, in the face of it, so, by means of*;
- Evidence: *as evidence*;
- Nonvolres: *as a result, then, therefore, thus, so*;
- Nonvolcause: *consequently, then, thus, for this reason*;
- List: *and, both, or, together with, moreover, also*;
- Contrast: *however, but, on the other hand, although*;
- Elaborate: *still*;
- Explain: *and, explaining*;
- Sequence: *then, and, before, after*.

Adotou-se também, com algumas modificações para este trabalho, o método de Marcu para compor segmentos maiores a partir de segmentos menores, estabelecendo-se relações de *ordem* e *agrupamento* entre as unidades informativas relacionadas retoricamente. A Tabela 3 mostra essas relações, as quais são interpretadas da seguinte forma:

- No caso da *ordem*, números próximos a 1 indicam que o núcleo de uma relação retórica é realizado antes do satélite, enquanto números próximos a 0 indicam o contrário;
- No caso do *agrupamento*, números próximos a 1 indicam segmentos que são agrupados como uma única sentença; números próximos a 0.5 indicam segmentos que são agrupados como sentenças adjacentes; números próximos a 0 indicam segmentos que são agrupados em parágrafos diferentes.

Tabela 3 – Ordem e agrupamento das relações retóricas

Relação retórica	**Ordem**	**Agrupamento**
purpose1	0.30	0,50
purpose2	0.30	0,50
means	0.70	0,50
justify1	0.85	0.50
justify2	0.85	0.50
evidence	0.80	0.50
nonvolres	0.86	0.50
nonvolcause	0	0.50
list	0.98	1
contrast	0.98	1
elaborate	0.97	0.50
evaluate	1	0
explain	0.67	0.75
sequence	0.98	0.50
background	0.03	0.24

A Figura 11 mostra um exemplo de *template* para a relação *means*. Esse *template*, caso aplicado, pode gerar, para dois segmentos quaisquer X e Y relacionados retoricamente por *means*, a sentença *X by means of Y.*

MEANS	
Ordem	núcleo antes do satélite
Agrupamento	sentenças adjacentes
Marcadores discursivos	*considering this, in the face of it*, etc

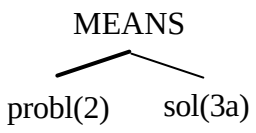
Figura 11 – *Template* para a relação *means*

Na implementação do DMSumm, além dos *templates* para o inglês, foram definidos também *templates* para o português do Brasil, já que foram feitos estudos de caso para as duas línguas. A definição dos *templates* para o português se deu simplesmente pelo mapeamento do que foi feito para o inglês, com pequenas alterações

quando necessário, principalmente no que diz respeito à questão de ordem entre os segmentos. Essa foi uma abordagem simplificada para a obtenção dos *templates* em português, já que um estudo de corpus seria necessário para a definição sistemática dos *templates*.

Tanto para o inglês quanto para o português, foram definidos 15 *templates* que observam as restrições de ordem e agrupamento da Tabela 3, não permitindo, portanto, variações dessa natureza na produção dos sumários. O único fator variável é o marcador discursivo que pode ser introduzido, dada a lista de marcadores para cada relação retórica.

Aplicando-se o *template* da Figura 11 ao plano de texto abaixo, por exemplo, obtemos o texto que segue (que já é um possível sumário para o texto *Using Computers in Manufacturing*):



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.

A próxima seção descreve o ambiente computacional do DMSumm, mostrando seu funcionamento por meio do exemplo discutido nessa seção.

6. O Sistema DMSumm

6.1. Escopo do Sistema

O DMSumm foi desenvolvido a partir de estudos de caso, destinando-se, em seu estágio atual, a usuários especialistas que conheçam o modelo de discurso utilizado, já que a interpretação textual, processo responsável por extrair o significado e, logo, a mensagem-fonte de um texto-fonte, não é automática (vide Seção 2).

Conhecendo o modelo de discurso, o usuário pode interferir nas decisões de sumarização por meio das interfaces do DMSumm. O usuário pode, por exemplo, alterar em tempo de execução os dados da mensagem-fonte, escolher e combinar diferentes heurísticas durante o processo de seleção de conteúdo e determinar o número de sumários que o sistema irá gerar e a complexidade deles. Desta forma, são as próprias interfaces que permitem a análise e crítica do sistema e, por fim, a análise do próprio modelo de discurso. As interfaces do DMSumm são mostradas na subseção a seguir.

6.2. Implementação

Conforme detalhado em Pardo (2002), a implementação do DMSumm é dividida em duas partes: o engenho de inferência e a interface com o usuário. O engenho de inferência corresponde ao sumarizador da Figura 8 propriamente dito, e foi implementado em PROLOG. A interface com o usuário, por sua vez, constitui-se no recurso pelo qual o usuário fornece os componentes da mensagem-fonte para o engenho de inferência e determina os parâmetros dos processos de sumarização, sendo esta parte do sistema implementada em DELPHI. Ela é constituída de três módulos específicos para cada um dos processos de sumarização, isto é, a seleção de conteúdo, o planejamento textual e a realização lingüística. Além disso, a interface oferece meios de visualização das estruturas de representação intermediárias (bases de conhecimento e planos de texto) e dos dados de entrada e saída do sistema (textos-fonte, sumários).

A Figura 12 mostra a interface para o processo de seleção de conteúdo. Nesta interface, o usuário deve indicar a base de conhecimento desejada, as heurísticas para podá-la e sua ordem de aplicação. Nesse exemplo, foram selecionadas a base de conhecimento referente ao texto *Using Computers in Manufacturing* (Figura 4, Seção 3) e a heurística *excluir detalhes de entidades*. A caixa de texto ao lado da heurística

indicaria, caso houvesse mais de uma heurística selecionada, a ordem de aplicação de cada uma delas. Selecionando-se o botão *executar*, o processo de seleção de conteúdo produz a base de conhecimento podada. A Figura 13 mostra essa base.

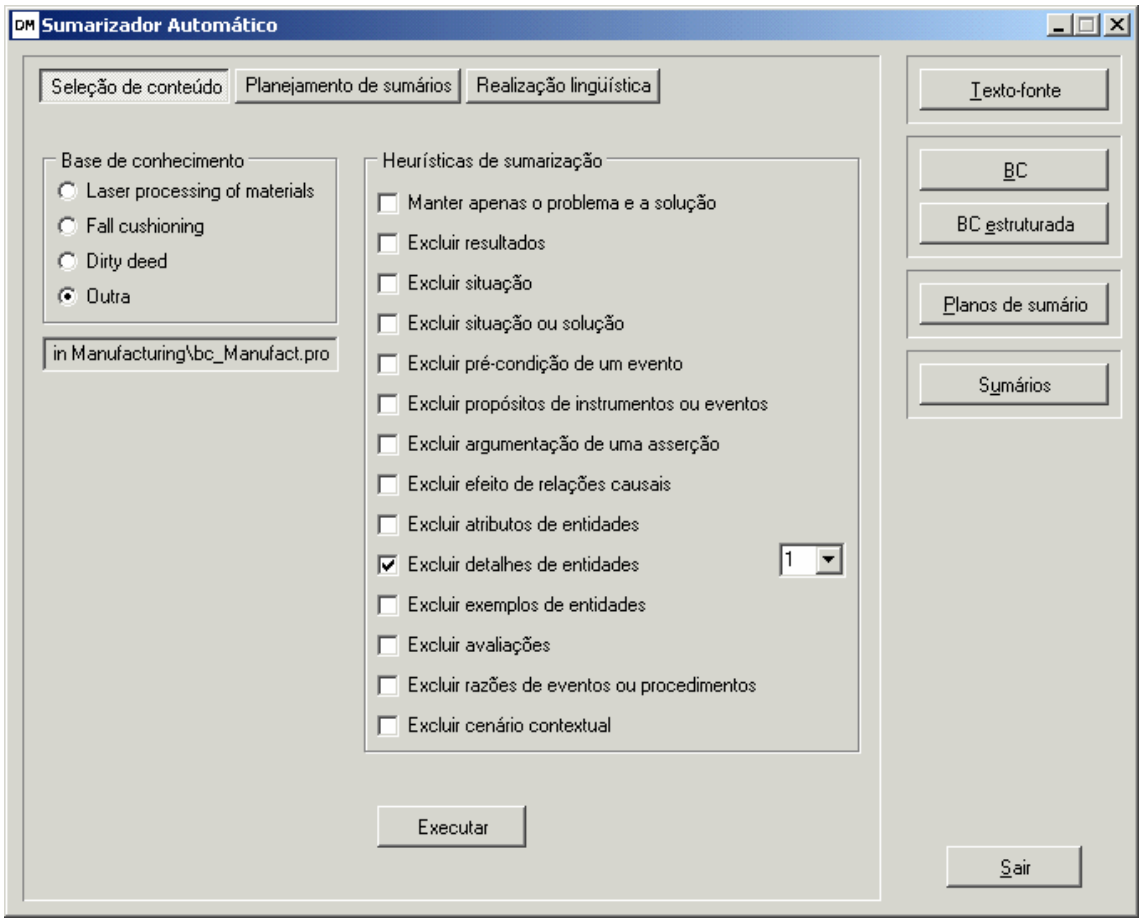


Figura 12 – Interface da seleção de conteúdo

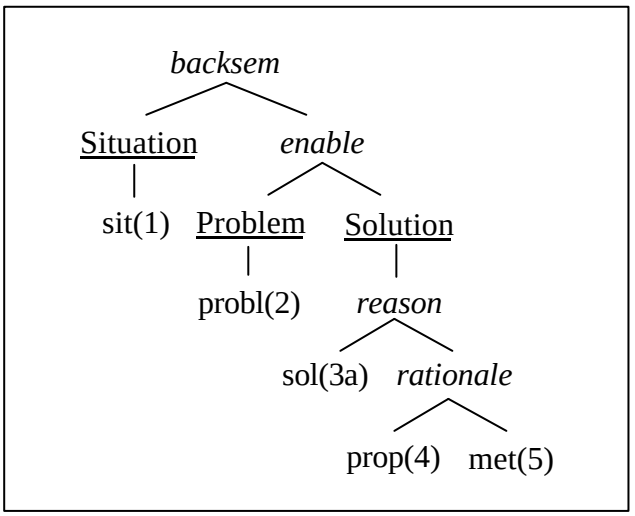


Figura 13 – Base de conhecimento podada

A interface para o planejamento textual é mostrada na Figura 14. Primeiramente, o usuário deve indicar a base de conhecimento do sumário a ser gerado. Para o texto ilustrado, esta será a base da Figura 13. Além disso, para completar a mensagem-fonte, o usuário deve fornecer os demais dados de entrada, isto é, o objetivo comunicativo (no exemplo, *relatar*) e a proposição central (no exemplo, *probl(2)*). Como essa interface se destina a usuários especialistas, permitindo que ele explore variações discursivas na geração de sumários, outras informações devem ainda ser fornecidas, de caráter profundo e dependente do modelo básico do DMSumm, quais sejam: o grupo de operadores de plano, o grupo de intenções, o nível de profundidade dos planos de texto que se deseja (no exemplo, *qualquer* profundidade) e o número de planos a serem gerados (no exemplo, *todos* os possíveis). Uma vez preenchidas todas as opções de customização do planejamento textual, basta mandar executa-lo (botão *executar*). A Figura 15 mostra todos os planos de texto possíveis gerados pelo DMSumm.

É importante notar que as opções de *nível de profundidade* e *número* de planos a serem gerados permitem ao usuário ter certo controle sobre a geração dos sumários. Quanto mais planos o usuário desejar, maior variedade de sumários será acessível. O nível de profundidade, por sua vez, é um fator particularmente interessante, pois permite ao usuário determinar, de certa forma, a complexidade dos planos e, conseqüentemente, dos sumários também, de forma que, quanto maior a profundidade de um plano, mais complexo ele será.

Como se pode perceber, alguns dos planos gerados na Figura 15 contêm o mesmo conteúdo proposicional, diferindo apenas na estrutura retórica (por exemplo, planos 1 e 4). Esse tipo de variedade é útil, por exemplo, em casos em que são necessárias diferentes estratégias retóricas em sumários com mesmo conteúdo. Vale notar que o plano 8 seria o que Hoey (1983) propôs, inicialmente, como correspondente a um sumário mínimo para um texto-fonte cuja estrutura fosse a do modelo Problema-Solução. Entretanto, no DMSumm, aceita-se, por exemplo, um plano de texto que veicule somente a proposição central, como é o caso do plano 10. A hipótese aqui é a de que a audiência a que se destina o sumário compartilha conhecimento profundo sobre o assunto do texto-fonte e, logo, basta mencionar a proposição central, que a comunicação estará estabelecida, isto é, a mensagem-fonte pretendida pelo escritor poderá ser compartilhada com o leitor.

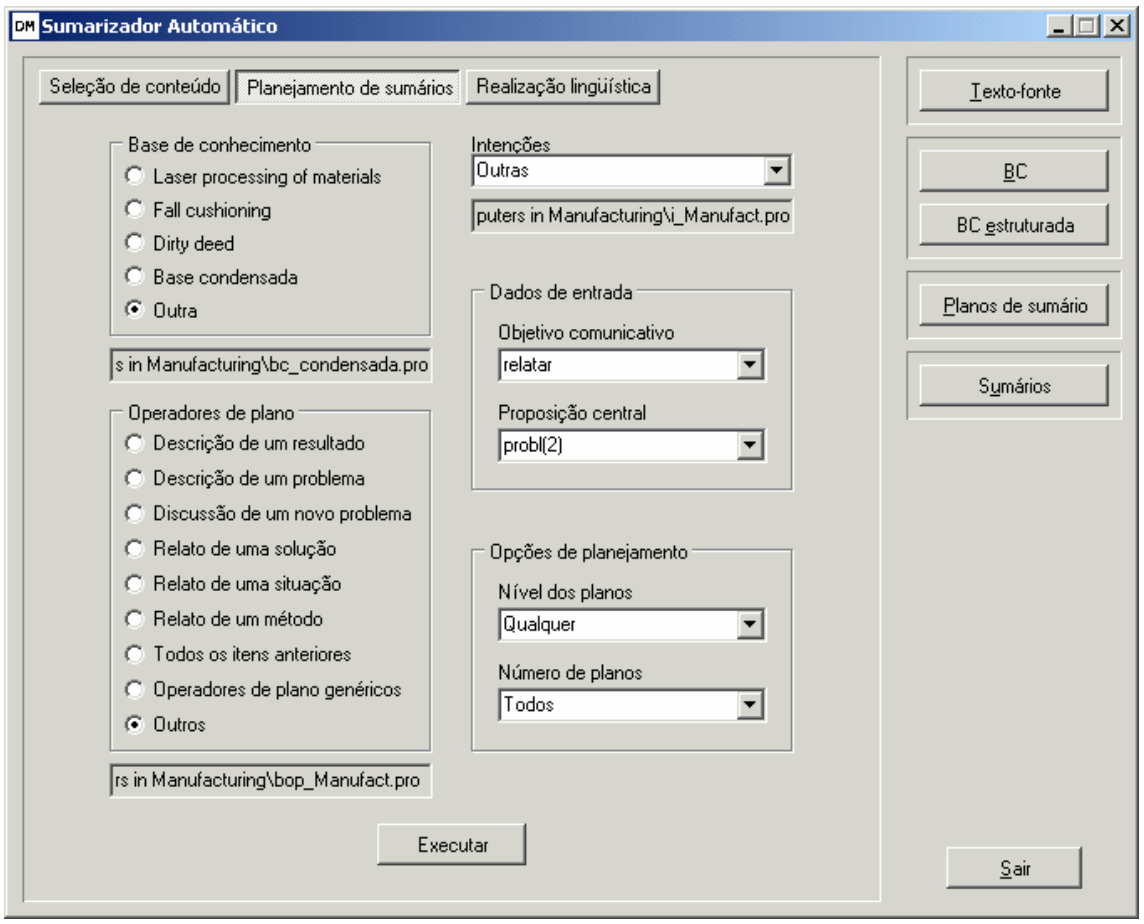


Figura 14 – Interface do planejamento textual

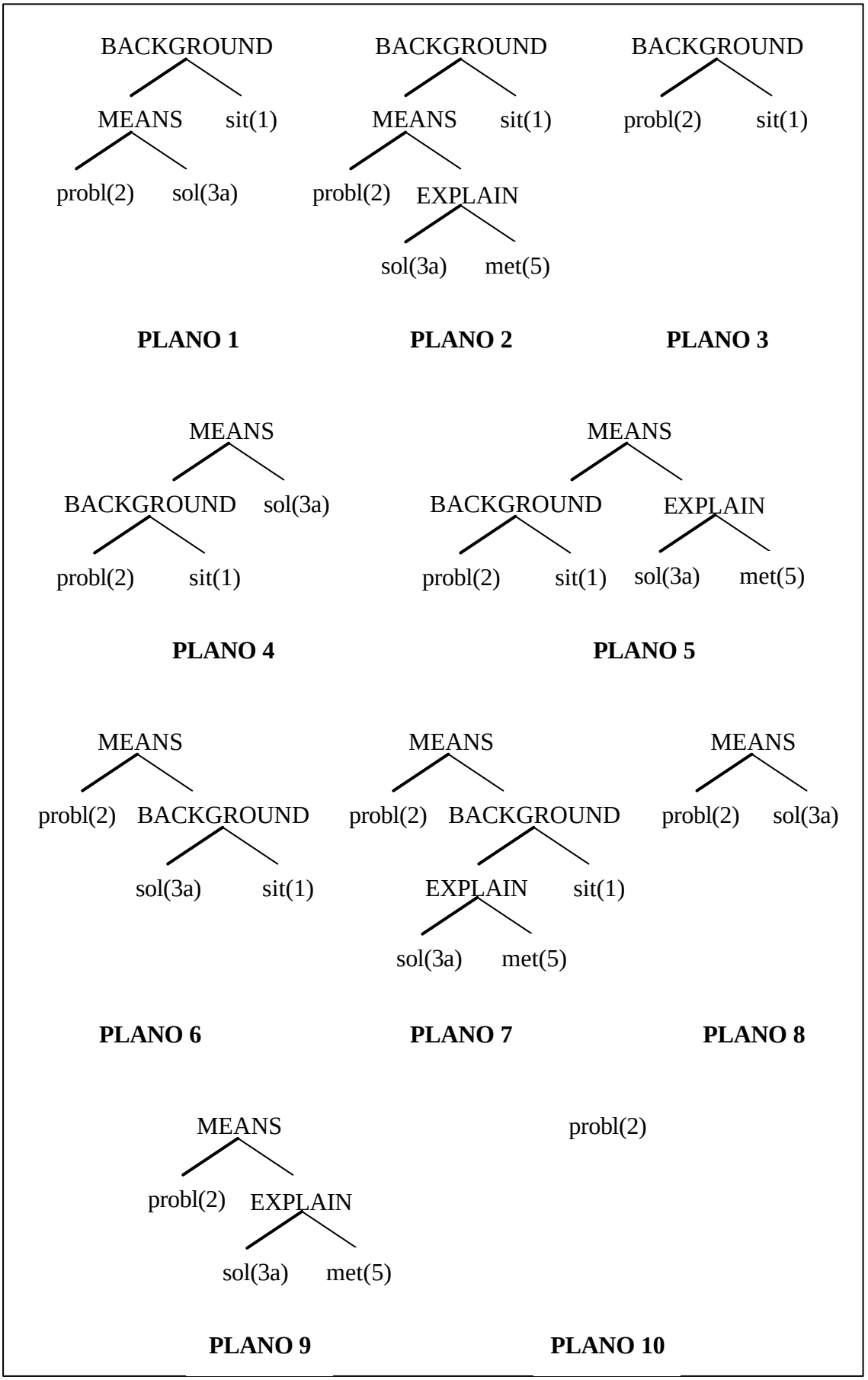


Figura 15 – Planos de texto

A Figura 16 mostra a interface para a realização lingüística. O usuário deve indicar o arquivo de planos de texto (gerado pelo processo de planejamento textual) e o texto da base de conhecimento correspondente (segmentado conforme os segmentos da base de conhecimento), para que os *templates* possam ser preenchidos. A Figura 17 mostra os sumários gerados para o exemplo adotado. A ordem dos sumários segue a ordem dos planos produzidos automaticamente (Figura 15).

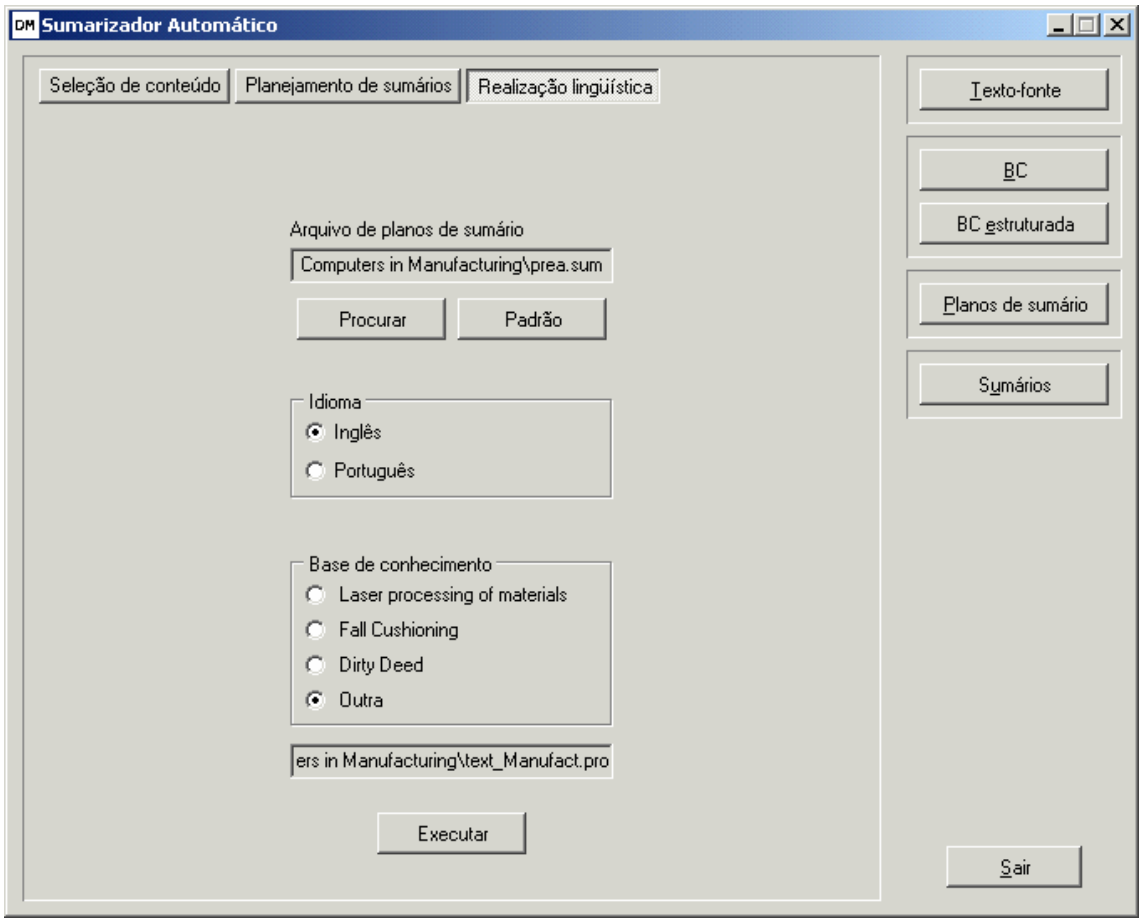


Figura 16 – Interface da realização lingüística

SUMÁRIO 1

Whether you regard computers as a blessing or a cure, 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.

SUMÁRIO 2

Whether you regard computers as a blessing or a cure, 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.

SUMÁRIO 3

Whether you regard computers as a blessing or a cure, 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.

SUMÁRIO 4

Whether you regard computers as a blessing or a cure, 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.

SUMÁRIO 5

Whether you regard computers as a blessing or a cure, 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.

SUMÁRIO 6

The general level of understanding of the power and weaknesses of computers among manufacturing managers is dangerously low. Considering this, whether you regard computers as a blessing or a cure, the fact is that we are all becoming more and more affected by them.

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.

SUMÁRIO 7

The general level of understanding of the power and weaknesses of computers among manufacturing managers is dangerously low. Considering this, whether you regard computers as a blessing or a cure, the fact is that we are all becoming more and more affected by them.

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.

SUMÁRIO 8

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.

SUMÁRIO 9

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.

SUMÁRIO 10

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

Figura 17 – Sumários

7. Avaliação do DMSumm

Esta seção relata a avaliação do DMSumm e de seus resultados, com especial enfoque nas premissas 3 e 4 deste trabalho, isto é, satisfação do objetivo comunicativo e preservação da proposição central, para validar o modelo de discurso de Rino (1996a). Os experimentos realizados basearam-se principalmente nos trabalhos de Sparck Jones e Galliers (1996), White et al. (2000) e Mani (2001).

Para os experimentos, foi utilizado um corpus de 10 introduções (com uma média de 530 palavras cada) de dissertações e teses de computação em português do Brasil, chamado *Thesis Corpus*, o qual é um recorte do *CorpusDT* (Feltrim et al., 2001). Esse corpus foi escolhido pelas seguintes razões: a) seus textos, do gênero científico, seguem o modelo Problema-Solução; b) os textos possuem os sumários autênticos, isto é, sumários escritos pelo próprio autor do texto, utilizados para comparação com os sumários produzidos automaticamente pelo DMSumm. Os experimentos são relatados a seguir.

7.1. Experimento 1: Identificação da Idéia Principal dos Textos-Fonte

Este experimento serviu como base para a realização do experimento seguinte[3]. Nove juízes, lingüistas computacionais e falantes nativos do português do Brasil, foram incumbidos de selecionar uma sentença que melhor representasse a idéia principal de cada texto do *Thesis Corpus*. Assumiu-se, para cada texto, que somente a sentença mais votada corresponderia à idéia principal.

A maioria dos juízes (60%) indicou a *solução* (de acordo com o modelo Problema-Solução) do texto como sendo a idéia principal; nos outros casos, o *método*, o *problema*, o *resultado* e uma *proposição genérica* foram indicados (somente uma vez cada no restante dos textos). Isso deixa evidente que: a) a macro-estrutura Problema-Solução teve um papel importante na determinação da idéia principal dos textos; b) considerando que frases indicativas são freqüentemente utilizadas no gênero do corpus, elas podem ter sido a base para a maioria dos julgamentos, reforçando trabalhos anteriores em extração de tópicos principais de textos (por exemplo, Paice, 1981; Hoey, 1983; Mani, 2001). Exemplos dessas frases indicativas no *Thesis Corpus* são *"este trabalho investiga..."*, *"neste trabalho de pesquisa apresenta-se..."*, *"tornou-se essencial o..."*, *"nessa dissertação são discutidos..."*, *"esse trabalho motiva-se em..."*, *"o objetivo*

[3] A elaboração desse experimento foi feita por Camilla Brandel Martins e Alice Picon Espina.

deste artigo é...", "*esse trabalho consistiu em...*", "*esse trabalho tem por objetivo investigar...*" e "*o trabalho reportado foi realizado com a finalidade de...*".

Os resultados desse experimento foram utilizados para a determinação da proposição central de cada texto do corpus, visando o próximo experimento.

7.2. Experimento 2: Avaliação dos Sumários Automáticos

Dez juízes, também lingüistas computacionais e falantes nativos do português do Brasil, julgaram os sumários apresentados para cada texto do corpus. O objetivo deste experimento foi medir a qualidade dos sumários, o que foi feito segundo uma série de parâmetros, conforme relatado abaixo.

Foram apresentados aos juízes 4 sumários por texto: 3 automáticos e 1 autêntico. Os sumários automáticos foram completamente gerados pelo DMSumm e correspondem, em média, a 40 % dos textos-fonte. Foram duas as tarefas dos juízes para cada texto do corpus: identificar o sumário autêntico (Tarefa 1) e classificar cada sumário de acordo com pontos de decisão (White et al., 2000) (Tarefa 2), que são os parâmetros do julgamento. Foram dois os pontos de decisão considerados neste experimento: *preservação da idéia principal* e *textualidade*, cujos valores são os indicados na Tabela 4.

Tabela 4 – 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

A seguir, serão descritos os resultados dessas avaliações.

7.2.1. Identificação do Sumário Autêntico

Na Tarefa 1, a maioria dos juízes identificou os sumários autênticos. Segundo a análise realizada, esse resultado se deve, principalmente, às seguintes razões: a) os sumários autênticos normalmente apresentam mais conteúdo do que o apresentado nas suas fontes (conforme discutido posteriormente na subseção 7.3) e b) eles possuem estruturas sintáticas mais ricas que os sumários gerados automaticamente, dado o

componente limitado de realização lingüística do DMSumm e a própria natureza da sumarização humana. Como exemplo, abaixo é mostrado o sumário autêntico e um sumário automático para o texto 2 do Apêndice A. É possível notar que o sumário autêntico é mais articulado que o automático.

SUMÁRIO AUTÊNTICO

Aplicações de Processamento de Língua Natural (PLN), como revisores ortográficos, gramaticais e tradutores, geralmente precisam consultar extensos dicionários contendo, por exemplo, informações morfossintáticas de várias centenas de milhares de palavras de uma língua. Autômatos finitos, largamente utilizados na construção de eficientes analisadores léxicos para compiladores, apresentam-se como ótimos candidatos para o problema de representação desse tipo de léxico de língua natural. Este trabalho investigou os métodos de codificação de léxicos utilizando autômatos finitos, as técnicas de minimização de autômatos determinísticos acíclicos e as estruturas de dados necessárias para uma representação compacta, resultando em um sistema computacional eficiente e versátil de representação de grandes léxicos de língua natural. Um léxico de mais de 430.000 palavras da língua portuguesa pode ser convertido em um autômato de menos de 220Kb, em menos de 5 minutos usando um computador doméstico.

SUMÁRIO AUTOMÁTICO

A representação de grandes dicionários de língua natural, principalmente nos casos em que se trabalha com vários milhões (ou dezenas de milhões) de palavras, é um interessante problema computacional a ser tratado dentro da área de Processamento de Língua Natural. Em face disso, autômatos finitos, largamente usados na construção de compiladores, são excelentes estruturas para representação de dicionários. Os autômatos finitos permitem acesso direto às palavras e seus possíveis atributos (gênero, número, grau, etc). Esse trabalho tem por objetivo investigar métodos eficientes de construção, representação e minimização de autômatos finitos de grande porte (dezenas de milhões de estados). Como resultado, será possibilitada a criação de um formato otimizado capaz de ser inserido diretamente no autômato finito, com o menor impacto possível no tamanho final do autômato. Será desenvolvido um protótipo para efeito de testes e confirmação de eficiência do sistema.

7.2.2. Avaliação da Qualidade dos Sumários Automáticos com Base em Dois Pontos de Decisão

Com relação à Tarefa 2, o Gráfico 1 sintetiza o julgamento dos juízes, mostrando a média das classificações dos sumários automáticos para cada texto (linhas sólidas). Para comparação, os sumários autênticos estão representados pela linha pontilhada. Seguindo White et al. (2000), a medida *means* foi calculada com base na média das classificações, resultando na nota 3 como nota mínima aceitável para se ter um sumário satisfatório. Como se pode ver, a maioria dos sumários automáticos está acima da média de aceitabilidade e, portanto, são considerados satisfatórios. Pode-se perceber também que o julgamento dos sumários automáticos e autênticos não apresenta discrepância significativa: 7 dos 10 sumários automáticos estão próximos dos correspondentes autênticos (referentes aos textos 2, 4, 6, 7, 8, 9 e 10). Entretanto, vale notar que somente em um dos casos o automático está praticamente no mesmo patamar que o autêntico (texto 10), enquanto que, nos demais casos, a qualidade dos autênticos é sempre superior. A maior discrepância entre os sumários automáticos e o sumário autêntico ocorreu para o texto 3, discrepância esta que será analisada nas próximas subseções.

O Gráfico 2 mostra a distribuição dos julgamentos dos sumários automáticos em função das notas, com 62% dos sumários acima de *means*[4].

Procurando validar esse experimento, foi calculada a medida *Kappa* (Carletta et al., 1997), a qual mede a confiabilidade dos julgamentos dos juízes. Como resultado, chegou-se ao valor 0.78, que indica que a concordância entre os juízes em seus julgamentos é *substancial*, validando, dessa forma, o experimento realizado e seus resultados.

[4] Os sumários que estão em *means*, ou seja, 22% dos sumários, são considerados *aceitáveis*, mas não satisfatórios, pois podem ser melhorados. Diferentemente deste trabalho, White et al. consideram que sumários em *means* também são satisfatórios.

Gráfico 1 – Média das notas dos sumários por texto

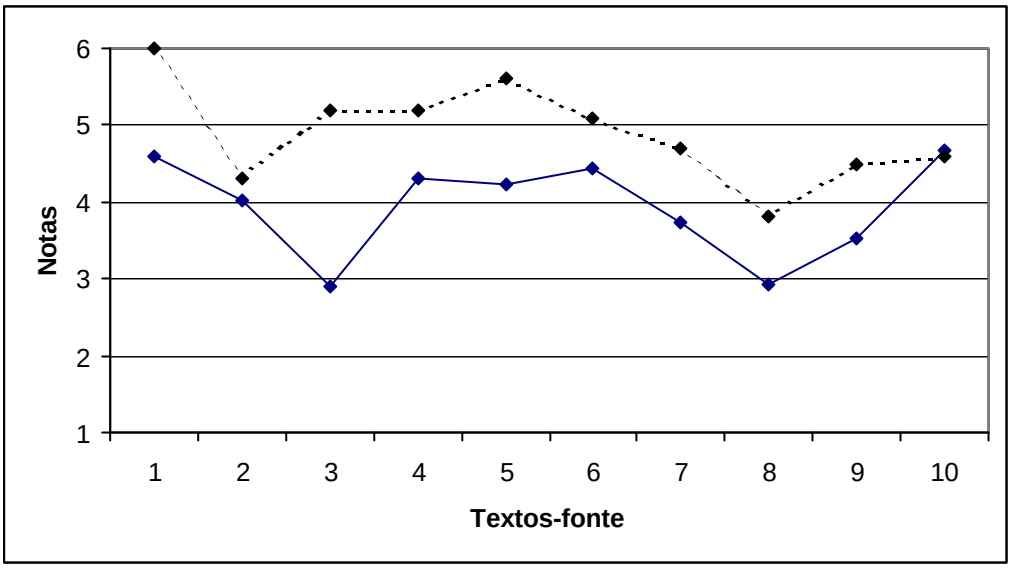
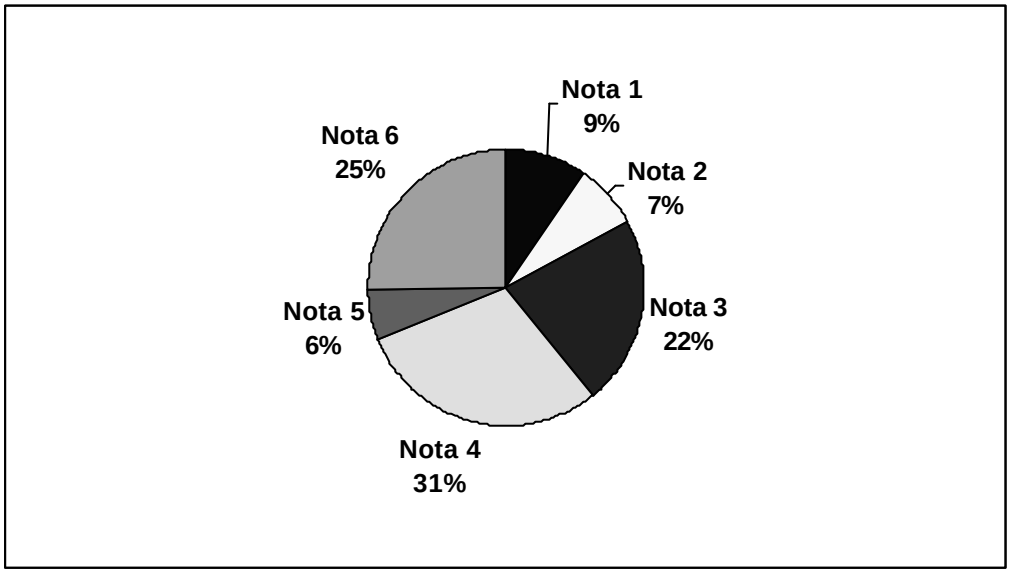


Gráfico 2 – Distribuição dos julgamentos dos sumários automáticos por nota



Com base nos experimentos acima, foi realizada uma análise mais detalhada, considerando os pontos de decisão isolados, conforme descrito a seguir.

7.2.3. Avaliação da Textualidade e da Preservação da Idéia Principal de Forma Isolada

Considerando somente a textualidade como indicador de qualidade, proposta também por Mani (2001), a maioria dos sumários automáticos manteve a textualidade (67%) dos textos-fonte (vide Gráfico 3). Entretanto, comparados aos autênticos, dos quais 90% mantiveram a textualidade, essa taxa é considerada baixa. Isto é parcialmente

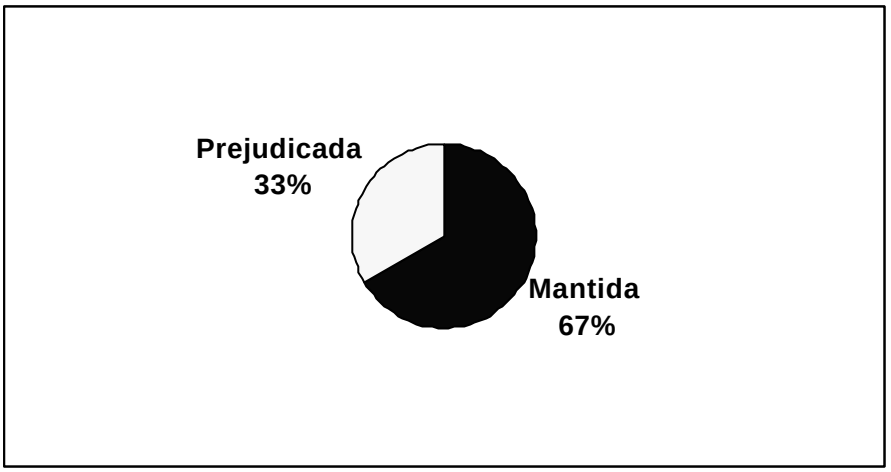
devido à *performance* do realizador lingüístico. Como evidência desse problema, considere o sumário a seguir gerado automaticamente pelo DMSumm para um dos textos do experimento. Os problemas de coesão são aparentes:

- como o plano de texto referente a este sumário possui muitas relações retóricas *elaborate* aninhadas, vários marcadores discursivos *ainda* foram inseridos em seqüência;
- o marcador *e*, por ser um marcador muito fraco (isto é, não indica claramente sua função discursiva, servindo para marcar praticamente qualquer tipo de relação de discurso), prejudica a leitura do sumário, pois passa a impressão ao leitor de que duas partes do texto sem conexão foram justapostas.
- o termo *itens de informação*, por exemplo, é exaustivamente repetido no decorrer do sumário, pois não se trata as questões de co-referência na abordagem simplista do realizador lingüístico.

Durante todo o ciclo de vida do software, são produzidos muitos itens de informação. ***Ainda****, todos os itens de informação, relativos ao software que está sendo construído ou modificado, constituem a Configuração de Software,* ***e****, os itens de informação precisam ser controlados durante todo o ciclo de vida do software.* ***Ainda****, o Gerenciamento de Configuração de Software é uma disciplina que ajuda a solucionar esse problema, pois envolve o desenvolvimento de padrões e procedimentos para administrar um software evolutivo.* ***Ainda****, devido à complexidade de suas tarefas, o gerenciamento de configuração de software necessita de ferramentas automatizadas que possam torná-las um pouco mais simples.* ***Ainda****, o RCS (Revision Control System) é uma ferramenta para gerenciamento de configuração de software.*

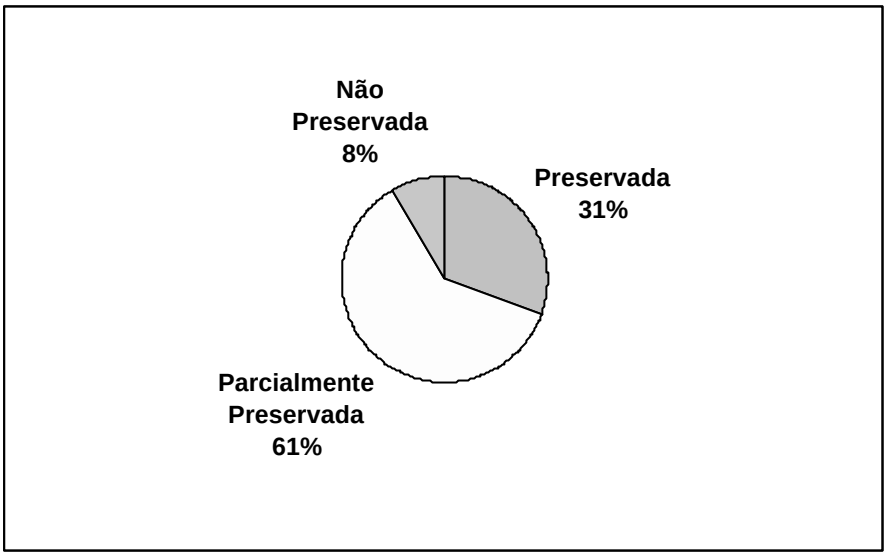
O RCS para UNIX foi alvo de um estudo detalhado e constatou-se que o RCS é um sistema difícil de ser usado, principalmente devido à sua interface, constituída apenas de comandos de linha. ***Em face disso****, o objetivo deste artigo é apresentar uma interface amigável para o sistema RCS, seguindo as noções básicas de Projeto de Interface.* ***Como resultado****, pretende-se possibilitar que o sistema seja utilizado com mais freqüência e interesse, através de uma interface mais simpática e fácil de utilizar, e além disso divulgar os importantes conceitos de gerenciamento de configuração de software, os quais se tornarão mais visíveis através da interface amigável.*

Gráfico 3 – Distribuição dos sumários automáticos de acordo com a textualidade



Além da questão do realizador lingüístico, o problema da textualidade também se deve ao fato de que assumir a proposição central como sendo uma única unidade informativa do texto-fonte pode ser uma restrição muito forte para a sumarização automática, pois, em alguns casos, informações esparsas podem ser importantes para a preservação da proposição central, ou podem, mesmo, fazer parte da própria proposição central, conforme discute O'Donnel (1997). Corroborando isso, durante a realização do primeiro experimento (subseção 7.1), no qual os juízes tinham que identificar a proposição central dos textos, de fato, em muitos casos, mais de uma sentença foi indicada como representando a idéia principal do texto. Porém, isso não foi considerado no DMSumm, dado que, neste trabalho, assume-se que a proposição central é uma informação elementar. Tal problema também é apontado pela distribuição dos sumários automáticos em relação à preservação da idéia principal (Gráfico 4): 61% dos sumários preservaram somente parcialmente a idéia principal e 31% a preservaram totalmente. Em contraste, todos os sumários autênticos preservaram a idéia principal.

Gráfico 4 – Distribuição dos sumários automáticos de acordo com a preservação da idéia principal



Segundo as análises realizadas, a maior discrepância entre os sumários automáticos e o sumário autêntico do texto 3 (vide Gráfico 1) deve-se tanto ao realizador lingüístico limitado, quanto ao fato da proposição central ter sido tomada como pontual, ou seja, representada somente por uma única sentença do texto-fonte. Em particular, para o texto 3, três sentenças foram quase igualmente voltadas pelos juízes como representando a idéia principal do texto. Abaixo são mostrados, respectivamente, o texto-fonte, o sumário autêntico e um sumário automático para esse texto. Mesmo sem ler o texto-fonte, é possível perceber as diferentes informações principais consideradas nos sumários: enquanto o autêntico focaliza a implementação em si de um módulo de redes para o sistema ASiA, o automático focaliza a motivação para tal implementação. Com relação à realização lingüística do sumário automático, percebe-se, por exemplo:

- a construção de um parágrafo muito grande (primeiro parágrafo) e de um parágrafo muito curto (último parágrafo), dada as decisões de agrupamento nos *templates*;
- a progressão temática ruim entre as sentenças do primeiro parágrafo[5];
- o marcador *como resultado* sendo utilizado inadequadamente, ou seja, o fato expresso pela sentença à qual ele pertence não é exatamente um resultado do que é apresentado pelo sentença anterior. Isso espelha uma das fragilidades, neste trabalho, da abordagem por *templates* para a realização, ou seja, a incapacidade de se selecionar adequadamente qual marcador discursivo utilizar.

[5] Progressão temática refere-se à forma como os tópicos (temas) são introduzidos em um texto. Se o tópico muda bruscamente, diz-se que a progressão temática está ruim.

- repetições demasiadas do termo *módulo de redes*.

TEXTO-FONTE

Os sistemas computacionais e as redes de computadores têm apresentado uma rápida evolução, se tornando cada vez mais rápidos e mais complexos. As técnicas, acopladas com o acréscimo de demanda por maior eficiência e menor tempo para disseminação da informação, conduzem a tecnologias mais rápidas. Existe um crescimento explosivo na utilização, desenvolvimento e pesquisa em redes de comunicação e computadores de alto desempenho.

Esse crescimento tem sido estimulado em grande parte pelos avanços nos sistemas computacionais e projetos de rede, protocolos, serviços, gerenciamento, arquitetura e aplicações.

Devido ao grande avanço, tanto em termos tecnológicos quanto de utilização das redes de computadores, tornou-se essencial o desenvolvimento de um estudo sobre o desempenho dessas redes. Além da avaliação de desempenho é importante ainda a validação do sistema a ser utilizado. Dentre as formas de avaliação e validação de sistemas, a utilização de modelagem de sistemas computacionais e redes de computadores se destacam. Um estudo da modelagem e simulação é necessário quando o projetista de sistemas computacionais precisa tecer comparações entre diversas alternativas de projeto para encontrar a abordagem de melhor relação custo/benefício. Não existem motivos para projetar e implementar novos sistemas que não tenham desempenho e que não podem competir com sistemas já existentes no mercado.

A utilização da técnica de modelagem também é necessária quando um gerente de um centro de computação precisa comparar diversos sistemas para decidir qual é o melhor para um dado conjunto de aplicações. A avaliação de desempenho de um sistema existente é também importante desde que auxilie na determinação de quão bem o sistema está, assim como quais melhoramentos são necessários para aumentar seu desempenho. As abordagens mais comuns para avaliar o desempenho de sistemas computacionais e redes de computadores são a modelagem analítica e a simulação.

A solução analítica permite a obtenção de resultados mais precisos, porém, à medida que o modelo se torna mais complexo, aumenta-se o grau de dificuldade da resolução por métodos analíticos. Além disso, alterações no modelo implicam em modificações em todas as equações e o recálculo das mesmas. A simulação, por outro lado, possui facilidades para modificações no modelo uma vez que apenas o programa deve ser alterado.

A simulação é uma ferramenta bastante útil para avaliação de desempenho de sistemas, porém, a sua utilização não é muito fácil, pois implica em aprendizado de linguagens de simulação, de modelagem de sistemas e de análise dos dados. Desse modo, para facilitar o uso da simulação existem ambientes de simulação automáticos que afastam o usuário da tarefa árdua de transcrição de um modelo em um programa de simulação.

No Grupo de Sistemas Distribuídos e Programação Concorrente do ICMC - USP está em desenvolvimento um Ambiente de Simulação Automático (ASiA) que permite ao usuário a utilização de simulação sem necessitar de conhecimento das diferentes áreas envolvidas no desenvolvimento e na análise dos resultados gerados por uma simulação, tais como: linguagens de simulação, estatística, etc. O ASiA fornece facilidades gráficas para o usuário desenhar o seu modelo através de redes de filas, e a partir de parametrizações desse modelo, o ambiente gera e executa o programa de simulação.

Devido a crescente importância que as redes de computadores tem adquirido, é essencial uma ferramenta de simulação de redes de computadores para o ASiA. A importância do sistema para simulação de redes do ASiA está relacionada com a grande flexibilidade apresentada pelo ASiA e pela possibilidade de se integrar a simulação de redes com outras facilidades disponíveis nesse sistema, tais como: simulação de arquiteturas de computadores; utilização de ferramentas para análise da saída; utilização de ferramentas para visualização dos resultados, etc.

Nessa dissertação são discutidos o projeto e a implementação de um módulo de simulação de redes de computadores para o ASiA. O módulo de redes permite que o usuário simule redes de computadores sem necessitar de conhecimentos específicos em redes de filas (que é a abordagem utilizada pelo ASiA) nem simulação. Essa ferramenta torna a simulação de redes uma tarefa mais simples pois fornece subsídios, tais como modelos prontos que o usuário pode utilizar de maneira mais fácil, podendo até modificar os parâmetros desses modelos prontos para obter resultados dessa rede com os parâmetros fornecidos pelo usuário. Desta forma, o objetivo deste trabalho é estender o ASiA para que este ofereça os recursos necessários para a simulação de redes de computadores e implementar uma versão inicial do módulo de redes que poderá ser facilmente estendido, através da inserção de novos modelos.

Um requisito inicial para o desenvolvimento do módulo de redes é o conhecimento de simulação de sistemas e dos parâmetros e dos elementos envolvidos no processo de simulação, além da análise dos dados obtidos.

SUMÁRIO AUTÊNTICO

Esta dissertação de mestrado descreve a implementação de um módulo de simulação de redes de computadores para o ASiA (Ambiente de Simulação Automático). Este módulo permite ao usuário a simulação de redes já definidas (com a possibilidade de alteração dos parâmetros de entrada) ou definir novas redes através da utilização dos recursos da barra de ferramentas.

Para aumentar a gama de sistemas que o usuário pode modelar foram implementados e incluídos na barra de ferramentas novas funcionalidades que permitem a modelagem de sistemas mais complexos.

Este trabalho apresenta também uma revisão bibliográfica sobre simulação, redes de computadores e ambientes de simulação.

SUMÁRIO AUTOMÁTICO

Os sistemas computacionais e as redes de computadores têm apresentado uma rápida evolução, se tornando cada vez mais rápidos e mais complexos. ***Como resultado****, a utilização da técnica de modelagem é necessária quando um gerente de um centro de computação precisa comparar diversos sistemas para decidir qual é o melhor para um dado conjunto de aplicações. O módulo de redes permite que o usuário simule redes de computadores sem necessitar de conhecimentos específicos em redes de filas (que é a abordagem utilizada pelo ASiA) nem simulação,* ***e****, além do módulo de redes tornar a simulação de redes uma tarefa mais simples pois fornece subsídios, tais como modelos prontos que o usuário pode utilizar de maneira mais fácil, podendo até modificar os parâmetros desses modelos prontos para obter resultados dessa rede com os parâmetros fornecidos pelo usuário.*

Dentre as formas de avaliação e validação de sistemas, a utilização de modelagem de sistemas computacionais e redes de computadores se destacam e, devido ao grande avanço, tanto em termos tecnológicos quanto de utilização das redes de computadores, tornou-se essencial o desenvolvimento de um estudo sobre o desempenho de redes e é importante a validação do sistema a ser utilizado.

Nessa dissertação são discutidos o projeto e a implementação de um módulo de simulação de redes de computadores para o ASiA.

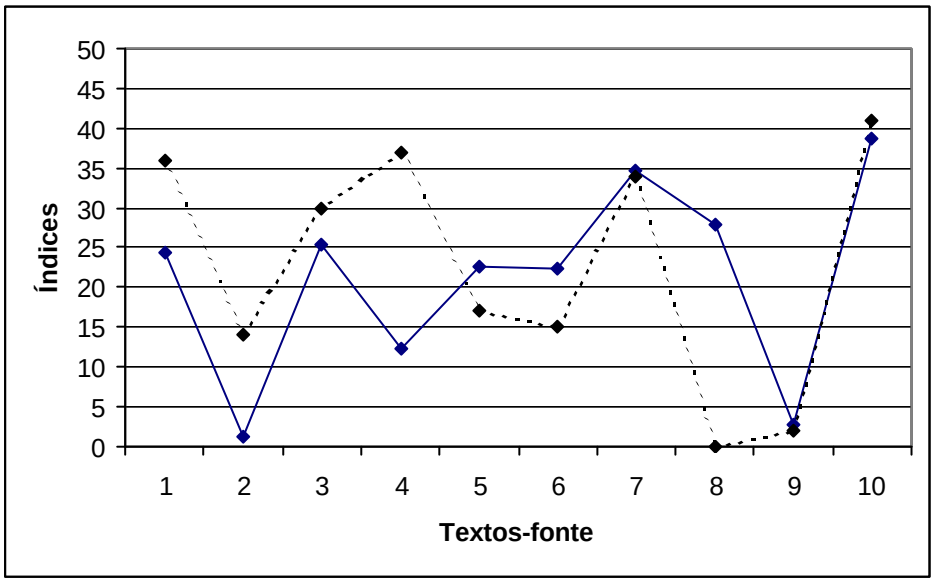
7.2.4. Avaliação da Proximidade entre Sumários e Textos-Fonte

Mani (2001) também sugere que a legibilidade dos sumários pode indicar o quão próximos estes estão de seus textos-fonte. Assim, se os sumários mantiverem o mesmo grau de facilidade/dificuldade de leitura dos textos-fonte, eles podem indicar, em

princípio, seu grau de correspondência com os textos-fonte. Entretanto, é importante ressaltar que, de acordo com o próprio Mani, não se sabe em que nível a legibilidade realmente reflete essa correspondência entre sumários e texto-fonte. De qualquer modo, optou-se, neste trabalho, por verificar tal correspondência também.

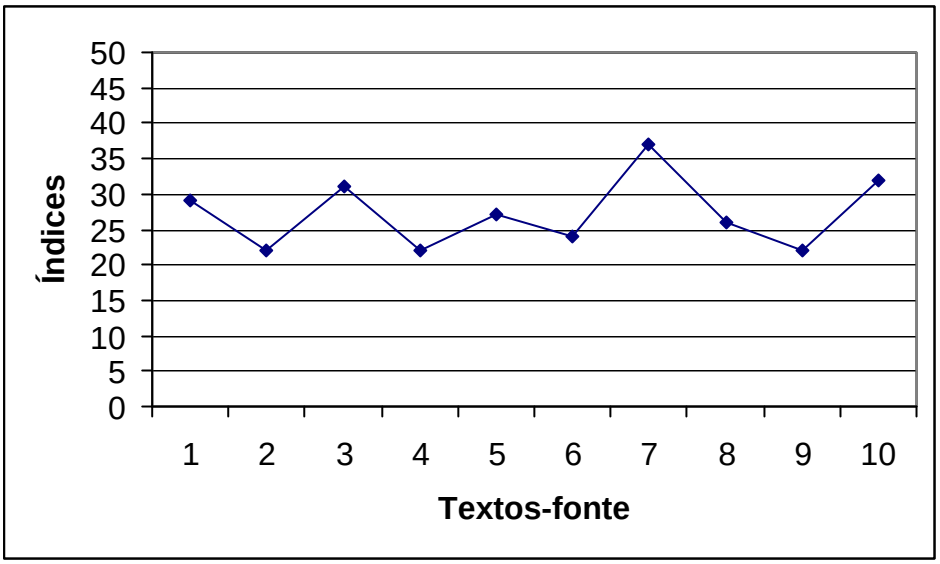
Utilizando o índice Flesch (1948) adaptado para o português do Brasil (Martins et al., 1996), o índice de legibilidade dos sumários (autênticos e automáticos) foi calculado, conforme mostra o Gráfico 5. O Gráfico 6 mostra a legibilidade dos textos-fonte correspondentes. Quanto maior o índice de legibilidade, mais fácil é a leitura do texto. Segundo os gráficos, todos os sumários foram considerados difíceis para a leitura[6], assim como os textos-fonte.

Gráfico 5 – Índice de legibilidade médio dos sumários para cada texto-fonte



[6] O intervalo 25-50 indica textos difíceis; 0-24 indica textos muito difíceis.

Gráfico 6 – Índice de legibilidade dos textos-fonte

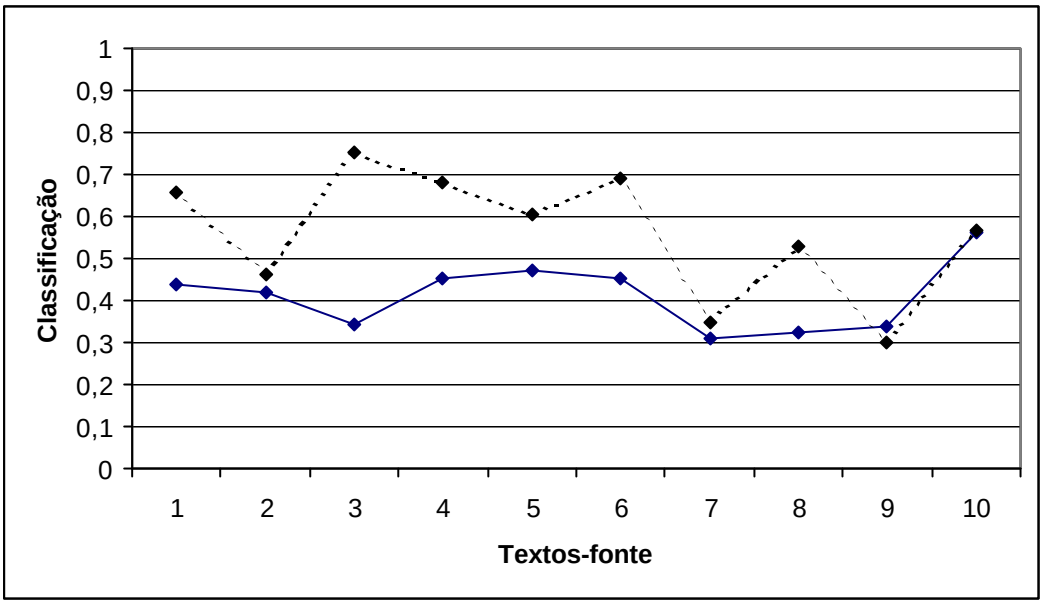


7.2.5. Avaliação da Informatividade dos Sumários

O Experimento 2 também permitiu a comparação entre os sumários, medindo-se a *"informatividade semântica"* deles, isto é, o quanto de informação apresentada no sumário autêntico 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. Por sua vez, quanto maior a classificação (o máximo é 1), mais informativo é o sumário.

O Gráfico 7 mostra a média das classificações dos sumários para cada texto-fonte. Grandes discrepâncias, quando se comparam os sumários automáticos e os autênticos, ocorrem somente para o texto 3, o qual já foi analisado na subseção 7.2.3. Por outro lado, os sumários automáticos relativos aos textos 2, 7, 9 e 10 são tão ou mais informativos quanto os autênticos. As outras classificações, próximas de *means* (0,41), mostram que o DMSumm produz sumários razoáveis, mas que ainda podem ser melhorados. Para os sumários autênticos, *means* foi calculada em 0,55.

Gráfico 7 – Informatividade semântica média dos sumários por texto-fonte



Abaixo é mostrado o sumário autêntico e um automático do texto 2 para exemplificar a proximidade entre eles em termos da informatividade semântica: são aproximadamente do mesmo tamanho e apresentam quase o mesmo conteúdo em relação ao texto-fonte.

SUMÁRIO AUTÊNTICO

Neste trabalho de mestrado foi desenvolvido o projeto de uma máquina paralela dedicada para solução de sistemas de equações lineares. Este é um problema presente em uma grande variedade de aplicações científicas e de engenharia e cuja solução torna-se uma tarefa computacionalmente intensiva, à medida em que o número de incógnitas aumenta. Implementou-se uma Arquitetura Sistólica, conectada numa topologia em anel, que mapeia métodos de solução iterativos. Essa classe de arquiteturas paralelas apresenta características de simplicidade, regularidade e modularidade que facilitam implementações em hardware, sendo muito utilizadas em sistemas de computação dedicados à solução de problemas específicos, os quais possuem como requisitos a grande demanda computacional e a necessidade de respostas em tempo real. Foram adotadas metodologias e ferramentas avançadas para projeto de hardware que aceleram o ciclo de desenvolvimento e para a implementação foram utilizados circuitos reconfiguráveis FPGAs (Field Programmable Gate Arrays). Os resultados de desempenho são apresentados e discutidos, indicando que a abordagem e metodologia adotada é viável e eficiente para solução deste tipo de problema.

SUMÁRIO AUTOMÁTICO

A Comunidade de Computação de Alto Desempenho tem observado que é necessário desenvolver soluções personalizadas para seus problemas, e, explicando, soluções genéricas muitas vezes não são adequadas, pois sacrificam muito o desempenho quando aplicadas a um caso específico de interesse. A evolução das metodologias de projeto de hardware, apoiadas em poderosas ferramentas de software que aceleram o ciclo de desenvolvimento, e especialmente o surgimento de dispositivos reconfiguráveis como os FPGAs (Field-Programmable Gate Arrays) abriram um novo horizonte entre os extremos da computação de finalidade geral e o hardware dedicado. Como evidência, hoje, é possível desenvolver rapidamente um projeto de sistema digital empregando-se novas metodologias como linguagens de descrição de hardware (HDLs), ferramentas de síntese lógica e simulação, entretanto, na busca de alto desempenho, a abordagem mais adotada é a utilização de técnicas de programação concorrente para paralelização de algoritmos a serem executados em máquinas paralelas de propósito geral. Neste trabalho de pesquisa apresenta-se "Uma Arquitetura Sistólica para Solução de Sistemas Lineares implementada com os circuitos FPGAs".

7.3. Precisão e Cobertura dos Sumários Automáticos

Foram calculadas, ainda, as medidas de precisão (*precision*), cobertura (*recall*) e *f-measure* dos sumários automáticos em relação aos autênticos. É importante ressaltar que um sumário autêntico pode apresentar mais informação do que a presente no texto-fonte correspondente, pois o autor pode se basear em sua experiência prévia para produzir o sumário, não se apegando ao conteúdo do texto-fonte propriamente dito (Mushakoji, 1993; Rino, 1996a). Para lidar com isso, costuma-se retirar do sumário autêntico as sentenças que apresentam informações que não estão no texto-fonte, resultando em um sumário ideal chamado *gold standard* (Kupiec et al., 1995). Neste trabalho, verificou-se que a inclusão de informação extra nos sumários autênticos do *Thesis Corpus* ocorre minimamente, não afetando significativamente os resultados obtidos. Desta forma, não se utilizou os *gold standards* para essa avaliação.

No contexto da sumarização automática, as medidas de precisão, cobertura e *f-measure* 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.

Na média, o DMSumm produziu sumários com 44% de precisão e 54% de cobertura, com uma *f-measure* de 48%. O desvio padrão, tanto para a precisão como para a cobertura, foi de 21%.

7.4. Avaliação do Modelo de Discurso do DMSumm

Até este momento, foi apresentada a avaliação do DMSumm baseada em sua *performance*, isto é, em função dos sumários gerados automaticamente e, logo, com influência dos três processos clássicos da geração automática de textos (vide Figura 8, Seção 5).

As subseções seguintes discutem os níveis de representação do modelo de Rino. O objetivo foi avaliar somente a proposta de estruturação profunda do discurso. Em particular, o foco dessa avaliação está na representação semântica (da base de conhecimento), na intencional e na retórica (dos planos de texto), de forma que seja possível investigar a necessidade de cada nível no modelo de Rino (e seu grau de complexidade inerente) e identificar problemas existentes e soluções possíveis.

7.4.1. Nível Semântico

Essa avaliação foi fruto da observação de problemas durante o processamento do DMSumm. Por exemplo, no modelo original (Rino, 1996a), não havia um meio de se distinguir entre a geração das relações retóricas *list* e *contrast*, pois, no mapeamento profundo, a partir de uma única relação semântica (*listSem*) e da mesma relação intencional, podia-se gerar qualquer uma dessas relações retóricas:

Relação semântica	**Relação retórica**	**Relação intencional**
listSem(X,Y)	list(X,Y) contrast(X,Y)	X symm Y

Isto implicou a geração de planos de texto incoerentes. Para corrigir o problema, refinou-se a definição de tais relações. Assim, *listSem* foi desmembrada nas relações *simSem* e *difSem*:

Relações semânticas	Relações retóricas	Relações intencionais
simSem(X,Y)	list(X,Y)	X symm Y
difSem(X,Y)	contrast(X,Y)	X symm Y

Dessa forma, eliminou-se a ambigüidade retórica proveniente do modelo de discurso original.

7.4.2. Nível Retórico

Nesta etapa, a avaliação foi feita por dois juízes, lingüistas computacionais e falantes nativos do português do Brasil, especialistas em RST. Pediu-se aos juízes que julgassem a adequação das relações retóricas dos planos do texto produzidos automaticamente pelo DMSumm, tendo em mãos os textos-fonte correspondentes aos planos. O objetivo desta avaliação foi verificar o quão completo é o conjunto de relações retóricas do modelo de Rino e se os planos de texto eram coerentes, ou seja, se o mapeamento profundo feito para a geração das relações retóricas dos planos foi correto.

Os juízes concordaram que os planos de texto espelhavam a estrutura retórica dos textos-fonte, descartando, a princípio, possíveis problemas no mapeamento entre as relações profundas do discurso. Entretanto, no que diz respeito à completude do conjunto de relações retóricas de Rino, os juízes sentiram falta de relações que expressassem ações *volitivas* no modelo, já que o modelo discursivo contém somente relações que expressam ações *não volitivas*, isto é, *nonvolres* e *nonvolcause*. Uma das formas de se resolver isso é generalizando essas relações retóricas para *result* e *cause*, respectivamente, ou introduzindo um novo par de relações para representar ações volitivas, por exemplo, *volres* e *volcause*, diferenciando-as das relações não volitivas, durante o planejamento textual, por meio das restrições intencionais.

Conforme já dito no decorrer desta dissertação, esse problema surgiu pelo fato do modelo de discurso de Rino ter sido desenvolvido com base em estudo de corpora, não privilegiando, portanto, todas as análises existentes.

7.4.3. Nível Intencional

Nesta fase, a avaliação objetivou verificar a real necessidade do nível intencional no modelo de discurso, já que este nível atribui ao modelo um grau de complexidade ainda maior. Para verificar isto, foi feito um experimento adicional, no qual se suprimiu as relações intencionais do modelo de discurso. Dessa forma, o mapeamento de relações semânticas e intencionais em relações retóricas foi simplificado, passando a ser um mapeamento direto de relações semânticas em relações retóricas.

Todo o repositório de operadores de plano foi redefinido, retirando-se do campo *constraints* dos operadores as restrições intencionais. Com base nesse novo repositório, vários planos de texto foram gerados. Verificou-se, pela análise desses planos, que houve geração de planos incoerentes. Exemplos de planos incoerentes ocorreram, por exemplo, em casos em que relações *causa-efeito* foram invertidas, produzindo relações onde o *efeito* gerou a *causa*. Isso ocorreu porque são as intenções as responsáveis por escolher os segmentos discursivos que serão os núcleos e os satélites das relações retóricas, já que, na organização semântica, não se considera o que é nuclear ou não. Como exemplo disso, considere o mapeamento abaixo sem a nível intencional:

Relação semântica	**Relação retórica**
cause(X,Y)	nonvolres(Y,X)

De acordo com esse mapeamento, caso se tenha a base de conhecimento da Figura 18 referente ao texto abaixo (já segmentado), no qual o fato de *ter chovido* (causa) fez a *grama crescer* (efeito), os planos de texto 1 e 2 da Figura 19 poderiam ser gerados. Entretanto, no plano 2, o *efeito* gera a *causa*, consistindo, portanto, em um plano incoerente.

1. Ontem choveu muito.
2. A grama do jardim da minha casa cresceu.

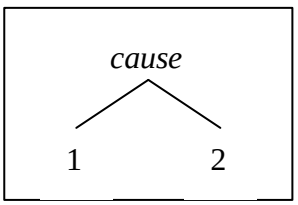


Figura 18 – Base de conhecimento com causa e efeito

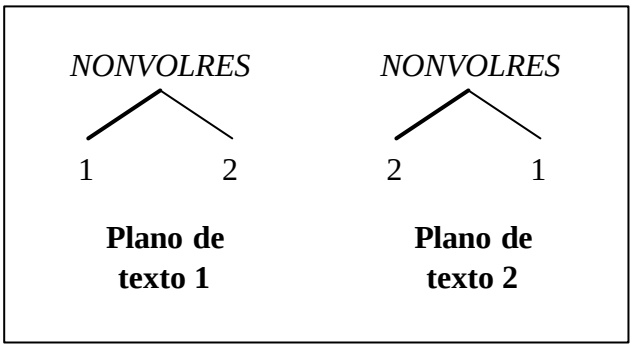


Figura 19 – Planos de texto com causa e efeito

Em suma, enquanto as relações semânticas da base de conhecimento determinam possíveis relações retóricas, as relações intencionais indicam se essas relações são viáveis, e, caso sejam, elas determinam o que será informação nuclear ou não. Portanto, conclui-se que o nível intencional é necessário para a estruturação coerente do discurso.

7.5. Resultados da Avaliação

A partir das avaliações apresentadas nesta seção, da *performance* do DMSumm, em função da avaliação dos sumários por ele produzidos, e do modelo de discurso subjacente à geração de sumários, conclui-se, principalmente, que o modelo de discurso utilizado é válido e viável computacionalmente para a estruturação coerente de discurso.

Com a avaliação, mostrou-se: a) a funcionalidade e necessidade da interação entre as relações profundas do modelo de discurso; b) a produção de sumários satisfatórios, mas que ainda podem ser melhorados pelo aprimoramento do sistema, principalmente no que diz respeito ao realizador lingüístico, e por meio da extensão do próprio modelo de discurso de Rino.

Este trabalho dá margem para uma vasta exploração nos campos da sumarização e da geração automática de textos. Pode-se delinear várias contribuições potenciais deste trabalho, assim como limitações. Essas características são discutidas a seguir.

8. Considerações Finais

O trabalho desta dissertação consistiu 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 profundas do discurso, isto é, relações semânticas, intencionais e retóricas. O objetivo principal consistiu na validação do modelo de discurso utilizado.

Por ser um gerador automático de sumários, os dados de entrada do DMSumm são adquiridos de forma manual a partir de um texto-fonte que se quer sumarizar. Neste trabalho, assume-se como premissa que é possível automatizar este processo. Além dessa, outras premissas, decorrentes do modelo de discurso, são estabelecidas: a) todo texto possui um objetivo comunicativo e uma proposição central associados a ele e b) 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.

O DMSumm e seu modelo discursivo subjacente foram avaliados, validando-se o modelo de discurso. Várias limitações, contribuições e trabalhos futuros foram identificados, conforme relatado nas próximas subseções.

8.1. Limitações deste Trabalho

Este trabalho apresenta algumas limitações deixadas aparentes durante o decorrer desta dissertação. No que diz respeito ao sistema computacional, tem-se:

- o DMSumm não é um sumarizador de fato, pois sumariza uma mensagem que *corresponde* à interpretação do texto;
- complementarmente, como o processo de interpretação não é automático, o usuário do DMSumm deve ser especialista para poder montar a mensagem-fonte;
- o realizador lingüístico do DMSumm é muito simples, fazendo uso de *templates* e produzindo somente *canned texts.*

Com relação ao modelo de discurso de Rino, tem-se a dependência do modelo Problema-Solução, a saber: a) o texto-fonte a ser sumarizado deve apresentar a estrutura

Problema-Solução e b) as relações intencionais se estabelecem entre as etiquetas das proposições elementares do texto-fonte, as quais se dão de acordo com suas funções na macroestrutura textual. Além disso, esse modelo requer conhecimento altamente especializado para a montagem da base de conhecimento, o que é custoso, quer em termos de projeto e desenvolvimento, quer em termos de produção de dados para o processamento computacional, já que não se tem o processo de interpretação automática.

As limitações referentes ao sistema computacional constituem alguns dos trabalhos futuros, conforme detalhado na subseção 8.3. A última limitação, referente à dependência do DMSumm do modelo Problema-Solução, tornar o sistema independente dessa restrição é desejável no futuro. Entretanto, caso se substitua ou altere esse modelo, o modelo intencional deverá ser alterado também, já que, na proposta atual, ele é essencial e dependente da forma como as informações do domínio se relacionam. Idealmente, busca-se um modelo intencional irrestrito, o que não é possível, dado o atual estado da arte da sumarização automática. No entanto, apesar dessa limitação, é importante ressaltar que o modelo Problema-Solução é suficientemente abrangente, aparecendo em textos de diferentes gêneros e domínios. Além disso, é muito conveniente, neste trabalho, definir a macroestrutura textual em termos desse modelo, já que algumas das heurísticas da seleção de conteúdo permitem a poda da estrutura do texto-fonte de maneira relativamente simples e direta por meio da exclusão dos componentes desse modelo.

Caso se opte por excluir o modelo Problema-Solução, é necessário definir que modelo o substituirá de forma adequada para que: a) as heurísticas da seleção de conteúdo ainda possam manipular a macroestrutura textual e b) seja possível determinar os segmentos de discurso que contribuem entre si e, conseqüentemente, que contribuem para o discurso como um todo, para a definição das intenções subjacentes ao discurso, de forma a assegurar as premissas e hipótese deste trabalho. Provavelmente, outro modelo igualmente ou mais complexo do que o proposto deva ser utilizado diante do nível de abstração considerado no processamento do discurso, o qual se aproxima mais do conhecimento extralingüístico inerente ao processo de comunicação do que do conhecimento lingüístico, o qual é mais claramente formalizado e explícito. Esta questão é uma questão em aberto na Lingüística Computacional de um modo geral: o nível de complexidade de sistemas automáticos, como o proposto aqui, aumenta à

medida que se buscam mecanismos para expressar e/ou manipular as características mais abstratas do processo de comunicação.

8.2. Contribuições deste Trabalho

Muitas são as contribuições deste trabalho para o estado da arte da sumarização e da geração automática de textos, destacando-se:

- implementação de um modelo de discurso inédito, para o qual não existem equivalentes na literatura;
- implementação de parte substancial de um sumarizador automático profundo, simbólico, que, embora limitado pela falta de um interpretador automático do texto-fonte a sumarizar e pela adoção de um realizador lingüístico bastante simplificado, contempla as fases de *transformação* e *síntese* de um sistema profundo de sumarização automática, nos moldes delineados por Sparck Jones (vide Figura 2, Seção 1) e pode ser completado com o acréscimo do módulo de interpretação e refinado com a troca do módulo de realização lingüística por outro mais elaborado, do ponto de vista lingüístico;
- exploração da geração dirigida por objetivos, delineando uma hierarquia de contribuição de objetivos a partir de um objetivo comunicativo inicial que permite gerar uma estrutura textual coerente.
- aprimoramento da rede sistêmica de objetivos comunicativos que permitirá a extensão do modelo no futuro, pela inclusão de novos objetivos, e a exploração exaustiva de estratégias comunicativas na estruturação discursiva;
- desenvolvimento de repositórios genéricos de operadores de plano e relações intencionais, nos moldes do modelo de discurso de Rino, que, a princípio, podem ser aplicados para a estruturação discursiva em domínio aberto;
- validação do modelo de Rino, a qual é evidenciada pela potencialidade do DMSumm na geração de sumários coerentes, que preservam as supostas proposições centrais das mensagens-fonte. Essa potencialidade é aparente nas avaliações apresentadas na seção anterior, muito embora os resultados indiquem um largo espaço para melhora, a qual é possível de obtenção, já que as causas da *performance* mediana são identificáveis e solucionáveis.

8.3. Trabalhos Futuros

O DMSumm pode ser estendido e aprimorado de diversas formas, dadas suas limitações e possibilidades não exploradas, como será delineado a seguir.

Inicialmente, como uma das limitações mais evidentes, o realizador lingüístico do DMSumm é extremamente simples, já que é baseado em *templates*. Um realizador lingüístico mais robusto deve ser capaz de lidar com dependências contextuais e com a coesão do sumário em uma representação ainda intermediária, não textual, como descrições funcionais ou mesmo lógica de predicados. Pode-se utilizar, por exemplo, uma gramática de unificação funcional (FUG – *Functional Unification Grammar* – Elhadad, 1993). O custo de se desenvolver um realizador lingüístico para o português é alto, mas pode trazer grandes benefícios, visto que não se tem nada feito em relação a isso para o português do Brasil. Além disso, como o modelo de discurso do DMSumm é independente de língua, pode-se produzir (ou mesmo utilizar) realizadores lingüísticos para outras línguas naturais, o que constitui um fator importante no contexto multilingual atual.

A interpretação automática de textos-fonte, que visa produzir a mensagem-fonte para o DMSumm, constitui um dos trabalhos futuros principais e também um dos principais gargalos do sistema atual. Já foi proposto um trabalho de doutorado para investigar como ela pode ser efetuada, em continuidade a este trabalho de mestrado. Uma das possíveis abordagens, neste caso, pode ser a compilação de dados a partir de corpora do português, buscando-se identificar marcadores discursivos e estruturas textuais[7] que caracterizem organizações textuais específicas. Com base em trabalhos que procuraram explicitar a relação entre retórica, intenções e semântica (por exemplo, o próprio modelo de discurso de Rino (1996a), Marcu (2000b), Moser e Moore (1996), Moore e Pollack (1992)), pode-se desenvolver, ainda, formas de correspondência e/ou extração das relações profundas do discurso de um texto, no estágio de análise de sistema.

Com relação ao modelo de discurso do DMSumm, este pode, ainda, ser expandido por meio da incorporação de outros objetivos comunicativos, podendo-se, assim, estender também a rede sistêmica de objetivos para maior abrangência a textos de outros gêneros. A extensão do modelo de discurso pode ser feita com base no estudo detalhado de outros corpora, identificando-se características/objetivos diferentes dos

abordados pela modelagem atual. Na literatura, há várias propostas de clusterização de objetivos comunicativos segundo gêneros textuais. Ao acrescentar outros objetivos comunicativos, a tendência será, assim, aumentar o potencial do sistema para tratar textos de gênero flexível e para investigar, com o auxílio do sistema, como tais objetivos podem se relacionar na construção do enredo de segmentos discursivos.

Outro aprimoramento do DMSumm envolve a incorporação de um modelo de usuário, de modo que possíveis sumários sejam dedicados a interesses específicos, por exemplo, casos em que o usuário é especialista no assunto do texto a sumarizar e deseja um sumário sem qualquer tipo de detalhe. Muitos trabalhos já exploraram modelagem de usuário, destacando-se, principalmente, o trabalho de Paris (1993), especificamente voltado para a geração textual. Isto também já foi previsto por Rino em seu modelo original. Como exemplo, Rino diz que as restrições sobre o perfil do usuário podem constituir condições adicionais para se definir a contribuição entre segmentos discursivos. Na especificação do sistema, essas restrições seriam incluídas no campo *constraints* do operador de plano responsável por fazer a ponte com o modelo de usuário.

A incorporação ao sistema de mecanismos mais eficientes para avaliação é considerada uma extensão importante. Por exemplo, interfaces para usuários *juízes* de diferentes graus de conhecimento – desde o juiz leigo na Lingüística Computacional, mas falante da língua natural em foco, até o juiz especialista, a exemplo do que foi feito na avaliação descrita na Seção 7. A interface permitiria, por exemplo, que todos os dados de teste fossem disponibilizados na tela, em tempo real de processamento, e que o juiz procedesse à sua avaliação, registrando no sistema suas respostas. Dependendo do tipo de experimento, tais respostas poderiam ser computadas automaticamente, consistindo em uma grande contribuição para um módulo de avaliação sistemática nos moldes necessários ao DMSumm ou ao sistema de sumarização completo, no futuro.

O DMSumm também pode ser utilizado em aplicações distintas da sumarização automática, por exemplo, no contexto de diálogos, como a proposta original de Moore e Paris (1993), cuja estrutura discursiva é utilizada para manter a conversação com o usuário, permitindo que o sistema tenha controle sobre o que dizer e que possa se recuperar em condições em que o usuário não é capaz de entender algo (condições em que houve falhas na comunicação). Ainda nessa linha, o DMSumm, ou alguns de seus

---

[7] Essa etapa remete ao trabalho de Winter (1977) sobre o uso de vocabulários distintos para o reconhecimento de padrões textuais.

recursos (como o repositório de operadores de plano), poderia ser acoplado a um ambiente de auxílio à escrita.

Como trabalhos a serem realizados em curto prazo, destacam-se a verificação da aplicabilidade do sistema a outros corpora, o aprimoramento do sistema para suportar o modelo do usuário e a incorporação ao sistema de mecanismos mais eficientes de avaliação. Em médio prazo, estima-se a possibilidade de trocar ou aprimorar o realizador lingüístico do DMSumm e expandir o modelo de discurso de Rino para outros objetivos comunicativos. Em longo prazo, visualiza-se a disponibilidade de um interpretador automático para o português do Brasil e a utilização dos recursos do DMSumm em outros contextos, como o contexto de diálogos e sistemas de auxílio à escrita.

O potencial para a realização desses projetos depende diretamente do conhecimento e capacidade de uma equipe e dos recursos disponíveis para o processamento de línguas. O NILC[8] – Núcleo Interinstitucional de Lingüística Computacional – formado por pesquisadores da Universidade de São Paulo/São Carlos, da Universidade Federal de São Carlos e da UNESP/Araraquara, é um centro totalmente capacitado nesses sentidos. Dentre os projetos do NILC, relacionam-se a este trabalho e aos trabalhos futuros sugeridos nesta subseção, respectivamente, o projeto EXPLOSA (EXPLOração de métodos diversos para a Sumarização Automática) e o AMADEUS, o qual é um ambiente de auxílio a escrita. O EXPLOSA, especificamente, abrange vários trabalhos de sumarização, incluindo este. Esses trabalhos dizem respeito à sumarização extrativa (da abordagem superficial), sumarização baseada em interlíngua e avaliação de sumarizadores.

[8] http://www.nilc.icmc.sc.usp.br

Apêndice A – Modelo Problema-Solução

Neste apêndice são descritas as etiquetas do modelo Problema-Solução. Também são mostrados exemplos de estruturas Problema-Solução de textos autênticos em português (retirados do *Thesis Corpus*) (Feltrim et al., 2001).

A.1. Etiquetas do Modelo Problema-Solução

As unidades informativas de estruturas Problema-Solução são etiquetadas de acordo com sua função na macro-estrutura do discurso. As possíveis etiquetas são descritas a seguir:

- *sit*: expressa a situação, conhecimento de mundo, para algum evento;
- *aspectofsit*: expressa algum aspecto da situação, detalhando-a;
- *probl*: expressa o problema sobre o qual se fala;
- *sol*: expressa a possível solução para o problema;
- *res*: expressa os resultados provenientes da resolução de um problema ou da ocorrência de algum evento;
- *evalres*: expressa a avaliação dos resultados apresentados;
- *newprobl*: expressa um novo problema a partir da resolução do problema em foco;
- *neweval*: expressa uma nova avaliação ou avaliação suplementar;
- *needsol*: expressa algo que precisa ser solucionado;
- *proc*: expressa um processo em andamento;
- *explan*: expressa uma explicação sobre algum evento ou processo apresentado;
- *action*: expressa uma ação em curso;
- *function*: expressa a função de algo ou alguém;
- *poseval*: expressa uma avaliação positiva de um evento ou resultado;
- *negeval*: expressa uma avaliação negativa de um evento ou resultado;
- *evalnonprobl*: expressa uma avaliação de algo que ocorre independentemente do problema apresentado;
- *prop*: expressa uma proposição que não pode ser caracterizada pelas etiquetas anteriores.

A.2. Extração da Estrutura Informativa de Textos-Fonte

TEXTO 1

[1] Durante todo o ciclo de vida do software, são produzidos muitos itens de informação. [2] Todos esses itens, relativos ao software que está sendo construído ou modificado, constituem a "Configuração de Software". [3] A configuração de software é a base para o entendimento de software desenvolvido e modificado por outras pessoas, e portanto deve estar sempre atualizada. [4] Para isso, os itens de informação precisam ser "controlados" durante todo o ciclo de vida do software.

[5] O "Gerenciamento de Configuração de Software" é uma disciplina que ajuda a solucionar esse problema, pois envolve o desenvolvimento de padrões e procedimentos para administrar um software evolutivo. [6] No entanto, devido à complexidade de suas tarefas, o gerenciamento de configuração de software necessita de ferramentas automatizadas que possam torná-las um pouco mais simples. [7] O RCS (Revision Control System) é uma dessas ferramentas.

[8] O RCS é uma ferramenta de domínio público, desenvolvida para ambientes de Workstations e proposição central. [9] O RCS para UNIX foi alvo de um estudo detalhado e constatou-se que o RCS é um sistema difícil de ser usado, principalmente devido à sua interface, constituída apenas de comandos de linha. [10] Os usuários, principalmente aqueles não familiarizados com o sistema, encontram extrema dificuldade em utilizar e explorar toda a sua funcionalidade, pois apesar de haver poucos comandos, cada comando possui várias opções associadas, que também podem ser combinadas para obter resultados diferentes. [11] Outro problema do sistema é que os resultados das operações ficam "soltos" na tela, dificultando o seu entendimento.

[12] Assim o objetivo deste artigo é apresentar uma interface amigável para o sistema RCS, seguindo as noções básicas de Projeto de Interface. [13] Possibilitando que o sistema seja utilizado com mais freqüência e interesse, através de uma interface mais "simpática" e fácil de utilizar, e além disso divulgar os importantes conceitos de gerenciamento de configuração de software, os quais se tornarão mais "visíveis" através da interface amigável.

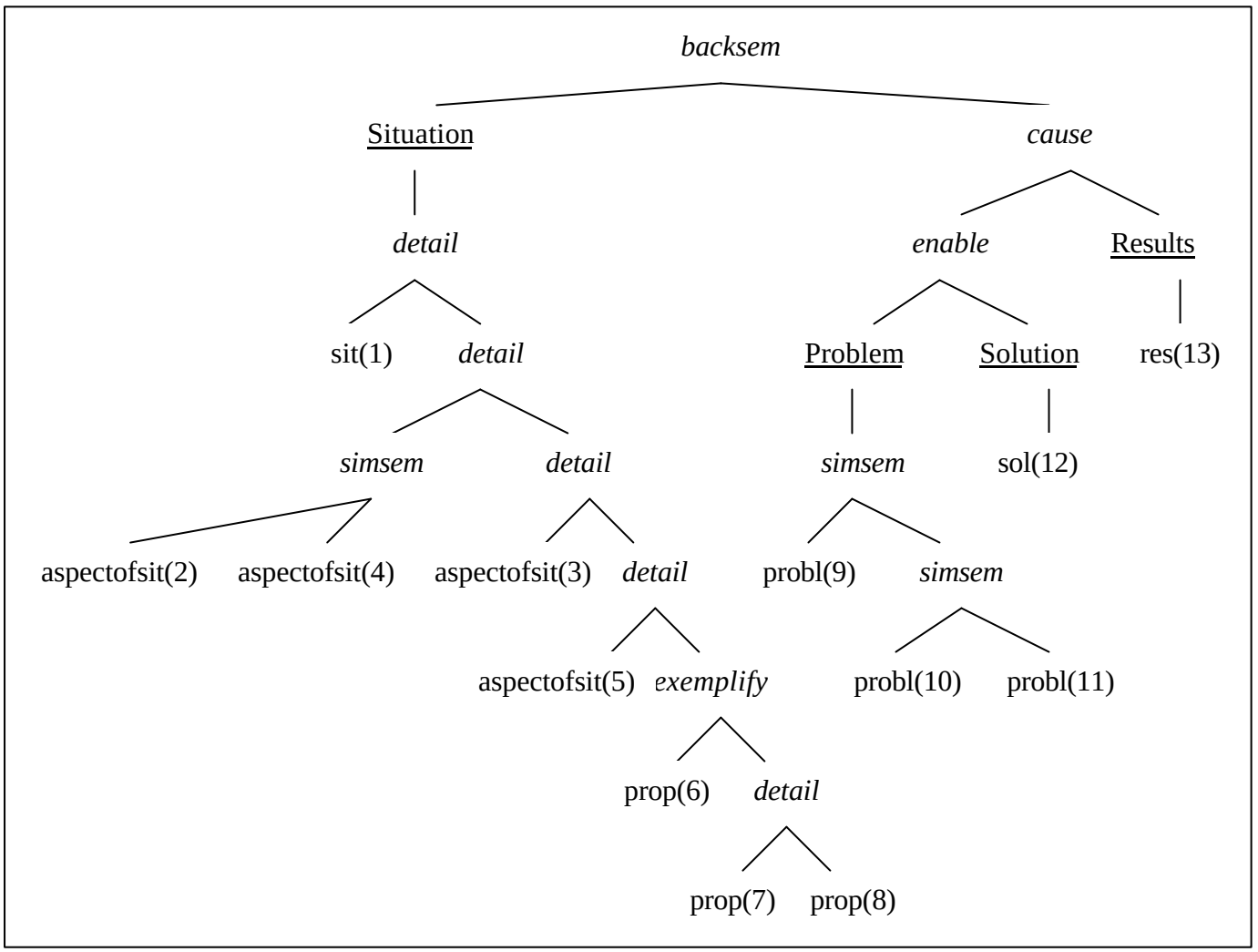


Figura 20 – Base de conhecimento do Texto 1

TEXTO 2

[1] A representação de grandes dicionários de língua natural, principalmente nos casos em que se trabalha com vários milhões (ou dezenas de milhões) de palavras, é um interessante problema computacional a ser tratado dentro da área de Processamento de Língua Natural (PLN). [2a] Autômatos finitos (AFs), largamente usados na construção de compiladores, são excelentes estruturas para representação desses dicionários, [2b] permitindo acesso direto aos às palavras e seus possíveis atributos (gênero, número, grau, etc).

[3] Um dicionário contendo mais de 430.000 palavras da língua portuguesa sem atributos, cuja representação em formato texto ocupa mais de 4.5Mb, pode ser convertido em um AF compactado de apenas 218Kb, conforme será apresentado (o processo de conversão para esse caso levou 4:17Seg, rodando em um Pentium166/Linux).

[4a] Esse trabalho tem por objetivo investigar métodos eficientes de construção, representação e minimização de AFs de grande porte (dezenas de milhões de estados), [4b] a fim de propor um conjunto de estruturas e algoritmos para criação de um sistema eficiente e versátil de representação de grandes léxicos de língua natural.

[5a] Outro objetivo é o de investigar métodos de representação para os atributos das entradas léxicas, [5b] de forma a permitir a criação de um formato otimizado capaz de ser inserido diretamente no AF, com o menor impacto possível no tamanho final desse último.

[6] Será também desenvolvido um protótipo do sistema, para efeito de testes e confirmação de eficiência do sistema.

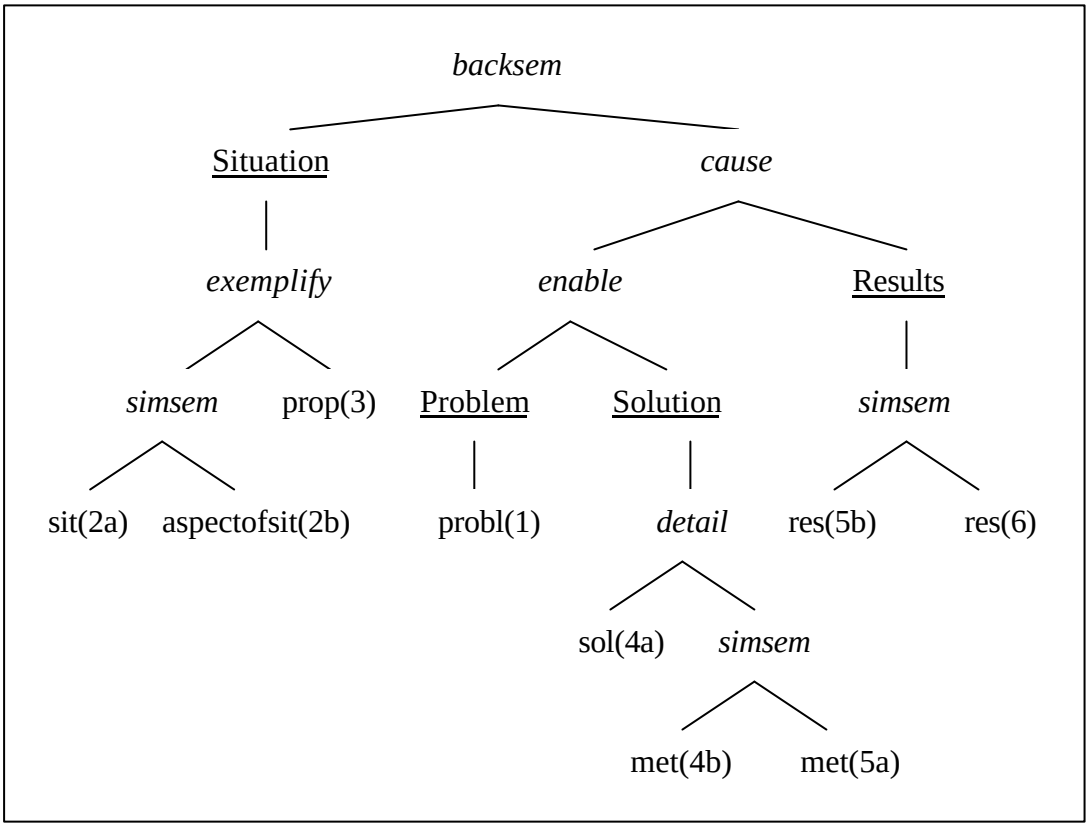


Figura 21 – Base de conhecimento do Texto 2

Apêndice B – Relações Semânticas

As relações semânticas utilizadas neste trabalho são mostradas e exemplificadas a seguir com trechos de textos autênticos em português (retirados do *Thesis Corpus*) (Feltrim et al., 2001), onde X e Y indicam segmentos discursivos.

ENABLE(X,Y): X fornece as condições para a ocorrência de um evento Y, mas não gera diretamente o evento. No exemplo abaixo, a ocorrência de [1] permite a realização de [2], isto é, a existência do problema permite a busca de uma solução.

[1] *Os métodos de solução de sistemas lineares envolvem cálculos matriciais que são computacionalmente intensivos.* [2] *Para solucionar esse problema, uma arquitetura sistólica é implementada.*

enable

[1] [2]

RATIONALE(X,Y): Y expressa o meio pelo qual se pode alcançar X. No exemplo, através da abordagem de avaliação sugerida em [2], pode-se realizar [1].

[1] *A avaliação de desempenho de um sistema existente é importante desde que auxilie na determinação de quão bem o sistema está, assim como quais melhoramentos são necessários para aumentar seu desempenho.* [2] *As abordagens mais comuns para se avaliar esse desempenho são a modelagem analítica e a simulação.*

rationale

[1] [2]

PROOF(X,Y): Y expressa a prova de uma asserção X. No exemplo, a asserção [1] é evidenciada pela [2].

[1] *A evolução das metodologias de projeto de hardware abriu um novo horizonte para a computação de finalidade geral e o hardware dedicado.* [2] *Hoje é possível desenvolver rapidamente um projeto de sistema digital.*

proof

[1] [2]

CAUSE(X,Y): expressa uma causa X e um efeito Y. No exemplo a seguir, a fato em [1] causa [2].

[1] *Os sistemas computacionais e as redes de computadores têm apresentado uma rápida evolução, se tornando cada vez mais rápidos e mais complexos.* [2] *Devido a esses avanços, tornou-se essencial o desenvolvimento de um estudo sobre o desempenho de redes.*

cause

[1] [2]

SIMSEM(X,Y): expressa fatos (X e Y) que possuem similaridade semântica. No exemplo, tanto [1] quanto [2] descrevem problemas do sistema RCS para UNIX.

[1] *O RCS para UNIX é um sistema difícil de ser usado, principalmente devido à sua interface, constituída apenas de comandos de linha.* [2] *Os usuários do RCS encontram extrema dificuldade em utilizar e explorar toda a sua funcionalidade, pois apesar de haver poucos comandos, cada comando possui várias opções associadas, que também podem ser combinadas para obter resultados diferentes.*

simsem

[1] [2]

DIFSEM(X,Y): expressa fatos (X e Y) opostos semanticamente. No exemplo, o segmento [1] contrasta com o [2].

[1] *Linguagens de programação de alto nível permitem ao programador uma maior naturalidade na forma de programar.* [2] *Entretanto, essas linguagens são mais lentas que linguagens de baixo nível durante sua execução.*

difsem

[1] [2]

ATTRIBUTE(X,Y), DETAIL(X,Y) e EXEMPLIFY(X,Y): Y fornece, respectivamente, atributos (características), detalhes e exemplos de X. No exemplo, [2] detalha [1].

[1] *Dentre as diversas áreas que estão em expansão atualmente, podemosdestacar a Internet.* [2] *A Internet é o meio digital através do qual pode-se fazer qualquer tipo de serviço.*

detail

[1] [2]

EVALSEM(X,Y): Y fornece a avaliação de X. No exemplo, [2] avalia [1].

[1] *Os circuitos FPGAs são os responsáveis pela evolução na área de hardware nos últimos tempos.* [2] *Esses circuitos mostraram-se de grande utilidade, produzindo resultados melhores do que o esperado, apesar de ainda serem considerados caros para aquisição e uso em larga escala.*

evalsem

[1] [2]

REASON(X,Y): Y fornece a razão para um evento X. No exemplo, [2] explica o porquê do fato em [1].

[1] *Na busca de alto desempenho, a abordagem mais adotada é a utilização de técnicas de programação concorrente para paralelização de algoritmos a serem executados em máquinas paralelas de propósito geral.* [2] *A construção de arquiteturas dedicadas tem sido uma abordagem pouco explorada, devido ao elevado custo de projeto e tempo de desenvolvimento.*

reason

[1] [2]

SEQUENCE(X,Y): expressa eventos ou fatos (X e Y) que acontecem em seqüência. No exemplo, o evento em [1] deve ser seguido pelo evento em [2].

[1] *Para melhorar a qualidade, é necessário, primeiro, conhecer a situação real em que se encontra a organização em matéria de produção de software, isto é, fazer um diagnóstico do processo de software corrente.* [2] *A partir dos resultados da avaliação, elabora-se um plano de ação que leve à implementação de melhorias.*

sequence

[1] [2]

BACKSEM(X,Y): Y fornece uma introdução (cenário contextual) para X. O fato apresentado em [1] fornece o cenário para [2].

[1] *Ao longo dos anos, pesquisadores nas áreas de Computação Gráfica e Visualização Científica vêm desenvolvendo técnicas computacionais para transformar dados em imagens.* [2] *Esses dados precisam ser manipulados por alguma ferramenta para a geração de gráficos.*

backsem

[1] [2]

Apêndice C – Relações Retóricas

As relações retóricas do modelo de Rino (e que são utilizadas no DMSumm) são definidas abaixo, onde N indica o núcleo da relação, S o satélite, E o escritor e L o leitor.

PURPOSE

Restrições sobre N	N apresenta uma atividade
Restrições sobre S	S apresenta uma situação não realizada ou resolvida
Restrições sobre N+S	A atividade em N permite realizar/resolver a situação em S
Efeito	L reconhece que a atividade em N permite realizar a situação em S

MEANS

Restrições sobre N	N apresenta uma situação não realizada ou resolvida
Restrições sobre S	S apresenta uma atividade
Restrições sobre N+S	A atividade em S permite realizar/resolver a situação em N
Efeito	L reconhece que a atividade em S permite realizar a situação em N

JUSTIFY

Restrições sobre N	Nenhuma
Restrições sobre S	Nenhuma
Restrições sobre N+S	O E acredita que a compreensão de S permite aumentar a disposição de L para aceitar a asserção apresentada em N
Efeito	A disposição de L para aceitar a asserção em N é aumentada

EVIDENCE

Restrições sobre N	L pode não estar suficientemente convicto da asserção apresentada em N
Restrições sobre S	E acredita que L acredita ou pode facilmente acreditar em S
Restrições sobre N+S	A compreensão de S aumenta a crença de L sobre N
Efeito	L aumenta sua crença na asserção apresentada em N

NONVOLRES

Restrições sobre N	Nenhuma
Restrições sobre S	Apresenta uma situação não volitiva
Restrições sobre N+S	N apresenta uma situação que causa a situação apresentada em S
Efeito	L reconhece que a situação em N pode causar a situação em S

NONVOLCAUSE

Restrições sobre N	Apresenta uma situação não volitiva
Restrições sobre S	Nenhuma
Restrições sobre N+S	S apresenta uma situação que causa a situação apresentada em N
Efeito	L reconhece que a situação em S pode causar a situação em N

LIST

Restrições sobre N	Relação multinuclear; todos os núcleos são igualmente relevantes
Restrições sobre S	Sem satélite
Efeito	L reconhece que o grupo de proposições em lista compartilha um conjunto de características

CONTRAST

Restrições sobre N	Relação multinuclear; todos os núcleos são igualmente relevantes
Restrições sobre S	Sem satélite
Efeito	L reconhece que as asserções de cada núcleo estão em relação de conflito

ELABORATE

Restrições sobre N	Nenhuma
Restrições sobre S	Nenhuma
Restrições sobre N+S	S apresenta detalhes adicionais sobre a situação ou algum elemento apresentado em N
Efeito	L reconhece que S fornece detalhes adicionais sobre N

EVALUATE

Restrições sobre N	Nenhuma
Restrições sobre S	Nenhuma
Restrições sobre N+S	S apresenta uma asserção que expressa a inclinação do E em relação à situação apresentada em N
Efeito	L reconhece que a asserção apresentada em S avalia a situação apresentada em N

EXPLAIN

Restrições sobre N	Nenhuma
Restrições sobre S	Nenhuma
Restrições sobre N+S	S apresenta uma razão para a interpretação da situação apresentada em N
Efeito	L reconhece que a asserção apresentada em S é uma razão para a existência ou ocorrência da situação apresentada em N

SEQUENCE

Restrições sobre N	Relação multinuclear; todos os núcleos são igualmente relevantes
Restrições sobre S	Sem satélite
Efeito	L reconhece a sucessão entre os núcleos

BACKGROUND

Restrições sobre N	Nenhuma
Restrições sobre S	Nenhuma
Restrições sobre N+S	S aumenta a capacidade do L de compreender a situação/asserção apresentada em N
Efeito	A capacidade de L para compreender N é aumentada

Apêndice D – Definição dos Operadores de Plano

Com base na Tabela 1 (subseção 5.2.1.1), os operadores de plano podem ser definidos conforme um dos padrões abaixo. Essas definições são feitas da mesma forma que em Moore e Paris (1993).

1. Operadores *inform*: são operadores terminais que inserem uma proposição do discurso no plano de texto.

Name	descrever-informando
Header	descrever(X)
Effect	know-about(reader,X)
Constraints	Nil
Nucleus	**inform(reader,X)**
Satellite	Nil

2. Operadores objetivo comunicativo-objetivo comunicativo: são operadores que alteram o objetivo comunicativo (objetivo comunicativo) em expansão, indicando que há outros segmentos discursivos, com outros objetivos comunicativos associados, que devem contribuir para o contexto atual.

Name	descrever-relatando
Header	**descrever(X)**
Effect	know-about(reader,X)
Constraints	Nil
Nucleus	**relatar(X)**
Satellite	Nil

3. Operadores objetivo comunicativo-EM: são operadores cujo núcleo deriva um objetivo comunicativo e o satélite um estado mental (EM).

Name	relatar-por-know
Header	relatar(X)
Effect	know(reader,X)
Constraints	not know(reader, X), isa(X,Y,enable), sp(Y,X), dom(X,Y)
Nucleus	**relatar(X)**
Satellite	**know(reader,purpose1(X,Y))**

4. Operadores objetivo comunicativo-*cohererel*: são operadores cujo núcleo deriva um objetivo comunicativo e, ao mesmo tempo, inserem no plano de texto uma relação retórica (indicada pelo predicado *cohererel*).

Name	identificar-por-elaborate
Header	know(reader,elaborate(X,Y))
Effect	know(reader,elaborate(X,Y))
Constraints	not know(reader,elaborate(X,Y)), isa(X,Y,detail), sup(Y,X)
Nucleus	**relatar(Y)**, **cohererel(elaborate(X,Y))**
Satellite	nil

Referências Bibliográficas

Baxendale, P.B. (1958). Machine-made index for technical literature – an experiment. *IBM Journal of Research and Development*, Vol. 2, pp. 354-365.

Black, W.J. and Johnson, F.C. (1988). A Practical Evaluation of Two Rule-Based Automatic Abstraction Techniques. *Expert Systems for Information Management*, Vol. 1, No. 3. Department of Computation. University of Manchester Institute of Science and Technology.

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.

Cawsey, A. (1993). Planning Interactive Explanations. *Int. Journal of Man-Machine Studies*, Vol. 38, pp. 169-199.

Edmundson, H. P. (1969). New Methods in automatic extracting. *Journal of the ACM*, Vol. 16, pp. 264-285.

Elhadad, M. (1993). *Using argumentation to control lexical choice: a unification-based implementation*. PhD. thesis, Computer Science Department, Columbia University.

Feltrim, V.D.; Nunes, M.G.V.; Aluísio, S.M. (2001). *Um corpus de textos científicos em Português para a análise da Estrutura Esquemática*. Série de Relatórios do NILC. NILC-TR-01-4.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, Vol. 32, pp. 221-233.

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

Hobbs, J.R. (1985). *On the Coherence and Structure of Discourse*. Technical Report CSLI-85-37, Center for Study of Language and Information, Stanford University.

Hoey, M. (1983). *On the Surface of Discourse*. George Allen & Unwin Ltd.

Hovy, E. (1988). *Generating Natural Language under Pragmatic Constraints.* Lawrence Erlbaum Associates Publishers, Hillsdale, New Jersey.

Hutchins, J. (1977). On the Structure of Scientific Discourse. In *UEA Papers in Linguistics*, Vol. 5, pp. 18-39. University of East Anglia.

Hutchins, J. (1987). Summarization: Some problems and Methods. In Jones (ed.), *Meaning: The frontier of informatics*. Cambridge. London, pp. 151-173.

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.

Jordan, M,P. (2000). A Pragmatic/Structural Approach to Relevance. *Technostyle*, Vol. 16, No. 2, pp. 47-67.

Knight, K. and Marcu, D. (2000). Statistics-Based Summarization – Step One: Sentence Compression. In *The 17th National Conference of the American Association for Artificial Intelligence*, Austin, Texas.

Kupiec, J.; Pedersen, J.; Chen, F. (1995). A trainable document summarizer. *ACM SIGIR*, pp. 68-73.

Larocca Neto, J.; Santos, A.D.; Kaestner, A.A.; Freitas, A.A. (2000). Generating Text Summaries through the Relative Importance of Topics. In the *Proceedings of the International Joint Conference IBERAMIA/SBIA*, Atibaia, SP.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development*, Vol. 2, pp. 159-165.

Maier, E. and Hovy, E. H. (1991). A Metafunctionally Motivated Taxonomy for Discourse Structure Relations. In the *Proceedings of the 3rd European Workshop on Language Generation*. Innsbruck, Austria.

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

Mani, I. and Maybury, M. T. (1999). *Advances in automatic text summarization*. MIT Press, Cambridge, MA.

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

Marcu, D. (1996). Building Up Rhetorical Structure Trees. In *The Proceedings of the Thirteenth National Conference on Artificial Intelligence*, Vol. 2, pp. 1069-1074. Portland, Oregon.

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

Marcu, D. (1998a). Improving Summarization through Rhetorical Parsing Tuning. In the *Proceedings of The Sixth Workshop on Very Large Corpora*, pp. 206-215. Montreal, Canada.

Marcu, D. (1998b). To build text summaries of high quality, nuclearity is not sufficient. In the *Working Notes of the AAAI-98 Spring Symposium on Intelligent Text Summarization*, pp. 1-8. Stanford, California.

Marcu, D. (2000a). *The Theory and Practice of Discourse Parsing and Summarization*. The MIT Press. Cambridge, Massachusetts.

Marcu, D. (2000b). Extending a Formal and Computational Model of Rhetorical Structure Theory with Intentional Structures à la Grosz and Sidner. In the *Proceedings of the 18th International Conference on Computational Linguistics (COLING'2000)*, Saarbrueken.

Martins, T.B.F.; Ghiraldelo, C.M.; Nunes, M.G.V.; Oliveira Jr., O.N. (1996). *Readability Formulas Applied to Textbooks in Brazilian Portuguese*. Notas do ICMSC-USP, Série Computação.

Martins, C.B.; Pardo, T.A.S.; Espina, A.P.; Rino, L.H.M. (2001). *Introdução à Sumarização Automática*. Relatório Técnico RT-DC 002/2001, Departamento de Computação, Universidade Federal de São Carlos.

Maybury, M.T. (1992). Communicative Acts for Explanation Generation. *Int. Journal of Man-Machine Studies 37*, pp. 135-172.

McKeown, K.R. (1985). *Text Generation: Using Discourse Strategies and Focus Constraints to Generate Natural Language Text*. Cambridge University Press.

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.

Moore, J. D. and Pollack, M. E. (1992). A problem for RST: the need for multi-level discourse analysis. *Computational Linguistics*, Vol. 18, No. 4, pp. 537-544.

Moore, J.D. (1995). *Participating in Explanatory Dialogs: Interpreting and Responding to Questions in Context*. The MIT Press. Cambridge, Massachusetts.

Moser, M. and Moore, J. D. (1996). Toward a synthesis of two accounts of discourse structure. *Computational Linguistics*, Vol. 22, No. 3, pp. 409-419.

Mushakoji, S. (1993). Constructing ‘Identity’ and ‘Differences’ in Original Scientific Texts and Their Summaries: Its Problems and Solutions. In the *Proceedings of Summarizing Text for Intelligent Communication*, B. Endres-Niggemeyer, J. Hobbs and K. Sparck Jones (eds). Dagstuhl, Germany.

O'Donnell, M. (1997). Variable-Length On-Line Document Generation. In the *Proceedings of the 6th European Workshop on Natural Language Generation*, Gerhard-Mercator University, Duisburg, Germany.

Paice, C. D. (1981). The automatic generation of literature abstracts: an approach based on the identification of self-indicating phrases. *Information Retrieval Research*. Butterworth & Co. (Publishers).

Pardo, T.A.S. (2002). *Descrição do DMSumm: um Sumarizador Automático Baseado em um Modelo Discursivo.* Relatório Técnico NILC-TR-02-02, São Carlos.

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.

Paris, C.L. (1993). *User Modelling in Text Generation*. St. Martin’s Press, New York.

Reiter, E. and Dale, R. (2000). *Building Natural Language Generation Systems.* Cambridge, University Press.

Rino, L.H.M. (1996a). *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. (1996b). A sumarização automática de textos em português. In *Anais do II Encontro para o Processamento Computacional de Português Escrito e Falado*, pp. 109-119. Curitiba - PR.

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.

Sampson, G. (1987). Probabilistic models of analysis. In G. Leech, R. Garsude and G. Sampson (eds.), *The computational analysis of English*, pp. 16-29. Longman. Harrow.

Sparck Jones, K. (1993). *Discourse Modelling for Automatic Summarising*. Technical Report No. 290. University of Cambridge.

Sparck Jones, K. (1999). Automatic Summarizing: factors and directions. In I. Mani and M. Maybury (eds.), *Advances in automatic text summarization*, pp. 1-12, The MIT Press.

Sparck Jones, K. and Galliers, J. R. (1996). Evaluating Natural Language Processing Systems. *Lecture Notes in Artificial Intelligence*, Vol. 1083.

Teufel, S. and Moens, M. (1999). Argumentative classification of extracted sentences as a first step towards flexible abstracting. In I. Mani and M. Maybury (eds.), *Advances in automatic text summarization*, MIT Press.

Van Dijk, T.A. (1979). Recalling And Summarizing Complex Discourse. In Burghart, W. and Hölker, K., *Text Processing Textverarbeitung*. Berlin, Walter de Gruyter.

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.

Winter, E.O. (1979). Replacement as a Fundamental Function of the Sentence in Context. In *Forum Linguitics*, Vol. 4, No. 2, pp. 95-133.