Universidade de São Paulo - USP
Universidade Federal de São Carlos - UFSCar
Universidade Estadual Paulista - UNESP

Sobre Geração e Sumarização de Textos

Lucia H. Machado Rino
Maria das Graças V. Nunes

NILC-TR-05-13

NOTAS DIDÁTICAS DO ICMC-USP (No. 67)

Outubro 2005

Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional
NILC - ICMC-USP, Caixa Postal 668, 13560-970 São Carlos, SP, Brasil

Sobre geração de textos e sumários[1]

Lucia Helena Machado Rino[2] e Maria das Graças Volpe Nunes[3]

Núcleo Interinstitucional de Lingüística Computacional (NILC)
Caixa Postal 668 – 13560-970 – São Carlos – SP – Brazil
lucia@dc.ufscar.br, gracan@icmc.usp.br

Resumo. *Este texto apresenta as características da geração e sumarização automáticas de textos. Inicialmente, são discutidas as características dessas tarefas quando executadas por humanos. Em seguida, é apresentada a arquitetura básica de um sistema de geração de textos e seus componentes são definidos. Finalmente, são apresentados e discutidos alguns métodos extrativos (estatísticos) e profundos (lingüisticamente motivados) de sumarização automática.*

1. Introdução

A geração, ou produção, de textos, de um modo geral, é uma atividade comum na vida de qualquer pessoa de nível de escolaridade médio ou superior. Textos são, se não um objeto principal de trabalho, um instrumento auxiliar para atualização ou comunicação em qualquer esfera profissional ou social. Com o crescente uso da Internet, os textos condensados, aqui chamados de *sumários*, tiveram sua função – *transmitir ou comunicar o que é importante* – evidenciada, devido à incapacidade de as pessoas digerirem o grande volume de informações disponível em sua íntegra. Assim, os sumários, hoje, existem também como objetos autônomos de comunicação, em relação a suas correspondentes fontes, no sentido de que eles próprios servem aos objetivos do leitor, o qual não busca a informação completa, mas assume a forma condensada como suficiente para se informar e se manter atualizado. Essa perspectiva é comum na maioria dos processos de transmissão de informação atuais: temos sumários nos telejornais, nos jornais, nas revistas e na própria Internet. Na esfera acadêmica, elaboramos sumários em nossas salas de aula e em nossos escritos; utilizamo-los em nossas buscas bibliográficas e em nossos pareceres sobre trabalhos de nossos colegas e alunos. Podemos dizer que, inconscientemente, estamos sempre sumarizando, quer oral, quer textualmente.

---

[1] Este texto foi produzido em 2002

[2] Centro de Ciências Exatas e de Tecnologia, Departamento de Computação, Universidade Federal de São Carlos (UFSCar).

[3] Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo (USP).

De modo geral, a sumarização, como um processo de geração textual, importa da geração suas principais características. Entretanto, é mais específica do que a geração, por envolver restrições particulares. Neste capítulo, apresentaremos os principais aspectos de ambos os processos, primeiramente baseados na geração e sumarização humanas, para, depois, apresentar as abordagens automáticas. Tratando-se da geração textual, a perspectiva adotada aqui é a do *produtor* do texto, i.e., do *escritor*. Como veremos, essa perspectiva é importante, pois é ela quem direcionará a modelagem automática, para a especificação dos principais processos de tomada de decisão.

A sumarização automática começou a ser explorada no final da década de 50, quando se utilizavam, sobretudo, técnicas estatísticas de extração de conhecimento lingüístico dos textos-fonte [Luhn, 1958; Edmundson, 1969]. Entretanto, devido aos resultados insatisfatórios e à impossibilidade técnica de aprimorá-los (por limitações de *hardware* e *software*, mas também de conhecimento específico para a modelagem dos processos correspondentes), a área ficou estagnada até a década de 1980, quando os computadores passaram a ser de uso geral, suas memórias baratearam e recursos lingüísticos expressivos se tornaram disponíveis para o processamento textual. Sobretudo com as idéias de Chomsky (1965), esses recursos começaram a ser explorados para a modelagem computacional.

O interesse pela geração textual, por sua vez, ocorreu um pouco mais tarde: a produção automática de textos completos se concentrou em áreas muito específicas, tais como a geração de textos a partir de cenários retóricos particulares [Hovy, 1988] ou aquela visando, muitas vezes, usuários com perfis particulares [McKeown, 1985; Paris, 1993]. Em sua grande maioria, até muito recentemente tais trabalhos se desenvolveram, sobretudo, de forma acadêmica. Isso ainda ocorre com a geração textual, mas já não é mais o caso da sumarização automática: é possível, hoje, utilizar ferramentas de sumarização automática quer na Internet (AltaVista, etc.), quer em ambientes de edição de textos (p.ex., o AutoResumo do MS Word ™).

Em relação à terminologia, é comum nos referirmos à geração ou sumarização textuais automáticas simplesmente como *Geração de Textos* e *Sumarização Automática*, respectivamente. Salvo quando estivermos falando do processo humano, essa terminologia será a adotada.

O capítulo está organizado da seguinte maneira: na Seção 2, apresentamos as principais características e problemas da geração textual e sumarização humanas. Passamos, então, ao contexto automático, agora voltado à Geração de Textos (Seção 3) e Sumarização Automática (Seção 4): apresentamos suas principais características, distinguindo, no segundo caso, duas metodologias: a empírica e a fundamental, a primeira restrita à sumarização automática extrativa. Finalmente, tecemos algumas considerações sobre ambos os processos automáticos, remetendo à literatura da área (Seção 5). Além das referências feitas no texto, no final do capítulo o leitor encontrará uma bibliografia complementar, classificada por alguns tópicos relevantes da área.

2. A geração e sumarização textuais: principais características

Nesta seção, introduzimos alguns conceitos fundamentais da tarefa *humana* de produção textual, a partir dos quais elaboraremos nossa perspectiva de Geração de Textos e Sumarização Automática.

2.1. O discurso

Destacamos aqui alguns aspectos que nós, humanos, mesmo inconscientemente, levamos em consideração ao elaborar um texto. Para a tarefa automática correspondente, tais fatores também são relevantes. Entretanto, veremos que algumas alternativas são exploradas, buscando minimizar a complexidade de modelá-los, mas cuidando para manter o potencial comunicativo e a qualidade lingüística. Essas propriedades para a produção do discurso são descritas por alguns conceitos elementares, conforme relacionamos a seguir.

2.2. Conceitos elementares do discurso

Os conceitos aqui descritos buscam elucidar as decisões pertinentes à modelagem automática na Geração de Textos e na Sumarização Automática.

- *Comunidade de discurso:* grupo de usuários ou participantes no contexto em que a comunicação se dá, envolvendo, na produção textual, o escritor e o leitor.
- *Discurso:* unidade de informação usada na comunicação verbal, envolvendo aspectos lingüísticos e sócio-culturais dos significados e valores da comunidade de discurso [Grosz and Sidner, 1986; Kress, 1989]; composição de segmentos informacionais representada por uma coleção estruturada de orações [Hovy, 1993].
- *Domínio do discurso:* assunto sobre o qual o discurso versa.
- *Objetivo comunicativo:* intenção fundamental do discurso.
- *Gênero discursivo:* conjunto de características que regem a produção do discurso, envolvendo o propósito comunicativo e os interesses da comunidade de discurso, os quais delineiam a estrutura esquemática do discurso, influenciando e restringindo as escolhas de conteúdo e estilo textual. São exemplos de gênero: poesia, prosa, ficção, escrita científica, etc.
- *Estrutura profunda do discurso:* representação estruturada e proposicional do discurso, que considera os conceitos referentes ao assunto, ao gênero, aos objetivos comunicativos e à comunidade de discurso.
- *Estrutura superficial do discurso:* manifestação lingüística da estrutura profunda.
- *Tipo textual:* modo como o discurso é apresentado na forma escrita. Textos persuasivos, expositivos, descritivos, etc., são exemplos textuais típicos.
- *Tópico do discurso:* aquilo sobre o que se fala, entidade sobre a qual o escritor pretende elaborar sua mensagem.
- *Proposição central do discurso:* núcleo ou hipótese geral inicial da unidade textual, sobre a qual se tece o encadeamento lógico proposicional do discurso.
- *Idéia central do texto:* expressão lingüística da proposição central do discurso.
- *Coerência:* propriedade lógica do discurso, que garante sua *trama* e resulta na transmissão do significado ao leitor.

- *Coesão:* propriedade de se unir componentes textuais, pelo uso de termos relacionados semanticamente, referências, elipses e conjunções; expressão superficial da coerência.
- *Textualidade:* propriedade de o texto ser coerente e coeso.

2.3. A comunicação

Consideramos dois extremos: aquele em que a comunicação, mesmo que falha do ponto de vista lingüístico, é efetiva do ponto de vista comunicativo e aquele que resulta em textos primorosos do ponto de vista lingüístico, mas cujo poder comunicativo é baixo. Assim, um objeto comunicativo, para nós, é um texto cujo poder comunicativo se encontra em algum ponto entre tais extremos. Entretanto, a determinação desse ponto não segue métricas objetivas e tampouco encontramos, na literatura, sugestões claras para determinar os fatores que interferem na sua qualidade. Para a modelagem automática, isso se torna um problema. É comum, assim, limitar a geração textual a, p.ex., um domínio (p.ex., futebol) ou gênero (jornalístico), os quais, por sua vez, delineiam uma comunidade de discurso específica. De um modo geral, cada fator limitante implica restrições em todos os níveis da modelagem e em todos os processos correspondentes.

2.4. As escolhas discursivas

Em sua tarefa de escrita, o escritor "constrói" sua perspectiva comunicativa elegendo recursos discursivos e lingüísticos que não transgridam a textualidade, i.e., que lhe permitam achar o *ponto médio* entre coerência e coesão e, portanto, transmitir sua mensagem. Assim, um texto satisfatório é aquele que permite ao leitor recuperar, no mínimo, sua *idéia central*, percebendo as intenções do escritor. Permeando essa perspectiva, há a flexibilidade com que o escritor organiza suas idéias e utiliza os recursos da língua natural para expressá-las, fazendo-o buscar um inter-relacionamento particular entre as unidades informativas selecionadas. Veremos que essa dinâmica entre diversos fatores que, inter-relacionados, interferem na qualidade de textos humanos ou automáticos, também é importante na geração de resumos.

2.5. A geração de resumos

A sumarização textual, como tarefa de produção de textos, obedece, de um modo geral, às mesmas características delineadas acima para a produção de um discurso. Entretanto, ela se distingue por uma restrição fundamental: a de *transmitir a mensagem essencial, de forma concisa*. Atreladas a essa restrição estão várias características a serem observadas: a existência de um *texto-fonte*, a partir do qual se extrai o sumário, a distinção entre informações *supérfluas*, *complementares* e *essenciais*, devendo as supérfluas ser eliminadas no sumário e as complementares, utilizadas para garantir a coerência na transmissão das essenciais. Novamente, é importante ressaltar a complexidade e diversidade características dessa tarefa: distinguir esses *graus de importância* das informações depende dos interesses do escritor, para a transmissão de sua mensagem ao leitor. Neste caso, é importante, ainda, distinguir *quem é* nosso escritor: devido à natureza da tarefa, temos, basicamente, dois tipos de escritor: o sumarizador profissional, i.e., aquele que toma um texto-fonte produzido por outra pessoa e produz o sumário, muitas vezes sem nem mesmo ter conhecimento do domínio, e o próprio autor do texto-fonte, conhecedor do assunto e *experienciador* ou

vivenciador do mesmo contexto discursivo. Essa distinção entre os perfis dos sumarizadores humanos levou à convenção das tarefas hoje chamadas *professional* e *author summarizing*, cujas particularidades interferem significativamente na modelagem computacional, como veremos adiante.

Ainda podemos distinguir vários tipos de sumários: resenhas de notícias jornalísticas, sinopses do movimento da bolsa de valores, sumários de textos novelísticos, extratos de livros científicos, resumos de previsões meteorológicas, etc. Cada um desses tipos envolve pressuposições e características diversas, assim como conteúdos e correspondência com suas fontes de teores variados. Por exemplo, um determinado autor de um sumário jornalístico pode considerar que um título espetacular para um texto que descreve um acidente automobilístico envolvendo uma personalidade conhecida deva ser "Acidente mata Fulano de Tal". Para o mesmo evento, outro autor pode priorizar o desfecho trágico do evento, ignorando a vítima. Neste caso, seu sumário pode ser "Acidente termina em tragédia".

Variações dessa natureza são comuns na produção humana de sumários. Podemos argumentar que, no primeiro exemplo, o autor pressupôs que, ao mencionar a vítima, sua mensagem surtiria um efeito espetacular sobre os leitores, compelindo-os a ler o texto correspondente. No segundo exemplo, podemos supor que o autor ignora a vítima proposital ou casualmente, ou por resolver não chocar explicitamente seus leitores ou por desconhecer a proeminência da vítima. Vale notar, no entanto, que ambas as formas podem ser associadas – e serão – ao mesmo texto descritivo do acidente envolvendo 'Fulano de Tal'. Ambos os exemplos de sumários, neste caso, apresentam ainda uma característica bastante peculiar: a de se considerar que títulos – e, portanto, formas não textuais – podem também sumarizar seus correspondentes textos.

3. A geração automática

Ao contrário da interpretação, a geração de textos consiste em produzir textos mono ou multi-sentenciais, em língua natural, a partir de um conjunto de elementos de conteúdo e de objetivos de comunicação. Em muitas aplicações de Processamento de Língua Natural, no entanto, a geração de língua natural é feita de uma maneira bastante simplificada, em que os textos são construídos pela justaposição de partes (ou segmentos) textuais pré-determinadas (e, neste caso, já definidas durante a fase de projeto do sistema). Outras vezes, esquemas de texto, conhecidos como *canned texts*, são "preenchidos" de forma a compor o texto final. Neste caso, os esquemas são também pré-definidos, mas possuem uma parte variável que somente pode ser determinada em tempo de processamento. Apesar das limitações inerentes a essas técnicas, para muitas aplicações elas se mostram bastante satisfatórias. É o caso, p.ex., de respostas a consultas a bases de dados, que são geralmente simples e, portanto, não exigem um processamento mais sofisticado (e caro!).

Da mesma forma, a gramática de geração, ao contrário da de interpretação, procede a partir das funções dos elementos conceituais do texto para produzir a estrutura textual e, portanto, seus elementos lingüísticos. Assim, decisões sobre o vocabulário, os constituintes sintáticos e a própria forma da sentença são de responsabilidade do gerador, que deve procurar atingir os objetivos de comunicação desejados. Dependendo do objetivo comunicativo, freqüentemente podemos decidir por um conjunto finito e, muitas vezes, simples, de alternativas de regras gramaticais para expressar o conteúdo informacional.

Eventualmente, se a aplicação permitir, podemos adotar apenas um padrão sintático. Por exemplo, se a aplicação envolver apenas proposições declarativas, o sistema pode produzir apenas sentenças na voz ativa ou na voz passiva. De modo similar, podemos escolher o padrão de cada um dos componentes sentenciais. Neste caso, o processo se torna dependente das especificações de entrada que, fornecidas adequadamente, fazem com que a tarefa de geração seja mais simples do que a de interpretação. Este é o caso, p.ex., de sistemas que têm a função exclusiva de transmitir informações constantes em uma base de dados (sistema de consultas a bases de dados com interface em língua natural, como já exemplificado anteriormente), isto é, de sistemas cuja função comunicativa principal seja a declarativa ou informativa. Entretanto, para outras aplicações, a delimitação do poder de geração visando à simplificação do sistema pode prejudicar o resultado, dado que a Geração de Textos não implica somente a manipulação do conteúdo informacional, mas também a manipulação dos aspectos comunicativos, segundo as intenções do escritor. Este é o caso, p.ex., da tradução automática, que exige a correspondência mais fiel possível entre o texto-fonte e o texto-alvo, ou da sumarização automática, cuja dependência do texto-fonte já foi delineada acima e será mais bem explorada na próxima seção. Para casos dessa natureza, temos, portanto, um grau de complexidade igualável, se compararmos um sistema de geração com um sistema de interpretação.

Classicamente, a Geração de Textos pode ser considerada como um processo de três passos. A Fig. 1 mostra esquematicamente os componentes desse gerador típico [Matthiessen and Bateman, 1991], cujos processos são descritos a seguir.

Seleção do conteúdo: esse processo tem a função de selecionar os itens de conhecimento que deverão fazer parte do texto. Por exemplo, numa aplicação de geração de respostas em língua natural a consultas a uma base de dados, isso equivale a extrair, do registro selecionado da base, os itens de dados que comporão a resposta (p.ex., o nome, a idade e o RG de um funcionário). O Modelo do Usuário pode determinar a quantidade de informação necessária na resposta. Diz-se, então, que essa fase determina, num processo comunicativo, *o que dizer*.

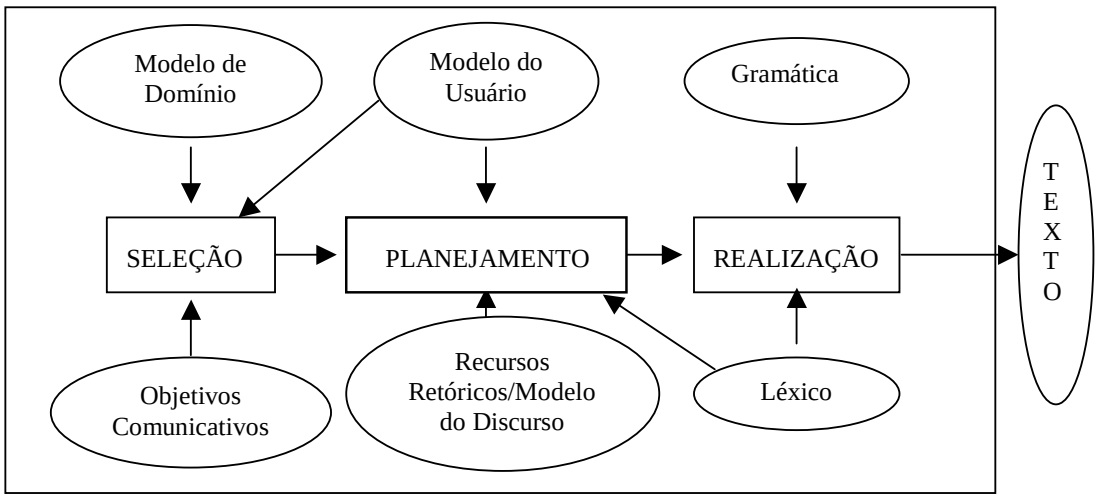


Figura 1. Fases principais de um gerador de textos

Planejamento do texto: esse componente, também chamado de *componente estratégico*, é responsável por planejar a comunicação. É nessa fase que se decide *quando dizer* o que foi selecionado na fase anterior. A entrada para o planejador pode ser bastante variada e depende da aplicação que contém o gerador. Por exemplo, o conteúdo informacional pode estar representado em forma de registros de uma base de dados, de uma tabela de registros, de proposições lógicas, etc. Nessa fase, o conteúdo deve ser organizado para uma melhor apresentação textual. Isso implica a adição de especificações retóricas na determinação do inter-relacionamento entre as informações, de sua seqüência e de sua expressão na língua natural escolhida. O planejador produz uma forma intermediária do texto, chamada de *plano do texto*. Os formalismos de representação do plano do texto diferem entre si quanto ao aspecto do plano que privilegiam. Alguns privilegiam a estrutura retórica do texto, como a *Rhetorical Structure Theory*, RST [Mann and Thompson, 1988]. Outros privilegiam os aspectos pragmáticos, como o gerador Pauline [Hovy, 1988].

Realização do texto: esse processo, também chamado de *componente tático*, componente lingüístico ou gerador de superfície, é responsável pela *realização* gramatical do plano de texto produzido pelo planejador. Nesta fase, podemos ter dois subprocessos distintos: a *determinação* dos itens lingüísticos, propriamente dita, e sua *linearização*, i.e., a "planificação" da estrutura textual pela seqüencialização de tais itens na forma textual, produzindo um encadeamento de sentenças válidas na língua em foco. As contribuições desse componente para o processo de geração envolvem as seguintes decisões lingüísticas e decisões sobre o conhecimento do domínio e do discurso: escolha de vocabulário; escolha do estilo do texto (p.ex., prosa, diálogo, etc.); escolhas léxicas, morfológicas e sintáticas adequadas para expressar conteúdo e estrutura textual; escolha de figuras de discurso para manifestar apropriadamente as intenções do escritor (questão de foco, ênfase, etc.); escolhas que garantam a coesão do discurso, i.e., a fluidez do texto (p.ex., o uso de marcadores de seqüencialização das informações); escolhas que garantam a coerência do discurso, i.e., que expressem o inter-relacionamento retórico/semântico desejado (p.ex., o uso de uma marca de contraste entre componentes textuais que devem ser contrastados); e, finalmente, decisões de linearização, p.ex., ordenação das informações, concordância gramatical, etc.

Normalmente, a distinção entre as fases de planejamento e realização é vantajosa porque provê pelo menos dois níveis de abstração, de modo que detalhes que são relevantes para o realizador possam ser ignorados pelo planejador. Por exemplo, a decisão sobre qual determinante usar no contexto de uma proposição não é uma consideração apropriada no momento em que se escolhe uma estratégia para convencer o leitor dessa proposição. Mais ainda, se a interface entre esses dois níveis for cuidadosamente especificada, parece possível construir um componente estratégico geral que possa ser usado para uma grande variedade de aplicações, mesmo que o componente tático seja variável (p.ex., quando se deseja obter um sistema de geração multilingual).

Podemos considerar três modos distintos de interação entre os processos de um gerador automático: (a) o *seqüencial* (ou *geração em pipeline*)**,** em que os três processos ilustrados são estritamente seqüenciais e, portanto, a atuação de cada módulo não interfere na do outro; (b) o *intercalado* (ou *interleaved generation*), em que os módulos executam suas funções de modo intercalado, intercomunicando-se entre si à medida que cada processo necessita tomar decisões que envolvem outras esferas de conhecimento (p.ex.,

decisões sintáticas dependentes do conhecimento do usuário) - neste caso, a intercomunicação ocorre por demanda, i.e., somente quando um módulo acusa a necessidade de outras informações que não são de sua responsabilidade e (c) o *combinado* (ou *merged generation*), em que os processos executam todas as tarefas sem que seja possível distingui-las ou modularizá-las.

4. A sumarização automática

As possibilidades de sumarização automática delineadas na Seção 2 indicam um grande problema: a busca de um modelo de sumarização automática adequado, para que os resultados automáticos reflitam a diversidade de sumários sem que estes percam sua interdependência com os textos-fonte correspondentes. É importante notar que duas características são essenciais nesse contexto: a) sumários remetem, necessariamente, a textos originários dos mesmos; b) sumários devem ser construídos de modo a não haver perda considerável do significado original, apesar de conterem menos informações e poderem apresentar diferentes estruturas, em relação a suas fontes. Assim, sumários são textos produzidos a partir de textos, podendo servir, principalmente, de indexadores ou substitutos dos mesmos. Essa distinção levou à sua classificação como *indicativos* ou *informativos*, respectivamente.

As principais diferenças entre sumários indicativos e informativos são as seguintes: os indicativos não podem substituir os textos-fonte, pois não necessariamente preservam o que aqueles têm de mais importante, em termos de conteúdo e estrutura. Ao contrário, transmitem somente uma vaga idéia do texto original, podendo, inclusive, apresentar uma forma não textual (uma lista de itens, por exemplo). Os informativos, ao contrário, podem substituir os textos-fonte, pois contêm todos os seus aspectos principais. Neste caso, eles são chamados autocontidos, i.e., eles dispensam a leitura do texto-fonte. Outra diferença reside na avaliação de sua funcionalidade e qualidade. Sumários indicativos podem ser utilizados na classificação de documentos bibliográficos, de um modo geral, indicando seu conteúdo e agilizando o acesso a suas informações relevantes. Por meio deles, o leitor de uma enciclopédia, p.ex., pode ir direto ao volume ou página que trata do assunto que lhe interessa. Dessa forma, a utilidade desses sumários é mais clara e sua função mais limitada do que as dos sumários informativos, permitindo uma avaliação mais robusta de sua funcionalidade e qualidade. Por outro lado, os sumários informativos apresentam uma relação mais complicada com seu texto-fonte. Do seu objetivo dependerá bastante a avaliação sobre o quanto ele atende às necessidades do usuário. De um modo geral, é menos complexo produzir automaticamente os indicativos do que os informativos. Entretanto, ambos podem servir a diversas aplicações, ressaltando-se, especialmente, a área de Recuperação de Informação, muito importante nos dias de hoje.

Além da diferença funcional, os sumários também são comumente classificados pelo modo como são obtidos. Sparck Jones (1993a) classifica-os como *extracts* (aqui chamados de *extratos*) ou *abstracts* (os nossos *sumários*, propriamente ditos), fazendo a correspondência com o que ela chama, respectivamente, de *extração textual* e *condensação de conteúdo* [Sparck Jones, 1997]. A extração textual consiste em se produzir textos pela seleção e simples justaposição de segmentos textuais inteiros (em geral, correspondendo a sentenças). Essa reprodução pode ser totalmente fiel ao texto-fonte ou sofrer somente pequenas modificações, motivo pelo qual um extrato *consiste inteiramente de material copiado do texto-fonte*. O problema, neste caso, é identificar, no texto-fonte, os segmentos

que sejam relevantes para transmitir sua idéia principal. Métodos estatísticos são explorados com esse fim, dando origem à *Sumarização Automática Extrativa*. A condensação de conteúdo, por sua vez, requer a identificação do que é relevante e sua subseqüente reestruturação, em uma tarefa de *reescrita* e *fusão* de vários conceitos do texto-fonte em um número menor de conceitos [Hovy and Lin, 1997]. Um sumário pode conter, assim, *algum material que não se encontra no texto-fonte.* Neste caso, é necessário lançar mão de métodos simbólicos e modelos computacionais de Geração de Textos bastante complexos, que refletem substancialmente o processamento lingüístico humano e a maior parte dos problemas da própria Geração de Textos. Esses métodos remetem à *Sumarização Automática Fundamental*. As seções seguintes apresentam ambas as abordagens, com especial ênfase em alguns métodos particulares de sumarização automática.

4.1. A sumarização automática extrativa

Métodos extrativos de sumarização automática consistem em a) identificar segmentos textuais relevantes para compor um sumário; b) extrair do texto-fonte as unidades mínimas de significado que incluam tais segmentos; c) justapor cada uma dessas unidades, resultando no sumário final. Em geral, os segmentos textuais relevantes são indicados por palavras ou elementos-chave mais complexos, cuja identificação é feita pela análise da distribuição de freqüência das palavras no texto-fonte, pela análise da própria palavra, quando esta carregar algum significado – caso especial das palavras de classe aberta, i.e., substantivos, verbos, advérbios e adjetivos – ou mesmo pela sua localização no texto.

A seguir são apresentadas algumas técnicas de sumarização extrativa, bem como estudos de casos para o português do Brasil: as técnicas baseadas em palavras-chave e características textuais superficiais e as baseadas em Mineração de Textos (*Text Mining*).

4.1.1. Técnicas baseadas em palavras e outros elementos-chave

Nesta seção, daremos exemplos de alguns métodos ou sistemas associados à abordagem superficial ou extrativa, sem sermos exaustivos. O leitor deve consultar a bibliografia sugerida para leitura complementar.

4.1.1.1. Método das palavras-chave

Este método pressupõe que as idéias principais de um texto possam ser expressas por algumas palavras-chave, que aparecem com maior freqüência no texto conforme as idéias vão sendo desenvolvidas [Luhn, 1958]. A idéia é, então, determinar a distribuição estatística das palavras-chave do texto e, a partir de sua freqüência, extrair as sentenças que as contenham, agrupando-as de forma a constituir um sumário, na ordem em que aparecem originalmente.

Podem-se privilegiar, como palavras-chave, os nomes (substantivos e adjetivos). Por exemplo, na sentença “A casa desmoronou”, a palavra “casa” seria mais representativa que o verbo “desmoronou”, porque indicaria o tópico sobre o qual se fala. Caso o autor decidisse mudar o foco para o evento propriamente dito, introduzido nesse exemplo pelo verbo “desmoronar”, seria necessário nominalizá-lo e, portanto, o substantivo continuaria sendo a forma-chave mais representativa do conteúdo textual. Assim, uma possível

progressão temática mudando o foco de “casa” para “desmoronar” seria construída pela sentença “O desmoronamento ocorreu devido aos ventos fortes”.

4.1.1.2. Método das palavras-chave do título

Este método é uma variação do anterior, em que somente o título do texto a ser sumarizado é considerado para se buscar as palavras-chave que irão nortear a sumarização. Caso o texto contenha subtítulo, também este poderia ser considerado para a extração das palavras-chave. Para aplicá-lo, parte-se do pressuposto de que o título do texto é bem formado e que as palavras componentes de um título têm maior probabilidade de serem representativas do tópico principal do texto correspondente [Edmundson, 1969].

Certamente é possível combinar este método com o anterior, simplesmente acrescentando as palavras do título ao conjunto de palavras-chave do texto.

4.1.1.3. Método da localização

A hipótese deste método é a de que a posição de uma sentença em um texto pode estar associada à sua importância no contexto. Por exemplo, a primeira e a última sentenças de um parágrafo podem conter suas idéias principais e, portanto, estas seriam consideradas para a produção de um sumário. Já em 1958, Baxendale mostrou que, em 85% de uma amostra de 200 parágrafos, a sentença tópica era a primeira e, em 7%, a última. Nos 8% restantes, as informações relevantes encontravam-se entre a primeira e a última sentenças de um mesmo parágrafo. Apesar do resultado aparentemente pouco significativo de 7% associado à última sentença, o autor considerou que esta também é importante para o sumário, pois constitui o elo de ligação com o parágrafo seguinte e, portanto, com a provável sentença tópica deste parágrafo. Assim, para que a coesão textual fosse mantida, Baxendale sugeriu que fossem incluídas em um sumário tanto a primeira quanto a última sentença de cada parágrafo.

4.1.1.4. Método das palavras sinalizadoras (*cue phrases*)

Similarmente aos dois métodos anteriores, este método também consiste em atribuir valores a sentenças de um texto, selecionando as de maior peso. O método faz uso de um dicionário, previamente construído, composto por palavras consideradas relevantes no domínio do texto. Por sinalizarem a importância dos segmentos textuais em que se encontram, tais palavras foram denominadas *cue phrases* [Paice, 1981]. Nesse método, pesos negativos são atribuídos a sentenças cujas palavras não constem do dicionário; pesos positivos são atribuídos, caso constem. Assim, o peso de uma sentença passa a ser a média ponderada entre valores negativos e positivos, conforme a especificação dicionarizada. Em um texto científico, por exemplo, as palavras *conclusões* e *resultados* serão, muito provavelmente, altamente significativas, estando presentes no dicionário. Sua ocorrência em alguma sentença implicará, conseqüentemente, o aumento da relevância da mesma. Textos de gêneros distintos teriam, correspondentemente, outros dicionários e, portanto, seus próprios marcadores da importância do conteúdo textual. Cada gênero, neste caso, delineia uma distribuição de marcadores distinta, a partir da qual a modelagem computacional pode resultar razoavelmente satisfatória para a sumarização automática.

Vale notar que, muito embora este método se assemelhe ao método das palavras-chave, o dicionário não é composto por palavras-chave ocorrentes no texto-fonte, mas por

palavras consideradas significativas como marcas da relevância de outros componentes textuais.

4.1.1.5. Método relacional

Também chamado de método adaptativo, o método descrito por Skorochodko (1971) é também considerado um método baseado em conhecimento. Ele utiliza uma representação gráfica do texto, na qual relações entre sentenças são criadas dependendo da relação semântica entre suas palavras. Sentenças semanticamente relacionadas a um grande número de outras sentenças, cuja supressão levaria a uma maior dificuldade no entendimento do texto, recebem um peso alto e, portanto, são mais prováveis de serem escolhidas para a composição do sumário. A relação semântica entre as sentenças é identificada superficialmente, por meio, p.ex., da ocorrência de substantivos comuns. Outras heurísticas de natureza similar também são definidas para se identificar as sentenças a serem selecionadas para compor o sumário.

4.1.1.6. Método da frase auto-indicativa

Este método faz uso de um indicativo explícito de quão significativa é a sentença a ser selecionada para compor o sumário. Uma frase auto-indicativa apresenta uma estrutura cuja ocorrência é freqüente no texto e indica explicitamente que a sentença se refere a algo importante sobre o assunto do texto. Exemplos de frases auto-indicativas são: *O objetivo deste artigo é investigar…*; *Neste artigo, é descrito um método para…*, etc. De acordo com esse método, as frases auto-indicativas permitiriam somente a produção de sumários indicativos, em oposição aos já descritos sumários informativos.

4.1.1.7. Método da idéia principal

Este método se baseia na premissa de que a idéia principal dos textos pode ajudar a construir extratos coerentes (Pardo et al., 2003a). Inicialmente, determina-se a sentença do texto que melhor expressa sua idéia principal (chamada *gist sentence*), o que é feito pela atribuição de uma pontuação a cada sentença, que corresponde à soma da freqüência de suas palavras, e posterior escolha da sentença com maior pontuação. A seguir, selecionam-se as sentenças do texto que complementam a informação da *gist sentence*, para, juntas, comporem o extrato final. Assume-se, neste método, que uma sentença complementa a *gist sentence* se elas possum palavras em comum.

Com esse método, sumários informativos são construídos. Vale notar que são possíveis diversas variações na forma de determinação da *gist sentence* e de suas sentenças complementares, podendo-se incorporar outros métodos extrativos, inclusive.

Um dos maiores problemas na aplicação dos métodos extrativos descritos é a produção de textos desconexos e/ou incoerentes, devido à justaposição de sentenças extraídas de seu contexto original. A resolução anafórica, p.ex., fica completamente comprometida se não forem consideradas todas as sentenças envolvendo referentes e referenciados. Para evitar ocorrências desse tipo, algumas providências podem ser úteis, como a seleção de sentenças anteriores ou vizinhas àquelas onde ocorrem os pronomes anafóricos. Porém, elas não são efetivas sempre, remetendo à necessidade de se considerar métodos mais informados para evitar a falta de coesão textual.

Experimentos de sumarização extrativa utilizando técnicas baseadas em palavras-chave, realizados com 18 textos científicos em português, extraídos de revistas da área de Computação, evidenciaram, entre outras coisas, a relação direta entre a eficácia do algoritmo de extração de palavras-chave e a eficácia do método de sumarização automática: as heurísticas envolvendo localização e palavras sinalizadoras mostraram-se muito pouco informativas [Pereira et al., 2002]. O estudo mostra também que, para a língua portuguesa, problemas de textualidade e proximidade com os textos-fonte são muito freqüentes nos extratos [Souza e Nunes, 2001].

4.1.2. Técnicas baseadas em Mineração de Textos

Mineração de Textos, ou *Text Data Mining* [Hearst, 1999] é um campo de pesquisa relativamente novo, cujas técnicas consistem em determinar relações existentes entre componentes textuais. Originalmente proposta para aplicações de Recuperação de Informação, a técnica de Text Mining explorada aqui é a estendida para a Sumarização Automática por Larocca Neto et al. (2000a), para a extração das sentenças mais relevantes de um texto.

Trabalha-se aqui com a noção da freqüência inversa de termos sentenciais (que, em geral, são palavras) e, logo, com a medida chamada TF-ISF (*Term Frequency – Inverse Sentence Frequency*), dada pela fórmula:

$$\text{TF-ISF}(p,s) = \text{TF}(p,s)*\text{ISF}(p)$$

sendo TF(p,s) o número de vezes em que a palavra p ocorre na sentença s, e ISF(p) a freqüência inversa da sentença, obtida pela fórmula

$$\text{ISF}(p) = \log(|S|/\text{SF}(p))$$

onde |S| é o número total de sentenças do texto e SF(p) é o número de sentenças em que a palavra p ocorre.

TF-ISF é similar à medida TF-IDF (*Term Frequency – Inverse Document Frequency*) de Salton e McGill (1983), para a Recuperação de Informação, sendo a noção de *documento* substituída pela de *sentença*. A idéia, aqui, é a de que uma palavra freqüente em um texto-fonte somente será significativa para representá-lo quando ela ocorrer em poucas sentenças, ou seja, quanto mais ela aparecer em sentenças diversas, mais sua importância será "diluída" no texto. Assim, para compor um sumário são selecionadas as sentenças com maior peso médio, relativo a TF-ISF, havendo um valor mínimo (*threshold*), definido pelo usuário, para o menor peso médio a considerar. Desse modo, esta proposta acrescenta outro grau de informatividade ao método simples das palavras-chave, fazendo com que, quanto maior a medida TF-ISF de uma palavra, mais representativas sejam as sentenças que a contêm. O método se torna, então, mais refinado que o método da palavra-chave, tendendo a produzir melhores resultados quanto à representatividade dos segmentos extraídos para compor os extratos. É por essa razão que Larocca Neto *et al* o exploram na sumarização automática de textos jornalísticos em inglês, com resultados preliminares bastante interessantes. Mais tarde, eles ainda o refinam [Larocca Neto et al., 2000b; Larocca Neto, 2002], considerando a segmentação do documento (i.e., a identificação de blocos de texto inter-relacionados semanticamente no TF) em tópicos, para determinar, agora, as sentenças para compor um extrato em função da importância relativa dos tópicos. A determinação dessas sentenças é baseada também na medida TF-ISF, agora em função

dos blocos textuais que, ao mesmo tempo em que indicam os tópicos mais relevantes, contêm as próprias sentenças.

Devido à perspectiva promissora de refinamento dos métodos extrativos iniciais, passamos a investigar este método também para sumarizar textos em português. Construímos, para tanto, um protótipo denominado *Text Mining Summarizer*, ou TMSumm [Martins et al., 2001; Espina e Rino, 2002]. Ilustramos, a seguir, a sumarização automática de um texto esportivo no TMSumm.

O texto de entrada é pré-processado de forma a uniformizar os caracteres, transformando-os em maiúsculos ou minúsculos (*case folding*), reduzir as palavras a seus radicais (*stemming*), e eliminar todas as palavras de classe fechada (*stopwords*), ou seja, pronomes, conjunções, artigos, preposições. O passo seguinte consiste no cálculo da freqüência dos componentes textuais, resultando no vetor V1, cujas coordenadas são dadas por pares no formato [palavra, (TF, SF)[4]]. V1 armazena, assim, os valores individuais de cada palavra na sentença. Para a sentença-exemplo [S1] "Diamantino arremessou a bola.", esses passos resultariam em S1':

[S1'] Diamantino arremess bola

com V1dado por:

V1= [[Diamantino, (1,8)], [arremess (1,1)], [bola, (1,3)]]

A seguir, constrói-se o vetor de pesos TF-ISF de cada palavra da sentença, V1', ilustrado abaixo com valores fictícios:

V1'= [[Diamantino, 0.60], [arremes, 0.35], [bola, 0.40]]

Calcula-se, em seguida, o peso médio de cada uma das sentenças, com base na freqüência inversa de seus componentes, para se determinar sua freqüência relativa no texto. Assim, o peso médio de S1, no exemplo, é dado por:

$$\text{Média (S1)} = \frac{\text{peso [Diamantino]} + \text{peso [arremes]} + \text{peso [bola]}}{\text{nº total de palavras da sentença}}$$

Para os valores ilustrados em V1', esse valor seria 0.45.

Para selecionar as sentenças a compor o extrato, bastaria agora calcular o limitante inferior de seus pesos médios: todas as sentenças que possuem peso médio maior ou igual ao limitante são selecionadas. Em geral, esse valor é calculado com base no fator de compressão do texto-fonte, que pode ser arbitrado pelo usuário. Seja V ($0 \leq V \leq 1$) esse fator de compressão e seja MAX ($0 \leq MAX \leq 1$) o maior peso médio de todas as sentenças do texto. Então o limitante inferior é dado por:

$$\text{Limitante} = V \times MAX$$

[4] *Term Frequency – Sentence Frequency*.

Comparando TMSumm com o AutoResumo do MS Word, para os textos de gênero esportivo selecionado, seu desempenho ultrapassou o do AutoResumo, quando consideradas taxas de compressão iguais [Espina e Rino, 2002].

Técnicas extrativas ainda mais robustas do que as apresentadas acima têm sido exploradas, em geral tendo como marco o trabalho de Kupiec et al. (1995), relativo a um método de extração estatístico e treinável, baseado em corpora de textos. Eles tratam a sumarização automática como um problema de classificação estatística: o objetivo é criar uma função que calcule a probabilidade de uma sentença ser incluída no extrato. Esse cálculo combina várias heurísticas encontradas na literatura, porém, introduzindo uma inovação: são considerados *"gold standards"* para o treinamento, definidos por especialistas humanos em sumarização textual. Teufel e Moens (1999) estendem essa idéia, adicionando a essa classificação probabilística a função retórica de cada sentença. A distribuição retórica do texto-fonte é baseada em sua macro-estrutura, i.e., nas categorias distintas de informação que caracterizam os segmentos mais genéricos do texto. Por exemplo, para os textos científicos sob análise, os macro-componentes podem incluir *problema, propósito, metodologia, resultados, conclusões, trabalho futuro*, etc. Para o português, especificamente, na linha de sistemas treináveis, baseados em algoritmos de Aprendizado de Máquina, há os trabalhos de Larocca Neto et al. (2002), Módolo (2003) e Pardo et al. (2003b), que buscam determinar as sentenças de um texto que devem ser incluídas no extrato correspondente utilizando características textuais diversas. Os dois primeiros trabalhos fazem uso de um classificador estatístico; o último utiliza uma rede neural do tipo SOM (*self-organizing map*) (Kohonen, 1982).

Ainda como outra linha de sumarização automática extrativa, Barzilay e Elhadad (1999) exploram a coesão lexical (i.e., o encadeamento de itens lexicais no texto) para construir cadeias léxicas que, quanto mais fortemente conectadas, mais chances têm de indicarem as sentenças significativas para compor o extrato. Nessa abordagem são utilizadas fontes robustas de conhecimento, tais como: a) a WordNet [Miller, 1995] – uma base de dados lexical que permite determinar a relação entre as palavras; b) um etiquetador morfológico – que associa etiquetas a cada palavra, indicando sua categoria morfológica; c) um *parser*, para identificar grupos nominais (envolvendo substantivos e adjetivos) e d) um algoritmo de segmentação textual, responsável por delimitar, no texto-fonte, os segmentos que indicam as cadeias léxicas mais fortes.

Propostas dessa natureza evidenciam a grande variedade de abordagens extrativas, várias delas recorrendo a técnicas de aprendizado e treinamento automáticos com base em grandes corpora de textos que tendem a ser mais robustas, quando comparadas aos métodos extrativos mais simples. É importante notar que elas sugerem a manipulação numérica, em geral estatística, de componentes textuais, considerando medidas que, *implicitamente*, incorporam características lingüísticas e a experiência de sumarizadores humanos. De fato, na tarefa de identificação e *cópia* de material dos textos-fonte para produzir os extratos, as métricas da sumarização automática extrativa modelam, sobretudo, aquelas adotadas pelos sumarizadores profissionais [Borko and Bernier, 1975; Cremmins, 1996], e, logo, estão próximas à tarefa de *professional summarizing*. Já a abordagem fundamental adota-as *explicitamente*, como veremos a seguir.

4.2. A sumarização automática fundamental

Do ponto de vista fundamental, a sumarização automática consiste na tarefa de se distinguir, do texto-fonte, o que é relevante, organizar tal conteúdo coerentemente e buscar sua expressão lingüística. Sparck Jones (1993b) distingue, para isso, três etapas básicas: a construção de uma representação do significado a partir do texto-fonte, a geração da representação do sumário correspondente e a sua síntese, ou realização lingüística, resultando no sumário, propriamente dito. Nessa perspectiva, é importante, então, determinar o conteúdo informacional do texto-fonte para compor o sumário e o contexto comunicativo desejado, da mesma forma que na comunicação humana. Três tipos de informação devem, assim, ser contemplados: o *lingüístico*, o *informativo* (ou de domínio) e o *comunicativo*, remetendo a questões semânticas e pragmáticas que aumentam a complexidade dos sistemas, devido à necessidade de modelagem desse conhecimento: é necessário haver uma linguagem de representação que possibilite o inter-relacionamento entre as unidades proposicionais (ou de significado) e engenhos de inferência capazes de interpretar o texto-fonte e gerar sua forma condensada correspondente. Vale notar que, nessa abordagem, uma vez construída a representação do significado do texto-fonte, os demais processos podem corresponder em maior ou menor grau àqueles da Geração de Textos (cf. Fig. 1, Seção 3), com a restrição de que cada uma de suas etapas considere os critérios de condensação da mensagem-fonte, para garantir a textualidade dos sumários resultantes e a correspondência com o texto-fonte.

Várias são as perspectivas dessa abordagem, todas elas buscando determinar as informações relevantes por meio de técnicas que refletem a modelagem discursiva em diferentes graus de profundidade. Nesse contexto, a *saliência* das informações de um texto-fonte é uma propriedade importante, definida como a medida de proeminência relativa dos objetos ou conceitos textuais[5]: aqueles com grande saliência são o foco de atenção no discurso e, logo, devem ser considerados na sumarização automática; os com baixa saliência são periféricos e, logo, são passíveis de exclusão. A seguir, comentamos brevemente dois trabalhos considerados representativos na linha do processamento profundo do discurso.

4.2.1. A sumarização automática a partir de segmentos salientes de estruturas discursivas

Visando a sumarização automática de um único documento, Marcu (1997a) propõe técnicas de segmentação do discurso para identificar o tópico e, a partir deste, estabelecer a saliência das informações relacionadas. A determinação das informações salientes é feita com base na estrutura retórica do texto, formalizada segundo a Teoria RST – *Rhetorical Structure Theory* (já mencionada na Seção 3). Assim, é preciso, primeiro, construir a estrutura retórica do texto-fonte, para, então, determinar o conteúdo e a forma de seus possíveis sumários. Considerando que, na RST, as relações retóricas são definidas mediante a assimetria de relacionamento proposicional no discurso, Marcu explora o fato de *núcleos* e *satélites* indicarem funções discursivas distintas, i.e., ele explora a própria *nuclearidade* da Teoria, a qual pode levar a diferentes graus de saliência (1997b).

[5] Definição de saliência de Boguraev e Kennedy (1997).

O cômputo da saliência dos componentes do discurso se baseia tanto na nuclearidade quanto em sua profundidade na estrutura RST: núcleos mais acima na estrutura serão mais importantes do que satélites ou outros núcleos mais profundos. A estrutura RST do Texto1 abaixo [Jordan, 1980, p. 225], p.ex., denominada PlanoRST1, é dada na Fig. 2 (ramos nucleares em negrito) [Pardo, 2002][6]. Cada proposição, neste caso, é delimitada pelos segmentos textuais numerados no Texto1 e ocupa uma folha da estrutura; os nós intermediários remetem às relações RST de Mann e Thompson. Nessa figura, as proposições são anotadas por etiquetas que remetem ao seu conteúdo informativo (*probl* = problema; *sol* = solução; *prop* = proposição genérica; *met* = método; *sit* = situação), mas esta não é condição inerente da modelagem retórica em foco.

Texto1: 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. Yet, in spite of this, the general level of understanding of the power and weaknesses of computers among manufacturing managers is dangerously low.

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

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

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

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

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

No PlanoRST1, as unidades mais salientes de cada segmento discursivo são indicadas junto aos nós intermediários. A ordem de precedência entre todas as proposições desse discurso é dada por 2>1>3a>3b>4>6>5 ('p1>p2' indica que p1 é mais importante que p2). Sumários do Texto1 podem, agora, ser construídos respeitando-se essa ordem: variando-se o número de segmentos a incluir, podemos ter os sumários 1 e 2 (Fig. 3), de diferentes tamanhos, para esse texto[7]. O Sumário 1 envolve somente a relação BACKGROUND entre *probl(2)*e

[6] ELABORATE indica uma relação em que o satélite apresenta detalhes adicionais sobre a situação ou sobre algum elemento do assunto apresentado no núcleo; EXPLAIN indica que o satélite da relação permite ao leitor aceitar o modo como o núcleo deve ser interpretado; PURPOSE identifica o satélite como a situação a ser realizada pela atividade apresentada no núcleo; BACKGROUND e MEANS se referem ao modo como o satélite apresenta o contexto ou a informação que permite aumentar a habilidade do leitor para a) entender um elemento do núcleo e b) entender como o evento ou situação apresentado no núcleo ocorre, respectivamente.

[7] A título de ilustração, os sumários foram construídos manualmente, modificando ligeiramente as construções lingüísticas correspondentes do Texto1, quando necessário. No processo automático, os planos

sit(1); o Sumário 2 envolve BACKGROUND e MEANS, esta entre *probl(2)*e *sol(3a)*). Ambos, no entanto, têm como proposição mais saliente do discurso a asserção do problema – *probl(2)*.

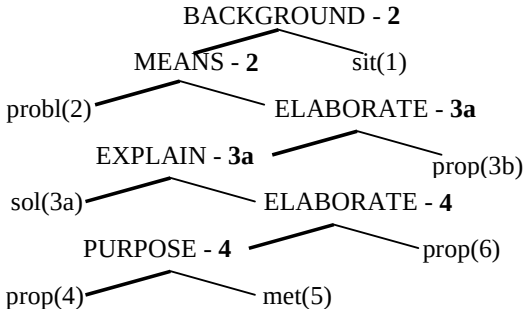


Figura 2. PlanoRST1 do Texto1

Do ponto de vista de independência de gênero e de generalidade, a abordagem de Marcu parece ser a mais consistente e efetiva atualmente. Marcu (1999) demonstra, ainda, que há uma correlação entre a nuclearidade de uma estrutura discursiva de um texto-fonte e a percepção que leitores têm sobre a importância de unidades textuais. Entretanto, sua proposta pressupõe a disponibilidade de estruturas RST para cada texto-fonte a sumarizar e, logo, requer um bom interpretador de língua natural, que gere suas estruturas profundas, retóricas. Isto seria um problema muito grande para a sumarização automática, se Marcu et al. (1999) não tivessem criado um ambiente completo de sumarização automática, incluindo ferramentas de *tagging* para anotação retórica dos segmentos de discurso e posterior construção de corpora de árvores retóricas. Entretanto, este trabalho se aplica somente à língua inglesa.

RST correspondentes seriam obtidos pela reestruturação do PlanoRST1 e, então, propriamente realizados lingüisticamente (vide Seção 3 para detalhes sobre cada um desses processos).

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

Sumário 2

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

Figura 3. Possíveis sumários do Texto1

4.2.2. A sumarização automática a partir do enredo semântico e intencional dos componentes discursivos

Também com base na nuclearidade da RST, propusemos um modelo fundamental de produção de discurso geral e independente de língua natural [Rino, 1996], mas bastante distinto do modelo de Marcu. Enquanto ele utiliza a RST como meio, nós a utilizamos como fim: o sistema de sumarização automática, chamado DMSumm (descrito em [Pardo and Rino, 2001; Pardo, 2002]) não parte do plano RST do texto-fonte, como ilustramos na seção anterior, mas produz um plano RST do sumário pretendido. A noção de nuclearidade, aqui, está na restrição de que a proposição central do discurso esteja na posição "mais nuclear" desse plano, i.e., que seja a folha mais à esquerda de uma árvore RST. Se, por um lado, não calculamos a saliência das proposições subjacentes ao texto-fonte, por outro lado, baseamo-nos na premissa de que a preservação da idéia central dos textos-fonte na sumarização automática é prioritária, sendo ela a responsável por um inter-relacionamento entre os segmentos discursivos que garanta a textualidade dos sumários resultantes e determine a contribuição de cada um deles ao discurso como um todo.

Nesse modelo, o plano RST do sumário é construído integralmente a partir de uma representação profunda do texto-fonte – a mensagem-fonte – constituída do objetivo comunicativo, da proposição central e de uma base informativa, ou base de conhecimento. A interpretação de um texto-fonte ainda não é automática. Assim, a mensagem-fonte é obtida, até o momento, de forma inteiramente manual e DMSumm se restringe aos processos de Geração de Textos clássicos (vide Fig. 1). Seu principal componente é o planejador textual, o qual visa identificar, na mensagem-fonte, os segmentos discursivos que permitam satisfazer o objetivo comunicativo e preservar a proposição central (planejamento dirigido por objetivos comunicativos). Esse modelo adota, assim, os três níveis de representação discursiva delineados por Sparck Jones (1993b) – o lingüístico, o informativo e o comunicativo.

O nível informativo está na representação da base de conhecimento: as unidades proposicionais do texto-fonte são semanticamente inter-relacionadas por sua função no domínio em foco. O modelo semântico é baseado em análise de corpora de textos científicos, que apresentam uma macro-estrutura similar àquela apontada por Teufel e Moens (vide Seção 4.1) e remetem ao Modelo Problema-Solução [Winter, 1977; Jordan, 1980]. As relações semânticas, na micro-estrutura da base de conhecimento, são definidas a partir dos macro-componentes, mas conectam as proposições subjacentes, conforme ilustra a Fig. 4[8]. Por exemplo, uma *solução* permite (*enable*) resolver um *problema*; a *situação* apresentada contextualiza (*backsem*) a apresentação do *problema*. Vale notar que, nessa base de conhecimento (do mesmo Texto1 ilustrado na seção anterior), as etiquetas das proposições, agora, são relevantes, pois indicam o tipo da informação, o qual, por sua vez, permite descrever o inter-relacionamento semântico entre duas unidades informativas.

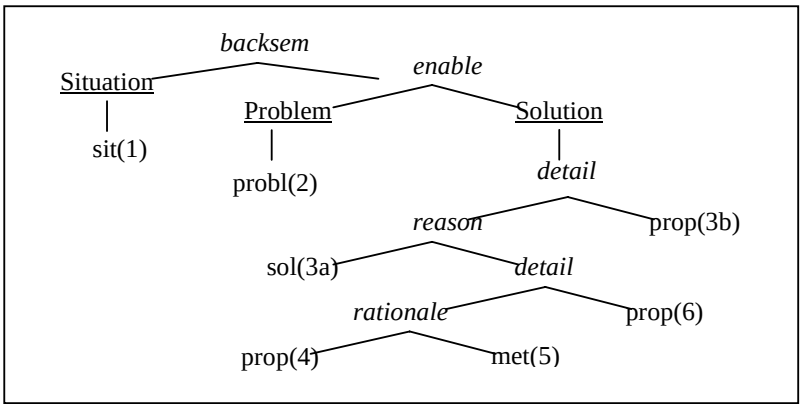


Figura 4. Base de conhecimento do Texto1

No nível comunicativo, usamos o modelo intencional da Teoria de Discurso de Grosz e Sidner – *Grosz and Sidner Discourse Theory*, ou GSDT [Grosz and Sidner, 1986]. Este pressupõe que as intenções subjacentes ao discurso (em nosso caso, intenções do E) têm uma função fundamental em sua estruturação, definindo sua coerência e até justificando sua existência. Cada segmento de discurso tem seu próprio objetivo comunicativo e a construção do discurso consiste em determinar como um segmento pode contribuir para a satisfação do objetivo comunicativo de outro segmento. De modo incremental, a contribuição local dos segmentos escolhidos para compor o discurso implicará a satisfação do objetivo comunicativo geral, i.e., do objetivo comunicativo da mensagem-fonte que se pretende transmitir. No nosso modelo, a definição dessa cooperação envolve os tipos de informação disponíveis na mensagem-fonte, remetendo à base de conhecimento. Por exemplo, para as relações GSDT no nível comunicativo, *dominates* (para *dom(Y,X)*, lê-se Y *dominates X*, ou X *contribui para Y*) e *satisfaction-precedes* (para *sp(X,Y)*, lê-se *X satisfaction-precedes Y*, ou *X deve preceder Y*),

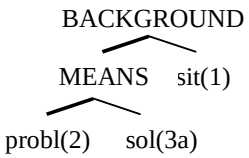
[8] As informações sublinhadas referem-se aos macro-componentes, as em itálico, às relações semânticas e as etiquetas das folhas, aos componentes do Modelo Problema-Solução.

*dom(sol,probl)*e *sp(probl,sol)* indicam, respectivamente, que a satisfação da intenção subjacente ao *problema* contribui para a satisfação da intenção subjacente à *solução* e a satisfação da intenção subjacente ao *problema* deve preceder a satisfação da intenção subjacente à *solução*.

Assim, em nosso modelo fundamental a definição intencional é independente da língua natural, mas dependente do modelo de domínio. Durante o planejamento textual, as informações desses dois níveis são combinadas, a fim de determinar as relações retóricas pertinentes, no nível lingüístico (o que levará à escolha da ordem de aparição de tais informações durante a realização lingüística). A regra de planejamento abaixo indica genericamente, p.ex., que as relações PURPOSE(sol(3a),probl(2)) ou MEANS(probl(2),sol(3a)) podem ser geradas a partir de *enable(sol(3a),probl(2))*, desde que a) *enable*, assim definida, seja uma relação da base de conhecimento, ou BC (para *enable(Y,X)*, lê-se *a existência de X fornece as condições para a ocorrência de Y*) e b) *sp(probl,sol)* e *dom(sol,probl)* se verifiquem no nível intencional.

$$enable(Y,X) \in BC\ ,\ sp(X,Y)\ ,\ dom(Y,X) \rightarrow PURPOSE(Y,X)\ ;\ MEANS(X,Y)$$

Um possível plano de sumário para o Texto1 (PlanoRST2) é ilustrado na Fig. 5, juntamente com sua realização lingüística em inglês (resolução totalmente automática do DMSumm, nos três processos da geração textual). Esse plano considera os segmentos 2, 1 e 3a como os mais relevantes do Texto1, sendo a proposição central = *probl(2)*. Observando o PlanoRST1 (Fig. 2), obtido ao aplicar a técnica de Marcu, 2 e 3a são também os segmentos mais relevantes. O PlanoRST2 poderia também ser gerado pelo sistema de Marcu, pois o PlanoRST1, estrutura retórica do Texto1, inclui BACKGROUND e MEANS, que coincidem com as escolhas retóricas do DMSumm (MEANS, p.ex., seria gerada pela regra de planejamento acima)



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

Figura 5. PlanoRST2 e sumário resultante

Os modelos fundamentais apresentados acima têm em comum a correspondência entre nuclearidade e importância de componentes do discurso, adotando critérios para sua identificação e inclusão nos sumários desejados baseados na RST. A diferença entre eles está no fato de o DMSumm incorporar, além dessa teoria, também a Teoria GSDT, considerando que o nível intencional é necessário para a produção de discursos coerentes e coesos. Tratando a sumarização como uma tarefa de seleção do conteúdo e sua completa reestruturação, a consideração dos níveis informativo, comunicativo e lingüístico se assemelha muito mais à tarefa de *author summarizing*, pela possível inclusão de conteúdo e

objetivos comunicativos não aparentes ou não considerados nos próprios textos-fonte a sumarizar.

5. Considerações finais

Apresentamos aqui algumas perspectivas de geração e sumarização automática de textos, neste caso envolvendo técnicas fundamentais e extrativas. Muito embora tenhamos destacado a sumarização automática, vimos que a maioria das decisões envolvidas na modelagem fundamental é comum à Geração de Textos, sendo as mais importantes relativas às restrições de existência de uma mensagem-fonte, de busca da satisfação de objetivos comunicativos e de produção de textos coerentes e coesos. As propostas de sumarização automática fundamental apresentadas se originaram, aliás, com as voltadas à Geração de Textos. Em ambas, a maior dificuldade está em se extrair a estrutura de significado do texto-fonte, a qual é, na maioria das vezes, dependente do gênero e da tipologia textual, levando a etapas muito difíceis de aquisição do conhecimento e processamento simbólico, razões pelas quais não se garante um processamento robusto e abrangente.

Modelos alternativos para a sumarização automática fundamental baseiam-se, com diversos graus de profundidade, na noção de que a relevância das sentenças é derivável de representações discursivas do texto-fonte. Vale notar que tais trabalhos avançaram, sobretudo, a partir da caracterização de Sparck Jones (1993b). Destacamos, particularmente, [Ono et al., 1994] e [Boguraev and Kennedy, 1997], por sua exploração da contigüidade e saliência de segmentos textuais.

Embora a abordagem extrativa seja mais robusta e abrangente do que a fundamental, muitas das técnicas envolvidas também dependem de fatores estruturais: muito embora os métodos extrativos se baseiem, prioritariamente, em modelos matemáticos, sua proposta inclui a manipulação de parâmetros com correspondentes lingüísticos ou discursivos. Exemplo disso está nas noções de identificação da posição das informações textuais, para determinar sua saliência ou o reconhecimento de cadeias lexicais mais conectadas, o qual remete a uma distribuição semântica do conhecimento do mundo, noção que extrapola os cálculos numéricos.

Outros modelos alternativos para a sumarização automática extrativa têm sido explorados recentemente, sobretudo a partir do trabalho de Kupiec et al. (1995) (vide Seção 4.1.2.). Seus *"gold standards"* para identificação de informações relevantes remetem às palavras *bônus* (p.ex., "melhor", "maior", "significativo") e *stigma* (p.ex., "dificilmente", "impossível"), de Rush et al. (1971), para indicar, respectivamente, as sentenças relevantes e não relevantes de um texto. Outra proposta interessante que, embora extrativa, baseia-se fortemente em noções lingüísticas, é a de Barzilay e Elhadad (1999), segundo a qual sentenças e conceitos importantes de um texto-fonte estão fortemente conectados em estruturas semânticas [Skorochodko, 1971; Lin, 1995; Salton et al., 1994]. Independentemente do nome ou método que se adote, a questão principal continua sendo explorar a saliência das informações, sob diversas perspectivas, buscando resultados mais satisfatórios.

Há problemas cruciais em ambas as abordagens: de um modo geral, as técnicas fundamentais procuram operar sobre o contexto discursivo, buscando a *reescrita* do sumário. Já as extrativas operam sobre palavras ou cadeias de palavras, com claro prejuízo

para a preservação do significado original, a generalização conceitual e a própria garantia de textualidade. Claramente, estratégias genéricas são preferíveis, o que privilegiaria a abordagem extrativa. Entretanto, esta freqüentemente leva a poucos sumários efetivos (i.e, com qualidade e poder comunicativo significativos).

Abordagens híbridas, que consideram que posições proeminentes de unidades textuais são dependentes de gênero e, logo, treinam seu sistema de sumarização automática em função de corpora de textos de um mesmo gênero, dentre outras, são também dignas de nota. Destacamos, aqui, a de Lin e Hovy (1997).

Finalmente, resta registrar que fizemos, aqui, um breve *recorte* da problemática relacionada à geração automática de textos e sumários, não havendo, de modo algum, esgotado as referências aos trabalhos relevantes da área, tampouco os tópicos associados à sumarização automática, dentre os quais destaca-se a avaliação dos sistemas de sumarização automática, tópico essencial, mas relativamente novo, cujas métricas ainda não atingem uma avaliação formal em larga escala.

6. Leituras Complementares

A bibliografia sobre sumarização automática em português é formada por poucas obras de referência (relatórios técnicos do NILC, em geral) e artigos de divulgação científica, notadamente os apresentados em conferências de Inteligência Artificial (SBIA) ou em encontros sobre o processamento da língua portuguesa (PROPOR), que aceitam o português como uma das línguas para apresentação dos trabalhos.

Em inglês, destacam-se, entre os textos introdutórios, principalmente a coletânea editada por Mani e Maybury (1999). Este volume sintetiza as principais abordagens de sumarização automática que foram adotadas ao longo dos anos, incluindo as clássicas e pioneiras propostas de Luhn (1958) e Edmundson (1969). Finaliza apontando novas áreas de exploração que ainda são atuais, muito embora o livro tenha sido publicado em 1999. Além disso, há uma seção exclusiva para os métodos de avaliação dos resultados automáticos e do próprio desempenho dos sumarizadores correspondentes, tópico de grande relevância atualmente. Essa obra completa, assim, o ciclo todo de assuntos relevantes para a pesquisa e desenvolvimento de aplicativos de sumarização automática. A maioria dos artigos coletados são reproduções fiéis dos originais, havendo, no início, uma referência às reproduções que vale a pena o leitor observar. As exceções incluem artigos originais, produzidos especialmente para observar o tema do livro – *Advances in Automatic Text Summarization*.

Outro livro bastante interessante é o de Mani (2001), que trata, particularmente, dos métodos de avaliação de sistemas de sumarização automática, aglomerando propostas que até muito recentemente eram adotadas de forma isolada na área (sendo várias destas apresentadas na coletânea anteriormente referida). São apresentados os principais fatores e as métricas que devem estar em foco para um processo de avaliação, resultando na possibilidade de o leitor se familiarizar significativamente com o problema.

Outros textos expressivos na área são os de McKeown e Radev (1995) e Paice (1990), os quais se referem à dependência da sumarização automática, de fatores de gênero e tipo textual. Já o texto de O'Donnell (1997a) discute o assunto sob a perspectiva de manipulação de estruturas RST, de modo similar ao indicado na Seção 4, porém, menos

formal. O'Donnell construiu uma ferramenta – a RSTTool (disponível no site http://www.wagsoft.com/RSTTool) – que permite a manipulação de estruturas RST visando a sumarização automática, por um usuário conhecedor da RST. Referências específicas para essa ferramenta são as de O'Donnell (1997b) e (2000).

Sobre avaliação de sistemas de processamento do português, de um modo geral, mas que remetem, certamente, à questão da sumarização automática, há o trabalho expressivo de Sparck Jones e Galliers (1996), que merece ser lido.

De modo mais geral, há várias outras obras relacionadas ao tópico deste capítulo, quais sejam: aquelas que versam sobre a Geração Automática de Textos, dentre as quais um dos melhores exemplares é a de Reiter e Dale (2000); as que versam sobre discurso, que incluem as de Couture (1985), Hobbs (1985), Levelt (1989), Stainton (1989), Swales (1990) e Moser e Moore (1993). Dentre estas, somente a última é de autoria de lingüistas computacionais, sendo as demais de cunho puramente lingüístico ou discursivo.

Referências Bibliográficas*

Barzilay, R. and Elhadad, M. (1999). "Using Lexical Chains for Text Summarization", In: Advances in Automatic Text Summarization, Edited by I. Mani and M. Maybury, The MIT Press, p.111-121.

Baxendale, P.B. (1958). "Machine-made Index for Technical Literature – An Experiment". IBM Journal of Research and Development, 2, p. 354-361.

Boguraev, B. and Kennedy, C. (1997). "Salience-Based Content Characterisation of Text Documents'. Proc. of the Intelligent Scalable Text Summarization Workshop,. ACL/EACL'97 Joint Conference. Madrid, Spain, p. 2-9.

Borko, H. and Bernier, C.L. (1975). Abstracting Concepts and Methods. Academic Press. San Diego, CA.

Chomsky, N. (1965). Aspects of the Theory of Syntax. MIT Press, Cambridge, MA.

Couture, B. (1985). "A Systemic Network for Analyzing Writing Quality", In: Systemic Perspectives on Discourse, Edited by James D. Benson and William S. Greaves, Ablex Publishing Corporation, Vol. 2, p. 67-87.

Cremmins, E.T. (1996). The Art of Abstracting. Information Resource Press. Arlington, Virginia.

Edmundson, H.P. (1969). "New Methods in Automatic Extraction", Journal of the Association for Computing Machinery, 16 (2), p. 264-285.

Espina, A.P. e Rino, L.H.M. (2002)*. Utilização de Métodos Extrativos na Sumarização Automática de Textos. Tech. Rep. NILC-TR-02-06. São Carlos, Março, 22p.

Grosz, B. and Sidner, C. (1986). "Attention, Intentions, and the Structure of Discourse", Computational Linguistics, Vol. 12, N. 3.

* As referências marcadas com '*' referem-se a trabalhos do NILC que podem ser encontrados em http://www.nilc.icmc.usp.br (Publications).

Hearst, M.A. (1999). “Untangling Text Data Mining”, Proc. of ACL’99, the 37th Annual Meeting of the ACL. University of Maryland, USA.

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

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

Hovy, E. (1993). “Automated Discourse Generation Using Discourse Structure Relations”, Artificial Intelligence 63, p. 341-385.

Hovy, E. and Lin, C.Y. (1997). “Automated Text Summarization in SUMMARIST”, Proc. of the Intelligent Scalable Text Summarization Workshop, ACL/EACL’97 Joint Conference. Madrid, Spain, p. 18-24.

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

Kress, G. (1989). Linguistic Processes in Socio-Cultural Practice. Oxford University Press.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, Vol. 43, pp. 59-69.

Kupiec, J.; Pedersen, J.O.; and Chen, F. (1995). “A Trainable Document Summarizer”, Proc. of the 18th ACM-SIGIR Conference, Association for Computing Machinery, SIG on Information Retrieval, p. 68-73.

Larocca Neto, J.; Freitas, A.A.; Kaestner, C.A.A. (2002). Automatic Text Summarization using a Machine Learning Approach. In the *Proc. of the XVI Brazilian Symposium on Artificial Intelligence*, *Lecture Notes on Compute Science*, No. 2507, pp. 205-215.

Larocca Neto, J. (2002). Contribuição ao Estudo de Técnicas para Sumarização Automática de Textos. Dissertação de Mestrado. PUC-PR, Curitiba, PR. Fevereiro.

Larocca Neto, J.; Santos, A.D.; Kaestner, C.A.; Freitas, A.A. (2000a). “Document Clustering and Text Summarization”, Proc. of the 4th Int. Conf. on Practical Applications of Knowledge Discovery and Data Mining, London, p. 41-55.

Larocca Neto, J.; Santos, A.D.; Kaestner, C.A.; Freitas, A.A. (2000b). “Generating Text Summaries through the Relative Importance of Topics”, Proc. of IBERAMIA/SBIA’2000. Atibaia, São Paulo.

Levelt, W.J.M. (1989). Speaking: From Intention to Articulation. The MIT Press.

Lin, C.Y. (1995). “Knowledge-based Automatic Topic Identification”, Proc. of the 33rd Annual Meeting of the ACL, Cambridge, MA, p. 308-310.

Lin, C.Y. and Hovy, E. (1997). “Identifying Topics by Position”, Proc. of the 5th Conference on Applied Natural Language Processing – ANLP’97, Washington, DC, p. 283-290.

Luhn, H.P. (1958). “The Automatic Creation of Literature Abstracts”, IBM Journal of Research and Development, 2 (2), p. 159-165.

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

Mani, I. and Maybury, M. (eds.) (1999). Advances in Automatic Text Summarization. The MIT Press.

Mann, W.C. and Thompson, S.A. (1988). "Rhetorical Structure Theory: Toward a Functional Theory of Text Organization", Text, 8 (3), p. 243-281.

Marcu, D. (1997a). "The Rhetorical Parsing of Natural Language Texts", Proc. of the ACL/EACL'97 Joint Conference, Madrid, Spain, p. 96-103.

Marcu, D. (1997b). "From Discourse Structures to Text Summaries", Proc. of the Intelligent Scalable Text Summarization Workshop, ACL/EACL'97 Joint Conference. Madrid, Spain, p. 82-88.

Marcu, D. (1999). "Discourse Trees are Good Indicators of Importance in Text", In: Advances in Automatic Text Summarization, Edited by I. Mani and M. Maybury, The MIT Press, p.123-136.

Marcu, D.; Amorrortu, E.; and Romera, M. (1999). "Experiments in Constructing a Corpus of Discourse Trees", Proc. of the ACL'99 Workshop on Standards and Tools for Discourse Tagging, Maryland, p. 48-57.

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

Martins, C.B.; Pardo, T.A.S.; Espina, A.P.; Rino, L.H.M. (2001)*. Introdução à Sumarização Automática. Rel. Técnico RT-DC 002/2001. Departamento de Computação, UFSCar. Abril. 38 p.

Matthiessen, C.M.I. and Bateman, J.A. (1991). Text Generation and Systemic-Functional Linguistics, Pinter Publishers, London.

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

McKeown, K. and Radev, D. (1995). "Generating summaries of multiple news articles", Proc. of the 18th Annual International ACM-SIGIR Conference on Research and Development in Information RetrievalSeattle, WA, p. 74-82.

Miller, G. (1995). "WordNet: A Lexical Database for English", Communications of the Association for Computing Machinery, 38 (11), pp. 39-41.

Módolo, M. (2003). SuPor: um Ambiente para a Exploração de Métodos Extrativos para a Sumarização Automática de Textos em Português. Dissertação de Mestrado. Junho. Departamento de Computação, UFSCar.

Moser, M. and Moore, J.D. (1993). "Investigating Discourse Relations", In: Intentionality and Structure in Discourse Relations, Edited by O. Rambow, Ohio State University. Ohio, USA. June, pp. 94-97.

O'Donnell, M. (1997a). "Variable-Length On-Line Document Generation", Proc. of the 6th European Workshop on Natural Language Generation, Gerhard-Mercator University, Duisbug, Germany.

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

O'Donnell, M. (2000). "RSTTool 2.4 – A Markup Tool for Rhetorical Structure Theory", Proceedings of the International Natural Language Generation Conference (INLG'2000), 13-16 June, Mitzpe Ramon, Israel, p. 253-256.

Ono, K.; Sumita, K.; and Miike, S. (1994). "Abstract Generation Based on Rhetorical Structure Extraction", Proc. of the International Conference on Computational Linguistics (COLING'94), Japan, p. 344-348.

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).

Paice, C.D. (1990). "Constructing Literature Abstracts by Computer: Techniques and Prospects", Information Processing and Management, 26 (1), p. 171-186.

Pardo, T.A.S.; Rino, L.H.M.; Nunes, M.G.V. (2003a). GistSumm: A Summarization Tool Based on a New Extractive Method. In N.J. Mamede, J. Baptista, I. Trancoso, M.G.V. Nunes (eds.), *6th Workshop on Computational Processing of the Portuguese Language - Written and Spoken*, pp. 210-218 (Lecture Notes in Artificial Intelligence 2721). Springer-Verlag, Germany.

Pardo, T.A.S; Rino, L.H.M.; Nunes, M.G.V. (2003b). NeuralSumm: Uma Abordagem Conexionista para a Sumarização Automática de Textos. *Anais do IV Encontro Nacional de Inteligência Artificial – ENIA'2003*. XXII Congresso Nacional da Sociedade Brasileira de Computação. Campinas – SP. Agosto.

Pardo, T.A.S. (2002)*. DMSumm: Um Gerador Automático de Sumários. Dissertação de Mestrado. DC/UFSCar. Março.

Pardo, T.A.S. and Rino, L.H.M. (2001)*. "A Summary Planner Based on a Three-Level Discourse Model", Proc. of the 6th NLPRS – Natural Language Processing Pacific Rim Symposium, National Center of Science, Tokyo, Japan. 27–29 November, p. 533-538.

Paris, C. (1993). User Modelling in Text Generation. Pinter Publishers.

Pereira, M.B.; Souza, C.F.R. e Nunes, M.G.V. (2002)*. "Implementação, Avaliação e Validação de Algoritmos de Extração de Palavras-Chave de Textos Científicos em Português", Revista Eletrônica de Iniciação Científica, SBC, Ano II, Vol.1, Num.1, Março.

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

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

Rush, J.E.; Salvador, R.; and Zamora, A. (1971). "Automatic Abstracting and Indexing. Production of Indicative Abstracts by Application of Contextual Inference and Syntactic

Coherence Criteria", Journal of American Society for Information Sciences, 22 (4), pp. 260-274.

Salton, G. and McGill, J.J. (1983). Introduction to Modern Information Retrieval. McGraw-Hill, New York.

Salton, G.; Allan, J.; Buckley, C.; and Singhal, A. (1994). "Automatic Analysis, Theme Generation, and Summarisation of Machine Readable Texts", Science, 264, p. 1421-1426.

Skorochodko, E.F. (1971). "Adaptive Method of Automatic Abstracting and Indexing", Information Processing, Vol. 2, pp. 1179-1182. North-Holland Publishing Company.

Souza, C.F.R. e Nunes, M.G.V. (2001)*. Avaliação de Algoritmos de Sumarização Extrativa de Textos em Português. Relatórios Técnicos do ICMC-USP (NILC-TR-01-9), 153, Outubro, 14p.

Sparck Jones, K. (1993a). Discourse Modelling for Automatic Summarising. Tech. Rep. No. 290. University of Cambridge, February.

Sparck Jones, K. (1993b). "What might be in a summary?" In: Information Retrieval, Edited by G. Knorz; J. Krause and C. Womser-Hacker, 93, Universitatsverlag Konstanz, June, p. 9-26.

Sparck Jones, K. (1997). "Summarising: Where are we now? Where should we go?" Proc. of the Intelligent Scalable Text Summarization Workshop, ACL/EACL'97 Joint Conference. Madrid, Spain, p. 1.

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

Stainton, C. (1989). Report on Genre and Genre Study, Research & Technology Project No. 322, Department of English Studies, University of Nottingham.

Swales, J.M. (1990). Genre Analysis: English in Academic and Research Settings. Cambridge University Press.

Teufel, S. and Moens, M. (1999). "Argumentative Classification of Extracted Sentences as a First Step Towards Flexible Abstracting", In: Advances in Automatic Text Summarization, Edited by I. Mani and M. Maybury, The MIT Press, p.155-171.

Winter, E.O. (1977). "A Clause-Relational Approach to English Texts: A Study of Some Predictive Lexical Items in Written Discourse", Instructional Science 6 (1), p. 1-82.