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

A Sumarização Automática com Base em Estruturas RST

Gislaine Fonseca Ribeiro
Lucia Helena Machado Rino

NILC-TR-02-05

Maio, 2002

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

Resumo

Neste relatório, apresentamos uma proposta de exploração de uma ferramenta de estruturação retórica – a RSTTool (*Rhetorical Structure Theory Tool*) – para a sumarização automática de textos. Consideramos que a todo texto corresponde uma estrutura profunda, aqui denominada *estrutura RST*, a partir da qual é possível construir a(s) estrutura(s) RST do(s) sumário(s) correspondente(s). Esse processo se dá pela identificação, na estrutura RST do texto-fonte, de segmentos proposicionais que podem ser considerados supérfluos. A Teoria RST permite, por meio da definição de suas relações RST, identificar tais segmentos e reestruturar o sumário após sua exclusão, resultando em novas estruturas RST, agora dos sumários. Por um processo de realização lingüística, tais estruturas podem dar origem a textos condensados, considerados os sumários dos textos originais.

Índice

Figuras

1. Introdução

A Sumarização Automática (SA) começou a ser explorada no final da década de 50, quando se utilizavam, sobretudo, técnicas estatísticas de extração de conhecimento lingüístico dos textos-fonte. Os trabalhos mais relevantes, nessa época, são os de Luhn (1958) que usava o recurso de identificação de palavras-chave para poder formar os sumários, e o de Edmundson (1969), que estendeu o de Luhn ao considerar o próprio título do texto e a identificação de frases indicativas para formar os sumários. O principal problema dessas propostas é que os sumários gerados, em geral, não continham informações relevantes, tampouco eram bem estruturados ou compreensíveis, por conter, por exemplo, referências anafóricas mal resolvidas, decorrentes da extração de sentenças isoladas do texto-fonte para formar o sumário. Devido à impossibilidade técnica de aprimorá-los, a SA 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. Nessa época, distinguiram-se dois modelos básicos de SA: o superficial e o profundo, sendo este baseado na modelagem de características lingüísticas para o Processamento de Línguas Naturais (PLN). Um exemplar clássico dessa abordagem é o modelo baseado na RST - *Rhetorical Structure Theory* (Mann and Thompson, 1987; 1988).

Nesse relatório, apresentaremos um estudo da RST no contexto da SA, fazendo uso de uma ferramenta de estruturação textual chamada RSTTool (O'Donnel, 2000). Uma breve descrição da RST é dada na Seção 2, seguida da descrição de algumas propostas de seu uso na SA (Seção 3). Na Seção 4, descrevemos a RSTTool, seguida de ilustrações de sua exploração na SA e de algumas considerações finais (Seção 5).

2. A Rhetorical Structure Theory

2.1 RST

Inicialmente destinada à interpretação de línguas naturais, a RST se tornou uma das teorias de discurso mais utilizadas na Geração de Língua Natural (GLN). A idéia central dessa teoria é a de que um texto tem uma estrutura *retórica* associada, composta por suas proposições elementares, inter-relacionadas por meio de relações retóricas. As proposições elementares do discurso são unidades mínimas de significado veiculado pelo texto e podem ser reconhecidas neste ao delimitar seus segmentos de significado; as relações retóricas são aquelas que indicam o relacionamento existente entre as proposições elementares do discurso. Por exemplo, a sentença [1] do Texto 1[1] ilustrado abaixo indica claramente a proposição central do discurso – a de que o medo determinava o modo como o personagem (Almir) agia. Por sua vez, os segmentos frasais [2] e [3] indicam, respectivamente, o fato de que poucas pessoas conhecem essa característica do personagem e o fato de que essa característica é verdadeira. Assim, temos três proposições distintas aqui, expressas por [1], [2] e [3]. Podemos, como leitores, reconhecer que tais proposições se relacionam nesse discurso e, ainda mais, podemos identificar que [2] e [3] constituem uma *elaboração* da afirmação expressa em [1]. Na RST, tal inter-relacionamento é expresso pela relação retórica ELABORATION[2], conforme ilustra a Figura 1.

[1] Cada segmento desse texto, numerado seqüencialmente, corresponde a uma proposição do discurso. Assim, indicamos o nível proposicional fazendo uso do próprio nível superficial, muito embora eles não sejam correspondentemente marcados.

[2] A definição das relações RST se encontra na Seção 2.2.

Texto 1

[1] Muitas das atitudes "corajosas" de Almir, o Pernambuquinho, eram ditadas pelo medo. [2] Poucos sabem disso, [3] mas é verdade. [4] Quem o via de punhos cerrados, dentes trincados, desafiando adversários mais fortes, não imaginava que, por trás da valentia, escondia-se o medo de parecer covarde.

[5] Certa vez ele foi suspenso por uma jogada violenta [6] que inutilizou Hélio, do América. [7] À medida que ia se aproximando o fim da suspensão, [8] Almir começou a queixar-se de uma estranha dor muscular na perna direita.

[9] Dr. Valdir Luz e todo o departamento médico do Vasco já não sabiam o que fazer para curar a inexplicável "distensão". [10] Acabou-se a suspensão, [11] mas permaneceu a dor. [12] Até que o técnico Yustrich chamou o jogador para uma conversa: [13] "Você não tem nada, garoto. [14] É o medo de que alguém vingue o Hélio [15] que faz você sentir a dor.".

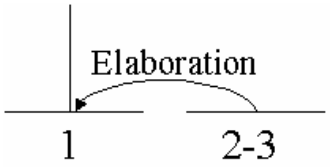

Figura 1: Relação RST ELABORATION entre proposições 1 & 2-3

Como essa figura indica, cada número representa uma proposição do discurso estruturado, que, no nosso texto exemplo, é indicada pelos segmentos textuais numerados. Cada relação RST é representada por um *arco direcionado*, sendo sua direção do satélite para o núcleo. Assim, a proposição na ponta da seta é sempre o núcleo. Relações que não são representadas por arcos direcionados, são relações multi-nucleares, cuja ilustração se encontra adiante. A definição e correspondente interpretação de uma estrutura RST como as ilustradas acima se encontram na Seção 2.2.

Estruturas retóricas completas, ou *estruturas RST*, por sua vez, dão origem a árvores binárias cujas folhas correspondem às proposições elementares e cujos nós internos, às relações retóricas (na seção 3.2 temos uma ilustração completa). Ainda, há uma hierarquização para a interpretação estrutural, indicada pela relação retórica: esta associa duas proposições com grau variado de *importância*, no contexto do discurso: uma proposição *nuclear* ou, simplesmente, um *núcleo*, é considerado mais central, ou mais relevante, do que seu *satélite* (daí a própria nomenclatura). No exemplo acima, a proposição correspondente à sentença [1] é o núcleo da relação ELABORATION, enquanto as duas proposições correspondentes aos segmentos [2] e [3], juntas, constituem seu satélite.

A RST é, assim, uma teoria que prescreve um conjunto de relações retóricas, a partir do qual se podem reconhecer os graus de relevância das informações de um discurso e representar sua estrutura hierarquicamente, mediante a delimitação de suas proposições elementares. Para determinar qual a estrutura retórica correspondente a um texto, é preciso, portanto, distinguir cada uma de suas proposições elementares, associando-as a um núcleo ou satélite de uma relação retórica, além de reconhecer a própria relação. Esta tarefa é uma tarefa de interpretação que nós, seres humanos, fazemos sem nos darmos conta de sua complexidade. Entretanto, em nosso estudo, precisamos interpretar o texto associando cada um desses componentes àqueles da própria RST, cuja descrição será dada adiante.

Além de identificar proposições pelo seu grau de importância, i.e., proposições nucleares ou satélites, também conseguimos identificar proposições que se encontram

em mesmo nível de importância. Devido a isso, a RST distingue as relações retóricas hipotáticas (que introduzem uma hierarquia de importância), das relações paratáticas (que introduzem igual hierarquia de importância). Exemplos de relações paratáticas são as relações multi-nucleares, já citadas, que relações que envolvem mais de duas proposições de mesmo nível informacional ou retórico. CONTRAST é um exemplo de relação RST multi-nuclear, cuja definição contrapõe as proposições envolvidas, como mostra a Figura 2, correspondente ao segmento textual S1 abaixo.

Segmento S1:

[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.

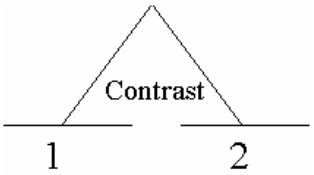

Figura 2: Relação CONTRAST, MULTI-NUCLEAR

Como ilustramos acima, a todo texto em língua natural é possível associar uma estrutura retórica por meio de um processo de interpretação. A RST é uma teoria que indica a) como reconhecer informações inter-relacionadas, distinguindo as nucleares e as satélites; b) como relacioná-las pela seleção de relações RST; c) como produzir uma representação estrutural, hierárquica, de um texto, aqui chamada de estrutura RST. Fazer uso dessa teoria não é trivial, pois envolve a distinção entre o que o escritor considera mais essencial para alcançar seu objetivo comunicativo, com seu discurso. O sucesso dessa distinção, pelo leitor, depende de sua observação empírica e subjetiva, atribuindo um grau de relevância maior ao que ele associa a um núcleo RST do que ao que ele associa a seu satélite correspondente. Essa noção de relevância pode ser entendida da seguinte forma, no contexto da RST: em geral, em uma estrutura RST, o núcleo expressa a informação que, se retirada, fará o texto resultante correspondente incoerente. Entretanto, ao retirar o satélite, embora o resultado contenha, potencialmente, menos informações (devido à exclusão da(s) proposição(ões) do satélite), ele continuará coerente. Esta é a característica que torna a RST muito interessante para a exploração da SA: ao distinguir informações essenciais (e, logo, nucleares) das complementares (e, logo, satélites), é possível elaborar mecanismos que modifiquem uma estrutura RST de um texto a ser sumarizado, produzindo a estrutura RST do seu sumário, sendo esta a representação retórica do texto condensado a gerar. Essa metodologia é descrita na Seção 3. Antes disso, apresentamos a seguir a definição das relações RST utilizadas neste projeto, ilustrando a construção de árvores RST.

2.2 Definição das Relações Retóricas Utilizadas

Abaixo se encontra a definição das relações retórica que foram usadas em todas as estruturas RST ilustradas nesse relatório. Têm-se: para núcleo (N), satélite(S), escritor(E) e leitor (L).

CIRCUNSTANCE

Restrições sobre o N	Nenhuma
Restrições sobre o S	S apresenta uma situação (não realizada).
Restrições sobre o N + S	S apresenta um conjunto de estruturas que cada L usa para interpretar a situação presente em N.
Efeito	L reconhece que a situação presente em S provê uma estrutura para a interpretação do N.

CONTRAST

Restrições sobre o N	Multi-nuclear
Restrições sobre os N	Não mais de dois núcleos; as situações presentes nos núcleos são: (a) compreendida como a mesma em muitos aspectos, (b) compreendidas como diferentes em poucos aspectos, (c) comparadas com respeito a uma ou mais diferenças.
Efeito	L reconhece as igualdades e as diferenças através da comparação que está sendo feita

ELABORATION

Restrições sobre o N	Nenhuma.
Restrições sobre o S	Nenhuma.
Restrições sobre o 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.

BACKGROUND

Restrições sobre o N	L não compreenderá N suficientemente antes de ter lido S.
Restrições sobre o S	Nenhuma.
Restrições sobre o N + S	S aumenta a habilidade do L de compreender N.
Efeito	A habilidade do L de compreender N aumenta.

NON-VOLITIONAL RESULT

Restrições sobre o N	Nenhuma.
Restrições sobre o S	Apresenta uma situação não proposital.
Restrições sobre o N + S	N apresenta uma situação que causa a situação presente em S.
Efeito	L reconhece que a situação em N pode ter causado a situação em S.

VOLITIONAL CAUSE

Restrições sobre o N	Apresenta uma situação proposital.
Restrições sobre o S	Nenhuma.
Restrições sobre o N + S	S apresenta uma situação que causa a situação presente em N.
Efeito	L reconhece a situação em S como a causa da situação em N.

SEQUENCE

Restrições sobre o N	Multi-nuclear.
Restrições sobre os N	Sucessão de acontecimentos entre as situações presentes nos N.
Efeito	L reconhece a sucessão de acontecimentos presentes nos N.

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.

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.

CONCESSION

Restrições sobre N	E possui uma consideração positiva sobre a situação em N.
Restrições sobre S	E não está alegando que a situação em S não existe .
Restrições sobre N + S	E percebe uma incompatibilidade aparente entre as situações presentes no N e no S.
Efeito	A consideração positiva do L sobre a situação em N aumenta.

SOLUTIONHOOD

Restrições sobre N	Nenhuma.
Restrições sobre S	Apresenta um problema.
Restrições sobre N + S	A situação presente em N é uma solução para o problema em S.
Efeito	L reconhece que a situação presente em N é uma solução para o problema presente em S.

2.3 Ilustração

A Figura 3 ilustra a estrutura RST completa do Texto 1 já apresentado na seção 2.1. Para recuperar o inter-relacionamento proposicional entre os segmentos indicados no Texto 1, basta recorrer à definição de cada relação RST, dada na seção anterior, e compará-la com o inter-relacionamento proposto na superfície textual, i.e., no próprio Texto 1, apresentado na Seção 2.1.

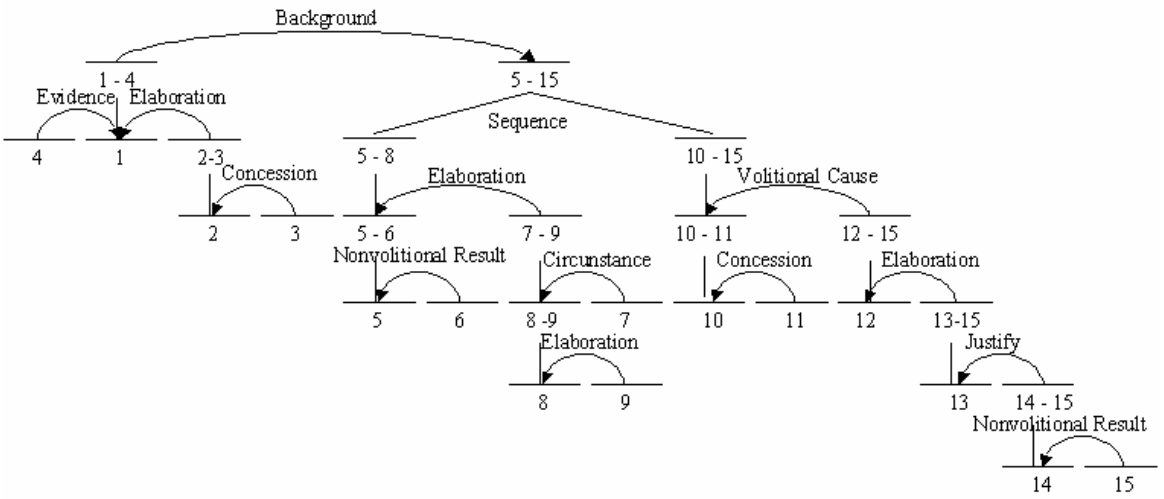

Figura 3: Estrutura RST do Texto 1

3. Propostas de exploração da RST na AS

O grande problema da SA é a busca de um modelo adequado de reconhecimento e manipulação de informações relevantes de um texto, para a produção de seu(s) correspondente(s) sumário(s). 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 (Rino, 1996). Essas características impõem restrições claras à SA, que, incorporadas a metodologias baseadas na RST, permitem delinear a forma como suas relações devem ser interpretadas, para a) determinar os segmentos proposicionais relevantes para compor um sumário; b) determinar o modo como esses segmentos irão se relacionar no texto final. São, basicamente, três as propostas de SA existentes na Lingüística Computacional que foram exploradas neste trabalho, como veremos a seguir.

3.1 Proposta de Sparck Jones

Com relação à modelagem discursiva como base para a sumarização automática, Sparck Jones (1993a) considera 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. Essas três etapas correspondem, assim, aos processos de *análise*, *transformação* e *síntese*, que, juntos, compõem a arquitetura genérica de um sistema de sumarização automática dada pela Figura 4 (Sparck Jones, 1999). 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; o processo de transformação é o processo principal da arquitetura de SA, pois ele é responsável pela geração da representação interna do sumário a partir da representação interna do texto-fonte. O processo de síntese, por sua vez, somente expressa em língua natural a representação interna condensada, produzindo o sumário, propriamente dito.

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 modelá-las. Ela 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.

Mais especificamente, com relação às relações retóricas de um texto, Sparck Jones (1993a) explorou em seu trabalho o fato do núcleo dessas relações fornecer 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. É necessário, assim, 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.

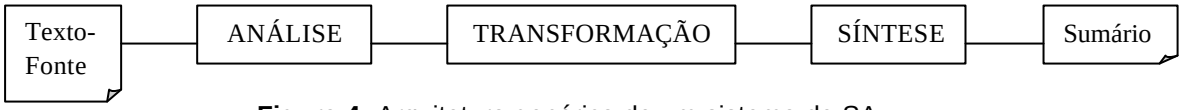

Figura 4: Arquitetura genérica de um sistema de SA

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 conceituais: aqueles com grande saliência são o foco de atenção do discurso e, logo, devem ser incluídos no sumário; os com baixa saliência são periféricos e, logo, são passíveis de exclusão.

Com base em modelo similar ao de Sparck Jones, Marcu (1997a) e Rino (1996) propõem modelos distintos de SA, conforme veremos a seguir. Também a RSTTool, descrita na Seção 4, segue a mesma base teórica de Sparck Jones.

3.2 Proposta de Daniel Marcu

Seguindo a proposta de SA de três passos de Sparck Jones, Marcu (1997a) propõe, como resultado da análise de um texto-fonte, uma representação RST, profunda, conceitual, a ser *transformada* na estrutura conceitual do sumário em construção. Assim, a SA de textos, segundo Marcu, corresponderá a sumarizar sua própria estrutura RST. A fase de síntese, neste caso, consistirá na realização lingüística da estrutura RST condensada em texto, propriamente dito, i.e., em sua expressão lingüística. Esse processo é ilustrado na Figura 5.

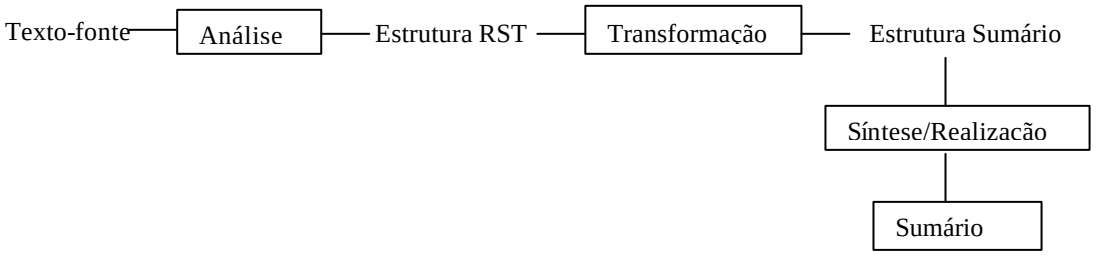

Figura 5: Cenário da SA na proposta de Marcu

A sumarização do Texto 2[3] abaixo (estrutura RST dada na Figura 6), segundo a proposta de Marcu, é descrita a seguir.

Texto 2: ***Using Computers in Manufacturing***

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

[3] In order to counteract this problem, the Manufacturing Management Activity Group of the IProdE is organising a two-day seminar on computers and manufacturing management. [4] The seminar will be held at the Birmingham Metropole Hotel at the National Exhibition Centre from 21-22 March 1979.

[5] It 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 learn of their successes and failures.

[7] The seminar will consist of plenary sessions followed by syndicates [8] where delegates will be arranged into small discussion groups.

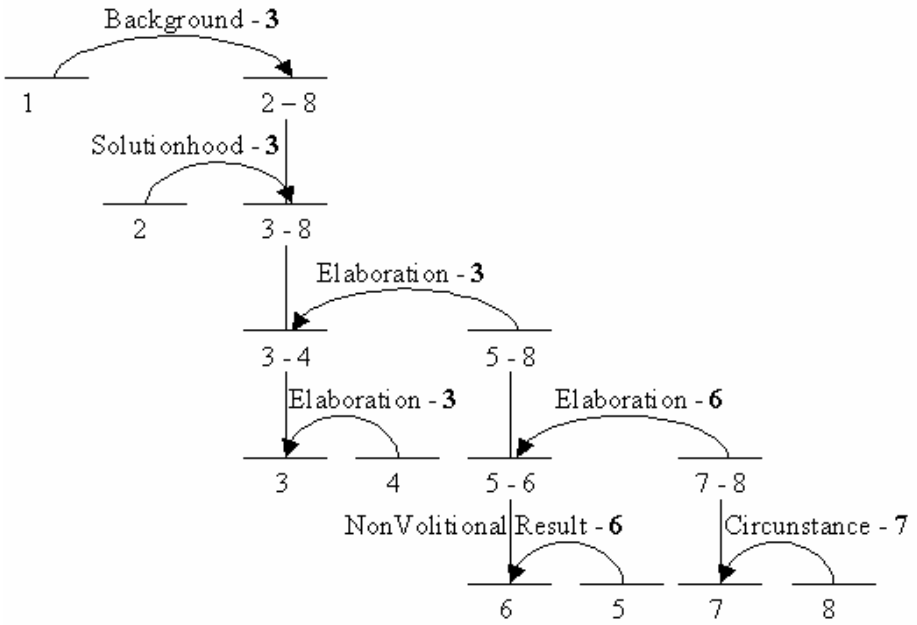

Figura 6: Estrutura RST do Texto 2

A qualidade das estruturas retóricas não é o único fator que pode determinar a melhor forma dos sumários. Uma melhor maneira de tirar vantagens das diferenças entre núcleo e satélite é explorar a noção de saliência que é associada com segmentos textuais, como proposto por Marcu (1997c). Se o conceito de saliência for aplicado a cada nó da estrutura retórica, uma ordenação parcial de todas as unidades importantes do texto pode ser induzida. A premissa atrás dessa abordagem é que unidades textuais que estão mais próximas à raiz da árvore são mais importantes do que aquelas encontradas mais abaixo. Esta observação permite considerar, assim, que sumários construídos com base na posição do segmento, possam ser melhores que aqueles construídos de outro modo.

[3] Texto extraído de Jordan (1980, p.225).

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.

Na estrutura RST do Texto 2, as unidades mais salientes de cada segmento discursivo são indicadas junto aos nomes das relações. A ordem de precedência entre todas as proposições desse discurso é dada por 3>1>2>5>4>7>6>8 ('p1>p2' indica que p1 é mais importante que p2).

Sumários do Texto 2 podem, agora, ser construídos respeitando-se essa ordem: variando-se o número de segmentos a incluir, podemos ter os sumários 1 e 2 (Figura 7), de diferentes tamanhos, para esse texto. O Sumário 1 envolve somente a relação SOLUTIONHOOD entre 2 e 3; o Sumário 2 envolve BACKGROUND entre 1 e 2, e SOLUTIONHOOD[4]. Ambos, no entanto, têm como proposição mais saliente do discurso a solução do problema – 3.

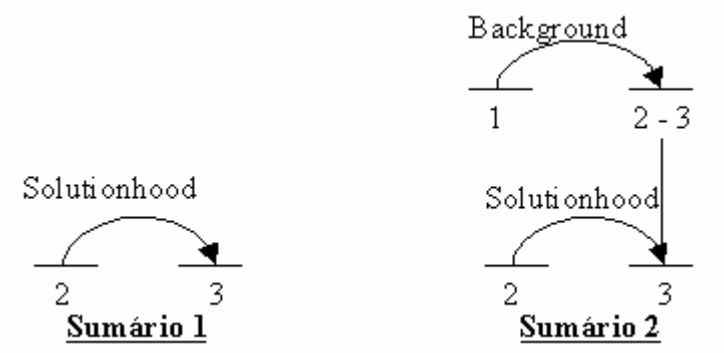

Figura 7: Estruturas RST dos Sumários 1 e 2

Possíveis sumários correspondentes a essas estruturas, resultantes de sua síntese, são ilustrados na Figura 8[5].

Sumário 1

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

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 problem is that the general level of understanding of the power and weaknesses of computers among manufacturing managers is dangerously low. 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 8: Sumários do Texto 2 a partir de suas estruturas RST

Com sua proposta, Marcu comprova que a estrutura retórica, mais especificamente, a estrutura RST, de textos pode ser usada eficientemente na SA, pois

[4] Ver seção 2.2 para obter mais detalhes sobre as relações retóricas.
[5] Sumários gerados por Pardo (2002).

permite determinar suas unidades proposicionais mais salientes. Claramente, a qualidade dos sumários finais é dependente do discurso e, portanto, da qualidade das estruturas RST geradas na fase de análise. Além disso, é dependente também da própria definição das relações RST, cuja semântica nem sempre é clara e não-ambígua. Sob essas considerações, Marcu (1998a) propõe sete métricas para se obter estruturas RST de boa qualidade para os sumários, que podem ser combinadas para um melhor resultado. Essas métricas são usadas na fase de transformação, para a geração da estrutura RST do sumário a partir da estrutura RST do texto-fonte, como seguem:

- Agrupamento de estruturas: uma suposição comum na maioria das teorias textuais, é que textos bons exibem uma estrutura de tópico bem definida, assim, nessa técnica, uma estrutura RST é considerada melhor que outra se apresentar um maior nível de estruturação que combine com os limites de tópicos do texto para o qual foi construído.
- Identificação de relações RST cuja realização implique o uso explícito de marcadores: um plano é considerado melhor que outro se contiver mais relações retóricas que possam ser explicitamente sinalizadas na superfície textual.
- Agrupamento retórico: supõe-se que sempre que existe uma relação retórica entre dois segmentos textuais, e essa relação também existe entre as unidades mais saliente associadas com esses segmentos, dessa forma, essa técnica mede o grau de similaridade entre unidades salientes associadas a dois segmentos textuais.
- Geração de uma estrutura RST do sumário baseada na forma da estrutura correspondente do texto-fonte: em muitos casos, pessoas escrevem textos de forma que as idéias mais importantes apareçam primeiro, quanto mais os escritores acrescentam, mais elaborados ficam os textos que apareceram antes, assim, considera-se que o processamento do texto-fonte acontece, geralmente, da esquerda para a direita, Marcu afirma que uma estrutura RST será melhor do que outra, para o sumário, o se ela for mais enviesada para a direita.
- Identificação de informações relevantes de uma estrutura RST a partir do título do texto-fonte (quando houver): as unidades proposicionais a compor a estrutura RST do sumário deverão ser aquelas associadas aos componentes do título do texto. Tais unidades estarão presentes na estrutura RST do texto-fonte e, assim, corresponderão às unidades mais salientes para compor a estrutura do sumário na transformação da estrutura RST do texto-fonte. Esta métrica pode ser justificada pela própria proposta de Edmundson (1969), que, de modo similar, utiliza as palavras do título do texto a sumarizar como chave para selecionar as sentenças mais relevantes do sumário (lembrar, entretanto, que esta proposta é superficial, no sentido de, simplesmente, extrair do texto-fonte as sentenças, sem alterá-las, enquanto a proposta de Marcu sugere a reescrita total do mesmo).
- Identificação de informações relevantes por sua posição: considerando que textos de gêneros com estruturas estereotipadas permitem identificar mais facilmente as sentenças relevantes por sua localização, Marcu atribui uma pontuação positiva para cada unidade textual que pertença ao começo ou ao final de um parágrafo, p.ex., e uma pontuação negativa, caso contrário. Assim, pelo cômputo global das pontuações dos segmentos proposicionais, Marcu assegura a seleção daqueles mais relevantes para compor o sumário. Esta métrica utiliza também a noção antiga de Luhn (1958), de adotar a localização como um fator determinante da identificação de segmentos textuais relevantes (novamente, na

proposta de Luhn, isto se refere a uma metodologia superficial, em oposição à metodologia profunda de Marcu).

- Identificação dos segmentos proposicionais mais fortemente conectados: uma heurística que é freqüentemente usada em sistemas de sumarização atuais é a de considerar importante às entidades mais altamente conectadas em estruturas semânticas mais ou menos elaboradas, assim, uma pontuação baseada em conexão é atribuída a cada plano, aquele que tiver uma pontuação maior é o melhor.

Outra característica importante que deve ser levada em consideração para se melhorar os sumários produzidos a partir de planos RST, é o modo como as relações são interpretadas. Por exemplo, satélites das relações ELABORATION e BACKGROUND podem ser removidos sem causar nenhum tipo de incoerência no texto, entretanto, outros satélites, como os introduzidos pelas relações VOLITIONAL e NON-VOLITIONAL CAUSE ou VOLITIONAL e NON-VOLITIONAL RESULT podem ser importantes, para o entendimento da mensagem. Sua importância será, assim, dependente do contexto, devendo sua exclusão ser analisada cuidadosamente, para a geração da estrutura RST do sumário correspondente (Macro, 1998b).

3.3 Proposta de Rino

Também com base na noção de nuclearidade da RST e seguindo o modelo delineado por Sparck Jones (Figura 4), Rino (1996) propõe um modelo de produção de discurso geral e independente da língua natural, mas bastante distinto do modelo de Marcu, no sentido de que seu sistema deve gerar uma estrutura RST do sumário pretendido, mas não a partir da estrutura RST do texto-fonte correspondente. O ponto de partida para a geração de sumário, neste caso, é uma representação da mensagem do texto-fonte, composta por um conjunto de três componentes: o objetivo comunicativo (OC), a proposição central (PC) e a base de conhecimento (BC), sendo esta uma representação do conteúdo informativo do texto-fonte, sem características discursivas, as quais são expressas justamente pela associação do OC e da PC a essa BC. Dessa forma, também são delineadas, nessa proposta, as três etapas antes delineadas na Figura 4, mas a SA tem ênfase prioritária nas fases de transformação e síntese, somente.

A noção de nuclearidade, na proposta de Rino, está associada à restrição de que a proposição central do discurso esteja na posição "mais nuclear" da estrutura RST do sumário, isto é, que seja a folha mais à esquerda de sua árvore RST.

A Figura 9 ilustra esse processo, descrito com mais detalhes em (Pardo and Rino, 2001; Pardo, 2002).

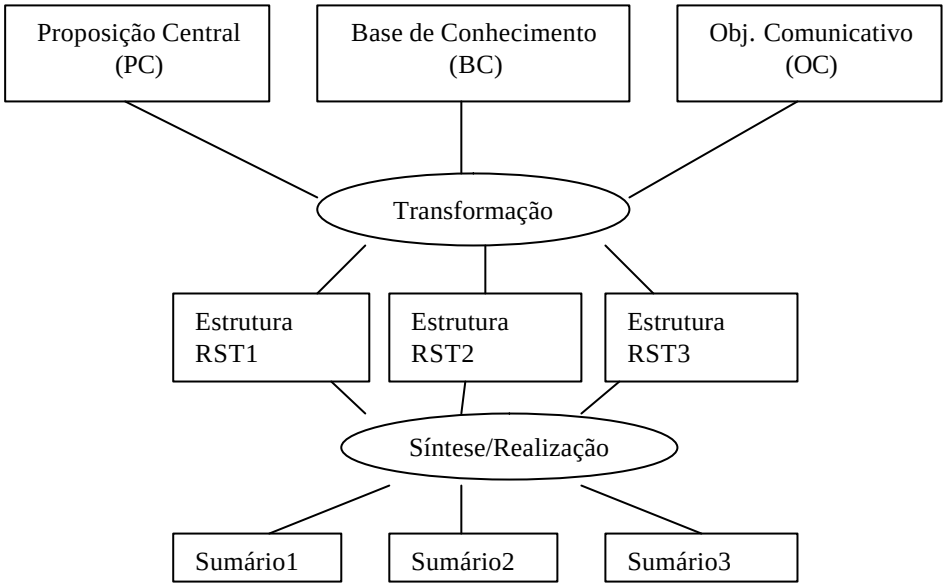

Figura 9: Cenário da SA proposta por Rino

4. A RSTTool

A RSTTool (O'Donnell, 2000)[6] é uma ferramenta robusta que facilita a exploração manual de estruturas retóricas, podendo ser usada tanto para a análise de textos reais, quanto para a obtenção de suas estruturas retóricas (neste caso, com base na RST) ou para sua sumarização. No estudo apresentado a seguir, ela é usada visando essa última característica. Por ser interativa, essa ferramenta admite somente um usuário especialista em RST e, para explorar recursos de SA, também conhecedor das técnicas de SA a partir de estruturas RST.

4.1. Descrição da RSTTool

4.1.1. Interface de Segmentação Textual

O primeiro passo necessário para se começar a análise de um texto utilizando a RST é determinar suas proposições elementares. A ferramenta RSTTool fornece uma interface que facilita essa tarefa. Neste caso, a segmentação pode ser feita de duas maneiras: a) automaticamente, quando a ferramenta considera sentenças do texto-fonte como unidades de significado correspondentes a proposições; b) manualmente, quando o analista RST pode usar a RSTTool para delimitar as proposições sugeridas pelo texto-fonte a partir de sua própria interpretação do mesmo. Dessa forma, há possibilidade de se gerar estruturas RST de diferentes maneiras.

4.1.2. Interface de Estruturação Textual

Para estruturar um texto, gerando sua estrutura RST, a interface de estruturação permite que o usuário conecte as proposições elementares que foram delimitadas anteriormente por meio de relações RST, resultando em sua estrutura retórica (exemplo dessa estrutura já foi dado na Seção 2.3). A ferramenta permite que o usuário indique se deseja utilizar uma relação mono ou multi-nuclear. Devido ao fato de as estruturas

[6] Manual disponível em http://www.wagsoft.com/RSTTool/index.html.

retóricas poderem se tornar muito complexas, a ferramenta disponibiliza a opção de agrupar sub-estruturas, ou sub-árvores RST, por meio de um único nó. Neste caso, este nó é complexo e, na verdade, representa um enredo de proposições inter-relacionadas retoricamente, que não estão representadas explicitamente na estrutura RST principal. Essa opção é dada pela ferramenta "Collapse/Expand", que, devido ao fato de estar diretamente relacionada a questões de SA, é detalhada na seção 4.2.

4.1.3. Interface de Edição de Relações

Nessa interface é possível editar o conjunto de relações retóricas. Relações podem ser adicionadas, excluídas e renomeadas. A ferramenta também disponibiliza conjuntos de relações prontos, entre eles o conjunto original de relações proposto por Mann & Thompson.

4.1.4. Interface de Estatísticas

Usando essa interface, é possível saber com que freqüências às relações são usadas e também o número de núcleos e satélites existentes em uma estrutura retórica. Essa interface é usada, em nosso projeto, durante a documentação, para sabermos quantas relações foram usadas, e quantas proposições há em cada estrutura RST.

4.2. O uso da opção "Collapse/Expand" na SA

Como esta opção permite "esconder" ou encapsular estruturas RST complexas sob um único nó, ela foi usada mais a fundo para a exploração de estruturas RST de sumários dos textos-fonte explorados neste projeto. A seguir, apresentamos um exemplo de sua utilização, para o Texto 3 segmentado em oito proposições (Figura 10). A partir dessa estrutura RST, podemos construir várias estruturas RST condensadas, ou seja, vários sumários poderão ser gerados, considerando a exclusão diversificada dos satélites indicados.

Texto 3[7]:

[1] Inusitadamente, [2] caso a partida de hoje termine empatada, [3a] a Taça Guanabara – [4] que esse ano não é o primeiro turno do campeonato – [3b] não terá um campeão.

[5] Nesse caso, o torneio será decidido quando os clubes se enfrentarem no quadrangular final do campeonato, no dia 17 de abril, [6] devido à Federação não ter colocado no regulamento um critério de desempate.

[7a] Os clubes haviam acertado, antes do torneio, que [8] caso houvesse um empate na final, [7b] seria o campeão quem tivesse a melhor campanha na primeira fase – o Vasco.

[7] Neste texto, a proposição 2 corresponde a "caso a partida de hoje termine empatada" e a proposição 3, a "a Taça Guanabara não terá um campeão." (composta por [3a] e [3b]). O mesmo se aplica a proposição 7, correspondente a "Os clubes haviam acertado, antes do torneio, que seria o campeão quem tivesse a melhor campanha da primeira fase – o Vasco." (composta por [7a] e [7b]) e a proposição 8 "caso houvesse um empate na final"

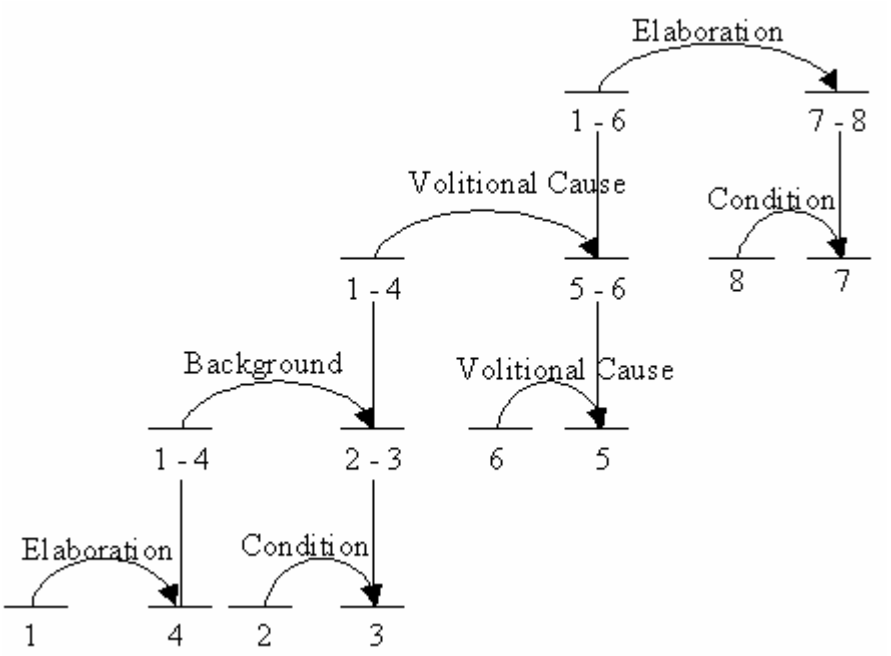

Figura 10: Estrutura RST do Texto 3

Para o Texto 3 pode ser observadas 6 estruturas RST condensadas se retiramos dois satélites de relações simples por vez, nesse caso proposições 1, 2, 6, e 8, a Figura 11 mostra uma possível estrutura RST condensada retirando os satélites 1 e 8, das relações ELABORATION e CONDITION, seu sumário é ilustrado na Figura 12[8].

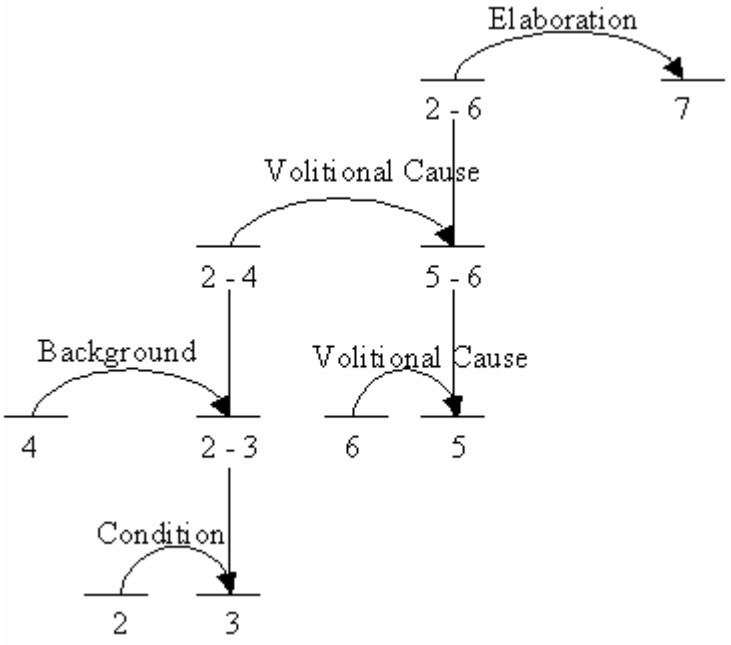

Figura 11: Estrutura RST condensada sem os satélites 1 e 8

[8] Este sumário foi gerado manualmente, a partir da estrutura RST correspondente (foram usadas as definições das relações RST – Seção 2.2 – e foram adequadas as formas superficiais, visando a conectividade entre as proposições).

Caso a partida de hoje termine empatada, a Taça Guanabara – que esse ano não é o primeiro turno do campeonato – não terá um campeão.

Nesse caso, o torneio será decidido quando os clubes se enfrentarem no quadrangular final do campeonato, no dia 17 de abril, devido à Federação não ter colocado no regulamento um critério de desempate.

Os clubes haviam acertado, antes do torneio, que seria o campeão quem tivesse a melhor campanha na primeira fase – o Vasco.

Figura 12: Possível sumário para a estrutura RST da Figura 11

É possível calcular, a partir da estrutura RST de um texto-fonte, quantos sumários é possível gerar somente pela exclusão dos satélites das relações RST, sob as premissas indicadas por Marcu, de que satélites podem ser excluídos para a SA, sem perda do significado essencial. Usando essa idéia, é possível combinar satélites base e também aqueles que estão a um nível acima na estrutura RST. A fórmula matemática de combinação foi utilizada para esse fim, resultando no número total de estruturas RST condensadas dado por:

$$C_{n,k} = \frac{n!}{k!\,(n-k)!},\ n>=k$$

para: n = número de satélites existentes na estrutura RST
k = número de satélites restantes depois da poda.
C = número de estruturas RST condensadas que podem ser formadas

Utilizamos essa fórmula para gerar vários sumários para um mesmo texto-fonte, como ilustrado anteriormente.

5. Considerações finais

Nesse relatório, foi apresentado um estudo da RST visando a SA, pelo uso da ferramenta de estruturação textual RSTTool. Algumas propostas do uso da RST na SA foram também apresentadas, especialmente em função de uma ferramenta particular – a "Collapse/Expand".

Como se pode notar, há várias propostas de exploração da Teoria RST para a SA, cada uma sob uma perspectiva diferente.

Nos experimentos realizados por Marcu, um texto é analisado e sua estrutura RST é construída. A partir dela, usando a noção de nuclearidade são construídos os sumários do texto-fonte, ou seja, Marcu usa a estrutura RST de um texto-fonte para a produção de estruturas RST de sumários, a partir das quais se podem gerar as formas superficiais, textuais, dos sumários, propriamente ditos.

Rino, por sua vez, não parte de estruturas RST de textos-fonte, mas contempla a geração de estruturas RST dos sumários correspondentes a textos-fonte. Portanto, diferentemente do modelo de Marcu, as estruturas RST de Rino são somente a saída de um sistema de SA.

Tanto uma proposta quanto a outra seguem a sugestão de Sparck Jones, de que são necessárias três etapas básicas na SA: a construção de uma representação do significado a partir do texto-fonte, a geração da representação conceitual do sumário correspondente e a sua síntese, ou realização lingüística, resultando no sumário, propriamente dito. Nos exemplos ilustrados neste relatório, concentramo-nos na fase de

transformação, que é a principal, nesse modelo. Os resultados ilustrados foram gerados manualmente, mas poderiam ter sido fruto de um processo automático de realização superficial das estruturas RST condensadas dos sumários.

Referências Bibliográficas

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

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

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

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

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

Marcu, D (1997a). *The rhetorical parsing, summarization, and generation of natural language texts*. Ph.D. Dissertation, Department of computer Science, University of Toronto.

Marcu, D. (1997b). From Discourse Structures to Text Summaries. *In the Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization,* pp. 82-88. Madrid, Spain, July 11.

Marcu, D. (1997c). The Rhetorical Parsing of Natural Language Texts. *In the Proceedings of the ACL'97/EACL'97,* pp. 96-103. Madrid, Spain, July 7-10.

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

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

Marcu, D. (1999). Discourse trees are good indicators of importance in text. In I. Mani and M. Maybury (eds.), *Advances in Automatic Text Summarization*, pp. 123-136. The MIT Press.

Marcu, D. (2000). The Rhetorical Parsing of Unrestricted Texts: A Surface-based Approach. *Computational Linguistics*, Vol. 26, No. 3, pp. 395-448. September.

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

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

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

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. Abril

Rino, L.H.M. e Nunes, M.G.V. (2002). Geração de textos e sumários. *In R.. Vieira. e V.L.S. de Lima (eds), Engenharia da Linguagem: uma introdução ao tratamento computacional da língua,* Cap. 3, Parte II: Aplicações (em publicação).

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 G. Knorz; J. Krause and C. Womser-Hacker (eds.), *Information Retrieval* 93, pp. 9-26. Universitatsverlag Konstanz, June.

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.