ESTRUTURAÇÃO RETÓRICA: UM MODELO DE REPRESENTAÇÃO DO CONHECIMENTO[*]

Thiago Ianez CARBONEL (DL-UFSCAR) [*]
Lucia Helena Machado RINO (DC-UFSCAR)[†]

RESUMO: Neste artigo apresentamos a Teoria de Estruturação Retórica (RST), que permite segmentar um texto em unidades elementares de significado, ou proposições. Em um texto coeso, essas unidades devem manter uma relação coerente entre si para a construção da tecitura textual. As relações RST definidas na Teoria permitem construir as estruturas RST subjacentes aos textos. Elas se aplicam tanto às unidades elementares, quanto a sub-estruturas RST, sendo, portanto, composicionais. A Teoria também indica como diferenciar informações nucleares e satélites, isto é, segmentos que tenham diferentes graus de relevância. Devido a essa característica, ela é usada para modelar processos na Sumarização Automática, como ilustramos neste artigo.

ABSTRACT: In this article we present the Rhetorical Structure Theory (RST), which allows segmenting a text into its elementary units, or propositions. In a cohesive text such units must be coherently inter-related for texture. The provided RST relations allow building the underlying RST structures of texts. They apply both to elementary units and RST sub-structures, being thus compositional. RST also signals how to differentiate nuclear and satellite information, that is, those segments that convey diverse degrees of relevance. Due to this, the theory is used to model computer processes in Automatic Summarization, as we shall see in this article.

1. Introdução

Com a crescente expansão das vias digitais de difusão da informação (em especial a Internet), tem sido cada vez maior a demanda por formas alternativas de condensação dos documentos distribuídos eletronicamente. A idéia de se produzir, a partir de um texto mais elaborado, uma versão condensada contendo apenas as informações centrais se coaduna perfeitamente com esta tendência global, especialmente quando esta tarefa pode ser automatizada.

A Sumarização Automática de textos (doravante SA) consiste na atividade de geração automática de uma versão condensada de um dado texto-fonte (Mani, 2001). Assim como na sumarização humana, um bom sumário deve preservar a idéia principal de seu texto-fonte, articulando em torno dela as informações centrais contidas no mesmo. Os sumários gerados atualmente, porém, ainda estão distantes dos resultados obtidos por humanos e várias características do atual estágio do processamento automático remetem a problemas que se devem às duas abordagens clássicas da SA: a baseada em conhecimento lingüístico – abordagem profunda – e a baseada em informações estatísticas ou empíricas – abordagem superficial.

A diferença entre as metodologias está na incorporação ou não de processos computacionais que façam uso de modelos reais de processamento de línguas naturais. Na abordagem profunda, modelos lingüísticos são utilizados para a seleção da informação contida no texto a fim de se gerar o sumário, por exemplo, modelos de estruturação do discurso (Van Dijk, 1979) ou modelos de progressão temática (Hutchins, 1987), que podem indicar graus diversos de relevância das unidades informativas do texto e, assim, permitir sua seleção para compor os sumários. Na abordagem superficial, os modelos são empíricos, matemáticos ou estatísticos, e possuem a vantagem de poder contar com processamento computacional mais independente do conhecimento lingüístico. Em outras palavras, as formas como os sistemas de SA lidam com o significado e com a representação do conhecimento textual variam em ambas as abordagens. Segundo Seno e Rino (2005), "a dicotomia entre as duas abordagens é crucial para a complexidade e computabilidade dos sistemas em construção". Se, por um lado, a abordagem superficial oferece ganhos em decorrência do aparato computacional já existente, o processamento computacional pode ser enriquecido com conhecimento profundo, possibilitando, assim, uma aproximação da mimetização da atividade humana de processamento lingüístico.

[*] Trabalho desenvolvido no contexto do Projeto ProCaCoSa (no. 50703020044), com apoio financeiro do CNPq.
* thiagocarbonel@gmail.com
† lucia@dc.ufscar.br

Embora a modelagem profunda seja mais complexa e, via de regra, leve a um processamento mais demorado e nem sempre viável, dependendo da aplicação em foco, ela é importante para assegurar critérios de textualidade, o que dificilmente se reproduz com a abordagem superficial. Muitos são os problemas encontrados nos sumários automáticos, que podem ser assim elencados: presença de informações redundantes ou irrelevantes, agramaticalidade e déficits de textualidade em geral, que remetem aos problemas distintos de coerência e coesão, A raiz destes problemas remete às duas abordagens: se superficial, não há conhecimento suficiente para tratar de alguns desses fenômenos; se profunda, o problema maior que se apresenta é a modelagem e representação do conhecimento, antes mesmo de se analisar a eficiência do sistema final. Dos sumarizadores disponíveis hoje, poucos utilizam níveis profundos de conhecimento lingüístico. Destacam-se, particularmente, os trabalhos de Marcu (1997a, 1997b, 2000) como os mais representativos na área, que utilizam a RST - *Rhetorical Structure Theory* (Mann & Thompson, 1987). Na mesma esteira, utilizando a RST aliada à determinação do domínio de acessibilidade referencial (*Veins Theory* – Cristea et al., 1998) com vistas à manutenção dos elos co-referenciais, podemos citar o RHeSumaRST (Seno, 2005).

Neste contexto insere-se o Projeto ProCaCoSa (PROcessamento de Cadeias de CO-referência para a Sumarização Automática de Textos em Português), que visa analisar e diagnosticar problemas causados à Sumarização Automática pela ocorrência de cadeias de co-referência não resolvidas durante a seleção e estruturação do conteúdo de sumários, ou perda da referência por conta de quebras de elos referenciais nos sumários, antes assegurados no texto-fonte. Para investigar esse tipo de déficit de textualidade, neste projeto utiliza-se o modelo da RST, quer na fase de estruturação dos textos completos, quer na fase de seleção e estruturação de seus sumários.

No texto que segue, o contexto de aplicação esboçado acima será apresentado na Seção 4. Antes, porém, descrevemos a RST na Seção 2 e, em seguida, apresentamos as propostas de exploração da RST em SA (Seção 3). Perspectivas futuras são apresentadas na Seção 5.

2. A Teoria RST

A RST é uma teoria relativamente nova que, apesar de desenvolvida para fins lingüísticos (Mann & Thompson, 1987), teve absorção expressiva na Lingüística Computacional, sobretudo para o processamento automático do inglês (Marcu, 1997a; Sporleder & Lascarides, 2005) e, mais recentemente, também do português (Pardo, 2005; Seno, 2005). Todavia, trabalhos recentes (Desiderato Antônio, 2004) apontam para um resgate desse modelo teórico dentro da Lingüística, devido ao seu potencial de estruturação textual.

A RST fundamenta-se no princípio de que um texto tem uma estrutura retórica subjacente à estrutura superficial. Através dessa estrutura retórica, é possível recuperar o objetivo comunicativo que o escritor do texto pretendeu atingir ao escrevê-lo. Embora esses pressupostos da RST remetam ao aspecto discursivo, seu uso se limita à Lingüística Textual, ou seja, discurso somente enquanto texto. Este pode ser segmentado em unidades mínimas de significado (ou conteúdo) denominadas EDUs – *Elementary Discourse Units* – que, necessariamente, mantêm relação entre si na construção textual. Suas relações são previamente definidas pelo modelo, cujo conjunto, apesar de não definitivo, pretende ser suficientemente amplo para cobrir os casos retóricos considerados. Assim, as relações RST são divididas em duas classes: hipotáticas e paratáticas (Marcu, 1997a). As relações hipotáticas inter-relacionam pares de EDUs que apresentam diferentes graus de importância, sendo uma nuclear e a outra satélite. Essas relações denominam-se mononucleares. As relações paratáticas inter-relacionam EDUs que apresentam o mesmo grau de importância e são denominadas relações multinucleares.

Como forma de ilustrar os dois tipos de relação RST, vejamos os exemplos apresentados nas Figuras 1 e 2, para as sentenças (A) e (B) respectivamente. Os números entre colchetes nessas sentenças indicam a delimitação de cada EDU e os ramos em negrito nas estruturas RST, a informação nuclear (em oposição à satélite). Podemos observar em (A) que a EDU 2 introduz o propósito (*purpose*) da EDU 1, sendo esta o núcleo e a outra, o satélite da relação PURPOSE. Em (B) temos uma relação de seqüência (*sequence*) entre os segmentos que não é hierárquica e, portanto, é multinuclear. As definições dessas relações na Teoria RST são apresentadas nas Figuras 3 e 4, respectivamente[1]. Optou-se, neste trabalho, por não traduzir os nomes da relações, haja vista a prática comum em outros trabalhos na área de Lingüística Computacional.

[1] Definições extraídas de Pardo (2005), são traduções das originais em inglês (N: núcleo; S: satélite).

Figura 1. Sentença (A) e sua estrutura RST

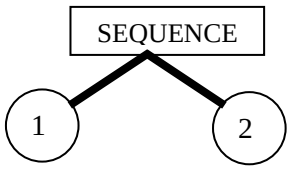
(B) Geraldo Alckmin visitou uma escola na periferia de São Paulo [1] e depois encontrou seus assessores no comitê central. [2]	

Figura 2. Sentença (B) e sua estrutura RST

Nome da relação: PURPOSE
Restrições sobre N: apresenta uma ação **Restrições sobre S**: apresenta uma situação não realizada **Restrições sobre N+S**: S apresenta uma situação que pode realizar N **Efeito**: o leitor reconhece que a atividade em N pode ser iniciada por meio de S

Figura 3. Definição da relação PURPOSE

Nome da relação: SEQUENCE
Restrições sobre os Ns: as situações apresentadas nos Ns são realizadas em seqüência **Efeito**: o leitor reconhece a sucessão temporal dos eventos apresentados

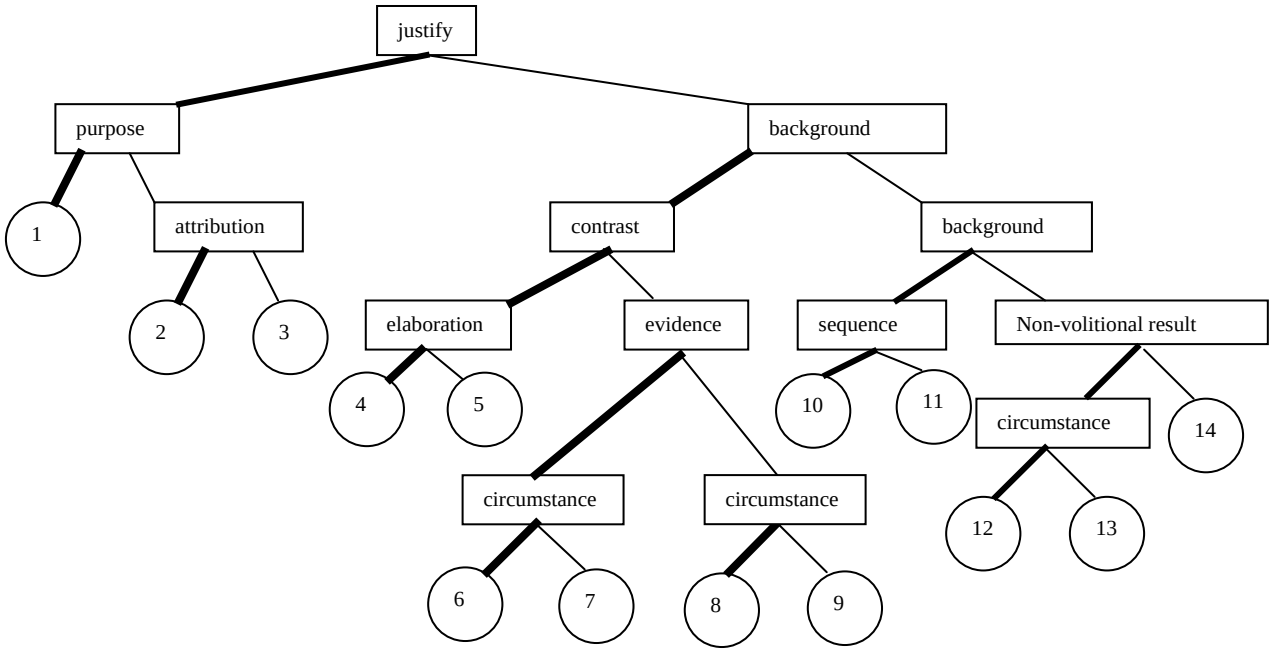
Como mostram os exemplos acima, a estruturação RST resulta em uma árvore e pode usar quaisquer relações do conjunto definido. Embora eles envolvam somente EDUs, isto é, unidades elementares do texto em foco, uma árvore RST é construída composicionalmente, ou seja, relações se estabelecem também entre subárvores RST, como mostra a estrutura RST para o Texto 1 exibida na Figura 5. Esse texto, extraído da Folha On-line (07/11/2006), foi analisado por um especialista em RST e sua estrutura foi construída manualmente, com o auxílio de uma ferramenta automática, a RSTTool (O´Donnel, 2000).

Figura 4. Definição da relação SEQUENCE

[1] O governo venezuelano proibirá o uso de árvores de Natal, imagens de Papai Noel ou botas vermelhas na decoração das repartições públicas. [2] O objetivo é erradicar uma tradição que tem origem nos Estados Unidos, [3] noticiou nesta terça-feira o jornal "Ultimas Noticias".
[4] O governo de Hugo Chávez quer [5] que as decorações natalinas sejam inspiradas na tradição do país, com presépios ou ornamentos feitos com a planta tropical conhecida como Flor do Natal.
[6] Há décadas [7] as comemorações natalinas na Venezuela combinam as duas tradições de presépio e pinheiro. [8] No dia 25 de dezembro, [9] as crianças ganham presentes tanto do Menino Jesus quanto do Papai Noel.
[10] O governo venezuelano mantém um forte confronto verbal com o americano [11] e sustenta um discurso permanente contra o imperialismo. [12] Em 1º de novembro passado, [13] o presidente Chávez antecipou o pagamento de três abonos de Natal para cerca de um milhão de funcionários públicos, aposentados e pensionistas, [14] o que provocou uma onda de compras natalinas.
[15] A Venezuela combina neste ano o ambiente natalino com o eleitoral, [16] com vista à eleição de 3 de dezembro próximo, [17] na qual Chávez pretende se reeleger para um novo mandato de seis anos.

Texto 1. O Natal na Venezuela.

Figura 5. Estrutura RST do Texto 1.



Inicialmente, os autores da teoria – Mann & Thompson (1987) – estabeleceram um conjunto de apenas 24 relações, as quais são exibidas na Figura 6, juntamente com sua nuclearidade. Outros autores, como Marcu (1997a), propõem conjuntos bastante diversos desse, remontando a mais de cem relações, o que torna um processo de análise bastante mais complexo. Como se pode notar na figura 5, algumas dessas relações têm significado evidente, já que são familiares a um falante competente, como a de propósito, evidência ou contraste. Entretanto, outras são mais obscuras, como ENABLEMENT e BACKGROUND. O leitor deve recorrer à obra de referência para entendê-las e recuperar seu contexto de uso.

Figura 6. Conjunto original de relações RST

Relação	**Mononuclear**	**multinuclear**	**Relação**	**Mononuclear**	**multinuclear**
ANTITHESIS	X		JUSTIFY	X	
BACKGROUND	X		MOTIVATION	X	
CIRCUMSTANCE	X		NON-VOLITIONA CAUSE	X	
CONCESSION	X		NON-VOLITIONAL RESULT	X	
CONDITION	X		OTHERWISE	X	
CONTRAST		X	PURPOSE	X	
ELABORATION	X		RESTATEMENT	X	
ENABLEMENT	X		SEQUENCE		X
EVALUATION	X		SOLUTIONHOOD	X	
EVIDENCE	X		SUMMARY	X	
INTERPRETATION	X		VOLITIONAL CAUSE	X	
JOINT		X	VOLITIONAL RESULT	X	

Para construir árvores RST como as ilustradas, o analista deve, primeiramente, reconhecer os tópicos principais de um texto e, assim, sua idéia principal, a fim de traçar o relacionamento retórico mais indicado entre os elementos macro-estruturais. Entretanto, a construção da estrutura se dá primeiramente pela segmentação do texto em EDUs e pelo seu relacionamento, ou seja, pela construção de subárvores simples. Desse modo, tanto a recuperação da macro quanto a recuperação da micro-estrutura são relevantes. Ao mesmo tempo, dados os pressupostos da Teoria RST, a tarefa de análise visa recuperar as intenções do escritor. Assim, o analista deve ser, ao mesmo tempo, um leitor competente e um especialista na representação do conhecimento, para elaborar sua tarefa de modelagem artificial a fim de produzir uma estrutura RST segundo os pressupostos da Teoria. O conhecimento requerido envolverá padrões de análise,

conhecimento morfossintático, reconhecimento de marcadores textuais e sua correspondência com as relações retóricas. Entretanto, a descoberta de padrões de análise é altamente dependente da apreensão da mensagem, ou idéia principal, do texto, assim como de sua organização macro-estrutural. É por esse motivo que o conhecimento do próprio analista, usuário da teoria e leitor competente, se torna essencial. Em geral, os padrões indicativos das relações RST são dependentes de gênero e domínio textual: para sua determinação é preciso reconhecer a ordem das proposições (sejam elas EDUs ou segmentos textuais mais complexos), a qual será determinante da nuclearidade, isto é, do reconhecimento das unidades que serão núcleos e satélites, assim como de seu relacionamento funcional (determinação da relação retórica, propriamente dita).

A anotação[2] RST consiste, portanto, na recuperação do mapeamento de uma situação em língua natural, previamente elaborado pelo escritor. Nesse sentido, a situação espelha o modo e as razões de haver usos particulares da língua natural. Assim, tem-se por hipóteses que: a) a língua natural e a situação discursiva levam ao *efeito do discurso no leitor* – aqui é possível descobrir, por exemplo, porque os usos particulares da língua natural podem ter sucesso ou falhar ante o leitor; b) o escritor deseja provocar, com seu texto, efeitos particulares no leitor; c) o texto é fundamentado, portanto, nas intenções do escritor, conforme evidenciado pelas figuras 3 e 4.

O que dificulta a estruturação RST é a variedade de estruturas textuais e, certamente, a ambigüidade das intenções de um produtor e mesmo do receptor. Além disso, as próprias definições das relações RST são ambíguas, podendo levar a várias árvores RST para um mesmo texto. Para a estruturação, são considerados ainda os tipos básicos de estrutura textual, a saber: estrutura holística, relacional e sintática. A estrutura holística indica as propriedades do gênero ou variedade textual, que, por sua vez, remetem à macro-estrutura textual; a relacional trata das estruturas linear (coesão) e reticulada (coerência) do texto. Por fim, a estrutura sintática simplesmente reflete a organização textual e discursiva. Claramente, esses tipos envolvem todas as condições a que se recorre intuitivamente para se recuperar o aspecto retórico de um texto.

Ao mesmo tempo em que a RST contempla essa variedade estrutural, ela também considera três aspectos funcionais distintos ao definir suas relações: o funcional, o apresentativo e o informativo (ou representacional). As relações funcionais são as relações de fato retóricas, ou discursivas, pois remetem aos efeitos que causam no leitor a partir dos objetivos do escritor, das suposições que ele faz sobre o leitor ou do uso de padrões específicos do domínio ou da audiência pressuposta. As apresentativas são as que servem de instrumento para a apresentação do assunto, ou conteúdo textual. Finalmente, as relações informativas espelham o meio (código ou instrumento) para transmitir a mensagem.

3. Propostas de exploração da RST na AS

Um dos grandes desafios da SA é a busca de um modelo adequado de representação do conhecimento lingüístico de um texto que possibilite a identificação da informação relevante para a produção de seu sumário. É importante notar que duas características são essenciais nesse contexto: a) sumários remetem, necessariamente, a textos-fonte; 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 determinar os segmentos proposicionais relevantes para compor um sumário e, ao mesmo tempo, determinar o modo como eles irão se relacionar no texto final. São basicamente três as propostas de SA significativas para este trabalho, como veremos a seguir.

3.1. Proposta de Sparck Jones

Sparck Jones (1993) considera três etapas básicas para a sumarização automática: 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. 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 identificação da informação relevante e geração da representação interna do sumário a partir da representação interna do texto-fonte. O processo de síntese, por

[2] Trata-se do termo usual para o processo de análise com vistas à produção de um outro texto, o anotado com informações retóricas.

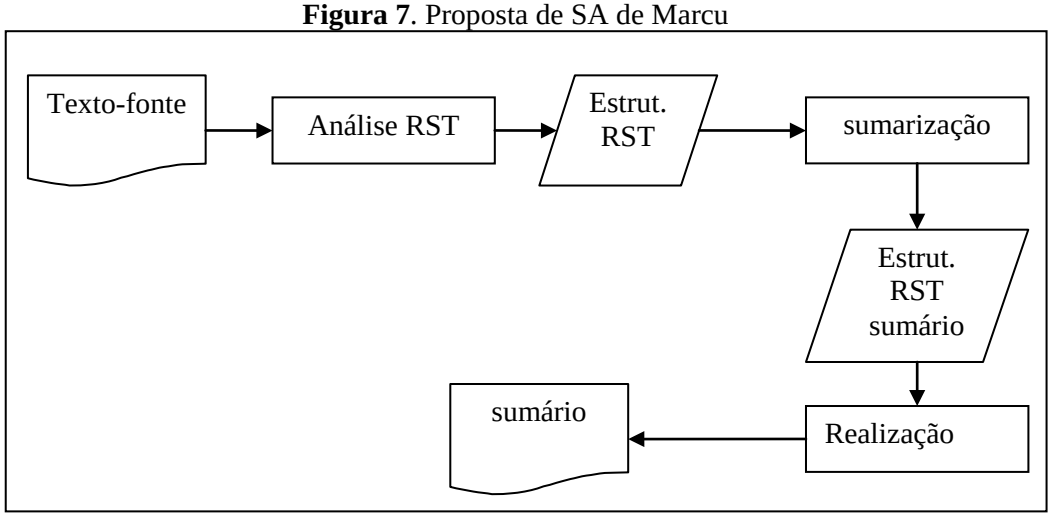
sua vez, somente expressa em língua natural a representação interna condensada, produzindo o sumário, propriamente dito. Três tipos de informação devem ser contemplados, segundo ela: o *lingüístico*, o *informativo* (ou de domínio) e o *comunicativo*, os quais remetem a questões semânticas e pragmáticas, indicando que todo texto tem um objetivo comunicativo que deve ser levado em consideração durante a geração de um sumário. Especialmente, essas informações servirão para indicar a importância da proposição central do texto na escolha das informações que comporão o sumário.

Sparck Jones explorou o fato de núcleos de relações RST fornecerem informações mais importantes que seus satélites, sendo úteis, portanto, para a identificação das unidades mais relevantes para compor um sumário. Entretanto, somente a escolha de informações nucleares não basta para produzir bons sumários e, assim, outros critérios para assegurar sua qualidade se fizeram necessários. Foi com base em sua proposta que os demais trabalhos se desenvolveram, embora explorando modelos distintos de SA.

3.2. Proposta de Daniel Marcu

Seguindo a idéia de SA em três passos de Sparck Jones, Marcu (1997a) propõe a representação RST de um texto-fonte como base para a estruturação do sumário. Assim, seu modelo de SA de um texto corresponde à sumarização de sua própria estrutura RST. A fase de síntese, neste caso, consiste na realização lingüística da estrutura RST. Esse processo é ilustrado na Figura 7. Para a escolha dos segmentos textuais do sumário, entretanto, Marcu agrega ao modelo estrutural da RST o modelo de saliência. A saliência de segmentos textuais é definida como a medida de sua proeminência relativa: segmentos com alta saliência são o foco de atenção no discurso e, logo, devem ser incluídos no sumário; segmentos com baixa saliência são periféricos e, logo, são passíveis de exclusão. Marcu usa a própria estrutura RST do texto-fonte para indicar o grau de saliência de todas as suas unidades. Desse modo, sua proposta de SA conjuga ambos os modelos, o da RST e o de saliência, buscando usar a estrutura retórica mais eficientemente do que simplesmente considerar a nuclearidade como fator determinante de escolhas informativas e estruturais para a SA.

Figura 7. Proposta de SA de Marcu



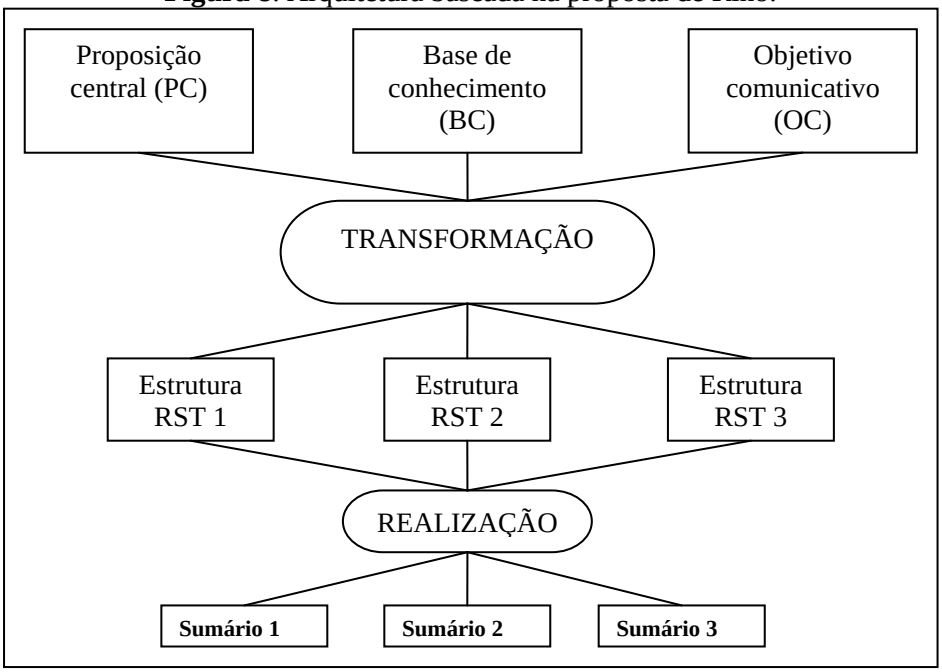
Mesmo essa proposta provou ser problemática, pois a qualidade dos sumários finais é dependente do discurso e, portanto, da qualidade das próprias estruturas RST, agora geradas automaticamente na fase de análise.

3.3. Proposta de Rino

Também com base na noção de nuclearidade da RST, Rino (1996) propõe um modelo de produção de discurso geral e independente da língua natural bastante distinto do modelo de Marcu, que agrega informações em níveis distintos do discurso explicitamente indicadas como complementos da estruturação retórica, quer do texto-fonte, quer do sumário correspondente. O ponto de partida para ambas as estruturas é a representação da mensagem do texto-fonte como um conjunto de três componentes explícitos: 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. A explicitação desses componentes remete às três etapas antes

delineadas por Sparck Jones. A noção de nuclearidade, neste caso, 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 8 ilustra esse processo, descrito com mais detalhes em (Pardo & Rino, 2001; Pardo, 2002).

Figura 8. Arquitetura baseada na proposta de Rino.



4. O contexto do Projeto ProCaCoSA: O sumarizador RHeSuma-2

A RST, conforme já vimos nas seções anteriores, oferece amplas vantagens ao processamento de textos com vistas à sumarização automática, tanto para a identificação de segmentos relevantes para compor um sumário, quanto para sua estruturação, desde que o texto-fonte também tenha sua árvore RST construída (e, portanto, tenha passado por um processo de análise). Além dessas vantagens, a existência de outros modelos de SA que fazem uso da RST, como os apontados, e de um analisador discursivo especialmente voltado ao português – o DiZer (Pardo, 2005) – também contribui para a continuidade da exploração de métodos ricos em conhecimento lingüístico, que é o contexto do Projeto ProCaCoSA.

Nesse projeto utilizamos a nuclearidade como critério para a SA, porém, complementando-a para a resolução de um problema bastante particular: o de manutenção de cadeias de co-referência em um sumário toda vez que um segmento anafórico for escolhido para compô-lo. Neste caso, a seleção exclusiva de núcleos também é insuficiente, pois pode introduzir a chamada quebra de cadeias de co-referência, resultando, assim, em problemas de tecitura. Uma forma de se tentar melhorar essa tecitura nos sumários gerados automaticamente é introduzir aqueles satélites reconhecidamente úteis para esse fim, ou seja, os satélites antecedentes de anáforas.

O modelo de sumarização automática do Projeto ProCaCosa (Carbonel et al., 2006) agrega no sistema RHeSuma-2 dois sistemas automáticos independentes: um gerador das estruturas RST dos textos-fonte, ou analisador retórico, chamado DiZer (Pardo, 2005) e um sumarizador dessas estruturas, o RHeSumaRST (Seno, 2005), seguindo a mesma arquitetura da Figura 7.

O DiZer tem como entrada o texto limpo, independente de uma forma de representação do conhecimento que opere como linguagem de interface entre o sistema e o texto, e como resultado do processamento, estruturas arbóreas RST. Podemos identificar no processo desempenhado pelo DiZer duas etapas de especial interesse, executadas seqüencialmente: i) a segmentação do texto fonte; e ii) a identificação das relações retóricas no texto segmentado.

O RHeSumaRST, por sua vez, é um sumarizador automático que opera somente sobre a representação RST e, portanto, é independente da forma superficial do texto-fonte. Ele gera estruturas RST condensadas a partir das árvores retóricas dos textos, com a proposta central de manter as cadeias de co-referência ante a escolha de uma EDU anafórica. O sistema apresenta três aspectos importantes: i) identifica as veias referenciais nas árvores de entrada, usando como modelo a Teoria das Veias (Cristea et al., 1998), para evitar

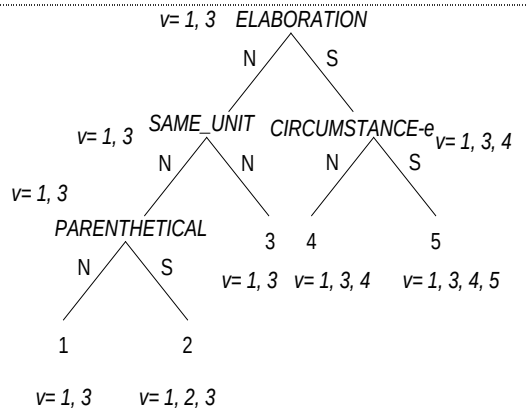
a quebra das cadeias referenciais; ii) calcula a saliência dos nós das árvores, usando o modelo de saliência de Marcu (1997a), para estabelecer uma ordem de saliência entre as EDUs e, assim, permitir que EDUs mais salientes sejam preferidas para compor os sumários; e iii) usa a estrutura retórica para estabelecer o vínculo entre unidades salientes e suas veias.

Em suma, o RHeSumaRST identifica, para cada EDU, o conjunto de outras EDUs que possam ser suas complementares, particularmente com foco em seu domínio de acessibilidade referencial. Heurísticas de reestruturação RST para os sumários são aplicadas considerando, inicialmente, as EDUs mais salientes e, posteriormente, sua complementação visando evitar a quebra de cadeias de co-referência.

A Figura 9 ilustra uma árvore RST com suas veias (v) anotadas (Seno, 2005). É importante ressaltar que a Teoria das Veias, aqui, não garante que essa anotação indique corretamente os possíveis candidatos a termos antecedentes de uma anáofra. Ou seja, a determinação das veias de uma árvore RST não é um processo de resolução anafórica automática, pois baseia-se somente nas informações estruturais da árvore. Tomando o segmento 4 do texto ilustrado nessa figura à guisa de exemplo – [***Os ganhos*** são atribuídos pela diretoria **da fábrica** à nova filosofia que vem sendo implantada **na empresa** desde outubro do ano passado] – vemos que a descrição definida [a fábrica] está em posição anafórica, tendo por antecedente [A empresa Produtos Pirata Indústria e Comércio Ltda], presente no segmento 1. A Teoria das Veias assegura que a veia do segmento 4 – v=1,3,4 – indica os possíveis antecedentes da anáfora: eles estarão nos segmentos 1, 3 ou 4 e somente nestes. Como o RHeSumaRST não resolve anáforas automaticamente, ele considera a hipótese de, uma vez saliente o segmento 4 para compor o sumário, também serão selecionados os demais segmentos de sua veia.

Figura 9. Texto 2 e sua estrutura RST anotada com veias

[1] **A empresa Produtos Pirata Indústria e Comércio Ltda**., de Contagem [2] (na região metropolitana de Belo Horizonte), [3] deverá registrar este ano ***um crescimento de produtividade*** nas suas áreas comercial e industrial de 11% e 17%, respectivamente. [4] ***Os ganhos*** são atribuídos pela diretoria **da fábrica** à nova filosofia que vem sendo implantada **na empresa** desde outubro do ano passado, [5] quando **a Pirata** se iniciou no Programa Sebrae de Qualidade Total.



O problema dessa abordagem é introduzir mais informação do que o necessário para impedir quebras de co-referência e, assim, ter sumários menos concisos ou, até, com mais problemas de tessitura, que é o que se pretende evitar. Além disso, o RHeSuma-2 também é dependente de gênero e domínio textuais, sobretudo devido ao DiZer, que toma por base marcas estruturais e lingüísticas para gerar as árvores RST dos textos-fonte. Conforme ilustram as figuras 10 e 11, na análise dos sumários produzidos a partir de um corpus de 47 textos jornalísticos curtos, de aproximadamente 200 palavras, alguns problemas foram evidenciados (Carbonel et al., 2006).

Figura 10. Texto 3

Homens e mulheres de 50 anos ou mais deveriam tomar uma aspirina por dia para reduzir o risco de ataques cardíacos e derrames, segundo estudo realizado pelo médico Peter Elwood, da Universidade de Cardiff, no Reino Unido.

O estudo, que será publicado no "British Medical Journal" , avaliou 2.500 homens por 25 anos e concluiu que as pessoas de meia idade têm risco alto o suficiente para se beneficiar do uso do medicamento. Elwood defende a ampliação do uso da aspirina.

Figura 11. Sumário do Texto 3

Homens e mulheres de 50 anos ou mais deveriam tomar uma aspirina por dia para reduzir o risco de ataques cardíacos e derrames,O estudo, que será publicado no British Medical Journal,

A partir da comparação do sumário com seu texto-fonte, podemos observar alguns problemas na estruturação do sumário. O segmento sublinhado na figura 6 [O estudo, que será publicado no British Medical Journal] corresponde a uma descrição definida em posição anafórica com relação a uma entidade presente no texto-fonte, mas ausente no sumário – [O estudo, que será publicado no British Medical Journal]. A existência desta anáfora não resolvida indica um problema de coesão referencial que, em última análise, implica incoerência, pois a informação necessária para a construção do sentido da anáfora (o assunto do referido estudo e o autor do estudo) não foi selecionada para compor o sumário.

Estes efeitos, como pudemos averiguar, ocorreram nos dois módulos do RHeSuma-2, o DiZer e o RHeSumaRST, e apresentaram ainda outras fragilidades do sistema. Frente a isso, então, inferimos algumas possíveis causas do desempenho referido: i) problemas de adaptação do DiZer ao tipo de texto a sumarizar; ii) problemas de segmentação, introduzidos pelas limitações do analisador morfossintático automático acoplado ao DiZer; iii) estruturas RST deficientes, resultado do DiZer que prejudicaram a performance do RHeSumaRST.

Os resultados, apesar de insatisfatórios, oferecem amplas perspectivas para o sistema. Na próxima seção, detalhamos estas perspectivas.

5. Perspectivas de avanço no desenvolvimento do sistema

O primeiro problema que fica evidenciado pelos resultados do referido experimento é a segmentação dos textos-fonte. Este processo é parte da arquitetura do DiZer, que possui acoplado a ele um analisador morfossintático automático para o português do Brasil (Aires et al., 2000). Esse etiquetador utiliza vinte etiquetas que representam, basicamente, as classes gramaticais tradicionais do português, possuindo precisão geral de 89%, o que, acreditamos, pode ser melhorado com a utilização de um novo analisador morfossintático, o PALAVRAS (Bick, 2000), que apresenta índices de precisão próximos de 98%.

A utilização do PALAVRAS no âmbito do Projeto ProCaCoSa não se justifica apenas pelo melhoramento da segmentação. Este *parser*, além da análise morfossintática, oferece também anotação semântica, o que poderá ser utilizado para, juntamente com a marcação das veias, elaborar heurísticas de sumarização com vistas à resolução anafórica. Nossa perspectiva neste sentido é que seja possível, através das etiquetas semânticas dos componentes lingüísticos de cada segmento, investigar padrões de relação entre referente e anáfora e, assim, determinar com maior precisão qual dos candidatos elencados pelo algoritmo das veias é, de fato, o termo antecedente da anáfora. Juntamente a esse estudo, estão sendo investigados critérios de reconhecimento de cadeias referenciais através da extração de modelos a partir de córpus anotados com cadeias (Coelho et al., 2006, Collovini & Vieira, 2006).

Cumpre, porém, ressaltar que a proposta da representação do conhecimento aqui apresentada é a investigação de possíveis relações entre as estruturas retórica (RST) e co-referencial de textos. Nós partimos da hipótese de que existe uma representação formal do texto que permite a identificação de informações que garantam o encadeamento referencial em sumários produzidos automaticamente. Ou seja, a partir de uma determinada representação da informação inserida no texto é possível identificar relações referenciais entre as unidades informativas de modo a possibilitar que sistemas automáticos gerem sumários sem quebras de elos co-referenciais. Nossa investigação caminha, então, no sentido de agregar a RST a este modelo de representação formal do texto – conforme a Teoria das Veias sinaliza ser possível – e, através de um estudo de córpus, determinar estratégias de construção textual, depreendendo das mesmas informações para a construção do referido modelo de representação formal.

6. Referências bibliográficas

BICK, E. *The Parsing System PALAVRAS: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. 2000. Tese (Doutorado) – Arhus University, Arhus, 2000.

CARBONEL, T.I.; SENO, E.M.; PARDO, T.A.S.; COELHO, J.C.; COLLOVINI, S.; RINO, L.H.M.; VIEIRA, R. A Two-Step Summarizer of Brazilian Portuguese Texts. In: WORKSHOP EM TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA – TIL'2006. Ribeirão Preto, 2006. *Anais do Workshop em Tecnologia da Informação e da Linguagem Humana – TIL'2006* Ribeirão Preto, 2006.

COELHO, J. C.; MÜLLER, V.; COLLOVINI, S.; VIEIRA, R. RINO, L. Resolving Portuguese Nominal Anaphora. In: WORKSHOP EM TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA – TIL'2006. Ribeirão Preto, 2006. *Anais do Workshop em Tecnologia da Informação e da Linguagem Humana – TIL'2006* Ribeirão Preto, 2006.

COLLOVINI, S; VIEIRA, R. Learning Discourse-new References in Portuguese Texts. In. VII Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada, Itatiaia, 2006. *Anais do Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada.* Itatiaia, 2006.

CRISTEA, D.; IDE, N.; ROMARY, L. Veins Theory: A Model of Global Discourse Cohesion and Coherence. In: COLING/ACL' 1998, Montreal, 1998. *Proceedings of the Coling/ACL' 1998*, Montreal, Canadá, 1998, p.281-285.

DESIDERATO ANTONIO, J. *Estrutura retórica e articulação de orações em narrativas orais e em narrativas escritas do português.* 2004. Tese (Doutorado) – Departamento de Lingüística, UNESP, Araraquara, 2004.

HUTCHINS, J. Summarization: Some problems and Methods. In: JONES. *Meaning: The frontier of informatics*. Cambridge: London, 1987, p. 151-173.

MANI, I. *Automatic Summarization*. Amsterdã: John Benjamins Publishing Co. 2001.

MANN, W.C.; THOMPSON, S.A. *Rhetorical Structure Theory: A Theory of Text Organization*. Relatório Técnico. ISI/RS-87-190, 1987.

MARCU, D *The rhetorical parsing, summarization, and generation of natural language texts.* 1997. Tese (Doutorado) - Department of computer Science, University of Toronto, Toronto, 1997a.

MARCU, D. From Discourse Structures to Text Summaries. In: ACL'97/EACL'97 WORKSHOP ON INTELLIGENT SCALABLE TEXT SUMMARIZATION. Madri, 1997. *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization.* Madri, Espanha, 1997b, p. 82-88.

O'DONNELL, M.. RSTTool 2.4 - A Markup Tool for Rhetorical Structure Theory. In: INTERNATIONAL NATURAL LANGUAGE GENERATION CONFERENCE (INLG'2000). Mitzpe Ramon, 2000. *Proceedings of the International Natural Language Generation Conference (INLG'2000)*. Mitzpe Ramon, Israel, 2000, p. 253-256.

PARDO, T.A.S.; RINO, L.H.M. A Summary Planner Based on a Three-Level Discourse Model. In: the NATURAL LANGUAGE PROCESSING PACIFIC RIM SYMPOSIUM. Tóquio, 2001. *Proceedings of the 6th NLPRS.* National Center of Science, Tóquio, 2001, p. 533-538.

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

PARDO, T.A.S. *Métodos para Análise Discursiva Automática*. 2005. Tese (Doutorado) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. São Carlos, 2005.
RINO, L.H.M. *Modelagem de Discurso para o Tratamento da Concisão e Preservação da Idéia Central na Geração de Textos*. 1996. Tese (Doutorado) – Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 1996.

SENO, E.R.M. *Especificação de Heurísticas de Sumarização de Estruturas RST com Base na Preservação dos Elos Co-Referenciais.* 2005. Dissertação (Mestrado) – Departamento de Computação, Universidade Federal de São Carlos, São Carlos, 2005.

SENO, E.R.M., RINO, L.H.M.. Co-referential chaining for coherent summaries through rhetorical and linguistic modeling. In: WORKSHOP ON CROSSING BARRIERS IN TEXT SUMMARIZATION RESEARCH – RANLP´05. Bulgária, 2005. *Proceedings of Workshop on Crossing Barriers in Text Summarization Research – RANLP´05*. Bulgária, 2005, p.70-75, 2005

SPARCK JONES, K. *Discourse Modelling for Automatic Summarising*. Relatório Técnico, Rep. No. 290. University of Cambridge, 1993.

VAN DIJK, T.A. Recalling And Summarizing Complex Discourse. In: BURGHART, W. AND HÖLKER, K. *Text Processing Textverarbeitung.* Berlim: Walter de Gruyter ed. 1979.