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

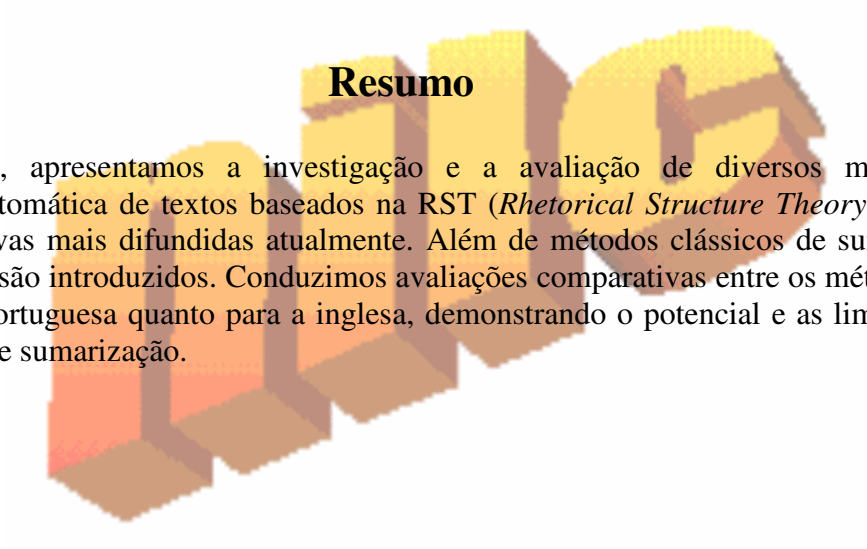
# Estudo e Avaliação de Métodos de Sumarização Automática de Textos Baseados na RST

Vinícius Rodrigues de Uzêda
Thiago Alexandre Salgueiro Pardo
Maria das Graças Volpe Nunes

**NILC-TR-07-07**

Agosto, 2007

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 a investigação e a avaliação de diversos métodos de sumarização automática de textos baseados na RST (*Rhetorical Structure Theory*), uma das teorias discursivas mais difundidas atualmente. Além de métodos clássicos de sumarização, novos métodos são introduzidos. Conduzimos avaliações comparativas entre os métodos tanto para a língua portuguesa quanto para a inglesa, demonstrando o potencial e as limitações da RST para fins de sumarização.

ÍNDICE

# 1. Introdução

Imensas quantidades de informação estão ao redor do ser humano todos os dias, e em constante crescimento. Não é possível digerir tanta informação assim, fazendo-se necessário um resumo, uma síntese da mesma, mas é impraticável fazer isto manualmente. A Sumarização Automática (SA), área em constante desenvolvimento, tem por finalidade criar sumários (ou resumos) através do uso do computador.

É impossível não se ver um sumário nos tempos atuais: sinopses de filmes, manchetes de jornais, resenhas de livros, por exemplo, podem ser considerados sumários. Dependendo da relação destes com seus textos-fonte, textos a partir dos quais foram gerados, podem ser classificados de diversas maneiras. A mais fundamental consiste na distinção entre extratos e *abstracts*. Um extrato é um conjunto de trechos do texto-fonte que é capaz de sintetizar as informações neste contidas, enquanto que em um *abstract*, um novo texto é criado, com todas as complicações impostas pela geração textual, a fim de expressar de forma mais concisa as idéias de seu texto-fonte.

Os métodos para se obter tais sumários, chamados métodos de sumarização, por sua vez, podem se dar de várias formas, mas duas se sobressaem (Mani, 2001). Os métodos superficiais fundamentam-se basicamente em estatísticas, requerendo pouca informação da língua, como dicionários, gramáticas e outras ferramentas do gênero, o que usualmente permite uma fácil migração entre diferentes idiomas. O GistSumm (*GIST SUMMarizer*) (Pardo et al., 2003) é um sistema de sumarização extrativo. O sistema consiste em encontrar a sentença que expressa a idéia central do texto através da freqüência com que suas palavras aparecem no mesmo, supondo que as palavras que mais se repetem devem expressar esta idéia. Este sistema considerou, para o texto da Figura 1.1 (traduzido da obra de Marcu, 2000), a sentença exibida na Figura 1.2 como a mais importante do texto. Após eleita a sentença mais significativa do texto-fonte, o sistema passa a selecionar as demais sentenças que apresentaram alto índice de repetição de palavras com a sentença escolhida anteriormente até se atingir a taxa de compressão desejada, que consiste na razão entre o número de palavras que comporão o sumário e o que compõe o texto-fonte. A Figura 1.3 mostra o sumário obtido para o mesmo texto com uma taxa de compressão de 50%.

Com sua órbita distante – 50% mais distante do sol que a Terra – e camada atmosférica fina, Marte experimenta condições climáticas gélidas. As temperaturas na superfície tipicamente encontram-se em torno de -60º Celsius (-76º Fahrenheit) no equador e podem descer a até -123ºC próximo dos pólos. Apenas o sol de meio-dia nas latitudes tropicais é quente o suficiente para derreter o gelo ocasionalmente, mas qualquer água obtida dessa forma evaporaria quase que instantaneamente por causa da baixa pressão atmosférica.

Apesar de a atmosfera possuir uma pequena quantidade de água, e nuvens de água e gelo algumas vezes se formarem, a maior parte do clima marciano envolve ventos de pó ou de dióxido de carbono. A cada inverno, por exemplo, uma tempestade de dióxido de carbono congelado ataca um pólo, e alguns metros dessa neve de gelo-seco se acumula enquanto dióxido de carbono previamente congelado evapora na outra calota polar. Entretanto, mesmo no pólo que está no verão, em que o sol permanece no céu o dia todo, as temperaturas nunca sobem o suficiente para derreter a água congelada.

Figura 1.1. Texto de exemplo

A cada inverno, por exemplo, uma tempestade de dióxido de carbono congelado ataca um pólo, e alguns metros dessa neve de gelo-seco se acumula enquanto dióxido de carbono previamente congelado evapora na outra calota polar.

Figura 1.2. Sentença eleita como mais importante pelo GistSumm para o texto da Fig. 1.1

Apesar de a atmosfera possuir uma pequena quantidade de água, e nuvens de água e gelo algumas vezes se formarem, a maior parte do clima marciano envolve ventos de pó ou de dióxido de carbono. A cada inverno, por exemplo, uma tempestade de dióxido de carbono congelado ataca um pólo, e alguns metros dessa neve de gelo-seco se acumula enquanto dióxido de carbono previamente congelado evapora na outra calota polar.

Figura 1.3. Sumário obtido através do sistema GistSumm para o texto da Fig. 1.1 com taxa de compressão de 50%

Como se pode perceber, a simplicidade do algoritmo usualmente acaba levando à escolha de sentenças mais longas como mais importantes, o que nem sempre é verdade. No texto da Figura 1.1, por exemplo, o texto fala sobre como o clima de Marte é frio, o que seria bem expresso pelo segmento "Marte experimenta condições climáticas gélidas", o que provavelmente levaria a um sumário muito diferente. Existe uma variação do método que, ao invés de simplesmente pontuar as sentenças pela presença de palavras freqüentes, ainda faz a média da pontuação com relação ao tamanho da sentença em questão, normalizando os resultados, mas ainda constatam-se outros problemas, como a presença de anáforas não resolvidas, problemas na coesão textual, entre outros, que foram analisados em (Balage et al., 2006). Estes problemas estão presentes na maioria dos sistemas que realizam apenas uma análise superficial, sendo quase impossível tratar tais falhas sem fazer uso de maiores informações.

A outra abordagem clássica é a profunda, em que são utilizadas técnicas formais e modelos lingüísticos para tentar obter melhores resultados, gerando-se sumários mais coerentes e informativos, mas aumentando a sua complexidade de desenvolvimento, já que esta pode ser diretamente relacionada com o nível de conhecimento lingüístico utilizado pelo sistema. Além disso, para tornar possível o desenvolvimento dessas ferramentas lingüísticas, optam-se por línguas e gêneros textuais, o que acaba por restringir o campo de atuação do sistema de SA.

O conhecimento lingüístico utilizado nesta abordagem pode ser classificado em 4 grandes níveis, que podem ser vistos como análises cada vez mais profundas da língua, a saber, a morfologia, a sintaxe, a semântica e o nível da pragmática/discurso. Tem-se evidenciado que, quanto mais avançado o nível de conhecimento utilizado, melhores resultados são obtidos em quase todas as áreas do Processamento de Línguas Naturais (PLN), área que pode ser subdividida em diversas outras, entre elas a de Sumarização, Tradução e Geração Textual automatizadas.

A proposta deste trabalho é avaliar alguns métodos que fazem uso do conhecimento discursivo para realizar a Sumarização Automática. Existem diversas teorias utilizadas a fim de representar este nível de conhecimento, sendo que algumas das principais que podem ser citadas são a teoria de Grosz e Sidner (1986) e, mais importante ainda, a de Mann e Thompson (1987), denominada *Rhetorical Structure Theory* (RST).

As teorias discursivas representam as características presentes nos segmentos de texto em pelo menos duas categorias: a primeira, intencional, em que se detecta a intenção do escritor ao expressar aquela sentença; e uma segunda, semântica, cuja finalidade é fazer com que o leitor reconheça a lógica subjacente àquele conjunto de sentenças, como o fato de uma pessoa ficar presa em um engarrafamento implicar no fato de ela chegar atrasada ao trabalho.

A teoria de Grosz e Sidner enfoca as relações intencionais, enquanto a RST abrange relações dos dois tipos acima descritos.

Neste trabalho, exploram-se e comparam-se várias técnicas de sumarização da abordagem profunda baseadas na RST, teoria mais utilizada para a representação discursiva, pela grande gama de aplicações em que a mesma pode ser utilizada, como avaliaram Taboada e Mann (2006). Vários autores propuseram técnicas de sumarização baseadas na estrutura retórica dos textos, mas não havia ainda sido feita uma análise comparativa exaustiva como a apresentada aqui. O principal objetivo deste trabalho é realizar esta avaliação, a fim de se constatar quais características discursivas mais influenciam na construção de um bom sumário.

Uma introdução a RST é apresentada na Seção 2. Os algoritmos de sumarização baseados na RST são explicado na Seção 3 e os resultados comparativos obtidos são expostos na Seção 4. Por fim, apresenta-se uma avaliação crítica dos resultados obtidos com este experimento.

# 2. *Rhetorical Structure Theory*

A *Rethorical Structure Theory* (RST) (Mann e Thompson, 1987) foi originalmente criada para descrever a estrutura de textos e a forma como a comunicação se dá. Desenvolvida a partir de estudos de textos editados ou preparados cuidadosamente de uma grande variedade de fontes, ela tem um status em Lingüística que é independente de seu uso computacional.

Esta teoria tem como objetivo explicar a coerência nos textos. Existem diversas definições de coerência, todas deixando evidente que, para um texto ser coerente, cada parte do mesmo deve exercer alguma função, deve ter alguma razão plausível para a sua presença no texto, relacionando-se com as demais partes.

Tendo por objetivo descrever os textos por meio da RST, é possível se obter uma gama de possíveis estruturas discursivas que representem cada texto, visto que diferentes interpretações de um texto são possíveis. Essas estruturas são compostas de relações discursivas, comumente chamadas relações de coerência pela literatura da área, que explicam porque o conteúdo de dois conjuntos de unidades textuais (orações, períodos, parágrafos etc.) encontram-se no mesmo texto. Por exemplo, o texto da Figura 2.1, já dividido em segmentos numerados, pode ser modelado na árvore de relações discursivas (também chamada árvore retórica) da Figura 2.2.

A RST apresenta ainda um conceito capaz de denotar a importância dos segmentos que são subordinados a cada relação: a nuclearidade. Os segmentos 1 e 2 do texto da Figura 2.1 estão ligados via uma relação de *Justify*, sendo 1 uma justificativa do fato expresso em 2. A relação *Justify* diz que a justificativa é menos importante para a compreensão do texto do que o fato justificado, portanto se representa essa informação dizendo-se que 1 é satélite e 2 é núcleo da relação em questão. Na Figura 2.2, as caixas pontilhadas representam satélites, enquanto que as linhas contínuas representam núcleos.

[Com sua órbita distante – 50% mais distante do sol que a Terra – e camada atmosférica fina,][1] [Marte experimenta condições climáticas gélidas.][2] [As temperaturas na superfície tipicamente encontram-se em torno de -60º Celsius (-76º Fahrenheit) no equador e podem descer a até -123ºC próximo dos pólos.][3] [Apenas o sol de meio-dia nas latitudes tropicais é quente o suficiente para derreter o gelo ocasionalmente,][4] [mas qualquer água obtida dessa forma evaporaria quase que instantaneamente][5] [por causa da baixa pressão atmosférica.][6]

[Apesar de a atmosfera possuir uma pequena quantidade de água, e nuvens de água e gelo algumas vezes se formarem,][7] [a maior parte do clima marciano envolve ventos de pó ou de dióxido de carbono.][8] [A cada inverno, por exemplo, uma tempestade de dióxido de carbono congelado ataca um pólo, e alguns metros dessa neve de gelo-seco se acumula enquanto dióxido de carbono previamente congelado evapora na outra calota polar.][9] [Entretanto, mesmo no pólo que está no verão, em que o sol permanece no céu o dia todo, as temperaturas nunca sobem o suficiente para derreter a água congelada.][10]

Figura 2.1. Texto da Figura 1.1 dividido em unidades textuais

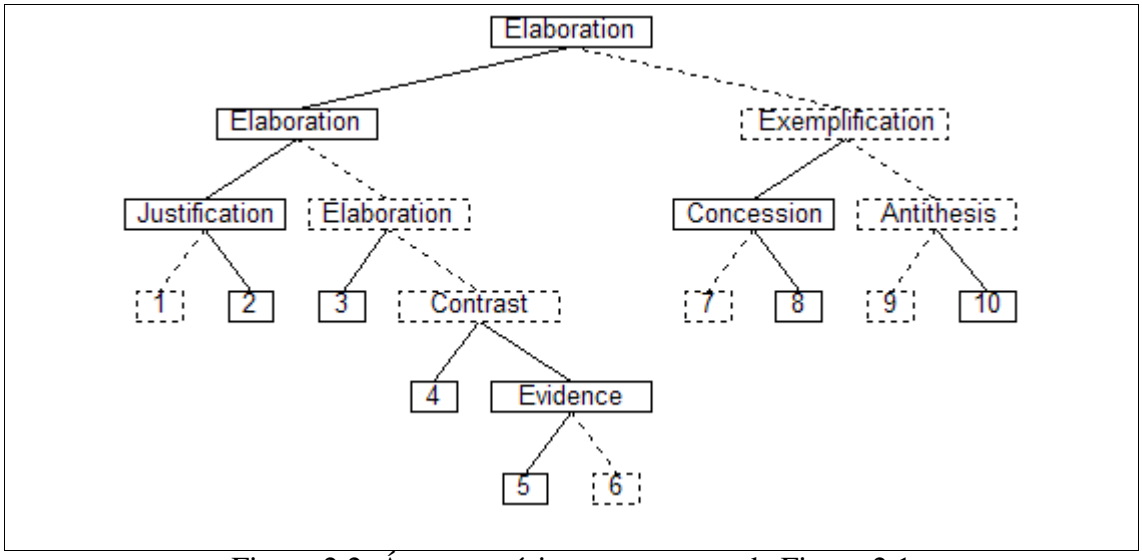



Figura 2.2. Árvore retórica para o texto da Figura 2.1

A RST originalmente propôs um conjunto de 32 relações, exibidas nas Tabelas 2.1 a 2.3. Estas relações podem ser classificadas em mononucleares ou multinucleares. As primeiras, mais comuns, unem dois segmentos, em que um tem maior importância do que o outro, sendo denominado núcleo da relação. Nas relações multinucleares, dois ou mais segmentos de igual importância são ligados, sendo todos considerados núcleos. É um exemplo de relação multinuclear a relação *Contrast*, já que duas sentenças que se contrapõem dificilmente vão se sobressair uma em relação a outra.

As relações podem ainda ser divididas em semânticas, responsáveis por criar a estrutura básica do texto, e em intencionais, cujo objetivo é causar algum efeito sobre o leitor. Assim, por exemplo, a relação *Elaboration* é considerada uma relação semântica, pois simplesmente une o conteúdo de dois segmentos, em que um provê detalhes adicionais ao outro. Por outro lado, a relação *Evidence* recebe a classificação de intencional, já que uma das sentenças unidas por ela, em que o leitor acredita, servirá para aumentar sua crença na outra.

Existem variações do conjunto de relações retóricas utilizadas nos diferentes trabalhos na área, como no presente trabalho, que fez uso de um conjunto de 164 relações. Na teoria original, foram propostas 15 relações mononucleares semânticas, 10 mononucleares

intencionais e 7 multinucleares semânticas. As Tabelas 2.1, 2.2 e 2.3 mostram essas divisões. O conjunto de relações utilizadas nesse trabalho é apresentado no Apêndice A.

Tabela 2.1. Relações Mononucleares Semânticas

| | | | | |
|---|---|---|---|---|
| *Circumstance* | *Evaluation* | *Non-volitional Cause* | *Purpose* | *Unless* |
| *Condition* | *Interpretation* | *Non-volitional Result* | *Solutionhood* | *Volitional Cause* |
| *Elaboration* | *Means* | *Otherwise* | *Unconditional* | *Volitional Result* |

Tabela 2.2. Relações Mononucleares Intencionais

| | | | | |
|---|---|---|---|---|
| *Antithesis* | *Concession* | *Evidence* | *Motivation* | *Restatement* |
| *Background* | *Enablement* | *Justify* | *Preparation* | *Summary* |

Tabela 2.3. Relações Multinucleares Semânticas

| | | | |
|---|---|---|---|
| *Conjunction* | *Disjunction* | *List* | *Multinuclear Restatement* |
| *Contrast* | *Joint* | *Sequence* | |

De acordo com a teoria, é possível identificar uma relação entre o conteúdo expresso por 2 segmentos textuais através de algumas características que as definem: as restrições sobre o núcleo, sobre o satélite e sobre ambos, além da intenção do escritor com elas. As Figuras 2.3, 2.4 e 2.5 mostram as características de algumas relações, seguidas de exemplos em que estas se aplicam na Figura 2.6. Na última, os segmentos textuais são separados com os colchetes, e duas marcas, N ou S, se apresentam sobrescritas, denotando, respectivamente, se o dado segmento é núcleo ou satélite da relação que existe ente eles.

**Relação:** *Elaboration*
**Restrições sobre o núcleo:** Nenhuma
**Restrições sobre o satélite:** Nenhuma
**Restrições sobre ambos:** O satélite apresenta detalhes adicionais sobre a situação ou algum elemento de importância sobre o assunto que é apresentado no núcleo ou acessível pelo núcleo através de inferência em uma ou mais das formas listadas a seguir. Na lista, o núcleo apresentaria a função do primeiro membro de qualquer par:

- Conjunto :: Membro
- Abstração :: Instância
- Todo :: Parte
- Processo :: Passo
- Objeto :: Atributo
- Generalização :: Especificação

**Intenção de escritor:** O leitor reconhece que o satélite provê detalhes adicionais ao núcleo. O leitor identifica a qual elemento os detalhes são providos.

Figura 2.3. Características que definem a relação *Elaboration*

**Relação:** *Evidence*
**Restrições sobre o núcleo:** O leitor pode não acreditar no núcleo em um grau satisfatório ao escritor
**Restrições sobre o satélite:** O leitor acredita no satélite ou o acha verossímil
**Restrições sobre ambos:** A compreensão do satélite pelo leitor aumenta a credibilidade do núcleo para o leitor
**Intenção de escritor:** A credibilidade do núcleo para o leitor é aumentada

Figura 2.4. Características que definem a relação Evidence

**Relação:** *Contrast*
**Restrições sobre cada par de núcleos:** Não mais de dois núcleos; as situações apresentadas nesses dois núcleos são:
- Similares em vários aspectos
- Diferentes em alguns aspectos
- Comparadas quanto a uma ou mais dessas diferenças

**Intenção de escritor:** O leitor reconhece a comparabilidade e a(s) diferença(s) contidas na comparação que está sendo feita

Figura 2.5. Características que definem a relação Contrast

**Exemplo da relação *Elaboration*:**
[Eu adoro colecionar carros clássicos.]$^{N}$ [Meu favorito é o Alfa Romeo Spider de 1968.]$^{S}$

**Exemplo da relação *Evidence*:**
[João é culpado,]$^{N}$ [encontraram suas digitais na arma.]$^{S}$

**Exemplo da relação *Contrast*:**
[Os animais andam,]$^{N}$ [mas as árvores se enraizam no solo.]$^{N}$

Figura 2.6. Exemplos de segmentos cujos conteúdos apresentam relações *Elaboration*, *Evidence* e *Contrast*

Com esta teoria em mãos, a gama de aplicações que podem se utilizar dela é imensa, como constatou Taboada e Mann (2006). Um de seus usos foi para a verificação de coerência de textos no processo de geração textual, mas também pode ser aplicada para outras áreas:

- Em Sumarização Automática, com experimentos realizados por Marcu (2000), O'Donnell (1997) e Ono et al. (1994), por exemplo;
- Na área de Tradução automática, evidenciados nos estudos de Ghorbel et al. (2001) e Marcu et al. (2000), entre outros.
- Em estudos lingüísticos que desejavam obter comparações ou generalizações entre diferentes idiomas, como mostram os trabalhos de Cui (1986), Kong (1998) e Ramsay (2001);
- Em pesquisas sobre pragmática, como as realizadas por Azar (1999) e Carenini e Moore (2000);
- No desenvolvimento de ferramentas de auxílio à escrita humana, por exemplo, as desenvolvidas por Bouwer (1998).

Para a área de SA, utilizando-se a relação de importância entre os segmentos textuais proveniente da nuclearidade das relações retóricas de um texto, vários métodos de medição da importância desses segmentos foram propostos. Com essas medidas, é possível se obter uma ordenação parcial das sentenças, de acordo com sua relevância no texto. Ordenação parcial é

um conceito matemático que se aplica na teoria de conjuntos: quando elementos de um conjunto qualquer podem possuir valores repetidos, é impossível se definir uma ordem precisa de seus elementos. Por exemplo, a Figura 2.7 mostra um conjunto de números inteiros e a ordenação parcial de seus elementos.

Isto ocorre em todos os métodos de sumarização que fazem uso da RST, pois é praticamente impossível garantir que cada segmento de texto receba uma pontuação distinta na avaliação de sua importância. O sumário gerado é afetado pela freqüência com que ocorrem essas pontuação iguais, pois não há uma forma de se decidir com certeza quais dos segmentos empatados devem pertencer ao sumário, no caso de não haver espaço para todos.

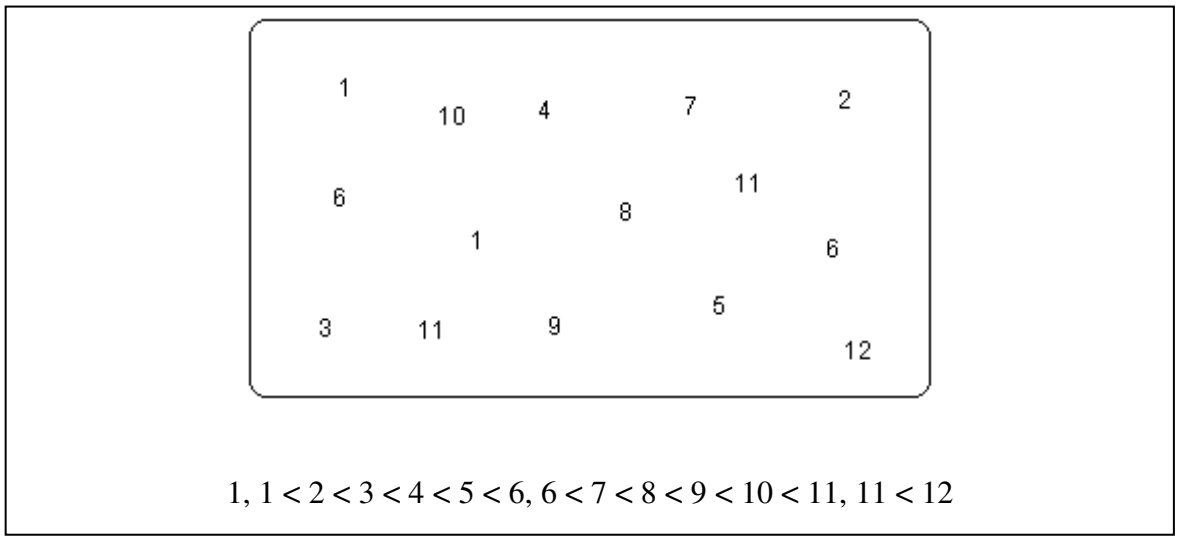



Figura 2.7. Um conjunto de números inteiros e sua ordenação parcial

A partir disso, algoritmos de SA consistem em tomar uma porcentagem dos trechos mais importantes e compor um sumário. Uma das grandes vantagens sobre outros métodos é a portabilidade do algoritmo (em geral, ele não é dependente de nenhuma língua em particular ou de domínio de textos). Os resultados também são bem atrativos. O principal problema no uso automatizado da RST é a obtenção das estruturas retóricas dos textos-fontes, mas não está no âmbito deste trabalho tratar disso, sendo que vários outros trabalhos já foram realizados nessa área, como (Marcu, 2000) e (Pardo, 2005). Para a avaliação dos métodos, empregaram-se um conjunto de estruturas retóricas geradas manualmente, para as línguas português e inglês, conforme descrito na Seção 4.

A seguir, os métodos de sumarização são descritos.

# 3. Descrição dos Métodos

Os algoritmos aqui descritos são métodos de se gerar uma ordenação parcial das unidades textuais com relação a sua importância no texto. Neste trabalho, as unidades textuais adotadas foram orações, pois os córpus utilizados no processo de avaliação foram anotados dessa forma, como se discute na próxima seção. No processo de se obter esta ordenação, cada oração recebe uma pontuação por sua localização na árvore retórica e a classificação de importância surge através da ordenação destas pontuações.

O sistema de SA tem como entrada a árvore retórica de cada texto, além do conjunto de orações que o compõe. Ainda é necessário especificar uma taxa de compressão, número entre 0 e 1 que representa a proporção entre o tamanho do sumário desejado e do texto original (em função do número de palavras).

Segue, então, uma descrição de cada método avaliado, mostrando qual a lógica utilizada e a função de pontuação em si. Para cada um dos algoritmos, segue-se um exemplo do percurso realizado no processo de pontuação do segmento 5 (meramente ilustrativo) do texto da Figura 2.1 com a estrutura mostrada na Figura 2.2. Além disso, mostra-se a ordenação obtida pelo método, bem como um sumário com taxa de compressão de 60%, já que para tal valor algumas diferenças se tornaram perceptíveis entre os diferentes métodos.

### *3.1. Ono et al. (1994)*

Um dos métodos mais simples, baseia-se simplesmente na nuclearidade das relações.

Percorre-se a árvore em profundidade em busca de cada segmento. A pontuação inicial é a profundidade da árvore e cada vez que se passa por um satélite no caminho, reduz-se a pontuação em uma unidade.

A Figura 3.1.1 mostra que a raiz da estrutura recebeu a pontuação 6, e que esta foi sendo decrementada a cada vez que se encontrou um satélite no percurso até a folha que representa a sentença 5. Na Figura 3.1.2, tem-se a ordenação parcial das sentenças do texto da Figura 2.1, e, finalmente, na Figura 3.1.3 vemos um sumário conforme descrito na introdução desta seção, produzido através do método proposto por Ono et al..

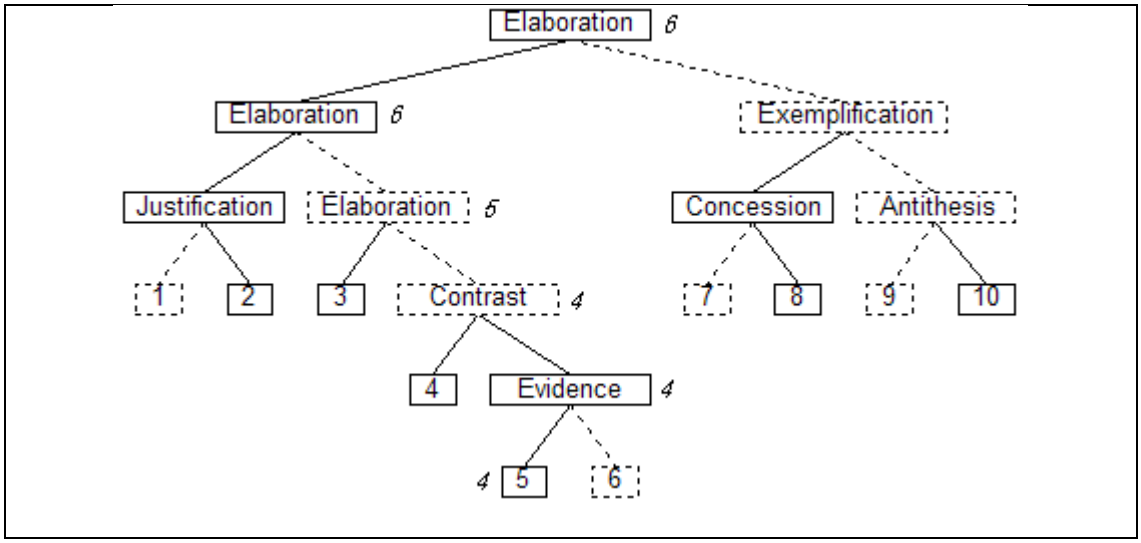



Figura 3.1.1. Percurso de pontuação do segmento 5 segundo Ono et al.

2 > 1, 3, 8 > 4, 5, 7, 10 > 6, 9

Figura 3.1.2. Ordenação parcial das orações da estrutura retórica da Figura 2.2 segundo Ono et al.

Com sua órbita distante – 50% mais distante do sol que a Terra – e camada atmosférica fina, Marte experimenta condições climáticas gélidas. As temperaturas na superfície tipicamente encontram-se em torno de -60º Celsius (-76º Fahrenheit) no equador e podem descer a até -123ºC próximo dos pólos. Apenas o sol de meio-dia nas latitudes tropicais é quente o suficiente para derreter o gelo ocasionalmente.

A maior parte do clima marciano envolve ventos de pó ou de dióxido de carbono.

Figura 3.1.3: Sumário para o texto da Figura 2.1 segundo Ono et al.

As principais vantagens deste método são sua simplicidade e o fato de não requerer qualquer informação além da árvore em si. Em compensação, ocorrem muitos segmentos com mesma

pontuação, o que prejudica o processo de seleção das sentenças que comporão o sumário, como descrito na seção anterior.

### *3.2. O'Donnell (1997)*

Além de utilizar a nuclearidade, este método tem como ponto central o fato de levar em conta a importância individual de cada relação.

Percorre-se a árvore em profundidade em busca de cada segmento. A pontuação inicial é 1 e cada vez que se passa por um satélite no caminho, multiplica-se a pontuação pelo fator de importância da relação atual (valor entre 0 e 1).

O principal problema na implementação deste método advém do fato dos valores das relações terem que ser definidos empiricamente. Assim, neste trabalho, de forma empírica, classificaram-se as relações em quatro grupos de importância, cada grupo recebendo uma pontuação (estes valores podem ser encontrados no Apêndice A). O método original possui valores diferentes dos utilizados para cada relação, o que certamente influenciou no resultado dos testes. Mais detalhes sobre isso serão dados na Seção 4.

Apesar da complexidade adicional de se determinar esses fatores de importância de cada relação, os resultados obtidos são muito bons: com valores razoavelmente distintos, a ocorrência de empates na pontuação das sentenças já cai bastante. Através dos resultados obtidos (discutidos na Seção 4), acredita-se que, com valores determinados de forma mais apropriada, os resultados podem melhorar consideravelmente. A grande desvantagem é que esses valores são dependentes do gênero textual a ser sumarizado e da língua em que este se encontra (em menor importância). Apesar disso, medidas genéricas podem ser utilizadas com bons resultados.

A Figura 3.2.1 mostra que a raiz da estrutura recebeu a pontuação 1, e que esta, a cada vez que se encontrou um satélite no percurso até a folha que representa a sentença 5, foi multiplicada por um fator entre 0 e 1 que sinaliza a importância daquela relação. Na Figura 3.2.2, tem-se a ordenação parcial das sentenças do texto da Figura 2.1, e, finalmente, na Figura 3.2.3 vemos um sumário conforme descrito na introdução desta seção, produzido através do método proposto por O'Donnell.

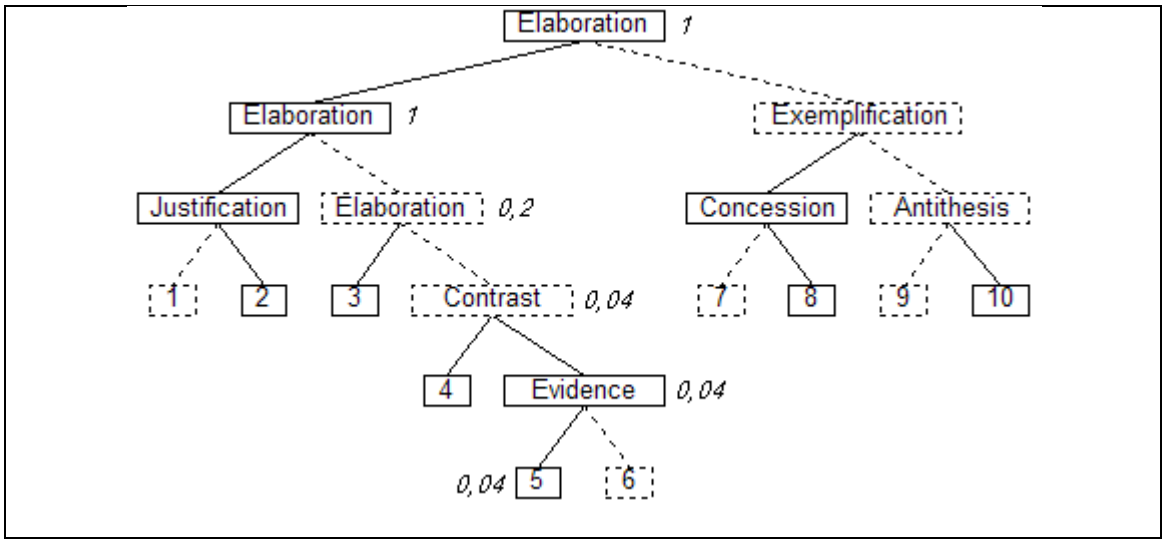



Figura 3.2.1. Percurso de pontuação do segmento 5 segundo O'Donnell

2 > 1 > 3, 8 > 7 > 4, 5, 10 > 9 > 6

Figura 3.2.2. Ordenação parcial das orações da estrutura retórica da Figura 2.2 segundo O'Donnell

Com sua órbita distante – 50% mais distante do sol que a Terra – e camada atmosférica fina, Marte experimenta condições climáticas gélidas. As temperaturas na superfície tipicamente encontram-se em torno de -60º Celsius (-76º Fahrenheit) no equador e podem descer a até -123ºC próximo dos pólos.

Apesar de a atmosfera possuir uma pequena quantidade de água, e nuvens de água e gelo algumas vezes se formarem, a maior parte do clima marciano envolve ventos de pó ou de dióxido de carbono.

Figura 3.2.3. Sumário para o texto da Figura 2.1 segundo O'Donnell

### *3.3. Marcu (2000)*

Para entender este método, o conceito de promoção de um segmento textual precisa ser esclarecido. Através da RST, é possível perceber que cada relação possui um conjunto de segmentos que lhe é mais importante, aquelas que são acessíveis a partir daquele nó passando-se apenas por núcleos. Este conjunto é denominado conjunto de promoção daquela dada relação na estrutura retórica. A Figura 3.3.1 exibe a mesma árvore da Figura 2.2 mostrando, para cada relação, qual é o seu conjunto de promoção.

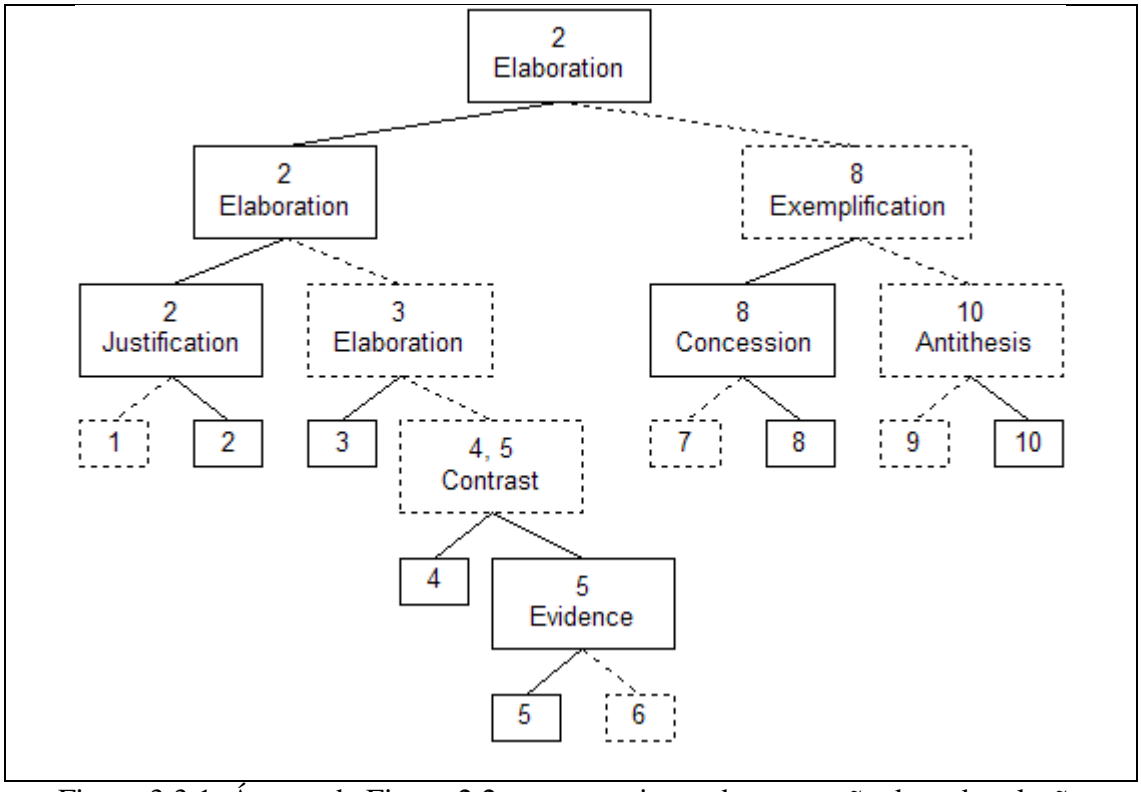



Figura 3.3.1. Árvore da Figura 2.2 com o conjunto de promoção de cada relação

Marcu propõe que a importância de cada segmento do texto pode ser obtida a partir da proximidade com que cada um deles passa a pertencer ao conjunto de promoção de uma relação.

Assim, percorre-se a árvore em profundidade em busca de cada segmento com pontuação inicial igual à profundidade da árvore. Se o segmento em questão pertence ao conjunto de promoção do nó atual, a pontuação é a atual. Se não, decrementa-se a pontuação atual de 1 para cada nível descido.

A Figura 3.3.2 mostra que a raiz da estrutura recebeu a pontuação 6, e que, a cada nó atravessado em que a sentença 5 não pertencia ao seu conjunto de promoção, este valor foi decrescido em uma unidade. Na Figura 3.3.3, tem-se a ordenação parcial das sentenças do texto da Figura 2.1, e, finalmente, na Figura 3.3.4 vemos um sumário conforme descrito na introdução desta seção, produzido através do método proposto por Marcu.

Este método possui uma complexidade menor que o de O'Donnell, já que depende apenas da estrutura retórica utilizada, como em Ono et al., mas já apresentando menor ocorrência de empates. Os resultados obtidos com este método são muito bons, mas ainda podem ser melhorados, como veremos a seguir.

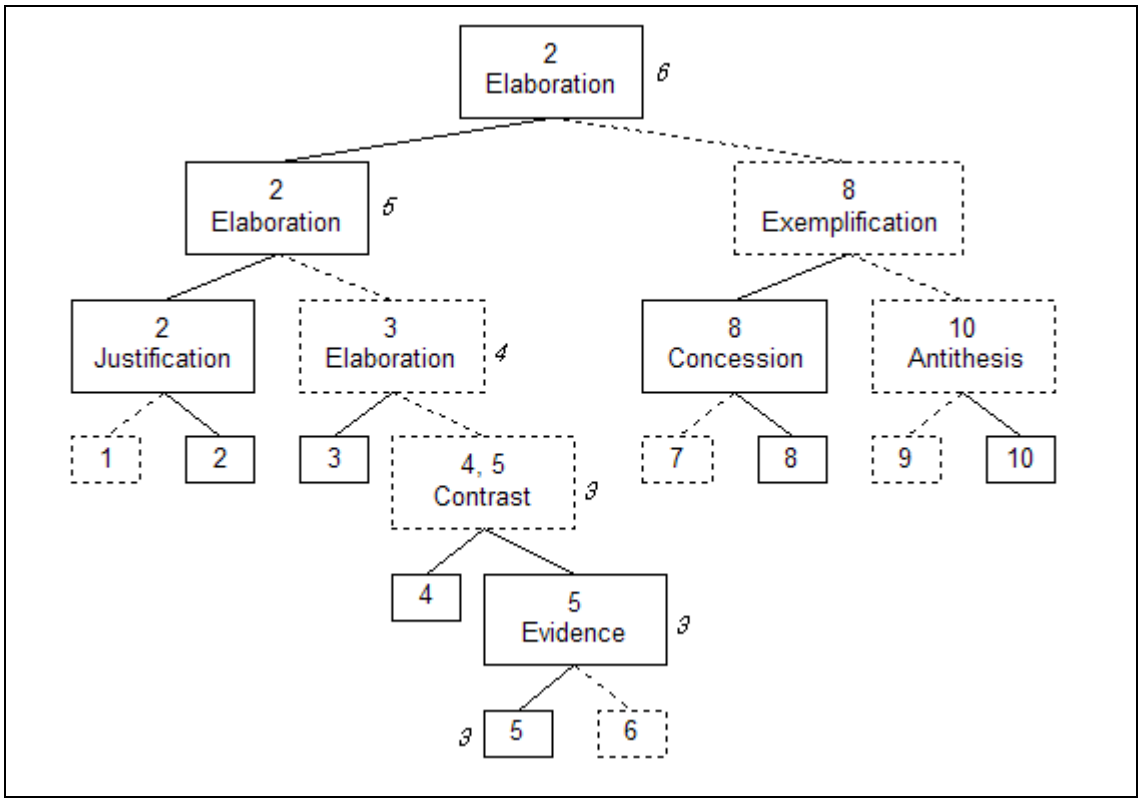



Figura 3.3.2. Percurso de pontuação do segmento 5 segundo Marcu

2 > 8 > 3, 10 > 1, 4, 5, 7, 9 > 6

Figura 3.3.3. Ordenação parcial das orações da estrutura retórica da Figura 2.2 segundo Marcu

Com sua órbita distante – 50% mais distante do sol que a Terra – e camada atmosférica fina, Marte experimenta condições climáticas gélidas. As temperaturas na superfície tipicamente encontram-se em torno de -60º Celsius (-76º Fahrenheit) no equador e podem descer a até -123ºC próximo dos pólos.

A maior parte do clima marciano envolve ventos de pó ou de dióxido de carbono. Entretanto, mesmo no pólo que está no verão, em que o sol permanece no céu o dia todo, as temperaturas nunca sobem o suficiente para derreter a água congelada.

Figura 3.3.4. Sumário para o texto da Figura 2.1 segundo Marcu

### *3.4. Marcu Aperfeiçoado (1998a, 1998b)*

Marcu, baseada em uma idéia informalmente proposta por Hovy, propôs uma modificação em seu método, anteriormente descrito: basicamente, adiciona-se à pontuação obtida pelo

processo anterior o número de níveis em que o dado segmento pertence ao conjunto de promoção das relações no caminho em busca da oração.

Na Figura 3.4.1, além da pontuação final, como exibido para os métodos anteriores, coloca-se também os valores obtidos pelo método convencional de Marcu e pela função de aperfeiçoamento.

Assim como o método descrito no tópico acima, a Figura 3.4.1 mostra que a raiz da estrutura recebeu a pontuação 6, e que, a cada nó atravessado em que a sentença 5 não pertencia ao seu conjunto de promoção, este valor foi decrescido em uma unidade. Num passo seguinte, vemos que a sentença 5 se encontra em conjuntos de promoção por dois níveis da árvore, sendo este valor acrescido na pontuação final desta sentença. Na Figura 3.4.2, tem-se a ordenação parcial das sentenças do texto da Figura 2.1, e, finalmente, na Figura 3.4.3 vemos um sumário conforme descrito na introdução desta seção, produzido através do método proposto e aperfeiçoado por Marcu.

Com estas alterações, ainda têm-se uma simplicidade considerável, além de se obter resultados ainda melhores, com pontuações repetidas ainda menos freqüentes.

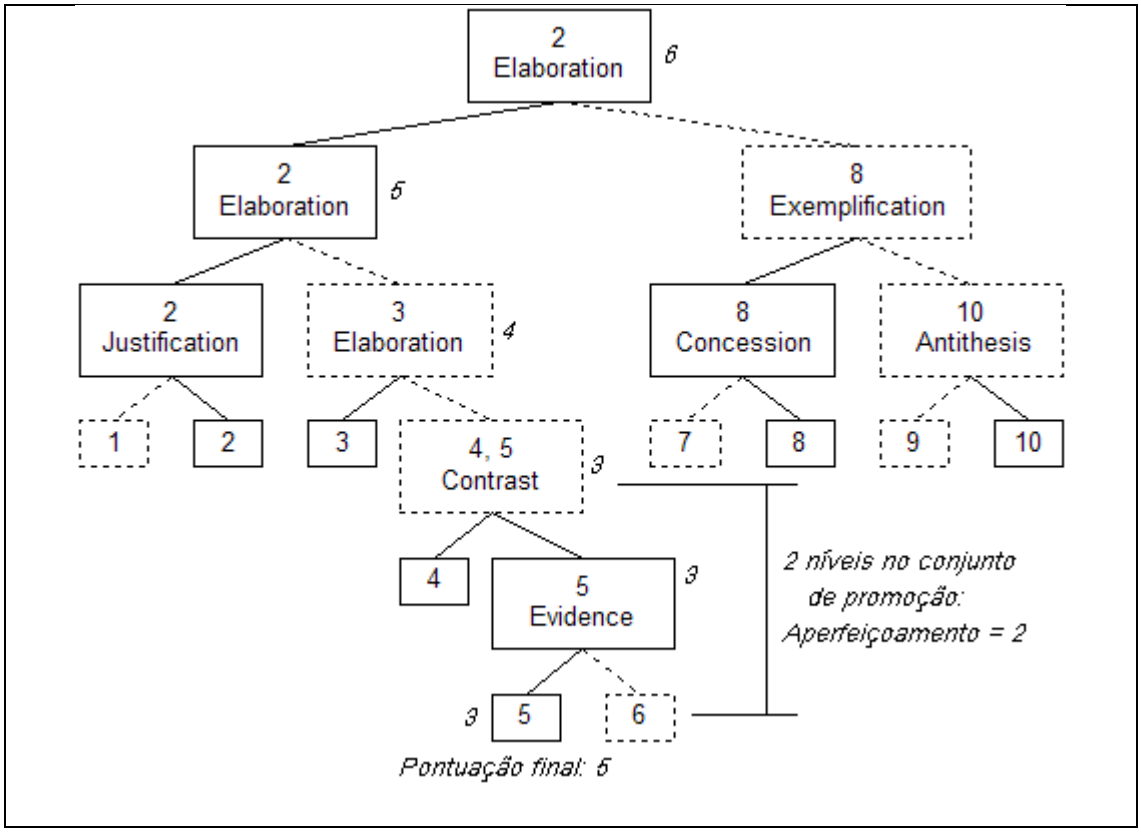



Figura 3.4.1. Percurso de pontuação do segmento 5 segundo Marcu aperfeiçoado

| 2 > 8 > 3, 5, 10 > 4 > 1, 7, 9 > 6 |
|---|

Figura 3.4.2. Ordenação parcial das orações da estrutura retórica da Figura 2.2 segundo Marcu aperfeiçoado

Marte experimenta condições climáticas gélidas. As temperaturas na superfície tipicamente encontram-se em torno de -60º Celsius (-76º Fahrenheit) no equador e podem descer a até -123ºC próximo dos pólos. Mas qualquer água obtida dessa forma evaporaria quase que instantaneamente

A maior parte do clima marciano envolve ventos de pó ou de dióxido de carbono. Entretanto, mesmo no pólo que está no verão, em que o sol permanece no céu o dia todo, as temperaturas nunca sobem o suficiente para derreter a água congelada.

Figura 3.4.3. Sumário para o texto da Figura 2.1 segundo Marcu aperfeiçoado

### *3.5. Uzêda*

Baseado nos métodos anteriores e extraindo de cada um dos demais conceitos que fossem considerados importantes no processo de pontuação das orações, foi proposto um novo método que seria também comparado entre os outros. Assim, levou-se em conta a proximidade com que o segmento passava a pertencer ao conjunto de promoção, como proposto por Marcu, a nuclearidade, como proposta por Ono et al., e um fator de importância dependente de cada relação, como fez O'Donnell.

A proposta final foi: percorre-se a árvore em profundidade em busca de cada segmento. Se o segmento em questão pertence ao conjunto de promoção do nó atual, a pontuação é a atual. Se não, decrementa-se a pontuação atual de um fator de desinteresse da relação em questão (o complemento do fator proposto por O'Donnell) para cada nível descido. Se ainda por cima, passar-se por um satélite no percurso, subtrai-se ainda mais uma unidade.

A classificação obtida por este método foi ainda melhor que a dos anteriores, apresentando muito poucos empates, mas, assim como o método de O'Donnell, os resultados são melhores dentro de um gênero específico de textos para os quais se determinam os fatores de desinteresse.

A Figura 3.5.1 mostra que a raiz da estrutura recebeu a pontuação 12. A cada nó atravessado em que a sentença 5 não pertencia ao seu conjunto de promoção, este valor foi decrescido em um fator de desinteresse, similar ao proposto por O'Donnell. Além disso, se o nó for um satélite, decrementa-se de mais uma unidade a pontuação. Na Figura 3.5.2, tem-se a ordenação parcial das sentenças do texto da Figura 2.1, e, finalmente, na Figura 3.5.3 vemos um sumário conforme descrito na introdução desta seção, produzido através do método proposto por Uzêda.

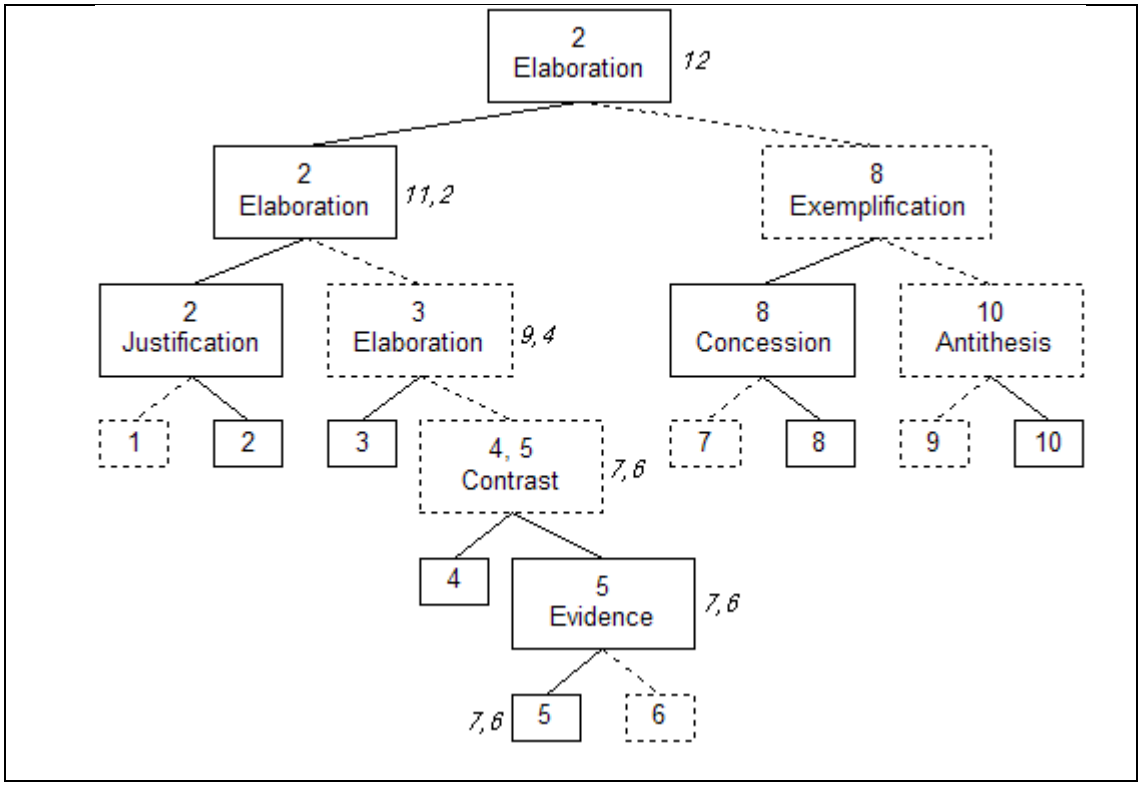



Figura 3.5.1. Percurso de pontuação do segmento 5 segundo Uzêda

2 > 8 > 3 > 1 > 10 > 7 > 4, 5 > 9 > 6

Figura 3.5.2. Ordenação parcial das orações da estrutura retórica da Figura 2.2 segundo Uzêda

Com sua órbita distante – 50% mais distante do sol que a Terra – e camada atmosférica fina, Marte experimenta condições climáticas gélidas. As temperaturas na superfície tipicamente encontram-se em torno de -60º Celsius (-76º Fahrenheit) no equador e podem descer a até -123ºC próximo dos pólos.

A maior parte do clima marciano envolve ventos de pó ou de dióxido de carbono. Entretanto, mesmo no pólo que está no verão, em que o sol permanece no céu o dia todo, as temperaturas nunca sobem o suficiente para derreter a água congelada.

Figura 3.5.3. Sumário para o texto da Figura 2.1 segundo Uzêda

Na seção seguinte, apresentam-se os resultados obtidos na avaliação dos métodos aqui descritos.

# 4. Avaliação

Foram utilizados dois conjuntos de textos jornalísticos, um em inglês e outro em português. Optou-se por fazer a avaliação em mais de uma língua a fim de se constatar a eficácia dos métodos em diferentes idiomas, visando-se verificar a independência destes.

Os textos para a língua portuguesa são do córpus de textos jornalísticos construído por Seno e Rino (2005), com sumários humanos com taxa de compressão de 70% e com estruturas retóricas geradas manualmente por apenas um especialista em RST. Foram utilizados 38 textos anotados deste córpus, utilizando um conjunto de 60 relações retóricas. Já os utilizados para a língua inglesa constam no *Rhetorical Structure Theory Discourse TreeBank* (Carlson et al., 2003) composto de texto jornalísticos com sumários, que não contém mais do que a raiz quadrada do número de sentenças do texto-fonte, e estruturas retóricas anotadas manualmente. Este último córpus foi anotado por uma equipe de lingüistas treinados em RST, e neste trabalho se utilizaram 30 textos marcados com um conjunto de 164 relações retóricas.

Além dos métodos acima descritos, foram avaliados dois outros algoritmos: o GistSUMM (Pardo et al., 2003) e outro sistema *baseline*, que seleciona as primeiras sentenças de cada texto até atingir a compressão desejada (denominado, daqui em diante, por firstSents), a fim de se ter uma base para compararem-se os resultados. Após a aplicação dos diferentes métodos, os sumários foram avaliados em relação aos sumários humanos (ideais) através da métrica ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) (Lin e Hovy, 2003), que será detalhada a seguir.

### *4.1. ROUGE*

Optou-se por utilizar uma medida automática, a fim de ter uma resposta imparcial, além de se ganhar tempo no processo de avaliação. Assim, adotou-se a ROUGE, pacote de medidas mais utilizado recentemente para a avaliação de sistemas de SA, que permitiria também a fácil reprodução desta avaliação, o baixo custo de se executá-la, se comparado com uma avaliação manual, consistência, evitando-se os erros humanos geralmente cometidos.

Baseada na medida BLEU (Papineni et al., 2001), fortemente utilizada para a avaliação de sistemas de tradução automática, a ROUGE usa a abordagem de co-ocorrência de n-gramas, que consiste em verificar a média de quantas vezes cada conjunto de n palavras adjacentes se repetem em cada texto a ser avaliado.

O pacote da ROUGE utilizado contém 5 medidas:

- ROUGE-1: equivalente a medida 1-grama, avalia a média de quantas vezes cada palavra aparece em cada um dos textos;
- ROUGE-2: similar a 2-grama, consiste em uma análise da freqüência com que cada par de palavras aparece em cada texto de entrada;
- ROUGE-3 e ROUGE-4, semelhantes às medidas 3-grama e 4-grama, respectivamente, são pouco utilizadas, visto que a repetição de conjuntos de 3 ou 4 palavras adjacentes é muito incomum;
- ROUGE-L: baseada no *Loguest Common Subsequence* (LCS), busca as maiores subcadeias comuns entre os dois textos, executando então uma avaliação similar a co-ocorrência de n-grama.

### *4.2. Resultados*

Os resultados obtidos são apresentados nas tabelas contidas nas figuras a seguir. As Tabelas 4.2.1 a 4.2.7 mostram os resultados de cada método para o português, enquanto as Tabelas

4.2.8 a 4.2.14 exibem os dados obtidos referentes à língua inglesa. Os valores determinados, conforme a medida ROUGE, vêm organizados em 5 diferentes categorias para cada método, de acordo com as métricas caracterizadas na seção anterior.

Tabela 4.2.1. Avaliação do sistema firstSents para a língua portuguesa

| **Método: firstSents** | **Cobertura** | **Precisão** | **F-Measure** |
|---|---|---|---|
| ROUGE-1 | 0,176040 | 0,620720 | 0,258810 |
| ROUGE-2 | 0,075650 | 0,292960 | 0,112930 |
| ROUGE-3 | 0,044760 | 0,191060 | 0,067880 |
| ROUGE-4 | 0,030820 | 0,139920 | 0,047060 |
| ROUGE-L | 0,158350 | 0,559010 | 0,232720 |

Tabela 4.2.2. Avaliação do sistema GistSUMM para a língua portuguesa

| **Método: GistSUMM** | **Cobertura** | **Precisão** | **F-Measure** |
|---|---|---|---|
| ROUGE-1 | 0,411580 | 0,507280 | 0,451250 |
| ROUGE-2 | 0,159620 | 0,197790 | 0,175360 |
| ROUGE-3 | 0,092320 | 0,114090 | 0,101250 |
| ROUGE-4 | 0,062090 | 0,076470 | 0,067960 |
| ROUGE-L | 0,372970 | 0,459610 | 0,408890 |

Tabela 4.2.3. Avaliação do algoritmo proposto por Ono et al. para a língua portuguesa

| **Método: Ono et al.** | **Cobertura** | **Precisão** | **F-Measure** |
|---|---|---|---|
| ROUGE-1 | 0,464320 | 0,503180 | 0,477910 |
| ROUGE-2 | 0,181760 | 0,198710 | 0,187780 |
| ROUGE-3 | 0,102490 | 0,112510 | 0,106110 |
| ROUGE-4 | 0,066800 | 0,073520 | 0,069270 |
| ROUGE-L | 0,430890 | 0,466780 | 0,443420 |

Tabela 4.2.4. Avaliação do algoritmo proposto por O'Donnell para a língua portuguesa

| **Método: O'Donnell** | **Cobertura** | **Precisão** | **F-Measure** |
|---|---|---|---|
| ROUGE-1 | 0,465400 | 0,500570 | 0,477390 |
| ROUGE-2 | 0,181850 | 0,196220 | 0,186700 |
| ROUGE-3 | 0,102440 | 0,110320 | 0,105060 |
| ROUGE-4 | 0,066720 | 0,071580 | 0,068320 |
| ROUGE-L | 0,432110 | 0,464860 | 0,443270 |

Tabela 4.2.5. Avaliação do algoritmo proposto por Marcu para a língua portuguesa

| **Método: Marcu** | **Cobertura** | **Precisão** | **F-Measure** |
|---|---|---|---|
| ROUGE-1 | 0,470350 | 0,491290 | 0,475250 |
| ROUGE-2 | 0,177070 | 0,186830 | 0,179650 |
| ROUGE-3 | 0,094530 | 0,099850 | 0,095950 |
| ROUGE-4 | 0,058880 | 0,062470 | 0,059900 |
| ROUGE-L | 0,437560 | 0,457230 | 0,442240 |

Tabela 4.2.6. Avaliação do algoritmo aperfeiçoado por Marcu para a língua portuguesa

| **Método: Marcu Aperfeiçoado** | **Cobertura** | **Precisão** | **F-Measure** |
|---|---|---|---|
| ROUGE-1 | 0,467830 | 0,500250 | 0,478690 |
| ROUGE-2 | 0,177330 | 0,192380 | 0,182740 |
| ROUGE-3 | 0,094330 | 0,102860 | 0,097470 |
| ROUGE-4 | 0,057580 | 0,063320 | 0,059750 |
| ROUGE-L | 0,433620 | 0,463630 | 0,443680 |

Tabela 4.2.7. Avaliação do algoritmo proposto por Uzêda para a língua portuguesa

| **Método: Uzêda** | **Cobertura** | **Precisão** | **F-Measure** |
|---|---|---|---|
| ROUGE-1 | 0,469880 | 0,491040 | 0,474380 |
| ROUGE-2 | 0,177150 | 0,187070 | 0,179600 |
| ROUGE-3 | 0,094110 | 0,099760 | 0,095580 |
| ROUGE-4 | 0,058050 | 0,061760 | 0,059070 |
| ROUGE-L | 0,436280 | 0,455740 | 0,440400 |

Tabela 4.2.8. Avaliação do sistema firstSents para a língua inglesa

| **Método: firstSents** | **Cobertura** | **Precisão** | **F-Measure** |
|---|---|---|---|
| ROUGE-1 | 0,379080 | 0,368400 | 0,351250 |
| ROUGE-2 | 0,110810 | 0,112760 | 0,105050 |
| ROUGE-3 | 0,044630 | 0,046110 | 0,042440 |
| ROUGE-4 | 0,020900 | 0,021490 | 0,019840 |
| ROUGE-L | 0,366960 | 0,356160 | 0,339920 |

Tabela 4.2.9. Avaliação do sistema GistSUMM para a língua inglesa

| **Método: GistSUMM** | **Cobertura** | **Precisão** | **F-Measure** |
|---|---|---|---|
| ROUGE-1 | 0,353130 | 0,381720 | 0,340780 |
| ROUGE-2 | 0,099010 | 0,113250 | 0,098330 |
| ROUGE-3 | 0,039760 | 0,046610 | 0,039540 |
| ROUGE-4 | 0,019300 | 0,024020 | 0,019490 |
| ROUGE-L | 0,336530 | 0,362710 | 0,324080 |

Tabela 4.2.10. Avaliação do algoritmo proposto por Ono et al. para a língua inglesa

| **Método: Ono et al.** | **Cobertura** | **Precisão** | **F-Measure** |
|---|---|---|---|
| ROUGE-1 | 0,490730 | 0,408530 | 0,422430 |
| ROUGE-2 | 0,189300 | 0,159510 | 0,164440 |
| ROUGE-3 | 0,096860 | 0,081810 | 0,084290 |
| ROUGE-4 | 0,055150 | 0,046090 | 0,047680 |
| ROUGE-L | 0,055150 | 0,045090 | 0,047680 |

Tabela 4.2.11. Avaliação do algoritmo proposto por O'Donnell para a língua inglesa

| **Método: O'Donnell** | **Cobertura** | **Precisão** | **F-Measure** |
|---|---|---|---|
| ROUGE-1 | 0,482920 | 0,408660 | 0,419350 |
| ROUGE-2 | 0,180300 | 0,157280 | 0,159490 |
| ROUGE-3 | 0,091580 | 0,080950 | 0,081520 |
| ROUGE-4 | 0,052000 | 0,046230 | 0,046320 |
| ROUGE-L | 0,471370 | 0,398910 | 0,409230 |

Tabela 4.2.12. Avaliação do algoritmo proposto por Marcu para a língua inglesa

| **Método: Marcu** | **Cobertura** | **Precisão** | **F-Measure** |
|---|---|---|---|
| ROUGE-1 | 0,478070 | 0,407130 | 0,416350 |
| ROUGE-2 | 0,172800 | 0,152540 | 0,153720 |
| ROUGE-3 | 0,082790 | 0,075360 | 0,074700 |
| ROUGE-4 | 0,043280 | 0,040390 | 0,039430 |
| ROUGE-L | 0,467360 | 0,397930 | 0,407010 |

Tabela 4.2.13. Avaliação do algoritmo aperfeiçoado por Marcu para a língua inglesa

| **Método: Marcu Aperfeiçoado** | **Cobertura** | **Precisão** | **F-Measure** |
|---|---|---|---|
| ROUGE-1 | 0,467790 | 0,409360 | 0,408400 |
| ROUGE-2 | 0,165190 | 0,147690 | 0,147240 |
| ROUGE-3 | 0,076630 | 0,066900 | 0,068270 |
| ROUGE-4 | 0,039630 | 0,033540 | 0,034970 |
| ROUGE-L | 0,458180 | 0,402130 | 0,400530 |

Tabela 4.2.14. Avaliação do algoritmo proposto por Uzêda para a língua inglesa

| **Método: Uzêda** | **Cobertura** | **Precisão** | **F-Measure** |
|---|---|---|---|
| ROUGE-1 | 0,486180 | 0,408780 | 0,419680 |
| ROUGE-2 | 0,181890 | 0,160870 | 0,161250 |
| ROUGE-3 | 0,090930 | 0,082780 | 0,081620 |
| ROUGE-4 | 0,050280 | 0,046560 | 0,045410 |
| ROUGE-L | 0,473860 | 0,399100 | 0,409450 |

A análise destas tabelas nos revela que o método de Marcu com aperfeiçoamento obteve o melhor resultado para o português, mas foi acompanhado de perto pelo método de Ono et al., que assumiu a melhor avaliação para o inglês. Isso causou certa surpresa, visto que este seria o método mais simples. Isso evidencia fortemente que a nuclearidade desempenha papel essencial na categorização da importância dos segmentos, ainda mais do que o quanto um dado segmento é elevado pela nuclearidade na estrutura discursiva, como faz Marcu.

Os métodos de O'Donnell e de Uzêda não ficaram muito distantes dos melhores resultados, podendo até mesmo superar os demais com melhores medições dos fatores de importância e de desinteresse, respectivamente. Para a língua inglesa, todos os métodos que utilizavam RST superaram, em todas as medidas, tanto o GistSUMM quanto o firstSents. Em compensação, para o português, os métodos de abordagem superficial obtiveram resultados excepcionais para precisão, mas não para cobertura, ficando atrás dos métodos profundos na F-Measure, demonstrando que há diferenças no estilo de discurso utilizado em cada uma das línguas, mas que os métodos que faziam uso da RST ainda conseguiam manter uma maior cobertura. Esse resultado também se deve, em parte, ao fato de somente um especialista em RST ter anotado o córpus em português, enquanto, para o inglês, ter havido mais de um anotador e concordância nas análises, tornando os dados mais robustos e confiáveis.

Com relação apenas à precisão, para o português, o método firstSents se sobressaiu com excelentes resultados, denotando que os textos neste língua tem a essência de seu conteúdo fortemente concentrado em seu início. Já para a língua inglesa, obteve melhor desempenho o método aperfeiçoado de Marcu, o que evidencia que para uma boa seleção de sentenças dignas de pertencerem ao sumário é necessário levar em conta a nuclearidade de cada sentença (dado pelo nível em que aquela sentença aparece no conjunto de promoção da estrutura retórica utilizada) e também o quanto cada sentença é nuclear no texto (medido constatando-se por quantos níveis aquela sentença permanece em conjuntos de promoção).

## 5. Conclusão

Através deste trabalho, constataram-se os bons resultados obtidos pelo uso da RST no processo de sumarização. Além disso, verificou-se que alguns aprimoramentos ainda podem ser feitos nesses sistemas, potencialmente levando-os a obter resultados superiores.

Como trabalhos futuros, pretendemos comparar os métodos estudados com outros métodos, profundas e superficiais.

## Agradecimentos

Este trabalho contou com apoio das agências de fomento à pesquisa FAPESP, CAPES e CNPq.

## Referências

Azar, M. (1999). Argumentative text as rhetorical structure: An application of Rhetorical Structure Theory. *Argumentation*, Vol. 13, N. 1, pp. 97-144.

Balage Filho, P.P.; Uzêda, V.R.; Pardo, T.A.S.; Nunes, M.G.V. (2006). *Estrutura Textual e Multiplicidade de Tópicos na Sumarização Automática: o Caso do Sistema GistSumm*. Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, no. 283. São Carlos-SP, Novembro, 18p.

Bouwer, A. (1998). An ITS for Dutch punctuation. In *Intelligent Tutoring Systems*, Vol. 1452, pp. 224-233.

Carenini, G. and Moore, J.D. (2000). A strategy for generating evaluative arguments, *Proceedings of the 1st International Conference on Natural Language Generation*, pp. 47-54. Mitzpe Ramon, Israel.

Carlson, L.; Marcu, D.; Okurowski, M.E. (2003). Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. In J. van Kuppevelt and R. Smith (eds.), *Current Directions in Discourse and Dialogue*, pp. 85-112. Kluwer Academic Publishers.

Cui, S. (1986). *A Comparison of English and Chinese Expository Rhetorical Structures.* Unpublished Master's thesis, UCLA.

Ghorbel, H.; Ballim, A.; Coray, G. (2001). ROSETTA: Rhetorical and Semantic Environment for Text Alignment. In P. Rayson, A. Wilson, A. M. McEnery, A. Hardie and S. Khoja (eds.), *Proceedings of Corpus Linguistics 2001*, pp. 224-233. Lancaster, UK.

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

Kong, K.C.C. (1998). Are simple business request letters really simple? A comparison of Chinese and English business request letters. *Text*, Vol. 18, N. 1, pp. 103-141.

Lin, C.Y. and Hovy, E. (2003). Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In *Proceedings of 2003 Language Technology Conference* (HLT-NAACL 2003), Edmonton, Canada.

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

Marcu, D. (1998a). Improving summarization through rhetorical parsing tuning. In the *Proceedings of the Sixth Workshop on Very Large Corpora*, pp. 206-215. Montreal, Canada.

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

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

Marcu, D.; Carlson, L.; Watanabe, M. (2000). The automatic translation of discourse structures. In the *1st Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL'00)*, Vol. 1, pp. 9-17. Seattle, Washington.

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

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

Ono, K.; Sumita, K.; Miike, S. (1994). Abstract generation based on rhetorical structure extraction. In the *Proceedings of the International Conference on Computational Linguistics (Coling-94)*.

Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.-J. (2001). *BLEU: a Method for Automatic Evaluation of Machine Translation.* IBM Research Report RC22176 (W0109-022).

Pardo, T.A.S. (2005). *Métodos para Análise Discursiva Automática*. Tese de Doutorado. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. São Carlos-SP, Junho, 211p.

Pardo, T.A.S.; Rino, L.H.M.; Nunes, M.G.V. (2003). GistSumm: A Summarization Tool Based on a New Extractive Method. In the *Proceedings of the 6th Workshop on Computational Processing of the Portuguese Language - Written and Spoken – PROPOR (Lecture Notes in Artificial Intelligence 2721)*, pp. 210-218. Faro, Portugal.

Ramsay, G. (2001). What are they getting at? Placement of important ideas in Chinese newstext: A contrastive analysis with Australian newstext. *Australian Review of Applied Linguistics*, Vol. 24, N. 2, pp. 17-34.

Seno, E.R.M. and Rino, L.H.M. (2005). *RHeSumaRST: Um sumarizador automático de estruturas RST*. Dissertação de Mestrado. Departamento de Computação. Universidade Federal de São Carlos. São Carlos, SP.

Taboada, M. and Mann, W.C. (2006). Applications of Rhetorical Structure Theory. *Discourse Studies*, Vol. 8, N. 4, pp. 567-588.

# Apêndice A: Conjunto de relações

As relações que aqui se encontram foram as utilizadas para a anotação retórica dos textos utilizados na avaliação dos métodos de sumarização. Para auxiliar a marcação das relações discursivas, em ambos os córpus foi utilizada a Daniel Marcu's *RST Annotation Tool* (que pode ser encontrada em http://www.isi.edu/~marcu/discourse/). Para mais detalhes a respeito das relações e seus significados, consultar o material de referência.

Estas relações foram empiricamente classificadas em quatro grupos e receberam fatores de importância em comum, conforme descrito na Seção 3.2. As Tabelas A.1, A.2, A.3 e A.4 representam cada um desses grupos, com seus fatores de importância 0.8, 0.6, 0.4 e 0.2, respectivamente.

Tabela A.1. Relações pertencentes à categoria ++ Importantes (fator de importância = 0.8)

| | | | |
|---|---|---|---|
| antithesis | List | purpose | reason-e |
| antithesis-e | manner | purpose-e | Reason |
| cause | manner-e | question-answer | result |
| cause-e | otherwise | question-answer-e | result-e |
| Cause-Result | otherwise-e | question-answer-n | Result |
| concession | Otherwise | question-answer-s | Same-Unit |
| concession-e | problem-solution | question-answer-n-e | Same-Unit-NS |
| condition | problem-solution-e | question-answer-s-e | Same-Unit-SN |
| condition-e | problem-solution-n | Question-Answer | Sequence |
| Contrast | problem-solution-s | statement-response-n | topic-drift |
| Disjunction | problem-solution-n-e | statement-response-s | topic-shift |
| Inverted-Sequence | problem-solution-s-e | Statement-Response | Topic-Drift |
| Joint | Problem-Solution | reason | Topic-Shift |

Tabela A.2. Relações pertencentes à categoria + Importantes (fator de importância = 0.6)

| | | | |
|---|---|---|---|
| comparison | Enablement | evaluation-n-e | nonrestrictive-relative-e |
| comparison-e | evaluation | evaluation-s-e | preference |
| Comparison | evaluation-e | Evaluation | preference-e |
| enablement | evaluation-n | means | relative-e |
| enablement-e | evaluation-s | means-e | restrictive-rel-e |

Tabela A.3. Relações pertencentes à categoria - Importantes (fator de importância = 0.4)

| | | | |
|---|---|---|---|
| Abstract | consequence-n-e | interpretation-n | summary |
| analogy | consequence-s-e | interpretation-s | summary-e |
| analogy-e | Consequence | interpretation-n-e | summary-n |
| Analogy | contingency | interpretation-s-e | summary-s |
| Attribution | contingency-e | Interpretation | summary-n-e |
| Author | evidence | justify | summary-s-e |
| Column-Title | evidence-e | justify-e | Summary |
| comment | explanation-argumentative | Parallel | Text |
| comment-e | explanation-argumentative-e | Proportion | TextualOrganization |
| Comment-Topic | Heading | restatement | Title |
| conclusion | hypothetical | restatement-e | Topic |
| conclusion-e | hypothetical-e | rhetorical-question | Topic-Comment |
| consequence-n | interpretation | SectionText | Topic-WA-Comment |
| consequence-s | interpretation-e | SectionTitle | |

Tabela A.4. Relações pertencentes à categoria -- Importantes (fator de importância = 0.2)

| | | |
|---|---|---|
| attribution | elaboration-part-whole | OTHERrel |
| attribution-e | elaboration-process-step | OTHERrel-e |
| attribution-n | elaboration-object-attribute | OTHERmultinuc |
| background | elaboration-general-specific | parenthetical |
| background-e | elaboration-additional-e | temporal-after |
| circumstance | elaboration-set-member-e | temporal-before |
| circumstance-e | elaboration-part-whole-e | temporal-sametime |
| definition | elaboration-process-step-e | temporal-after-e |
| definition-e | elaboration-object-attribute-e | temporal-before-e |
| elaboration | elaboration-general-specific-e | temporal-sametime-e |
| elaboration-e | example | TemporalSameTime |
| elaboration-additional | example-e | |
| elaboration-set-member | motivation | |