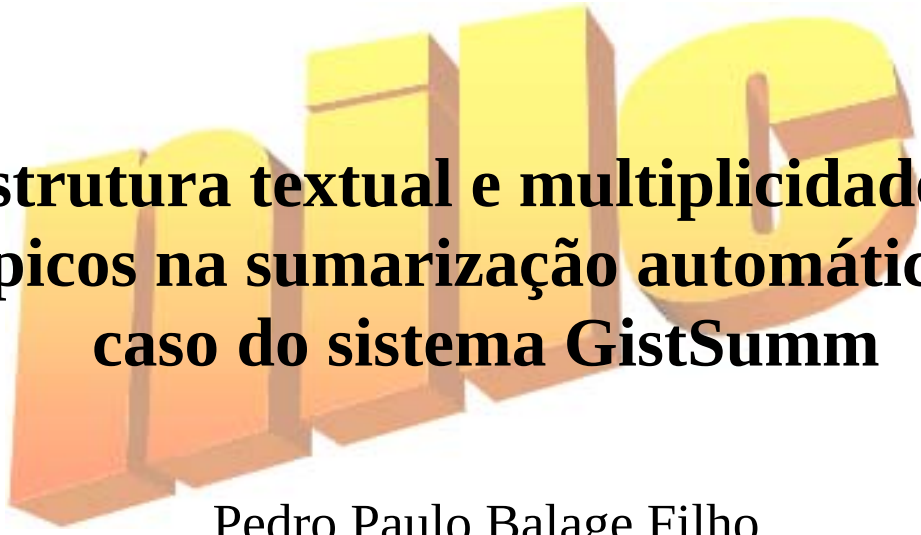
Universidade de São Paulo - USP



Estrutura textual e multiplicidade de tópicos na sumarização automática: o caso do sistema GistSumm

Pedro Paulo Balage Filho
Vinícius Rodrigues de Uzêda
Thiago Alexandre Salgueiro Pardo
Maria das Graças Volpe Nunes

NILC-TR-10-06

Novembro, 2006

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

Resumo

Apresenta-se, neste relatório, a descrição das modificações feitas no sistema GistSumm – *GIST SUMMarizer* (Pardo, 2002, 2005; Pardo et al., 2003), um sumarizador automático de textos. Após uma avaliação do sistema, alguns pontos foram indicados como passíveis de aperfeiçoamento: (a) segmentação sentencial, (b) reconhecimento e tratamento da estrutura textual e (c) detecção de mais de um tópico textual. As modificações do sistema para o estudo dessas melhorias são descritas neste relatório.

ÍNDICE

1. Introdução

A quantidade de informação disponível no atual momento é muito grande, impossível de ser apreendida em sua totalidade. Para amenizar o problema, costuma-se procurar versões menores, mais enxutas: resumos. Também chamados de sumários, tratam-se de versões reduzidas dos textos a que se referem e contêm as idéias principais dos mesmos (Mani, 2001). Há diversos tipos de documentos que são exemplos de sumários: previsões meteorológicas, sinopses de novelas, chamadas de notícias jornalísticas, resenhas e resumos de livros e teses.

Apesar de bastante úteis, a produção dos sumários é bastante trabalhosa, visto que é necessária a leitura e interpretação do texto para, então, perceberem-se suas idéias centrais. Por isso, hoje em dia, buscam-se formas automáticas de produzir esses resumos. A esta área de estudo dá-se o nome de Sumarização Automática de Textos. A sumarização automática caracteriza-se pela geração de um sumário através de métodos computacionais. Essa área tem se tornado proeminente devido à crescente demanda por informação, relacionada, principalmente, com o crescimento da Internet e o desenvolvimento de sistemas de informação que dela se aproveitam.

Tradicionalmente, definem-se duas formas de se abordar o problema da sumarização: a superficial e a profunda. Na primeira, utilizam-se métodos estatísticos e/ou empíricos para se obter o sumário. Mais simples de ser implementada, trata-se de uma grande área de pesquisa, mas pode produzir sumários com problemas de coesão e coerência. Na abordagem profunda, são utilizadas técnicas formais e modelos lingüísticos, o que aumenta a sua complexidade de desenvolvimento. Há, também, muitos sistemas de natureza híbrida, que utilizam técnicas das duas abordagens.

Existem dois tipos básicos de sumários: o extrato e o *abstract*. O extrato é um resumo produzido extraindo-se do texto-fonte frases que expressem as idéias principais, ao contrário do *abstract*, uma forma distinta de apresentar as mesmas idéias que o autor do texto desejava expor ao leitor. Os extratos são, em geral, produzidos por métodos da abordagem superficial. Os *abstracts*, por sua vez, podem ser produzidos na abordagem profunda. Os sumários podem ainda ser classificados como indicativos, informativos ou críticos (Mani e Maybury, 1999). Sumários indicativos apenas listam ou indicam o assunto principal dos textos-fonte; os informativos são autocontidos, isto é, possuem toda a informação essencial dos textos-fonte, dispensando a leitura destes; os críticos avaliam ou apenas comentam o conteúdo de suas fontes.

Para o português do Brasil, há, atualmente, diversos sumarizadores disponíveis, como apresentado por Rino et al. (2004). Um dos primeiros sumarizadores que surgiu para esta língua e que ainda é bastante utilizado é o GistSumm (GIST SUMMarizer) (Pardo, 2002, 2005; Pardo et al., 2003). Por se basear em um método superficial relativamente simples e produzir somente extratos, o sistema apresenta diversas limitações. Neste relatório, são descritas as adaptações promovidas no sistema, visando seu aprimoramento. Em particular, tratam-se os casos de reconhecimento e processamento automático da estrutura textual e de multiplicidade de tópicos em um texto.

Na próxima seção, o processo de sumarização do GistSumm é brevemente revisto. Na Seção 3, o GistSumm é avaliado, resultando nas adaptações efetuadas apresentadas na Seção 4. Por fim, comentários finais são feitos na Seção 5.

2. A sumarização no GistSumm

O GistSumm tenta simular a forma como a sumarização humana acontece. Inicialmente, procura-se pela idéia principal do texto-fonte para, então, complementá-la com informações adicionais relevantes.

Em sua versão inicial, o GistSumm possuía as seguintes características:

- realização de sumarização monodocumento, ou seja, para produção do sumário, um único texto-fonte é dado como entrada para o sistema;
- realização de sumarização extrativa, isto é, produção do sumário (ou extrato) de um texto-fonte pela seleção de sentenças inteiras do texto e posterior justaposição delas;
- pela razão anterior, é de natureza intersentencial, ou seja, não se realiza sumarização no interior das sentenças;
- produção de sumários genéricos, que são sumários voltados para uma audiência qualquer, sem interesses especificados.

Em sua segunda versão, novas funcionalidades foram adicionadas ao sistema:

- realização de sumarização multidocumento, em que vários textos-fonte são fornecidos ao sistema para produção de um único sumário;
- produção de sumários focados nos interesses da audiência, isto é, sumários que tentam, de alguma forma, responder perguntas ou apresentar fatos relevantes sobre um tópico especificado pelo usuário;
- realização de sumarização intra-sentencial, isto é, sumarização no interior das sentenças.

O processo de sumarização no GistSumm consiste, basicamente, em três etapas: segmentação sentencial, ranqueamento e seleção de sentenças.

Na segmentação sentencial, as sentenças do texto-fonte são identificadas por meio de regras simples baseadas na ocorrência de sinais de pontuação, como o ponto e os sinais de interrogação e de exclamação. Verifica-se, também, a presença de abreviaturas (por meio de uma lista de abreviaturas) para diferenciar o ponto que segue as palavras desta classe do ponto delimitador de sentenças.

O ranqueamento de sentenças consiste em atribuir pontuação às sentenças identificadas na etapa anterior e produzir um ranque destas sentenças. Essa etapa compreende vários passos, como delineados a seguir:

a) *case folding*: todas as letras das sentenças são transformadas em letras minúsculas;
b) *stemming*: por meio de um *stemmer*, as palavras do texto são substituídas pelas suas respectivas raízes (*stems*);
c) remoção de *stopwords*: as *stopwords* (que são palavras muito comuns e, portanto, irrelevantes para o processamento em questão) são removidas do texto;
d) pontuação de sentenças: a pontuação de sentenças pode ocorrer por um de dois métodos estatísticos simples: métodos *keywords* (Black & Johnson, 1988) ou o método *average keywords*, isto é, o método *keywords* com normalização em função do tamanho das sentenças (medido em número de palavras);
e) ranqueamento das sentenças em função da pontuação obtida no passo anterior, sendo que a sentença de maior pontuação é eleita g*ist sentence*, isto é, a sentença que melhor representa a idéia principal do texto.

Os passos (a), (b) e (c) são para fins de uniformização dos dados e para a produção de melhores resultados. Em relação ao passo (b), utiliza-se um *stemmer* que segue o modelo de Porter (1980); para o português, em particular, utiliza-se uma adaptação desse *stemmer* (Caldas Jr. et al., 2001). Sobre o passo (c), utiliza-se uma lista de *stopwords* (isto é, uma *stoplist*) para se identificar essas palavras.

Por fim, na etapa de seleção de sentenças, são selecionadas as sentenças que formarão o sumário. Selecionam-se as sentenças que: (a) contenham pelo menos um *stem* em comum com a *gist sentence* selecionada na etapa anterior e (b) tenham uma pontuação maior do que um *threshold*, que é a média das pontuações das sentenças. Por (a), procura-se selecionar sentenças

que complementem a idéia principal do texto; por (b), procura-se selecionar somente sentenças relevantes.

O número de sentenças selecionadas para formar o sumário, por sua vez, depende da taxa de compressão especificada pelo usuário do sistema. A taxa de compressão é uma medida que determina o tamanho do sumário em relação ao tamanho do texto-fonte.

Como já mencionado, algumas novas funcionalidades foram adicionadas ao GistSumm em sua segunda versão: sumarização intra-sentencial, multidocumento e focada em interesses do leitor.

A sumarização intra-setencial, quando requerida pelo usuário do sistema, é realizada em todas as sentenças pela exclusão das *stopwords*. Apesar das sentenças resultantes terem a legibilidade prejudicada, o tamanho das sentenças é significativamente reduzido.

Para a realização da sumarização multidocumento, todos os textos dados ao sistema são justapostos, como se fossem um único texto, e o processo tradicional de sumarização do GistSumm é realizado. Questões complexas da sumarização multidocumento não são tratadas, como o reconhecimento e eliminação de informações redundantes provenientes de diferentes textos e a ordenação temporal dos eventos relatados nos textos.

Para a produção de sumários focados nos interesses do usuário, utiliza-se uma sentença de consulta fornecida pelo usuário na qual o sistema irá procurar pela *gist sentence* que mais se assemelhe a essa sentença de consulta (em vez da sentença com maior pontuação). A busca desta *gist sentence* se dá pelo cálculo da medida do cosseno (Salton, 1989) entre as sentenças do texto-fonte (ou textos-fonte, no caso de sumarização multidocumento) e a consulta especificada. Com base nessa medida, a sentença mais próxima da consulta especificada é determinada e escolhida como *gist sentence*.

Para mais detalhes sobre o GistSumm e seu processo de sumarização, sugere-se a leitura das obras de referência. A próxima seção discute a avaliação do sistema e os problemas encontrados no método acima descrito a partir da análise de exemplos gerados.

3. Avaliação do método de sumarização do GistSumm

Como um estudo inicial, o método de sumarização do GistSumm foi aplicado em alguns textos para estimar a eficácia do mesmo. Foi utilizada uma amostra de textos de cunho geral retirada da Internet. Deu-se preferência a textos jornalísticos, prática comum em sumarização automática. Utilizou-se uma taxa de compactação entre 60 e 80% na obtenção dos sumários.

Com a análise dos resultados obtidos, detectou-se que:

- em muitos textos, a idéia principal não se apresentava contida na sua sentença de maior pontuação (*gist sentence*); percebeu-se que um conjunto um pouco maior de sentenças poderia transmitir a idéia principal do texto de forma mais adequada;
- em textos estruturados (como artigos científicos, por exemplo), muitas vezes a *gist sentence* escolhida não contemplava a idéia geral do texto; os sumários obtidos eram direcionados mais para a seção de onde era retirada a sentença principal, muitas vezes prejudicando as demais seções;
- utilizando-se o método de segmentação sentencial do GistSumm, alguns subtítulos eram considerados parte do segmento seguinte por usualmente não apresentarem sinal de pontuação.

Sistemas de sumarização que produzem extratos também possuem outros problemas mais conhecidos, como resolução de anáforas, coesão textual, etc. Esses problemas são campo para várias pesquisas atuais e de solução complexa. Portanto, para o propósito deste trabalho, buscou-se apenas identificar e propor soluções para os problemas mais triviais e marcantes. A análise desses problemas é abordada mais profundamente nas subseções seguintes.

3.1. Texto cuja idéia principal se encontra dispersa

Alguns textos dissertam sobre vários assuntos correlatos, mas nem sempre estes assuntos podem ser resumidos sob uma única sentença. O método tradicionalmente aplicado pelo GistSumm elege uma única *gist sentence*, podendo descartar teses sobre assuntos menos significativos, prejudicando o sumário.

No texto da Figura 3.1, o trecho inicial disserta sobre a atividade da Polícia Federal, o trecho com subtítulo "Prisões**"** trata das prisões que foram feitas durante o mesmo final de semana, enquanto que o último trecho explica o que é a Operação Toupeira. Cada trecho deste texto deveria ter uma influência no sumário final, já que cada um foca um assunto diferente.

PF recupera R$ 646 mil em ações contra o PCC

Dinheiro pode ser parte dos R$ 164,7 milhões roubados do BC de Fortaleza
Fabio Graner e Jocyelma Santana

PALMAS - A Polícia Federal encontrou na madrugada deste domingo R$ 196,6 mil na sede da fazenda Boa Sorte, em Pium, a 120 km de Palmas (TO). O dinheiro estava escondido numa caixa de isopor, sob um piso falso na sala da sede do imóvel. Com isso, chegou a R$ 646 mil o total recuperado pelo PF com a Operação Facção Toupeira. A suspeita é que o dinheiro faça parte dos R$ 164 milhões furtados em 2005 do Banco Central de Fortaleza.

Na fazenda no Tocantins, os policiais também prenderam Francisco do Nascimento Barbosa, o Chicão, 57 anos, que tem mandado de prisão preventiva expedido pela Justiça Federal do Ceará. A propriedade, com 47 alqueires, pertence a Raimundo Laurindo Barbosa Neto, preso na última sexta-feira, dia 1º, em Parnaíba (PI), acusado de participar do assalto ao Banco Central de Fortaleza. A fazenda estava sendo monitorada por agentes federais há cinco meses, segundo a superintendente da Polícia Federal no Tocantins, Neide Alves Almeida Alvarenga. "Nós sabíamos que a fazenda era de um dos participantes do assalto do Banco Central" , disse.

O dinheiro está na Superintendência da Polícia Federal em Palmas e deve ser periciado nesta segunda-feira, dia 4, para confirmar se as notas pertencem à mesma série das que foram retiradas do Banco Central. Só depois os valores serão transferidos para uma agência bancária.

Prisões

O Ministério da Justiça também informou que neste final de semana, além do dinheiro recuperado, mais duas pessoas foram presas em São Paulo, acusadas de envolvimento no assalto ao BC. Com isso, o total de criminosos capturados na operação subiu para 42.

A PF estuda transferir cinco líderes do PCC capturados na operação para o presídio federal de Catanduvas, no Paraná, segundo o Ministério da Justiça. Esses chefes foram capturados nos Estados do Rio Grande do Sul, São Paulo e Piauí. De acordo com o ministério, essa transferência não deve ocorrer nesta semana, porque a PF quer encerrar antes a primeira parte do inquérito, ouvindo os criminosos capturados.

Operação Toupeira

A Operação Facção Toupeira desbaratou uma ação que visava assaltar simultaneamente o Banrisul e a Caixa Econômica Federal de Porto Alegre. Além disso, prendeu, entre os 40 criminosos, dois acusados de serem líderes do PCC, como Lucivaldo Laurindo, o Torturado - suspeito de ter sido o mentor do assalto ao BC de Fortaleza e da tentativa frustrada no Rio Grande do Sul -, e Carlos Alberto da Silva, o Balengo - suspeito de ter liderado o seqüestro do jornalista da TV Globo, Guilherme Portanova, e do cinegrafista Alexandre Calado.

Figura 3.1. Texto com mais de uma idéia central

3.2. Textos estruturados

O tratamento do texto tradicionalmente realizado pelo GistSumm prejudica a extração de uma *gist sentence* que representasse a idéia principal em textos estruturados, como artigos científicos, por exemplo, que apresentam diversas seções. Nestes textos, a seção à qual pertence a *gist sentence* é privilegiada na extração das sentenças para formar o sumário. Em alguns textos observou-se que nem sempre a seção da onde se retira a sentença de maior pontuação é a responsável pela transmissão da idéia principal de um texto. Com isso, as sentenças de outras seções que podem conter informações importantes são penalizadas.

Como em muitos textos o objetivo das seções é explorar aspectos variados de um assunto, o tratamento do texto como uma estrutura única, sem preservar a divisão em seções, leva normalmente esses textos a terem sumários pouco abrangentes e ruins.

Na Figura 3.2, mostra-se um texto retirado do *site* do jornal *Folha de São Paulo*, exemplificando a análise. Note que o texto tem uma subseção intitulada *Gay pay-per-view*. O sumário obtido com esse texto refletiu apenas a segunda seção (*Gay pay-per-view*) de onde foi obtida a *gist sentence*, como mostra a Figura 3.3. O restante do texto ficou prejudicado, originando então um sumário desconexo.

O jovem que cresceu vendo a MTV desde 1990 e enjoou do perfil adolescente da emissora musical, tem a partir desta segunda-feira a opção de manter-se fiel a seus princípios de entretenimento, mas num canal um pouco mais adulto.

Entra em operação nesta segunda-feira o VH1, de variedades, cultura pop e música.

A faixa etária de público pretendida, de 25 a 49 anos, é a mesma atingida pelo canal-matriz norte-americano, ligado à própria MTV e à gigante Viacom.

O objetivo também é captar a audiência formada pela MTV Brasil, mas que, passado um tempo, quer algo mais.

Nos países latinos onde estreou – Argentina e México, por exemplo -, o canal começou devagar e foi subindo na audiência.

A expectativa é a mesma para a performance brasileira.

Pelo menos um terço da programação terá produção e profissionais brasileiros.

Rostos e vozes de Marisa Monte, Seu Jorge, Lulu Santos e Los Hermanos dividirão a tela com, entre tantos, Paul McCartney, Alannis Morisette, Madonna e a trupe dos Stones.

O restante dos programas são da matriz – exibidos com legendas.

Programas originais do canal tratam de temas como cinema, música, bastidores do mundo pop e detalhes da vida de astros e estrelas.

No Grande ABC, a Sky é a operadora disponível para assinantes.

É onde o VH1 pode ser sintonizado, pelo menos por enquanto.

"A Vivax também está na lista das operadoras que devem fechar com o canal, mas ainda faltam detalhes contratuais e a decisão pode sair ainda este ano", afirma Cristina Bandiera, diretora de marketing da Vivax.

Enquanto isso, na estréia do canal, Gisele Bündchen e Leonardo DiCaprio.

Eles aparecem como protagonistas em A Vida Glamourosa, produção feita enquanto eles ainda namoravam e que vai ao ar nesta segunda-feira, às 19h.

Trinta minutos depois do ex-casal, é a vez do programa mostrar o que o bad boy do hip hop P.Diddy – ex-Puffy Daddy – tem de glamour em sua vida.

All Access, às 20h, também estréia nesta segunda-feira tendo o mundo do rock – fama, intriga, moda – como mote principal.

Seriados e reality shows também fazem parte da programação, bem como filmes, tanto clássicos quanto inéditos.

Nesta terça-feira, às 21h, será exibido O Cantor de Jazz.

Na quinta-feira, às 19h, passa o documentário Beyonce Knowles, no bloco Driven, sobre a vida da cantora Beyoncé.

Da timidez da infância, ela passa pelo grupo Destiny's Child e é responsabilizada por sua

dissolução até se firmar solo com o álbum Survivor.

Uma história de sucesso e superação bem ao gosto norte-americano.

Na sexta, o cartaz é o cult Os Irmãos Cara-de-Pau.

Gay pay-per-view

Além do VH1, também entrou no ar pela Sky, na sexta-feira passada, o Logo TV, primeiro serviço pay-per-view da América Latina para o público gay.

Ambos são canais administrados pela Viacom Networks Brasil – responsável ainda pelo infanto-juvenil Nickelodeon – e farão parte da MTV Networks Latin América, divisão da Viacom.

A MTV Brasil é gerenciada no país pelo Grupo Abril, e será parceira das operações da Viacom Brasil no canal VH1.

Figura 3.2. Exemplo de texto estruturado

A faixa etária de público pretendida, de 25 a 49 anos, é a mesma atingida pelo canal-matriz norte-americano, ligado à própria MTV e à gigante Viacom.

O objetivo também é captar a audiência formada pela MTV Brasil, mas que, passado um tempo, quer algo mais.

"A Vivax também está na lista das operadoras que devem fechar com o canal, mas ainda faltam detalhes contratuais e a decisão pode sair ainda este ano", afirma Cristina Bandiera, diretora de marketing da Vivax.

Eles aparecem como protagonistas em A Vida Glamourosa, produção feita enquanto eles ainda namoravam e que vai ao ar nesta segunda-feira, às 19h.

Da timidez da infância, ela passa pelo grupo Destiny's Child e é responsabilizada por sua dissolução até se firmar solo com o álbum Survivor.

Além do VH1, também entrou no ar pela Sky, na sexta-feira passada, o Logo TV, primeiro serviço pay-per-view da América Latina para o público gay.

Ambos são canais administrados pela Viacom Networks Brasil - responsável ainda pelo infanto-juvenil Nickelodeon - e farão parte da MTV Networks Latin America, divisão da Viacom.

A MTV Brasil é gerenciada no país pelo Grupo Abril, e será parceira das operações da Viacom Brasil no canal VH1.

Figura 3.3. Sumário obtido pelo GistSumm para o texto da Figura 3.2

Neste exemplo, tem-se que as três últimas sentenças do texto, isto é, aquelas que compunham a seção *Gay pay-per-view*, são mais bem pontuadas e vão para o sumário. Neste caso, a sentença principal, a *gist sentence*, foi a penúltima frase: “Ambos são canais administrados pela Viacom Networks (...)”.

Em decorrência disso, pouco do restante do texto, que é maioria, foi inserido no sumário. Esse mesmo texto identifica uma posição mais centrada no assunto e genérica, o mesmo que o autor do texto gostaria de transmitir, enquanto a última seção da qual foi extraída todas as suas sentenças identifica apenas uma posição específica do assunto. Algo que o autor gostaria apenas de comentar em seu texto.

Textos estruturados de grande interesse são artigos científicos, teses e dissertações. Normalmente esse tipo de texto é caracterizado por seguir uma estrutura onde tópicos como Introdução, Métodos e Conclusão são só alguns dos exemplos que podemos encontrar nesse tipo de texto. A geração de sumários que consideram as seções presentes nesses textos são de grande interesse como resultado do trabalho. Na seção 4.4, comenta-se especificamente esse tipo de texto.

É importante notar a diferença entre as questões da *gist sentence* dispersa e da sumarização de textos estruturados. No primeiro caso, a idéia principal do texto é expressa em várias *gist sentences*, pertencentes ou não em seções diferentes. No segundo caso, a questão consiste em garantir que cada seção de um texto estruturado contribua para a formação do sumário.

3.3. Precauções pré-sumarização

A forma de tratamento do texto dada pelo GistSumm não compreendia o fato de que títulos e subtítulos de textos não costumam ser pontuados, além de não haver forma de lidar com nomes de empresas, produtos etc., que possuem pontuação (por exemplo, *Level Up!*, nome de uma empresa desenvolvedora de jogos, que, apesar de possuir esta exclamação em seu nome, não identifica um fim de período). Acredita-se que muitos desses detalhes devem ser considerados em uma etapa de pré-processamento, manualmente, preferencialmente, pois variam de forma muito imprecisa.

4. Aprimoramento do GistSumm

Após a etapa de avaliação dos resultados do GistSumm, algumas mudanças foram feitas no sistema, visando a melhorar seu desempenho. A Tabela 4.1 mostra, para cada problema identificado acima, a solução encontrada. Cada modificação será discutida mais profundamente nas subseções posteriores.

Problema	**Solução**
Necessidade de um pré-processamento do texto para melhorar a sumarização	Parcialmente solucionado: uma chamada externa ao SENTER (Pardo, 2006)
Falta de tratamento da estrutura textual	Identificação de seções do texto, aplicando o método convencional para cada seção identificada de forma individual
Texto cuja idéia principal está dispersa	Seleção de múltiplas *gist sentences*

Tabela 4.1. Soluções encontradas para alguns dos problemas do GistSumm

4.1. Segmentação textual

O sistema de delimitação de sentenças desenvolvido no NILC, SENTER (Pardo, 2006), foi proposto como alternativa para o sistema de segmentação sentencial do GistSumm. Optou-se, então, pela acoplagem do SENTER ao sistema, feito através de uma chamada do GistSumm antes de iniciar o processo de sumarização. Assim, no processo de sumarização, o texto que será processado é aquele resultante da saída do SENTER.

O uso do SENTER se mostrou mais eficaz que o método de segmentação original. Essa melhora também refletiu na resolução dos problemas que foram inicialmente levantados nesse aspecto.

4.2. Múltiplas *gist sentences*

A modificação no critério de seleção de sentenças deveria tornar possível que mais de uma sentença fosse indicada como *gist sentence*, visto que, por exemplo, em textos com estruturas internas bem definidas, cada estrutura pode ter suas próprias idéias centrais.

A Figura 4.3 mostra um texto utilizado para exemplo, extraído da *Folha On-line*.

Violência israelo-libanesa gera fuga em massa para a Síria

Publicidade
DANIELA LORETO da Folha Online

A mais recente onda de violência envolvendo Israel e Líbano desencadeou uma fuga em massa de brasileiros e outros estrangeiros que vivem no sul do território libanês em direção à Síria. Iniciada

ontem, a nova crise já deixou 53 mortos e 103 feridos no Líbano [incluindo quatro brasileiros da mesma família], e um morto e cerca de cem feridos em Israel.

A Síria fechou a entrada por sua fronteira com o Líbano devido ao grande número de pessoas que tentam deixar o país em direção a Damasco após a ofensiva militar israelense lançada ontem, segundo o cônsul-geral do Brasil em Beirute, Michael Gepp.

Segundo ele, um grande número de brasileiros estão entre os que tentam deixar o Líbano. "Há brasileiros em Beirute estão assustados com a situação e tentam sair do país", disse.

Procurado pela Folha Online, o consulado brasileiro em Damasco não confirmou a informação de que a fronteira com o Líbano esteja fechada, mas informou haver possibilidade de que isso realmente esteja acontecendo.

"Há muitos brasileiros e descendentes de árabes em férias que estão tentando deixar Beirute", afirmou a vice-cônsul brasileira em Damasco, Ana Maria Azevedo. "Há três quilômetros de congestionamento na região da fronteira e muita gente tenta deixar o Líbano. Os hotéis em Damasco estão lotados. Existe a possibilidade de que a fronteira tenha sido fechada, mas não é oficial", acrescentou.

Segundo ela, a Embaixada do Brasil em Damasco não foi procurada diretamente por brasileiros que tentam deixar o Líbano. "Tivemos contato apenas com o Consulado em Beirute".

De acordo informações de agências internacionais, cerca de 12 mil turistas deixaram o Líbano nesta quinta-feira.

Táxi

De acordo com o cônsul brasileiro em Beirute, muitos tentam escapar para Damasco usando táxi, mas a fronteira está fechada. "Não é que a Síria fechou as portas para o Líbano, mas o país não tem condições de absorver um fluxo imigratório tão grande".

Como exemplo, Gepp cita um brasileiro que pretendia ir de táxi até a capital síria para embarcar para Paris, onde seus netos o aguardavam. "No entanto, entrei em contato com as autoridades da imigração na Síria e me disseram que ele ficaria preso na fronteira, não teria onde dormir."

Gepp explica que o brasileiro resolveu permanecer em Beirute após ser informado de que não conseguiria entrar em território sírio. Segundo ele, um grupo de brasileiros que tentava deixar o país de ônibus pela cidade portuária de Tiro foram frustrados ao receber ordens para retornar. "Tiro é uma cidade histórica, que está isolada porque Israel destruiu quatro de suas pontes", afirmou.

O cônsul negou, no entanto, informações de que Israel teria bombardeado a estrada entre Beirute e Damasco. "Não tenho informações sobre um ataque deste tipo".

Neste momento, nem pela via aérea é possível sair do Líbano. "Os que foram para o aeroporto de Beirute na manhã desta quinta-feira ainda não conseguiram embarcar. Agora os vôos estão lotados", explica.

Segurança

Embora a tensão a região seja crescente, Gepp afirma que não há motivo para pânico entre brasileiros que estão no Líbano. "Recomendamos aos brasileiros no país que permaneçam em seus hotéis e mantenham contato com o consulado", disse.

Outra orientação dada por ele é que os brasileiros entrem em contato com as companhias aéreas que os transportaram ao Líbano para que elas os auxiliem a deixar o país. Segundo Gepp, o Consulado trabalha ao lado da embaixada brasileira e mantém contato com o Itamaraty.

"A embaixada passa as informações sobre a situação política, e o consulado, sobre o auxílio consular. O Itamaraty acompanha a situação, e deve ordenar uma retirada dos brasileiros se isso for necessário, mas não é o caso neste momento", afirmou.

Figura 4.1. Texto-exemplo

As figuras 4.2 e 4.3 mostram, respectivamente, a sentença eleita como *gist sentence* e o sumário obtido, com uma taxa de compressão de 60% (valor apenas para exemplificar) pelo GistSumm, fazendo uso do SENTER.

"Há três quilômetros de congestionamento na região da fronteira e muita gente tenta deixar o Líbano. Os hotéis em Damasco estão lotados. Existe a possibilidade de que a fronteira tenha sido fechada, mas não é oficial", acrescentou.

Figura 4.2. *Gist sentence* eleita pelo GistSumm com SENTER para o texto-exemplo

A Síria fechou a entrada por sua fronteira com o Líbano devido ao grande número de pessoas que tentam deixar o país em direção a Damasco após a ofensiva militar israelense lançada ontem, segundo o cônsul-geral do Brasil em Beirute, Michael Gepp.

Procurado pela Folha Online, o consulado brasileiro em Damasco não confirmou a informação de que a fronteira com o Líbano esteja fechada, mas informou haver possibilidade de que isso realmente esteja acontecendo.

"Há muitos brasileiros e descendentes de árabes em férias que estão tentando deixar Beirute", afirmou a vice-cônsul brasileira em Damasco, Ana Maria Azevedo.

"Há três quilômetros de congestionamento na região da fronteira e muita gente tenta deixar o Líbano. Os hotéis em Damasco estão lotados. Existe a possibilidade de que a fronteira tenha sido fechada, mas não é oficial", acrescentou.

Segundo ela, a Embaixada do Brasil em Damasco não foi procurada diretamente por brasileiros que tentam deixar o Líbano.

"Não é que a Síria fechou as portas para o Líbano, mas o país não tem condições de absorver um fluxo imigratório tão grande".

Embora a tensão a região seja crescente, Gepp afirma que não há motivo para pânico entre brasileiros que estão no Líbano.

"A embaixada passa as informações sobre a situação política, e o consulado, sobre o auxílio consular. O Itamaraty acompanha a situação, e deve ordenar uma retirada dos brasileiros se isso for necessário, mas não é o caso neste momento", afirmou.

Figura 4.3. Sumário gerado pelo GistSumm com SENTER para o texto-exemplo

A solução que foi colocada em prática foi considerar como *gist sentences* as sentenças centrais, detectadas, segundo a abordagem do GistSumm, por possuírem palavras que se repetissem com freqüência no texto, mas que fossem suficientemente diferentes umas das outras. Acredita-se, assim, que seria possível detectar sentenças significativas, mas que tratassem de assuntos distintos.

Para melhor entendimento do método utilizado, será necessário explicitar o que é a medida do cosseno. Trata-se de uma medida muito utilizada em Processamento de Línguas Naturais para se detectar a semelhança entre dois segmentos de textos. É calculada através da fórmula abaixo:

$$co\text{-}seno\ (sent1, sent2) = \frac{\sum_{palavra} (freq\ (palavra, sent1) \bullet freq\ (palavra, sent2))}{\sqrt{\sum_{palavra} (freq\ (palavra, sent1)^2) \bullet \sum_{palavra} (freq\ (palavra, sent2)^2)}}$$

em que *sent1* e *sent2* são as sentenças a serem comparadas e *freq* é uma função de dois parâmetros: o primeiro uma palavra e o segundo uma sentença, representando a freqüência com que a palavra ocorre na sentença. Assim sendo, o cosseno de duas sentenças idênticas será 1 e o de duas sentenças completamente díspares será 0.

Para a detecção de múltiplas *gist sentences*, cria-se uma lista com as sentenças de maior pontuação, mas cujas medidas do cosseno em relação às demais sentenças da lista são menores que um valor de base (no caso, o valor 0,1 foi adotado empiricamente). Cada uma dessas sentenças é, então, considerada *gist sentence*. Outros valores de base foram testados, mas praticamente sentença alguma seria selecionada com valores menores ou só haveria *gist sentences* no sumário. Visando melhorar a seleção das demais sentenças dignas de pertencer ao

sumário, percebeu-se que seria necessário calcular a pontuação de cada sentença em relação à *gist sentence* com que a sentença mais se assemelhasse. Então, ocorreu também uma nova adaptação: a pontuação de cada sentença seria recalculada, tornando-se o produto da antiga pontuação pela maior medida do co-seno obtida com as sentenças da lista de *gist sentences*. Mas, com isso, a pontuação das sentenças secundárias se tornou muito baixa, enquanto a das *gist sentences* seria excessivamente superior a das demais. Achou-se viável recalcular o *threshold*. Logo, a nova pontuação mínima para que a sentença entrasse no sumário foi calculada sem incluir a pontuação das *gist sentences*, já consideradas dignas de pertencerem ao sumário, diminuindo, assim, o valor dos *thresholds* e causando uma melhor seleção das demais sentenças.

Tomando-se o texto da Figura 4.1 como exemplo, as figura 4.4 e 4.5 mostram, respectivamente, as *gist sentences* e o sumário obtido pelo GistSumm após as modificações.

A mais recente onda de violência envolvendo Israel e Líbano desencadeou uma fuga em massa de brasileiros e outros estrangeiros que vivem no sul do território libanês em direção à Síria.

"Há três quilômetros de congestionamento na região da fronteira e muita gente tenta deixar o Líbano. Os hotéis em Damasco estão lotados. Existe a possibilidade de que a fronteira tenha sido fechada, mas não é oficial", acrescentou.

Segundo ele, um grupo de brasileiros que tentava deixar o país de ônibus pela cidade portuária de Tiro foram frustrados ao receber ordens para retornar.

"A embaixada passa as informações sobre a situação política, e o consulado, sobre o auxílio consular. O Itamaraty acompanha a situação, e deve ordenar uma retirada dos brasileiros se isso for necessário, mas não é o caso neste momento", afirmou.

Figura 4.4. Lista de *gist sentences* obtidas pelo GistSumm após modificações

A mais recente onda de violência envolvendo Israel e Líbano desencadeou uma fuga em massa de brasileiros e outros estrangeiros que vivem no sul do território libanês em direção à Síria.

A Síria fechou a entrada por sua fronteira com o Líbano devido ao grande número de pessoas que tentam deixar o país em direção a Damasco após a ofensiva militar israelense lançada ontem, segundo o cônsul-geral do Brasil em Beirute, Michael Gepp.

Segundo ele, um grande número de brasileiros estão entre os que tentam deixar o Líbano.

Procurado pela Folha Online, o consulado brasileiro em Damasco não confirmou a informação de que a fronteira com o Líbano esteja fechada, mas informou haver possibilidade de que isso realmente esteja acontecendo.

"Há três quilômetros de congestionamento na região da fronteira e muita gente tenta deixar o Líbano. Os hotéis em Damasco estão lotados. Existe a possibilidade de que a fronteira tenha sido fechada, mas não é oficial", acrescentou.

Segundo ela, a Embaixada do Brasil em Damasco não foi procurada diretamente por brasileiros que tentam deixar o Líbano.

Segundo ele, um grupo de brasileiros que tentava deixar o país de ônibus pela cidade portuária de Tiro foram frustrados ao receber ordens para retornar.

"A embaixada passa as informações sobre a situação política, e o consulado, sobre o auxílio consular. O Itamaraty acompanha a situação, e deve ordenar uma retirada dos brasileiros se isso for necessário, mas não é o caso neste momento", afirmou.

Figura 4.5. Sumário gerado pelo GistSumm após modificações

Em uma avaliação subjetiva, notaram-se algumas melhoras para alguns textos, mas isso pode não ser verdade sempre. Uma avaliação objetiva, com mais textos, deve ser realizada.

4.3. Tratamento da estrutura textual

Para tornar o sistema capaz de reconhecer e tratar a estrutura textual, buscou-se primeiramente entender como normalmente os textos estruturados são organizados. Com a análise de um conjunto de textos de diversos gêneros, concluiu-se que os textos se dividem em

blocos precedidos de títulos que os definem, os títulos de seção. Para ilustrar os casos encontrados, são exibidas 3 diferentes formas dessa divisão se apresentar nas Figuras 4.6 e 4.7.

(...) *um maremoto causou a morte de 280 mil pessoas em países do sul da Ásia e leste da África.*

Prevenção

Segundo a OMS e o Unaids, o Relatório Mundial sobre a Epidemia de Aids deste ano está centrado na prevenção do HIV.

(...) *O fluxo de operações estará direcionado para o atendimento do cliente.*

2 RELACIONAMENTO COOPERATIVO

A responsabilidade compartilhada (...)

Figura 4.6. Exemplos de subseções de um texto

Gripe *- A explicação oficial para os quatro dias de sumiço, quando deixou de comparecer a cerimônias oficiais* (...)

Leis *- Com Yeltsin parecendo estar na ante-sala da UTI, não se sabe o que pode acontecer de hoje para amanhã.* (...)

Figura 4.7. Exemplo de listagem de itens em um texto

Com a observação desses padrões, foi definido então o que seria considerado uma seção: um trecho de texto que vem precedido de um título, consistindo de uma sentença com no máximo 90 caracteres de comprimento (valor determinado empiricamente) e que não possuísse pontuação final.

Com a estrutura textual caracterizada, o sistema então foi modificado para que houvesse o reconhecimento das diversas seções e seus títulos. Dos exemplos mostrados anteriormente, o último caso não foi implementado devido as ambigüidades na sua estrutura. O uso de travessões precedidos de nomes também são bastante utilizados para indicar fala de um personagem em texto, tornando-se assim essa forma ambígua para o sistema. O texto da Figuras 4.8 retirado do Corpus TeMário (Pardo e Rino, 2003) do NILC serve de exemplo para ilustrar as seções identificadas pelo sistema.

Grandes cidades devem perder população => ***1ª seção***

Tendência nesta próxima década é de desconcentração populacional por uma melhor qualidade de vida => ***1ª seção***

VICTOR AGOSTINHO => ***2ª seção***
Da Reportagem Local => ***3ª seção***

Até o fim do século o mundo vai assistir (...) => ***3ª seção***

Figura 4.8. Texto dividido em seções

O texto a ser processado, então, será segmentado em seções de modo em que cada seção tenha o seu processamento em separado, com sua própria *gist sentence* e valores de pontuação de sentenças e limite de corte. Cada seção terá seu número máximo de palavras calculado a partir da

taxa de compressão, o que conserva a mesma válida para o texto em geral (Propriedade Distributiva da Matemática).

O sistema age como se cada seção fosse um texto diferente que ele deveria sumarizar, sendo, então, o sumário de um texto com seções o mesmo que o conjunto de sumários separados de cada seção.

O texto da Figura 3.2 foi processado no sistema com reconhecimento de tópicos. O sumário produzido é exibido na Figura 4.9 a seguir.

O jovem que cresceu vendo a MTV desde 1990 e enjoou do perfil adolescente da emissora musical, tem a partir desta segunda-feira a opção de manter-se fiel a seus princípios de entretenimento, mas num canal um pouco mais adulto.

O objetivo também é captar a audiência formada pela MTV Brasil, mas que, passado um tempo, quer algo mais.

Programas originais do canal tratam de temas como cinema, música, bastidores do mundo pop e detalhes da vida de astros e estrelas.

"A Vivax também está na lista das operadoras que devem fechar com o canal, mas ainda faltam detalhes contratuais e a decisão pode sair ainda este ano", afirma Cristina Bandiera, diretora de marketing da Vivax.

Eles aparecem como protagonistas em A Vida Glamourosa, produção feita enquanto eles ainda namoravam e que vai ao ar nesta segunda-feira, às 19h.

All Access, às 20h, também estréia nesta segunda-feira tendo o mundo do rock - fama, intriga, moda - como mote principal.

Ambos são canais administrados pela Viacom Networks Brasil - responsável ainda pelo infanto-juvenil Nickelodeon - e farão parte da MTV Networks Latin America, divisão da Viacom.

Figura 4.9. Sumário obtido pelo GistSumm modificado para o texto da Figura 3.2

Observa-se que nesse novo sumário há uma maior abrangência do assunto por todo o texto e não apenas na última seção como obtivemos anteriormente. Também se nota que as sentenças selecionadas para o sumário relativas à primeira seção estão mais relacionadas com a mensagem passada pelo texto-fonte. Percebe-se, assim, que o GistSumm com reconhecimento da estrutura textual produziu um sumário de melhor qualidade que o seu original.

Na análise geral dos sumários produzidos a partir da identificação e sumarização separada de seções, pode-se dizer que em média os sumários melhoraram, isto é, as sentenças melhor pontuadas e que, conseqüentemente, compunham o sumário, eram as que mais se aproximavam da idéia principal das seções. Também se detectou que sumários de textos com até duas seções não apresentavam diferença significativa em relação à sumarização tradicional.

Uma deficiência percebida foi que, em casos em que as seções possuam poucas sentenças, se a taxa de compressão for razoavelmente alta, a sentença melhor pontuada ultrapassa o limite de caracteres definido para aquela seção. Conseqüentemente, o sumário resultante dessa seção não possui nenhuma sentença, e um texto com muitas seções de poucas sentenças tende a ter seu conteúdo muito desprezado, portanto.

Uma avaliação objetiva deste método deve ser ainda realizada.

4.4. Sumarização de artigos científicos

Um outro ponto identificado para o qual um sistema de sumarização com reconhecimento e tratamento da estrutura textual seria de grande utilidade foi a sumarização de artigos científicos.

Em geral, para um artigo científico, deseja-se que o sumário aborde os aspectos principais das questões apresentadas em cada uma de suas seções. Isto é, que o sumário contenha um pouco de cada visão do artigo, como resumo, apresentação, metodologia e conclusão. Sem o

reconhecimento dessa estrutura o sumário desse artigo apresentaria pouca abrangência e seria focado dentro da seção de onde o sistema retornaria a *gist sentence*.

Devido a características próprias desse tipo de texto em foco, foram realizadas pequenas modificações em cima do GistSumm que já era capaz de reconhecer a estrutura textual. Essas modificações foram: (a) o sistema deverá manter o título da seção e (b) no mínimo uma sentença dentro de cada seção deverá compor o sumário final.

Essas modificações foram apresentadas depois de uma análise da categoria desses textos e de sugestões para sua melhoria. O item (a) é conseqüência do fato de que o sumário produzido deve manter a mesma estrutura do artigo científico para que o leitor possa analisá-lo devidamente. O item (b) é resultado da obrigação de satisfazer a taxa de compressão. Com essa obrigação, muitas seções pequenas onde a sentença principal continha mais palavras do que permitia a taxa, eram excluídas do sumário final. Como o objetivo é apresentar um sumário com um pouco de cada seção, foi definido este critério (b).

O exemplo da Figura 4.10 exibe um trecho do sumário do artigo científico *Uso de marcadores estilísticos para a busca na Web em português* de Rachel V. X. Aires e Sandra M. Aluísio, do NILC, publicado no X Simpósio de Teses e Dissertações da USP em 2005. Este sumário já apresenta as modificações sugeridas.

...

2 Marcadores de estilo

Em geral, através da freqüência dos marcadores de estilo em um texto, podemos tecer conclusões quanto a características como formalidade, elegância, complexidade sintática e complexidade lexical de um texto.

Alguns exemplos de marcadores estilísticos são: (i) marcadores relacionados a palavras, como expressões idiomáticas, expressões sofisticadas, terminologia científica, palavras formais e abreviaturas; (ii) marcadores sintáticos, como o número de palavras por frase, número de conjunções por frase, número de sentenças por parágrafo, proporção de verbos versus substantivos, porcentagem de verbos na terceira pessoa, porcentagem de orações subordinadas e proporção de adjetivos versus substantivos.

Textos formais seriam, por exemplo, caracterizados pelo grande uso de palavras formais e expressões sofisticadas, pela pouca freqüência de abreviações e expressões idiomáticas, por um número alto de palavras por frase, número pequeno de frases por parágrafo, um número alto de conjunções por frases, uma porcentagem alta de verbos na terceira pessoa e voz passiva predominante (Michos et al., 1996).

É formado por estatísticas baseadas em palavras, como número de palavras longas; estatísticas baseadas no texto como um todo, como número de frases e outras estatísticas, como número de advérbios de lugar.

O terceiro conjunto de marcadores é composto pelas 62 palavras mais freqüentes do corpus de necessidades: eliminando-se as stopwords, verbos auxiliares, advérbios, palavras relacionadas a domínios e agrupando algumas das palavras mais freqüentes como um único marcador.

3 Esquemas de classificação e Corpora

Neste trabalho, o enfoque desejado para os resultados de uma consulta pode ser selecionado de uma taxonomia de gêneros, de tipos textuais, de necessidades de busca ou de taxonomias binárias de necessidades personalizadas

O Lácio-Ref é um corpus aberto e de referência do português contemporâneo do Projeto Lácio-Web, composto de textos em português brasileiro, tendo como característica serem escritos respeitando a norma culta.

Entretanto, em sua versão atual, o corpus não contém textos do gênero de referência ou do gênero técnico-administrativo.

Em nossos experimentos com classificação em gêneros, utilizamos os textos dos gêneros disponíveis e reunimos os gêneros poesia, prosa e drama em um único supergênero Literário.

...

4 Resultados

Utilizamos um total de 44 algoritmos (Aires et al, 2004a): Naive Bayes, Naive Bayes Multinomial, Naive Bayes Updateable, Multilayer Perceptron, SMO, Simple Logistic, IB1, IBK, KStar, LWL , AdaBoostM1, Attributive Selected Classifier, Bagging, Classification via regression, CV parameter selection, Decorate, Filtered classifier, Logit Boost, Multiclass classifier, Multi Scheme, Ordinal class classifier, Raced incremental logit boost, Random committee, Stacking, Stacking C, Vote, FLR, HyperPipes, VFI, Decision Stump, J48, LMT, Random Forest, Random Tree, REP Tree, User classifier, ZeroR, Conjunctive Rule, OneR, Decision Table, Part, NNGe, Ridor, e JRIP.

Com a inclusão de mais gêneros do tipo Instrucional e a troca dos textos Informativos do corpus Lácio-Ref por textos jornalísticos da Web, a taxa de acerto foi menor, 94,87%, o que se deve a dois fatores: (1) o gênero instrucional já era problemático; com os experimentos com a versão original, os 3 textos instrucionais eram classificados erroneamente; e (2) trocamos 3.792 textos informativos por 150 textos da Web de diversas fontes (diversas seções de jornais, diversos jornais, de diferentes cidades e estados).

...

Figura 4.10. Sumário obtido com o GistSumm para artigos científicos

No exemplo da Figura 4.11, o mesmo artigo que acima é sumarizado através do GistSumm para textos científicos é sumarizado com o GistSumm original, que não leva em conta a estrutura textual. Foi utilizada uma taxa de compressão de 80%.

Alguns exemplos de marcadores estilísticos são: (i) marcadores relacionados a palavras, como expressões idiomáticas, expressões sofisticadas, terminologia científica, palavras formais e abreviaturas; (ii) marcadores sintáticos, como o número de palavras por frase, número de conjunções por frase, número de sentenças por parágrafo, proporção de verbos versus substantivos, porcentagem de verbos na terceira pessoa, porcentagem de orações subordinadas e proporção de adjetivos versus substantivos.

É formado por estatísticas baseadas em palavras, como número de palavras longas; estatísticas baseadas no texto como um todo, como número de frases e outras estatísticas, como número de advérbios de lugar.

Com a inclusão de mais gêneros do tipo Instrucional e a troca dos textos Informativos do corpus Lácio-Ref por textos jornalísticos da Web, a taxa de acerto foi menor, 94,87%, o que se deve a dois fatores: (1) o gênero instrucional já era problemático; com os experimentos com a versão original, os 3 textos instrucionais eram classificados erroneamente; e (2) trocamos 3.792 textos informativos por 150 textos da Web de diversas fontes (diversas seções de jornais, diversos jornais, de diferentes cidades e estados).

Figura 4.11. Sumário obtido com o GistSumm original

Como se observa, quando o sistema não leva em conta a estrutura textual, o sumário produzido perde o foco, resultando na escolha de sentenças que não transmitem a informação principal de cada seção. No exemplo acima, os dois primeiros parágrafos são oriundos da Seção 2 (Marcadores de Estilo). Nenhuma sentença foi obtida da Seção 3 e a última sentença é oriunda da Seção 4 (Resultados).

A seguir, algumas considerações finais são feitas.

5. Considerações finais

Os resultados apresentados ainda não são suficientes para uma avaliação precisa do método. O sistema deverá ser submetido a uma avaliação objetiva. Até o momento, pela análise dos resultados parciais, observou-se que os aprimoramentos propostos para o GistSumm obtêm bons resultados.

Em particular, a sumarização de textos científicos revelou-se um campo que pode ser muito auxiliado pelas modificações feitas no sistema. Uma outra sugestão de melhoria do sistema é a de utilizar o título de cada seção como tópico para a sumarização da própria seção. Isso deve ser investigado no futuro.

Referências

Black, W.J. and Johnson, F.C. (1988). A Practical Evaluation of Two Rule-Based Automatic Abstraction Techniques. *Expert Systems for Information Management*, Vol. 1, N. 3. Department of Computation. University of Manchester Institute of Science and Technology.

Caldas, Jr., J.; Imamura, C.Y.M.; Rezende, S.O. (2001). Evaluation of a stemming algorithm for the Portuguese language (in Portuguese). In the *Proceedings of the 2nd Congress of Logic Applied to Technology*, Vol. 2, pp. 267-274.

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

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

Pardo, T.A.S. (2002). *GistSumm: Um Sumarizador Automático Baseado na Idéia Principal de Textos*. Série de Relatórios do NILC. NILC-TR-02-13.

Pardo, T.A.S. e Rino, L.H.M. (2003). *TeMário: Um Corpus para Sumarização Automática de Textos*. Série de Relatórios do NILC. NILC-TR-03-09. São Carlos-SP, Outubro, 13p

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

Pardo, T.A.S. (2005). *GistSumm - GIST SUMMarizer: Extensões e Novas Funcionalidades*. Série de Relatórios do NILC. NILC-TR-05-05.

Pardo, T.A.S. (2006). SENTER: *Um Segmentador Sentencial Automático para o Português do Brasil*. Série de Relatórios do NILC. NILC-TR-06-01. São Carlos-SP, Janeiro, 6p.

Porter, M.F. (1980). An algorithm for suffix stripping. *Program*, Vol. 14, N. 3.

Rino, L.H.M.; Pardo, T.A.S.; Silla Jr., C.N.; Kaestner, C.A.; Pombo, M. (2004). A Comparison of Automatic Summarization Systems for Brazilian Portuguese Texts. In the *Proceedings of the 17th Brazilian Symposium on Artificial Intelligence* – SBIA (Lecture Notes in Artificial Intelligence 3171), pp. 235-244. São Luis-MA, Bazil. September, 29 - October, 1.

Salton, G. (1989). *Automatic Text Processing. The Transformation, Analysis and Retrieval of Information by Computer*. Addison-Wesley.