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

Sumarização Automática de Textos Científicos: Estudo de Caso com o Sistema GistSumm

Pedro Paulo Balage Filho
Thiago Alexandre Salgueiro Pardo
Maria das Graças Volpe Nunes

NILC-TR-07-11

Agosto, 2007

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 atividades realizadas no âmbito de iniciação científica em relação ao estudo e aprimoramento do sistema GistSumm – *GIST SUMMarizer*, um sumarizador automático de textos. Tais atividades consistiram de um estudo e adaptações promovidas no sistema, visando seu aprimoramento, avaliação das melhorias propostas e análise dos resultados, além de uma participação experimental no CLEF 2006 - *Cross-Language Evaluation Forum* – que é uma competição internacional de sistemas de perguntas e respostas. A principal adaptação promovida no sistema é o tratamento da estrutura textual durante a sumarização promovendo sumários de melhor qualidade, o que pôde ser comprovado através de avaliações subjetivas e automáticas do sistema aplicado a textos científicos.

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, ao crescimento da Internet e ao 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 atividades de âmbito de iniciação científica realizadas com este sumarizador. Tais atividades são: estudo e adaptações promovidas no sistema, visando seu aprimoramento; avaliação subjetiva e automática das melhorias propostas; e adaptação do sistema para uma participação experimental no CLEF 2006[1] – *Cross-Language Evaluation Forum* – que é uma competição internacional de sistemas de perguntas e respostas.

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. Na seção 5 é vista a metodologia da avaliação das adaptações realizadas e a análise dos seus resultados. Na seção 6 é apresentada a participação no CLEF. Alguns comentários finais são feitos na Seção 7.

[1] http://www.clef-campaign.org/

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 e 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 intrasentencial, 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 dos problemas de dispersão da idéia principal do texto e de segmentação sentencial são melhores apresentados em Balage Filho et al. (2006a). A análise do problema em textos estruturados, foco do trabalho, é abordada a seguir.

3.1. Textos estruturados

O tratamento do texto tradicionalmente realizado pelo GistSumm prejudica a extração de uma *gist sentence* que represente 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. 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 3.1, 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 3.2. 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 □mérica, 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 3.1. 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 3.2. Sumário obtido pelo GistSumm para o texto da Figura 3.1

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 considerem as seções presentes nesses textos é de grande

interesse como resultado do trabalho. Na Subseção 4.2, 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 a seções diferentes. No segundo caso, a questão consiste em garantir que cada seção de um texto estruturado contribua para a formação do sumário.

4. Aprimoramento do GistSumm

Após a etapa de avaliação dos resultados do GistSumm, foram feitas modificações no sistema para a melhoria dos problemas identificados. Em especial, foca-se, neste trabalho, no problema de falta de tratamento da estrutura textual, o qual foi solucionado através da identificação de seções do texto. Aplica-se, então, o método convencional para cada seção identificada de forma individual. Nas subseções a seguir, essa modificação é mais profundamente discutida.

4.1. 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 4.1 e 4.2.

(...) *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 4.1. 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 4.2. 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 possui pontuação final.

Com a estrutura textual caracterizada, o sistema foi modificado para que houvesse o reconhecimento das diversas seções e seus títulos. Dos exemplos mostrados anteriormente, o último caso não foi implementado devido às ambigüidades na sua estrutura. O uso de travessões precedidos de nomes também é bastante utilizado para indicar fala de um personagem em texto, tornando-se ,assim, uma forma ambígua para o sistema. O texto da Figuras 4.3, retirado do Corpus TeMário (Pardo e Rino, 2003), 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 4.3. 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 3.1 foi processado no sistema com reconhecimento da estrutura textual. O sumário produzido é exibido na Figura 4.4 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 4.4. Sumário obtido pelo GistSumm modificado para o texto da Figura 3.1

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 realizada pelo GistSumm.

Uma deficiência percebida foi que, em casos em que as seções possuem poucas sentenças, se a taxa de compressão for razoavelmente alta, a sentença melhor pontuada ultrapassa o limite de palavras 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.

Na seção 5, é descrita uma avaliação mais aprofundada a respeito das modificações realizadas no sistema.

4.2. 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 introdução, 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, opcionalmente, manter o título da seção e (b) no mínimo uma sentença de cada seção deve ser selecionada para o sumário.

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 em que 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 4.5 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 4.5. Sumário obtido com o GistSumm para artigos científicos

No exemplo da Figura 4.6, 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 4.6. 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.

Na seção a seguir é apresentada a metodologia da avaliação das adaptações realizadas e a análise dos seus resultados.

5. Avaliação e análise dos resultados

Para avaliação do GistSumm para textos estruturados, em comparação com o sistema GistSumm original, duas avaliações distintas foram conduzidas. A primeira avaliação foi baseada em um julgamento por um juiz humano sobre um córpus de 20 textos da língua portuguesa. Para a segunda avaliação foi utilizada a ferramenta ROUGE (Lin e Hovy, 2003) para avaliação automática de sumários e um córpus de 150 textos de língua inglesa. Os resultados obtidos em ambas as avaliações demonstram que as modificações realizadas no sistema contribuíram substancialmente para a melhoria de seus sumários em textos que possuem uma estrutura textual.

A avaliação subjetiva e seus resultados são discutidos na Subseção 5.1, enquanto a avaliação automática é discutida na Subseção 5.2.

5.1. Avaliação subjetiva

A avaliação do sistema de uma forma subjetiva foi realizada baseando-se em percepções comparativas entre o sistema GistSumm original e sua nova versão que conta com o reconhecimento da estrutura textual.

Para esta avaliação foram utilizados 20 artigos do Workshop de Teses e Dissertações do ICMC-USP (WTD-ICMC-USP) escolhidos de maneira aleatória. Os artigos desse evento são em geral textos curtos, de 7 ou 8 parágrafos, e que possuem a seguinte estrutura textual definida: Contexto, Problema, Propósito, Metodologia, Resultado.

Na geração dos sumários com o GistSumm original utilizou-se a sua sumarização padrão usando o método *keywords*. Para o GistSumm para textos estruturados, utilizou-se o mesmo método de ranqueamento de sentenças.

A opção de manter ao mínimo uma sentença por seção foi realizada devido à natureza dos textos, isto é, por serem textos com seções de tamanho reduzido. A utilização de uma taxa de compressão que já resultasse em um *threshold* menor do que o tamanho da sentença mais pontuada levaria a produção de uma seção sem nenhuma sentença. Como o texto em questão é de tamanho reduzido então era observado que uma taxa de compressão de 70% levava a um sumário sem sentenças de muitas seções.

Os sumários gerados pelo novo GistSumm foram avaliados com diversas opções de taxa de compressão. Constatou-se, empiricamente, que a taxa de compressão que mais se encaixaria naquela evidenciada pela opção de manter ao mínimo uma sentença por seção era de 40%. Assim

os dois sistemas foram configurados para produzirem seus sumários a uma taxa de 40%, o que, na prática, não levou a diferenças significativas no tamanho de seus sumários.

Os critérios avaliados comparativamente entre os sumários produzidos pelos dois textos foram: informatividade e textualidade. Essa análise foi realizada de forma subjetiva por um juiz humano. Quanto à informatividade de um sumário, pretende-se avaliar o quanto de conteúdo informativo original ele contém. Quanto a sua textualidade, pretende-se avaliar se o sumário mostra coesão e coerência entre suas sentenças. As respostas possíveis são: apresenta um resultado bom, razoável ou ruim.

A Tabela 5.1 abaixo exibe para cada sistema qual porcentagem de seus sumários foi enquadrada em cada categoria.

.

	GistSumm original			*Novo GistSumm*		
Critério/Desempenho	Bom	Razoável	Ruim	Bom	Razoável	Ruim
Informatividade	35 %	55 %	10 %	65 %	30 %	5 %
Textualidade	25 %	55 %	20 %	35 %	40 %	25 %

Figura 5.1. Distribuição dos critérios avaliados subjetivamente em cada sistema

Observamos a partir do quadro acima que a nova versão do sistema GistSumm foi comparativamente melhor na produção de sumários para este gênero de texto. Analisando o critério informatividade o sistema GistSumm novo teve um desempenho bem melhor que o original tendo 65 % dos sumários avaliados com nível bom, 30 % razoável e só 5 % foram considerados ruins quanto a este nível avaliado. O GistSumm original avaliado pelo mesmo critério apresentou apenas 35 % dos sumários com um nível bom, 55 % razoável e 10 % ruim. Nota-se, portanto que o GistSumm novo apresentou um desempenho superior quanto a informatividade avaliada.

Analisando o critério textualidade observamos que as diferenças não foram tão significativas quanto ao outro critério analisado, contudo os resultados se mantiveram favoráveis ao GistSumm novo. O desempenho apresentado por este foi de 35 % dos sumários considerados bons, 40 % razoáveis e 25 % ruins. O GistSumm original apresentou 25 % dos seus sumários avaliados com um nível bom de textualidade, 55 % com um nível razoável e 20 % com um nível ruim. Apesar de o novo GistSumm apresentar um aumento dos casos de sumários considerados com um nível ruim de textualidade ele também aumentou em uma proporção maior o número de casos considerados bons nesse critério o que leva a acreditar que tenha sido melhor que o GistSumm original.

Cabe aqui também acrescentar algumas avaliações gerais percebidas durante a realização da avaliação subjetiva. Em muitos textos analisados, muitas de suas sentenças carregavam sentidos aditivos ao texto e que não apresentavam um alto grau de articulação com as sentenças ao seu redor. Por isso muitos sumários que embora fossem compostos por sentenças distintas traziam ambos a mesma quantidade de informação sendo assim avaliados com o mesmo nível no critério informatividade.

Uma outra característica particular em relação ao tipo de texto utilizado pela avaliação é que são textos bem enxutos e que discorrem sucintamente sobre cada seção. Isso ocasionava que em muitas seções a escrita do primeiro parágrafo decorria do entendimento do título da seção, por exemplo, o parágrafo da seção objetivo começava com “Desenvolver um framework ...”, sem necessariamente introduzir que sobre o que se decorre é uma informação do objetivo do trabalho apresentado. Como os sumários não apresentavam os títulos das seções esse aspecto particular desse tipo tornou a avaliação quanto a textualidade difícil resultando em uma performance global não tão boa quando se evidenciou quanto a informatividade.

5.2. Avaliação automática através da ferramenta ROUGE

A ferramenta utilizada para a avaliação automática sobre o desempenho dos sumários produzidos pelo novo GistSumm foi a ROUGE. A ROUGE é um pacote de avaliação baseado na co-ocorrência de n-gramas entre textos que deseja-se analisar. Com ela é possível fazer a análise entre o desempenho de sumários com relação ao *abstract* do texto, considerado o sumário ideal. Baseando-se na contagem da co-ocorrência de n-gramas a ferramenta resultará uma pontuação para suas duas mediadas avaliadas: cobertura e precisão.

A ROUGE é uma ferramenta de avaliação já bem conceituada entre os pesquisadores dessa área, sendo utilizada como medida de avaliação na DUC (*Document Understanding Conferences*), a principal conferência de avaliação de sistemas sumarizadores, realizada anualmente.

O córpus utilizado para esta avaliação é uma compilação de 183 documentos da coleção *Computation and Language* (cmp-lg) disponibilizados como parte da *TIPSTER Text Summarization Evaluation Conference*. Os textos são artigos científicos em inglês de conferências patrocinadas pela *Association for Computational Linguistics* (ACL).

Os textos retirados do córpus passaram por uma filtragem aleatória onde foram escolhidos 150 documentos para avaliação. Antes de serem comparados através da ferramenta ROUGE esses textos passaram por um pré-processamento. Neste, fórmulas e os *abstracts* foram removidos do texto principal e os *abstracts* armazenados em arquivos a parte.

A análise dos textos e avaliação ocorreu da seguinte forma: em um primeiro passo, os sumários humanos, isto é, os *abstracts* removidos dos artigos, foram analisados e medidos quanto a sua taxa de compressão em relação ao texto do artigo. Essa taxa de compressão passou então por uma suavização, sendo reduzida até o menor múltiplo de 5 com limite superior a 90 %. Foi estipulado que os sumários não deveriam ser processados com uma taxa de compactação superior a 90 % pois isso resultaria na seleção de muitas poucas sentenças o que não garantiria a avaliação precisa dos sistemas.

A utilização dessa suavização ocorreu devido ao fato de que a taxa informada no sumarizador automático GistSumm é o limite inferior de compactação que o respectivo sumário deve conter. Assim, um texto com uma taxa nominal de x % de compressão sempre acaba gerando um sumário com uma taxa real de y > x % de compressão devido à imposição de que a seleção de sentenças para compor o extrato permaneça dentro dessa taxa. Desse modo, a suavização desta taxa prevê esse aumento natural na taxa de compressão durante o processamento pelo sumarizador, gerando um sumário mais fiel à taxa de compressão utilizada no *abstract*.

Para a comparação dos sumários gerados pelos dois sistemas, o GistSumm original e o novo GistSumm, foram geradas diversas situações de configuração dos dois sistemas para comparação. Cada sumário gerado nas situações avaliadas foram comparados dentro dos critérios da ROUGE com o abstract removido do texto principal, considerado o sumário ideal. As situações geradas foram:

1. Sumarização com o sistema GistSumm original na mesma taxa de compressão suavizada do *abstract*;
2. Sumarização com o novo GistSumm na mesma taxa de compressão suavizada do *abstract* e sem a opção de manter no mínimo uma sentença por seção;
3. Sumarização com o novo GistSumm na mesma taxa de compressão suavizada do *abstract* e com a opção de manter no mínimo uma sentença por seção;
4. Sumarização com o sistema GistSumm original na taxa de compressão suavizada analisada no sumário produzido no item 3;
5. Sumarização com o novo GistSumm na taxa de compressão suavizada analisada no sumário produzido no item 3 e sem a opção de manter no mínimo uma sentença por seção.

Essas 5 situações de configuração foram planejadas de modo que a avaliação automática realizada fique a mais imparcial possível e que também sejam observados em qual situação cada sistema produzirá um melhor resultado. Na situação 1 deseja obter um sumário com o mesmo tamanho do abstract para comparação direta com este. Na situação 2 deseja-se novamente um sumário no mesmo tamanho do *abstract* só que avaliando agora a performance do GistSumm novo.

Como se sabe que o sistema GistSumm novo sumarizará cada seção encontrada no texto dentro da taxa especificada de compressão então existe a grande possibilidade que seu sumário correspondente simplesmente omita seções onde a sentença mais pontuada ultrapassou a taxa de compressão definida. Assim seu sumário resultante apresentaria uma taxa de compressão bem maior que a do *abstract*. Para que se possa ponderar isso foram utilizados mais 3 situações a serem avaliadas.

Na situação 3 pretende-se minimizar esse problema obrigando o GistSumm novo a selecionar ao menos uma sentença para cada seção do texto. Na situação 4 pretende-se fazer uma avaliação comparativa entre o sumário do GistSumm original com o obtido pelo GistSumm novo obtido na situação 3. Para isso utiliza-se a mesma taxa de compactação nele avaliada.

Para fechar a avaliação analisando todos os possíveis resultados dessas duas variações, mantendo-se uma sentença mínima ou não, é realizada uma avaliação com o sistema GistSumm novo na taxa de compressão das duas situações anteriores e sem a opção de uma sentença mínima por seção.

Em todas as avaliações com o novo GistSumm, foi utilizado a opção de remover os títulos das seções do sumário final, gerando um sumário mais justo para comparação com o *abstract*.

As 5 avaliações exibidas acima foram também simuladas em 4 diferentes cenários. Cada cenário visava avaliar parâmetros dos sumarizadores e o desempenho de cada. Foram testados os resultados para os cenários variando o método de ranqueamento de sentenças entre K*eywords* e *Average-keywords*. Também se variou a utilização do texto completo para a sumarização e a utilização do texto sem as seções bibliografia, agradecimentos, notas de rodapé e referências, consideradas irrelevantes para a produção do sumário.

A partir desses cenários definidos foram executados os casos de testes na ferramenta ROUGE, utilizando-se de uma taxa de confiança de 95%. Os resultados são exibidos nas Tabelas seguintes, onde 0 indica que o sumário automático é diferente do sumário humano e 1 indica a proximidade máxima entre os dois. Quanto maior esse número, melhor o sumário automático é considerado. Foram utilizadas duas variações da ROUGE. A ROUGE-1, em que se faz uma comparação de palavras (unigramas) entre os sumários humanos e automáticos, e a ROUGE-L, em que se faz uma comparação de uma seqüência de palavras em comum entre os sumários humanos e automáticos. Para cada variação da ROUGE, são relatadas as tradicionais medidas de precisão, cobertura e medida-f.

Cenário 1: Método *keywords* e texto integral							
		X-ROUGE 1			X-ROUGE L		
Sistema	Situação avaliada	Cobertura	Precisão	Medida F	Cobertura	Precisão	Medida F
GistSumm original	1	0.54626	0.16727	0.24212	0.49045	0.14960	0.21672
Novo GistSumm	2	0.35212	0.22968	0.24373	0.30724	0.20174	0.21294
Novo GistSumm	3	0.62199	0.13460	0.21142	0.56475	0.12204	0.19175
GistSumm original	4	0.48729	0.14827	0.20931	0.58246	0.11355	0.18077
Novo GistSumm	5	0.53799	0.16411	0.23482	0.48237	0.14678	0.21020

Figura 5.2. Avaliação do Sistema GistSumm na fermenta ROUGE – Cenário 1

Cenário 2: Método *Average-keywords* e texto integral							
		X-ROUGE 1			X-ROUGE L		
Sistema	Situação avaliada	Cobertura	Precisão	Medida F	Cobertura	Precisão	Medida F
GistSumm original	1	0.47655	0.15193	0.21346	0.43962	0.13933	0.19611
Novo GistSumm	2	0.47553	0.21689	0.27369	0.43757	0.19922	0.25136
Novo GistSumm	3	0.54090	0.18603	0.26013	0.49962	0.17195	0.24030
GistSumm original	4	0.48729	0.14827	0.20931	0.45000	0.13607	0.19246
Novo GistSumm	5	0.48906	0.21295	0.27107	0.44992	0.19562	0.24899

Figura 5.3. Avaliação do Sistema GistSumm na fermenta ROUGE – Cenário 2

Cenário 3: Método K*eywords* e texto relevante							
		X-ROUGE 1			X-ROUGE L		
Sistema	Situação avaliada	Cobertura	Precisão	Medida F	Cobertura	Precisão	Medida F
GistSumm original	1	0.51929	0.18088	0.25383	0.46332	0.16057	0.22562
Novo GistSumm	2	0.32319	0.25111	0.24323	0.27960	0.21834	0.21065
Novo GistSumm	3	0.59796	0.15176	0.22863	0.53829	0.13626	0.20546
GistSumm original	4	0.59993	0.13799	0.21204	0.54733	0.12528	0.19274
Novo GistSumm	5	0.50802	0.18768	0.25351	0.45050	0.16630	0.22469

Figura 5.4. Avaliação do Sistema GistSumm na fermenta ROUGE – Cenário 3

Cenário 4: Método *average-keywords* e texto relevante							
		X-ROUGE 1			X-ROUGE L		
Sistema	Situação avaliada	Cobertura	Precisão	Medida F	Cobertura	Precisão	Medida F
GistSumm original	1	0.51920	0.17662	0.24904	0.47669	0.16161	0.22806
Novo GistSumm	2	0.46155	0.23449	0.28653	0.42291	0.21463	0.26218
Novo GistSumm	3	0.51903	0.20895	0.27848	0.47775	0.19254	0.25633
GistSumm original	4	0.52314	0.17541	0.24604	0.48048	0.16045	0.22534
Novo GistSumm	5	0.46878	0.23166	0.28408	0.42927	0.21221	0.26003

Figura 5.5. Avaliação do Sistema GistSumm na fermenta ROUGE – Cenário 4

A análise comparativa dos dados obtidos na avaliação pela ferramenta ROUGE leva a conclusão de que o novo GistSumm foi melhor para este gênero científico de texto. Também se pode afirmar que os resultados tornam-se mais expressivos na utilização do método A*verage-keywords.* Este método que não apresentava um desempenho satisfatório no GistSumm original (segundo avaliações anteriores) passa a ter agora melhores resultados na nova arquitetura de sumarização do novo GistSumm.

Observa-se também uma melhora nos valores de avaliação dos cenários com o texto relevante apenas, isto é, o texto sem seções de bibliografia, agradecimentos, notas de rodapé e referências.

Analisando apenas a medida-f, que é uma média harmônica entre cobertura e precisão, observamos que para o método A*verage-keywords* é estabelecida uma ordem de resultado que privilegia o sistema GistSumm novo. Nos cenários onde este método foi avaliado, isto é, os cenário 2 e 4, todos os resultados relacionados os sistema GistSumm novo se mostraram melhores que o sistema GistSumm original. O melhor resultado foi demonstrado pela situação avaliada 2 que obteve para o cenário 2 o resultado 0.27369 e para o cenário 4 o resultado 0.28653. O melhor resultado do GistSumm original foi para a situação avaliada 1 com o resultado 0.21346 para o cenário 2 e 0.24904 para o cenário 4.

Para o método *keywords*, apesar de menos expressivos, os resultados se mantêm em geral favoráveis ao novo GistSumm. Os dois cenários avaliados foram o cenário 1 e o 3. O sistema GistSumm original apresenta o melhor resultado com 0.24212 no cenário 1 e 0.25383 no cenário 3. Os três melhores em seguida são relativos ao GistSumm novo com o resultado mais expressivo sendo 0.24373 para a situação avaliada 2 no cenário 1 e 0.25351 para a situação avaliada 5 no cenário 3. O pior resultado em ambos os cenário ficou com a situação avaliada 4 referente ao GistSumm original. Observa-se que embora o GistSumm original tenha o melhor resultado, este em comparação como o melhor do GistSumm novo está muito próximo.

Cabe ressaltar, contudo que a situação 1 avaliada contém uma taxa de compressão diferente das demais situações. Nas situações 2 e 3 avaliadas, apesar de manterem nominalmente a mesma taxa de compressão, os sumários observam taxas reais de compressão diferentes devido aos critérios usados para sumarização. Na situação 2, não se requer uma sentença por seção, o que leva muitas seções que tiveram sua *gist sentence* acima da taxa de compressão a desaparecerem do sumário final, gerando uma taxa de compressão real muito maior. Na situação 3, é mantida uma sentença para cada seção, o que resulta em uma taxa de compressão real menor do que a avaliada em 1.

Uma outra análise pode ser feita a partir desses resultados. Lembremos que o método *keywords*, por não normalizar sua pontuação por sentença, acaba elegendo comumente as sentenças de maior tamanho no artigo. Por esse propósito e devido a separação do texto em seções, os resultados avaliados com o método *Average-keywords* são mais expressivos dentro da arquitetura de sumarização do novo GistSumm e para os textos científicos em questão.

6. Participação no CLEF

O CLEF – *Cross-Language Evaluation Forum* – é uma competição internacional de sistemas de perguntas e respostas. Nesta competição, são submetidas respostas às perguntas elaboradas com base em textos fornecidos pela competição e no processamento do sistema sobre um córpus. Os pesquisadores também são encorajados a elaborarem sistemas que analisem os textos e respondem em línguas distintas.

Apresentamos a seguir a aplicação de um sistema de sumarização para a modalidade de *Question Answering* (QA) em única língua do CLEF 2006 para textos em português. Dois conjuntos de respostas foram submetidos para o CLEF. Para o primeiro, nós submetemos o resultado do sistema de sumarização sem nenhum pós-processamento. Para o outro, nós aplicamos um filtro simples que desenvolvemos para encontrar no sumário produzido o pedaço da resposta necessária a competição. Esta segunda submissão foi realizada pelo aluno Vinícius Rodrigues de Uzêda e não será detalhada neste relatório. Para maiores detalhes do sistema e do filtro utilizado veja Balage Filho et al. (2006b). O desempenho de ambos os métodos no CLEF foi muito ruim, indicando que técnicas simples de sumarização sozinhas não são suficientes para a tarefa de perguntas e respostas.

O sistema de sumarização que usamos é descrito na próxima subseção. Nossos resultados no CLEF são reportados na subseção 6.2.

6.1. Adaptações do GistSumm o CLEF

Uma característica do sistema GistSumm original e também do GistSumm com o tratamento da estrutura textual é a possibilidade da produção de sumários focados nos interesses do usuário. Para isto, a *gist sentence* é escolhida como aquela com a maior pontuação em relação ao tópico específico, com esta correlação sendo mensurada pela medida do co-seno (Salton, 1989).

Para participação no CLEF, realizamos pequenas modificações no processamento dos sumários gerados pelo GistSumm com tratamento da estrutura textual. Para cada questão, retornamos a *gist sentence* mais pontuada encontrada utilizando-se a opção de sumarização focada nos interesses do usuário, com a questão a ser respondida como tópico específico da sumarização. Desta maneira, nós procuramos por possíveis respostas no texto que tenham uma boa correlação com a pergunta. Para rodar os experimentos do CLEF, a *gist sentence* era escolhida independentemente para cada questão de cada texto na base de dados do CLEF. A melhor *gist sentence* era então selecionada com sendo a resposta.

Os resultados obtidos para este sistema submetido são reportados na subseção seguinte.

6.2. Resultados e Discussão

Foi escolhido o idioma português para a participação no CLEF pelas seguintes razões: português é nossa língua nativa e, então, isto nos possibilita um melhor julgamento dos resultados; português é uma das línguas suportadas pelo GistSumm. A coleção de dados do Português para o CLEF contém textos de notícias do jornal brasileiro Folha de São Paulo e do jornal português Público, dos anos 1994 e 1995.

Os organizadores da modalidade QA (*Question Answering*) disponibilizaram 200 questões, as quais incluem as chamadas "questões factuais", "questões de definição", e "questões de listagem". Como definido pelo CLEF, questões factuais são questões baseadas em fatos, perguntando, por exemplo, por um nome de um personagem ou a localização; questões de definição são questões como "Qual/Quem é X?"; questões de listagem são questões que requerem uma lista de itens como resposta, por exemplo, "Quais cidades européias têm hospedado os Jogos Olímpicos?". Há diferentes quantidades de cada tipo de questão na modalidade QA: 139 questões factuais, 47 questões de definição e 12 questões de listagem.

A principal medida de avaliação usada pelo CLEF é a precisão. Para esta medida, juizes humanos têm que dizer, para cada questão, se a resposta foi certa, errada, não suportada (isto é, para respostas contendo uma informação correta, mas não suportada pelo texto provido), inexata (a resposta contém uma informação correta e o texto provido a suporta, mas a resposta é incompleta/truncada ou é mais longa do que a mínima quantidade de informação requerida), ou não foi julgada (para o caso que nenhum julgamento foi realizado para a resposta). Algumas variações da medida de precisão são a *Confidence Weighted Score* (CWS), a *Mean Reciprocal Rank Score* (MRRS) e a medida K1. Para leitores interessados, nós sugerimos a referência para as regras de avaliação do CLEF para estas definições de medidas. A Figura 6.1 mostra os resultados para o primeiro sistema submetido.

Precisão			CWS	MRRS	K1
Julgamento das questões	Factuais e Definição	Listagem	0	0	-0.6445
Certa	0	0			
Errada	179	9			
Não Suportada	7	0			
Inexta	2	3			
Não julgada	0	0			

Figura 6.1. Resultados para submissão do primeiro sistema ao CLEF

Pode-se observar que os resultados foram muito ruins. O desempenho do nosso sistema pode ser melhor visualizado no exemplo (dos dados do CLEF) na Figura 6.2. Neste exemplo, o sistema retornou as 6 melhores sentenças correlatas com a pergunta. Apesar da medida do co-seno ser eficiente quando nós desejamos identificar sentenças de um tópico para compor um sumário, nós podemos observar que isto não se mostra eficiente na resposta a perguntas.

Questão: Como se chama a primeira mulher a escalar o Evereste sem máscara de oxigênio?

Primeiras respostas selecionadas e suas mediadas do co-seno (em relação com a pergunta):

Co-Seno	Resposta
0.500000	Ou: Uma mulher como eu
0.500000	Como se chama?
0.500000	Como se chama?
0.472456	Em Maio deste ano, Hargreaves, de 33 anos, tinha se tornado a primeira mulher a escalar o Evereste, sozinha e sem a ajuda de oxigênio.

Resposta Selecionada: Ou: Uma mulher como eu

Figura 6.2. Exemplo de uma resposta submetida ao CLEF

Analisando o desempenho do sistema para o conjunto inteiro de perguntas e respostas submetido, nós pudemos identificar alguns pontos importantes que poderiam ser investigados no futuro:

- o nível de análise sentencial realizado pelo sistema de sumarização não é suficientemente apropriado para responder perguntas;
- apesar da medida do co-seno ser boa para construção de sumários orientados a tópicos, ela não é boa para procurar por respostas, pois ela tende a selecionar sentenças mais curtas que têm mais palavras em comum com a pergunta, ignorando sentenças mais longas que podem provavelmente conter a resposta;
- apesar da resposta não estar evidenciada nas melhores sentenças pontuadas, uma análise das outras sentenças mostram que a sentença pode ser encontrada em torno das 100 melhores sentenças pontuadas, em geral.

Após os experimentos com o CLEF, nós acreditamos que simples técnicas de sumarização não são suficientes para a tarefa de perguntas e repostas, mesmo se estes sistemas são bons no que eles fazem. Nós acreditamos que muito mais trabalho nesta direção é necessário.

7. Considerações finais

Neste trabalho, pudemos mostrar que simples sumarizadores extrativos podem conseguir melhores resultados quando melhorias diretas são incorporadas no processo de sumarização. Em especial, focando-se em melhorias no tratamento da estrutura textual, obtêm-se um sumário de melhor qualidade.

A nossa participação experimental no CLEF teve resultados muito ruins, mas que puderam direcionar possíveis pesquisas sobre a utilização de métodos de sumarização para a tarefa de perguntas e respostas em questão.

Agradecimentos

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

Referências

Balage Filho, P.P.; Uzêda, V.R.; Pardo, T.A.S.; Nunes, M.G.V. (2006a). *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.

Balage Filho, P.P.; Uzêda, V.R.; Pardo, T.A.S.; Nunes, M.G.V. (2006b). Using a Text Summarization System for Monolingual Question Answering. In the *Proceedings of the Cross Language Evaluation Forum 2006 Workshop* – CLEF. Alicante, Spain. September 20-22.

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.

Lin, C-Y. and Hovy, E.H. (2003). Automatic Evaluation of Summaries Using N-gram Cooccurrence Statistics. In the *Proceedings of Language Technology Conference* – HLT. Edmonton, Canada. 0May 27 - June 1.

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.

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.