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

A Coleção TeMário e a Avaliação de Sumarização Automática

Lucia Helena Machado Rino
Thiago Alexandre Salgueiro Pardo

NILC-TR-06-04

Janeiro, 2006

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

Resumo

Neste capítulo apresentamos as principais estratégias atualmente em uso em avaliações de sistemas de Sumarização Automática, ressaltando as dirigidas ao processamento de textos em inglês. Na tentativa de solucionar os problemas complexos dessa tarefa para a SA de textos em português, construímos a coleção TeMário. Finalmente, discutiremos possibilidades de avaliação de sumarizadores automáticos especialmente dedicados à língua portuguesa.

ÍNDICE

1. INTRODUÇÃO **2**

2. AS PRINCIPAIS ABORDAGENS PARA A CONSTRUÇÃO DE SISTEMAS DE SA **2**

3. A AVALIAÇÃO DE SISTEMAS DE SA **4**

4. TIPOS DE AVALIAÇÃO **6**

4.1. AVALIAÇÕES INTRÍNSECAS 6

4.2. AVALIAÇÕES EXTRÍNSECAS 9

4.2.1. Categorização de documentos *9*

4.2.2. Recuperação de Informação *9*

4.2.3. Perguntas e Respostas *9*

5. A CONSTRUÇÃO DO TEMÁRIO COMO UMA COLEÇÃO DE AVALIAÇÃO DE SISTEMAS DE SA **10**

6. DISCUSSÃO E CONCLUSÕES **11**

REFERÊNCIAS BIBLIOGRÁFICAS **12**

1. Introdução

As propostas de avaliação de sumarizadores automáticos de textos se apresentam de forma bastante variada atualmente. Há duas características fundamentais normalmente consideradas na análise dos sumários: seu nível de informatividade e sua qualidade. O nível de informatividade de um sumário é imprescindível, porque expressa o quão adequada é sua correspondência com seu texto-fonte: o sumário deve preservar, no mínimo, sua idéia central. Já sua qualidade remete ao fato de ser um *texto* e, assim, ter que apresentar as propriedades inerentes desse objeto lingüístico: minimamente, o sumário deve apresentar *textura*, isto é, deve ser coerente e coeso.

No contexto da manipulação computacional de textos, existem inúmeros problemas para se verificar ambas as características, sendo eles de diversas naturezas. Neste capítulo apresentamos as principais estratégias atualmente em uso para uma avaliação sistemática, assim como as necessidades e limitações dessa tarefa. Em sua grande maioria, as propostas significativas são originárias do interesse no processamento de textos em inglês. Na tentativa de solucionar os problemas complexos dessa tarefa, construímos uma coleção de textos em português (do Brasil), o TeMário, cuja utilidade será evidenciada. Finalmente, discutiremos possibilidades de avaliação de sumarizadores automáticos especialmente dedicados à língua portuguesa. Antes, contudo, faremos uma síntese das principais abordagens de SA atuais, para contextualização das questões de avaliação.

Neste texto, utilizaremos as abreviaturas **SA**, para Sumarização Automática, e **PB**, para português do Brasil. Quando a referência à língua incluir o português europeu, usaremos somente o termo português. Também é importante ressaltar que a SA considerada aqui se refere somente à SA mono-documentos, em contraposição à SA multi-documentos, outra área de crescente interesse nos últimos cinco anos, que estende a problemática da sumarização mono-documentos.

2. As principais abordagens para a construção de sistemas de SA

Se considerada a simulação da tarefa humana de sumarização, um sumário, como texto resumido, é fruto de um conjunto variado de processos, os quais incluem a compreensão e filtragem de material do texto-fonte, assim como a abstração, fusão e condensação desse conteúdo, típicos da *reescrita textual* [HovyLin97; SpaJo97]. Essa simulação caracteriza a abordagem *fundamental* de SA (*knowledge-rich approach*), a qual leva à produção de textos praticamente *independentes* em relação à estrutura e estilo de seus textos-fonte.

As propostas atuais na abordagem fundamental remetem à construção das estruturas discursivas dos textos-fonte, para a recuperação de uma representação conceitual que permita reconhecer (ou recuperar), por exemplo, as intenções do autor do texto e, assim, distinguir informações em níveis diferenciados de importância. Procura-se contemplar, diversos níveis de representação do conhecimento, destacando-se o lingüístico, o informativo (ou de domínio) e o comunicativo. Como resultado, são exploradas diversas linguagens de representação do discurso, as quais devem ser suficientemente expressivas para garantir o inter-relacionamento entre as unidades de significado (i.e., as proposições) subjacentes ao texto, assim como a manipulação das estruturas correspondentes por planejadores textuais e realizadores lingüísticos automáticos.

Alguns métodos que visam à determinação das informações relevantes de um texto-fonte a partir de sua modelagem discursiva, para sua inclusão na estrutura de um sumário, contemplam o cômputo da *saliência*, ou importância, das proposições [BogKen97]: proposições mais salientes seriam o foco de atenção pretendido pelo autor e, logo, devem ser consideradas na SA; as com baixa saliência seriam periféricas e, portanto, passíveis de

exclusão. Neste caso, a exclusão deve vir acompanhada da reestruturação de todo o discurso para a construção da estrutura discursiva do sumário, como propõe Marcu [Mar97a, Mar97b, Mar00] ao analisar a estrutura discursiva do texto-fonte. É importante notar que, nesta abordagem, as estruturas discursivas dos textos-fonte devem estar disponíveis (e são resultado da análise discursiva prévia). Particularmente, Marcu usa a Teoria RST – *Rhetorical Structure Theory* [MannTho88] com esse fim. As estruturas RST dos sumários são, portanto, resultado da *condensação* das estruturas originais. Uma variação dessa abordagem foi explorada por nós [Rino96; Par02a; ParRin02], que envolve não só a noção de nuclearidade, como também as relações intencionais [GroSid86] delineadas pelo suposto objetivo comunicativo do escritor, isto é, como as intenções subjacentes às proposições se relacionam e colaboram para a satisfação da intenção primária do texto. Essas abordagens pressupõem a existência de bons analisadores discursivos, já existentes para a língua inglesa [Mar00; CorOli98; Schil02]. O DiZer [ParEtAl04] é um sistema análogo para o PB. Adicionalmente, a construção dos sumários ainda depende de bons realizadores lingüísticos, inexistentes para o português.

A segunda abordagem, *empírica* (*knowledge-poor approach*, freqüentemente *corpus-based*), privilegia o tratamento estatístico ou relacional das informações textuais, buscando evidências lingüísticas por meio do reconhecimento de padrões derivados de suas informações morfossintáticas ou de distribuições numéricas associadas a elas. O aspecto relacional, neste caso, pode remeter a inúmeras formas na superfície textual, resultando na análise das associações entre os componentes textuais (cuja granularidade pode ser, simplesmente, a de itens lexicais, mas pode envolver segmentos tão longos quanto um parágrafo ou uma seção). Por exemplo, o encadeamento de cadeias lexicais na forma de cadeias co-referenciais, a sinonímia, a antonímia ou mesmo outras relações ontológicas. Na SA, sua identificação e composição automáticas dependem de modelos artificiais de reconhecimento do encadeamento lingüístico, os quais podem resultar em métodos apreendidos pelo sistema com base na lingüística de corpus. Várias técnicas são usadas com esse fim, que incluem propostas sofisticadas, como, por exemplo, a busca de radicais de palavras co-ocorrentes ou a construção das topografias textuais como representações temáticas dos textos-fonte, as quais permitam indicar os nós de termos tematicamente relacionados que melhor representem os principais tópicos do texto-fonte, assim como propostas mais simples, como a identificação do comprimento e posição das sentenças ou de informações-chave que não sejam redundantes no texto-fonte ou, ainda, a classificação de sentenças baseada no peso das palavras significativas.

Algumas dessas técnicas envolvem somente processamento estatístico; outras, manipulação de conhecimento lingüístico. No entanto, é crescente a busca por metodologias mistas, cujo processamento estatístico seja elaborado sobre repositórios lingüísticos robustos. A TIPSTER SUMMAC (*Text Summarization Evaluation*)[1] [ManiEtAl98], iniciativa ampla que visou, sobretudo, as novas tecnologias de avaliação de sumarização textual, deixa claro que técnicas comuns a outras áreas, especialmente à Recuperação de Informação e à Extração de Informação, devem ser consideradas para a SA. Nesta conferência foram apresentadas inúmeras tecnologias que envolveram, por exemplo, a identificação de tópicos textuais, pela identificação dos termos relacionados, usando-se um thesaurus ou a macro-estrutura do discurso de cada texto-fonte; a construção de representações temáticas com o auxilio de ontologias de domínio ou ontologias abertas da língua natural em foco; a identificação de informações-chave pela medida MMR (*Maximal Marginal Relevance*) [GolEtAl00] (também usada com objetivo oposto, de eliminar ou agrupar informações redundantes, para exclusão) [WhiCar02] ou pela TF-IDF [SalBuc88] modificada, para identificar sentenças singulares em um único texto, em vez de documentos singulares em uma ou mais coleções.

[1] http://www.itl.nist.gov/iaui/894.02/related_projects/tipster_summac/ (último acesso em junho/2004).

Esses métodos podem, ainda, se basear em informações resultantes de um pré-processamento textual. Por exemplo, a seleção de sintagmas nominais ou a identificação de nomes próprios, os quais certamente remetem a pistas lingüísticas sobre a importância dos segmentos. Essas operações são bastante favorecidas, hoje, com a etiquetagem prévia dos textos de interesse e com o uso de analisadores morfológicos ou sintáticos. Para a resolução co-referencial, por exemplo, buscam-se informações que normalmente sejam associadas a mesmos indivíduos, organizações ou eventos.

Por fim, todas essas técnicas visam à identificação de pistas contextuais e pistas de conteúdo para a produção de distribuições de importância dos segmentos textuais, para decidir quais deles devem ser considerados nos sumários em construção. Vale notar que a forma como as informações lingüísticas são manipuladas na abordagem empírica difere substancialmente da abordagem fundamental: na empírica não há a modelagem lingüística, porém, busca-se reconhecer os elos lingüísticos por meio do reconhecimento de "pistas superficiais". Embora essa forma de tratar a informação textual não seja nova – Edmundson [Edmun69] já a empregava, denominando seu método de "*seleção computacional de sentenças com maior potencial de transmitir ao leitor a substância do documento*" (tradução livre) – a inovação está em agregá-las em uma única estratégia computacional, na busca por um desempenho melhor, como propõe Mani [Mani01]. Nessa abordagem, embora a textura dos sumários finais também seja objetivada, prevalecem as decisões sobre seu nível de informatividade. Sobretudo, não existe qualquer preocupação com a reescrita, típica da simulação do processo humano. Por essa razão, a forma mais comum produzida automaticamente é o *extrato*, sumário composto unicamente pelos segmentos textuais *copiados e colados* do texto-fonte. Essa estratégia caracteriza, portanto, a SA *extrativa*.

3. A avaliação de sistemas de SA

A avaliação de sumarizadores automáticos depende da forma como seus dois métodos típicos – a identificação dos segmentos relevantes de um texto-fonte e a composição desses segmentos nos respectivos sumários – é elaborada. Como se viu na seção anterior, a abordagem fundamental, por exemplo, leva à produção de textos praticamente *independentes*, em relação à estrutura e estilo de seus textos-fonte. Por essa razão, sua correspondência com os mesmos não está necessariamente explícita na superfície textual, dificultando sua comparação e avaliação. Contrastando, assim, as formas de avaliação possíveis, os extratos podem ser avaliados de modo bem diferente das avaliações na abordagem fundamental, pois possuem segmentos inteiramente reproduzidos da própria superfície textual de seus textos-fonte. Automaticamente, basta, por exemplo, contar quantos, dentre os segmentos de um texto-fonte considerados relevantes, são incluídos no extrato. Se calculada, então, sua representatividade no extrato (em relação a todas as unidades textuais do extrato), temos a medida de precisão. Adicionalmente, se calculada sua representatividade em relação a todos os dados considerados relevantes, temos a medida de cobertura.

A dificuldade para o cálculo de medidas tão triviais quanto a precisão e cobertura está em se elaborar os dados de referência, isto é, os conjuntos de informações que serão usados para seu cômputo. Neste caso, essas informações devem ser julgadas de fato relevantes por leitores comuns ou por profissionais (lingüistas, educadores ou sumarizadores profissionais, por exemplo). A garantia de relevância desses dados ainda não pode ser automática. Ou seja, se a avaliação depender de sumários de referência automáticos, sua qualidade remeterá à própria qualidade dos sumarizadores automáticos que os produziram. Assim, passamos a ter um processo cíclico sem solução para a avaliação. A perspectiva reside, então, em considerar (a) dados produzidos manualmente, isto é, textos ou conjuntos de informações de autoria. Chamaremos de *autênticos* os dados produzidos pelo próprio autor dos textos-fonte que se

pretende sumarizar; de *profissionais* os dados produzidos por qualquer outro escritor competente. Assim, podemos ter sumários autênticos ou sumários profissionais como exemplos de sumários manuais que sirvam de referência para comparação com os sumários automáticos. Dados de referência que não sejam sumários manuais incluem, por exemplo, listas de palavras-chave, títulos vinculados aos textos-fonte ou qualquer outra forma não textual de informação que, geralmente indicada por seres humanos, seja, por hipótese, fiel ao texto-fonte e, portanto, representativa de seus conceitos principais, podendo servir de base para se elaborar um processo automático. Listas de palavras-chave, por exemplo, podem ser usadas para se determinar as sentenças mais representativas do texto-fonte, as quais poderão compor um extrato. Essa idéia é análoga à adotada no GistSumm [Par02b; ParEtAl03], sumarizador extrativo que constrói os extratos com base em uma sentença representativa da idéia central do texto fonte (a sentença *gist*).

Quando construídos dessa maneira, esses dados de referência são também chamados dados *ideais*: assume-se que, construídos ou indicados por leitores ou autores competentes, eles representam a melhor referência para a avaliação de sistemas de SA. São também considerados ideais quando são adequados à tarefa específica em foco. Um extrato ideal, por exemplo, é o melhor tipo de sumário para a SA extrativa, pois, por conter somente sentenças do texto-fonte, ele pode ser comparado diretamente com os sumários automáticos correspondentes.

Existem outras formas de se construir dados de referência a partir de dados manuais. Por exemplo, um extrato ideal pode ser produzido a partir de sentenças do texto-fonte que mais se assemelhem às sentenças de seu sumário autêntico (ou profissional) correspondente ou por sentenças do texto-fonte que sejam julgadas, por leitores, essenciais para compor um sumário desse texto. Entretanto, deve-se ressaltar que essas propostas de construção de dados de referência, por ser dependente do julgamento humano, também apresentam o problema de concordância para a própria avaliação: os juizes nem sempre concordam sobre as sentenças julgadas essenciais. Em geral, eles discordam tanto sobre as medidas de qualidade, quanto sobre as medidas de informatividade de um sumário [RathEtAl61; SaltEtAl97; LinHovy02]. Para garantir sua confiabilidade, no entanto, é comum agregar-se às respostas dos juizes a medida Kappa [SiCas88].

Em avaliações de sumarizadores, pode-se ter em mente várias facetas: (a) pode-se avaliar o desempenho computacional dos sistemas, isto é, sua utilização de memória, seu tempo de execução, a complexidade de seu algoritmo principal, etc.; (b) pode-se considerar sua usabilidade, isto é, a clareza de sua interface, o grau de intuição (dos possíveis usuários) necessário para seu uso ou, ainda, sua consistência e sua flexibilidade para possíveis grupos diferenciados de usuários; e, por fim, (c) pode-se avaliar se os resultados produzidos automaticamente são satisfatórios, sendo esta medida ainda dependente do contexto em que se insere a métrica de satisfação (que, por sua vez, remete aos conceitos de qualidade ou adequação de um sumário). Até o momento, as avaliações recaem pesadamente sobre (c), pois não há sistemas de SA suficientemente robustos que justifiquem as análises com foco em (a) ou (b).

Diversos desafios se destacam, dentre os quais [ManMay99]: (i) a identificação do que seria um resultado automático satisfatório – para um único texto-fonte, pode haver uma infinidade de sumários com perspectivas diversas. Por exemplo, para um leitor leigo no assunto do texto-fonte, um sumário que incluísse a contextualização do assunto poderia ser importante, enquanto que, para um leitor especialista, somente a informação nova poderia ser considerada. Essa variação tende a implicar taxas de compressão variadas. Para um sumário substitutivo de um texto-fonte, seu grau de informatividade deve ser bom, enquanto que, para um sumário indicativo, ele pode ser simplesmente uma lista de palavras. (ii) a determinação de uma taxa de compressão ideal – que pode variar, dependendo dos tipos de sumários

delineados em (i): em geral, quanto mais alta a taxa de compressão, mais informativo será o sumário (e, provavelmente, mais útil, mas menos comprimido). Vale notar que, aqui, o significado adotado para 'taxa de compressão' é o mesmo de Mani [Mani01] para *compression rate*: é a porcentagem de segmentos textuais *incorporados* ao sumário (e, assim, indica o tamanho do sumário). A inexistência de uma taxa ideal se deve ao fato de a informatividade depender do nível de conhecimento do usuário ao qual o sumário se destina, do tempo que ele dispõe para lê-lo e da utilidade do sumário, dentre outras coisas. Logo, essa relação de dependência não pode ser simplesmente fixada a qualquer modelo de SA. (iii) A forma como a qualidade e a informatividade de sumários automáticos podem ser avaliadas automaticamente, as quais, geralmente, são dependentes do julgamento humano, com todos os problemas decorrentes dessa dependência, como já citados. (iv) A identificação da situação e da forma de se utilizar o julgamento humano. Identificar o perfil adequado do juiz humano não é trivial, principalmente considerando-se que as avaliações admitem tantas facetas, implicando ainda um custo maior, de disponibilização de juízes humanos (e, quando necessário, especialistas em técnicas de sumarização), de robustez e abrangência das avaliações e, finalmente, de gerenciamento dos graus de subjetividade das respostas desses juizes.

Tentando minimizar essa dependência, existem várias propostas de estratégias e ferramentas para avaliação automática, algumas delas com desempenho significativo mesmo para a avaliação da qualidade dos sumários automáticos. Sua descrição será apresentada na próxima seção. Vale notar, ainda, que mesmo que haja avanços expressivos em relação a essas tarefas automáticas, a própria construção de seus corpora de referência ainda é uma tarefa árdua e cara, também dependente do esforço humano, como discutiremos na Seção 5.

4. Tipos de avaliação

Como podemos ver pelo exposto acima, não há métricas universais de avaliação de sumarizadores automáticos. É possível medir-se, por exemplo, o grau de utilidade de um sistema de SA, sua adequação a determinadas tarefas, a validade de sua metodologia, além de se avaliar as tradicionais textura e informatividade (lembrando que ambas também são associadas às medições anteriores). Várias métricas são reconhecidamente imprescindíveis, porém, na maioria das vezes, interdependentes, como evidenciaremos nesta seção.

Atualmente, a busca por avaliações automáticas robustas e consistentes se intensificou, implicando, sobretudo, grandes iniciativas de avaliação, que hoje podem se basear na avançada tecnologia de processamento textual e, em alguns casos, podem ser realizadas independentemente, sem o apoio de juizes humanos. Destaca-se, particularmente, a DUC (*Document Understanding Conference*)[2], cujos objetivos seguem similares aos da TIPSTER SUMMAC. Embora ambas representem o maior incentivo à avaliação de sistemas diversos, como os de Recuperação de Informação, Extração de Informação e SA, a DUC tem sido realizada regularmente e é considerada, hoje, a iniciativa mais importante de avaliação de sistemas de SA.

As principais formas de avaliação de sumarizadores automáticos são delineadas a seguir.

4.1. Avaliações intrínsecas

Nas avaliações intrínsecas, os aspectos principais a avaliar são o conteúdo e a qualidade. O conteúdo diz respeito à informatividade do sumário, isto é, o quanto de informação relevante do texto-fonte o sumário automático incorpora [Mani01], podendo haver variações sobre o texto-fonte; em relação à qualidade, verifica-se se o sumário automático observa as normas

[2] http://duc.nist.gov/ (último acesso em junho/2004).

gramaticais da língua em foco e as normas discursivas. Neste sentido, busca-se parametrizar sua textura, isto é, sua coerência e coesão [LinHovy03].

Para a avaliação de conteúdo, as principais medidas contemplam as seguintes formas de comparação: conteúdo do sumário com conteúdo de seu texto-fonte (esta medida é chamada *retention rate* por Mani [Mani01]); conteúdo do sumário com conteúdo de um sumário humano; conteúdo do sumário com conteúdo de um sumário ideal. Como já mencionamos na seção anterior, essas comparações só são possíveis a partir de corpora de sumários humanos e sumários ideais. Se considerada a abordagem empírica, sejam **Ni** o número de sentenças do sumário humano ou ideal (ambos considerados sumários de referência, serão referidos pelo termo mais genérico, sumários ideais), **Na** o número de sentenças do sumário automático e **Nc** o número de sentenças comuns ao sumário ideal e ao sumário automático. Neste caso, as principais medidas de comparação de conteúdo são assim definidas:

- Cobertura (C=Nc/Ni): indica a representatividade do sumário em relação a todos os dados considerados relevantes.
- Precisão (P=Nc/Na): indica o índice de sentenças relevantes no sumário automático, ou seja, a porcentagem de sentenças ideais que ele também apresenta.
- F-Measure (Fm=2*C*P/(C+P)): medida única de eficiência do sistema, indica que, quanto mais próxima de 1, maior será a capacidade do sistema em produzir sumários ideais.
- Utilidade [RadEtAl00]: como o próprio nome indica, esta medida visa expressar a utilidade do sumário automático. Ela é obtida manualmente, a partir de notas atribuídas por juízes humanos a cada sentença do texto-fonte, que indicam sua importância no texto. A partir dessa pontuação humana, selecionam-se as sentenças mais importantes para compor um sumário ideal e sua nota geral é calculada pela somatória das notas de suas sentenças. Como resultado, tem-se a *medida de utilidade ideal*. Para se avaliar um sumário automático, calcula-se de modo análogo sua medida de utilidade. Se ela for próxima à medida ideal, o sumário será considerado suficientemente informativo para ser útil.
- Conteúdo [SalMcGill83]: Embora algumas das medidas descritas antes contemplem o nível de informatividade, esta é mais genérica e lida com o grau de informatividade de um sumário mais diretamente. Também é mais diretamente aplicável a ambas as abordagens de SA (fundamental e empírica). Ela permite verificar a parcela de conteúdo do sumário ideal que é transmitida pelo sumário automático, porém, não considera simplesmente valores numéricos (número de sentenças ou tamanho dos sumários, por exemplo), mas classes de informatividade. Embora essa verificação possa ser auxiliada por processos automáticos, normalmente ela é subjetiva e necessita de julgamento humano para seu cômputo. Uma variação dessa medida faz uso da medida do co-seno [Salt89], que pode ser calculada automaticamente (podendo produzir resultados duvidosos, entretanto). Por essa medida, todas as sentenças dos sumários ideal e automático são representadas por vetores. O co-seno dos ângulos entre os vetores de cada sentença correspondente indica o grau de informatividade do sumário automático: quanto maior o valor resultante, menor a distância entre os vetores e, portanto, mais próximos em conteúdo estão os segmentos textuais dos dois sumários sob comparação. Raciocínio análogo pode ser estendido para verificar, por exemplo, a preservação do conteúdo do próprio texto-fonte no sumário correspondente, como fizeram Brandow et al. [BrandEtAl94].

Nos últimos anos, a avaliação intrínseca de sistemas de SA passou por grandes inovações que alteraram significativamente sua realização. Duas ferramentas de auxílio à avaliação se destacam: a SEE (*Summary Evaluation Environment*) [Lin01] e a ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) [LinHovy03]. A SEE permite que juízes humanos

avaliem de forma simples e intuitiva sumários automáticos, comparando-os a sumários ideais. Diversas medidas podem ser calculadas, então, a partir das respostas dos juízes: cobertura, precisão, completude de cada unidade de conteúdo do sumário em relação ao sumário ideal, fluência, gramaticalidade, coerência, informatividade e tamanho.

A ROUGE, baseada na métrica BLEU [PapiEtAl01] de avaliação de sistemas de tradução automática, é uma medida que verifica a cobertura de sumários automáticos em relação a um sumário ideal usando a co-ocorrência de n-gramas. Esse processo é realizado de forma completamente automática, apresentando um desempenho próximo às avaliações humanas, segundo os autores. Devido a seu desempenho e simplicidade surpreendentes, esta é a ferramenta de avaliação adotada na DUC'2004. Entretanto, o uso da ROUGE implica a construção de sumários manuais, os quais continuam demandando uma equipe de profissionais competentes.

Os critérios para se julgar a qualidade dos sumários variam muito, mas também envolvem o julgamento humano. Minel et al. [MinelEtAl97], por exemplo, pediram a juízes humanos que dessem notas a sumários observando os seguintes critérios: presença de referências anafóricas não resolvidas, não preservação da integridade de estruturas como listas e tabelas, falta de coesão entre as sentenças do sumário, presença de tautologias (definida como a repetição de um mesmo pensamento com palavras diferentes), etc. Saggion e Lapalme [SagLapal00] pediram que se observassem ortografia, gramática, a indicação clara do tópico do texto-fonte, o estilo impessoal, a concisão, legibilidade e facilidade de compreensão do sumário, além da presença de siglas seguidas de suas expansões. Pardo e Rino [ParRin02], seguindo os critérios de avaliação de White et al. [WhiteEtAl00], focalizaram a textura dos sumários e sua legibilidade.

É importante notar que as medidas de precisão, cobertura e utilidade também são importantes nas avaliações de métodos fundamentais. No entanto, torna-se necessário redefinir a forma de cômputo dessas medidas, já que não é mais possível, simplesmente, reconhecer padrões comuns entre os sumários automáticos e os dados de referência, sejam eles os textos-fonte ou os sumários ideais. Uma das possibilidades de avaliação intrínseca, neste caso, consiste em se comparar estruturas discursivas dos sumários automáticos com as correspondentes de textos-fonte ou, mesmo, com *estruturas discursivas ideais*, construídas de modo análogo ao da construção de sumários ideais (isto é, a partir de sumários autênticos ou profissionais). Note que essa proposta remete a uma avaliação *glass-box*, mais particularmente, do planejador de conteúdo dos sumários automáticos. Se o conteúdo for o foco, torna-se possível, assim, verificar a preservação dos conceitos-chave do texto-fonte, por sua ocorrência simultânea nas estruturas discursivas do próprio texto e de seu sumário. Se a qualidade for o foco, essa comparação pode ser feita entre as estruturas discursivas dos sumários automático e ideal. Já a avaliação *black-box* exige maior esforço, pois torna-se necessário comparar os resultados automáticos aos textos-fonte ou aos sumários ideais considerando-se que alterações substanciais entre eles podem ocorrer, devido à reescrita que, agora, permeia tanto o processo automático quanto os processos de construção dos sumários ideais. Uma das possíveis estratégias automáticas consiste, por exemplo, em se verificar se palavras-chave do texto-fonte, fornecidas pelo autor ou calculadas por algum "concordanceador", também podem ser palavras-chave do sumário automático ou se, pelo menos, estão presentes nele.

Para avaliações de sistemas fundamentais, Pardo e Rino [Par02a] ainda sugeriram outro tipo de avaliação: juízes humanos deviam dar notas a sumários automáticos de acordo com a preservação da idéia principal de seus textos-fonte. As notas, nesse caso, indicariam se o sumário preservava, preservava parcialmente ou mesmo se não preservava a idéia principal. Avaliação similar se aplicou também ao sistema extrativo GistSumm [Par02b] sem problemas, já que ela é totalmente dependente do juízo humano.

4.2. Avaliações extrínsecas

A avaliação extrínseca visa avaliar um sistema em uso, para a realização de alguma tarefa específica. Para a SA, ela pode envolver, por exemplo, os contextos de categorização de textos, recuperação de informação ou perguntas e respostas. Algumas dessas avaliações são descritas a seguir.

4.2.1. Categorização de documentos

A categorização de documentos, muito comum em *sites* de notícias e necessária para a catalogação de documentos em bibliotecas, visa atribuir uma categoria ou classe a um dado documento. Normalmente, essa atribuição é feita por juízes humanos. Na avaliação extrínseca de sumarizadores automáticos, esse contexto de categorização delineia o uso de sumários como único meio para se descobrir as categorias dos documentos correspondentes [ManiEtAl98]. Assim, os juízes humanos continuam com a responsabilidade de categorizá-los, porém, baseando-se somente na leitura dos sumários. Para completar a avaliação extrínseca, verificam-se ainda o tempo necessário para a realização da tarefa e a taxa de acerto dos juízes. Idealmente, espera-se que a taxa de acerto não degrade em relação à tarefa realizada de forma usual (isto é, lendo-se os textos-fonte em vez dos sumários) e que o tempo de realização da tarefa diminua. O objetivo dessa avaliação é, portanto, verificar se os sumários apresentam informação suficiente para a correta classificação dos textos-fonte, a partir da categorização dos próprios sumários.

4.2.2. Recuperação de Informação

Nesta tarefa, o objetivo é recuperar documentos que abordem um determinado tópico, como se costuma fazer em *sites* de busca de informação na web. Em geral, buscam-se os documentos de uma base de dados cujo tópico coincida com o tópico indicado pelo usuário. Na avaliação extrínseca de um sistema de SA, os sumários dos documentos da base em foco são gerados, os quais passam a servir de base para a busca, funcionando tipicamente como indexadores dos documentos. A busca pode ser automática ou manual, mas seu sucesso é avaliado por juízes humanos, que julgam se os sumários preservam os tópicos relevantes para que a busca seja bem sucedida. Calculam-se, assim, a taxa de acerto da recuperação de documentos e o tempo necessário para sua busca, como no caso da categorização. Novamente, espera-se que a taxa de acerto se mantenha e que o tempo de busca diminua. Além de Mani et al. [ManiEtAl98], outros pesquisadores adotam essa forma de avaliação, por exemplo, Tombros e Sanderson [TomSan98], Jing et al. [JingEtAl98] e Brandow et al. [BrandEtAl94].

4.2.3. Perguntas e Respostas

Em uma avaliação extrínseca de perguntas e respostas, tem-se por objetivo verificar se os sumários automáticos são informativos. Dada uma base de documentos, elaboram-se perguntas de múltipla escolha para cada texto e geram-se seus sumários automaticamente. Assume-se, neste caso, que as perguntas são bem elaboradas. Por fim, procede-se à avaliação propriamente dita, com base em respostas de juízes humanos. Essas respostas podem ser de três tipos: (i) os juizes devem responder às questões sem ler os textos-fonte nem os sumários; (ii) os juízes devem responder às questões somente lendo os sumários; (iii) eles devem responder às questões somente lendo os textos-fonte. A hipótese principal, neste caso, é que,

se os sumários forem suficientemente informativos, os juízes conseguirão responder às perguntas satisfatoriamente somente com base neles.

5. A construção do TeMário como uma coleção de avaliação de sistemas de SA

Do exposto anteriormente, conclui-se que corpora para avaliação ou mesmo modelagem e treino de sistemas de SA devem ser construídos com bastante cuidado, para contemplar as tarefas de avaliação em suas variadas formas. O gênero textual, neste caso, é uma das características mais importantes para a especificação de um corpus. Se as avaliações envolverem juízo humano, gêneros muito rebuscados podem dificultá-lo devido ao esforço necessário para o entendimento e apreensão dos conceitos principais dos textos envolvidos na tarefa. Por essa razão, o gênero mais comum utilizado nas grandes conferências de avaliação é o jornalístico. O fato de ser comum textos jornalísticos terem sua idéia principal explícita (em geral, em seu início) favorece o uso desses corpora.

Logo, textos jornalísticos são interessantes porque incorporam uma linguagem simples e abrangente, visando uma alta penetração na comunidade de leitores. Além disso, são relativamente curtos. Porém, os corpora correspondentes de dados de referência têm que ser integralmente construídos para as avaliações, a não ser nos casos em que elas envolvam, simplesmente, a comparação com seu título. Ainda assim, é preciso colecionar textos jornalísticos que sejam acompanhados pelos títulos, o que nem sempre é o caso. Os textos científicos já não apresentam o mesmo esforço de construção de dados de referência, principalmente quando consideramos artigos ou teses, por exemplo. Esse gênero é interessante porque os textos já são acompanhados por sumários de autoria. Além disso, eles possuem uma estrutura bastante padronizada, favorecendo as estratégias baseadas em localização de segmentos no texto-fonte ou em características macro-estruturais, por exemplo. No entanto, sua linguagem técnica ou mais elaborada restringe a comunidade de leitores e, assim, também a escolha de possíveis juizes humanos.

Outra preocupação fundamental que se deve ter na construção de corpora diz respeito à qualidade dos dados: é necessário garantir que os textos sejam de boa qualidade. Em geral, isso é assegurado, simplesmente, considerando-se dados sabidamente revisados. No entanto, mesmo o estilo da escrita livre pode introduzir complicadores para a modelagem e avaliação de sistemas de SA.

A coleção TeMário [ParRin03] (http://www.linguateca.pt/Repositorio/TeMario) foi escolhida pensando nesses problemas: seu corpus de textos-fonte contém 100 textos jornalísticos em PB, com tamanho médio de 613 palavras. Construída especialmente para avaliação de sistemas de SA, ela também contém sumários manuais e extratos ideais, como dados de referência. Assim, ela constitui uma coleção completa de treino ou teste para a SA. Os sumários manuais foram construídos por um sumarizador profissional e, portanto, são considerados também ideais, conforme nossa concepção apresentada na Seção 3. Os extratos ideais, por sua vez, foram construídos automaticamente, pelo GEI, **G**erador de **E**xtratos **I**deais (disponível em http://www.nilc.icmc.usp.br/~thiago), ferramenta baseada na similaridade de co-senos [Salt89], análoga à medida de conteúdo (vide Seção 4). Porém, no GEI ela é usada para se construir (e não avaliar) extratos, a partir de sumários manuais. Ela funciona da seguinte forma: para cada sentença do sumário manual, a sentença mais similar do texto-fonte é encontrada. Obtido o conjunto de sentenças similares, as mais representativas são selecionadas para compor o extrato ideal. Esse procedimento é baseado na premissa de que extratos ideais devem ser compostos por tantas sentenças quantas forem as de seus sumários manuais [ManBloe98].

6. Discussão e conclusões

Se, por um lado, o TeMário é bastante limitado por conter somente 100 textos, por outro ele tem um bom potencial de avaliação, justamente por ser agregado à coleção de sumários e extratos ideais. Um de seus maiores pontos frágeis está em seu corpus de extratos ideais: embora a técnica de similaridade de co-senos seja usada por vários sistemas proeminentes para o inglês, assumimos que o GEI tinha um bom desempenho e, assim, denominamos seus extratos de *ideais*. Certamente, os resultados de avaliações baseadas no TeMário, no momento, devem ser analisados considerando-se a interferência da qualidade desses extratos ideais. Entretanto, é importante lembrar que todos os sistemas que usam ferramentas complementares também têm sua qualidade dependente do desempenho das ferramentas. A nosso favor, temos os resultados de uma avaliação abrangente de sete sumarizadores automáticos de textos em PB, baseada somente no TeMário [RinoEtAl04]: as medidas de eficiência (*f-measures*) dos dois sistemas melhor classificados remontaram a, aproximadamente, 0,43, índice que, embora baixo, não difere significativamente dos índices de outros sistemas de SA voltados à língua inglesa, conforme discutimos em [RinoEtAl04].

Outra questão importante para as avaliações é a taxa de compressão utilizada. No TeMário, extratos ou sumários ideais foram construídos com uma taxa média de 30%, pois esta é a taxa aproximada dos sumários manuais. Interessante é registrar aqui a dificuldade da montagem desse corpus, que se deve tanto às concepções que profissionais possam ter em relação à tarefa que são solicitados a desenvolver, quanto a variações cronológicas. Por exemplo, para o profissional *educador*, construir um sumário manual de um texto com menos do que essa taxa seria corromper seu nível de informatividade e, certamente, sua qualidade. Esta foi justamente a acepção do profissional contratado para a construção do TeMário. Para um sumarizador profissional, várias possibilidades se abririam, dependendo do objetivo final de seus sumários. Já para um autor, o título de sua obra já é uma síntese e, portanto, um sumário (embora fragmentado), assim como o sumário de autoria de um trabalho científico, com diferente taxa de compressão.

Em relação às variações cronológicas, temos um sério problema: a demanda crescente por material disponível na *web* e a necessidade cada vez maior de se consumir velozmente a informação impõem variações freqüentes ao tamanho considerado ideal para os sumários automáticos. Assim é que, em um espaço de aproximadamente cinco anos, vimos a taxa de compressão cair de 30 para 10% e, mais recentemente, para exatas 10 palavras, como nas duas últimas DUC (sumários muito curtos, neste caso, com no máximo 75 bytes em formato linear). Considerando-se o contexto de SA multi-documentos, essa taxa é mais flexível (sumários curtos com, no máximo, 665 bytes para cada *cluster*), porém, ainda não ultrapassa a média de 90 palavras por sumário, medida que, para nosso estilo livre de escrita em PB, parece absolutamente inalcançável. Considerando-se tarefas muito específicas para as quais os sumários automáticos possam ser úteis, como a simples indicação da idéia principal do texto-fonte, essa restrição é justificável. Entretanto, ela implica severas limitações de projeto e desenvolvimento, incluindo também as de avaliação. Particularmente na DUC04, os sumários exíguos podem ser meras listas de palavras-chave. Por essa razão, sequer é considerada a possibilidade de que eles representem fatos completos; ao contrário, está em foco o compartilhamento de conteúdo entre os sumários automáticos e seus modelos manuais pelo simples compartilhamento de *referências* a entidades (pessoas, objetos, locais, eventos, etc.), cuja definição será construída manualmente, pelos assessores da conferência.

O problema principal, para o processamento do português, envolve também a complexidade cronológica e organizacional: as investigações na área de SA começaram mais de dez anos após as voltadas, sobretudo, ao inglês. Como sabemos, essa defasagem se deve ao fato de só recentemente termos ferramentas, recursos e corpora relativamente adequados a

várias tarefas do processamento de línguas naturais, sendo o TeMário o primeiro corpus que agrega também dados de referência construídos de duas formas distintas e, portanto, com potencial para vários tipos de experimentos. Ainda assim, a SA de textos em português é insignificante, se considerarmos o contingente de pesquisadores que trabalham na área, a inexistência de recursos lingüísticos robustos, assim como a dificuldade em se produzir "*gold standards*" [KupiEtAl95] que resultem em um diferencial considerável para a comunidade de Lingüística Computacional em geral. A solução para situar a SA de textos em português nesse amplo contexto parece ser, então, agregá-la às áreas afins, como a Recuperação e Extração de Informação, buscando uma forma comum de se produzir recursos e padrões que sirvam a ambas as áreas. Além disso, para que tenhamos sistemas de interesse real, seria interessante analisar seriamente a existência de características do português que sejam singulares, em relação às outras línguas mais exploradas para a SA, de modo a propor métricas diferenciadas de avaliação. Desse modo, nossos sistemas se tornariam mais úteis e mais dedicados aos interesses da própria comunidade luso-brasileira.

Outra forma de trazer maior motivação à SA seria, por exemplo, incorporar propostas de outras áreas do PLN, inovando no sentido de personalizá-las aos sumarizadores automáticos. Por exemplo, poderiam ser construídos alinhadores de sumários automáticos e sumários ideais, de modo a superar as dificuldades de avaliação geradas pelas diversas formas de produção de ambos os tipos de textos. Marcu, por exemplo, propõe um alinhador de textos e sumários usando o modelo noisy-channel [DauMarc02]. Esse tipo de ferramenta já está sendo explorado no NILC e, portanto, incorporá-la ao contexto de SA é uma perspectiva interessante. No mínimo, alinhadores como esse seriam suficientes para investigar a qualidade de nossa coleção TeMário: os extratos ideais produzidos pelo GEI poderiam ser comparados com os sumários manuais dessa mesma coleção, indicando se o corpus de extratos ideais disponibilizado é seguramente um corpus de referência. Um ambiente de apoio à SA também serviria de incentivo a colaborações entre diversos grupos de pesquisa. O SuPor [Modo03; RinMod04], por exemplo, pretende ser um ambiente dessa natureza, porém, ainda não foi explorado adequadamente devido à sua indisponibilidade pública.

Como conclusão, vale ressaltar que, para que os sistemas de SA de textos em português sejam equiparáveis àqueles que lidam com outras línguas, não basta somente replicarmos as técnicas e métricas de avaliação já exploradas, muito embora possamos concluir que algumas delas se aplicam até melhor, quando se tratam de sumários em português. O desenvolvimento da tecnologia de SA para nossa língua deve, assim, sofrer constantemente um questionamento sobre as necessidades e características particulares, além de nos levar a definir claramente os objetivos desejados para a SA, no contexto luso-brasileiro.

Referências bibliográficas

[BogKen97] Boguraev, B.; Kennedy, C. Salience-Based Content Characterisation of Text Documents. In I. Mani and M. Maybury (eds.), *Proc. of the Intelligent Scalable Text Summarization Workshop* (Madrid, Spain, 1997). ACL/EACL'97 Joint Conference. 1997. pp. 2-9.

[BrandEtAl94] Brandow, R.; Karl, M.; Rau, L.F. Automatic Condensation of Electronic Publications by Sentence Selection. *Information Processing & Management*, Vol. 31, N. 5 (1994): 675-685.

[CorOli98] Corston-Oliver, S. *Computing Representations of the Structure of Written Discourse*. PhD Thesis, University of California, Santa Barbara, CA, USA. 1998.

[DauMarc02] Daumé III, H.; Marcu, D. A Noisy-Channel Model for Document Compression. *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics*, ACL'02). (Philadelphia, July 2002): pp. 449-456.

[Edmun69] Edmundson, H.P. New Methods in Automatic Extracting. *Journal for Computing Machinery* 16 (2). (1969): pp. 264-285.

[GolEtAl00] Goldstein, J.; Mittal, V.; Carbonell, J.; Kantrowitz, M. Multi-document summarization by sentence extraction. In *Proc. of the ANLP/NAACL Workshop on Automatic Summarizatio*. (Seattle, WA, 2000).

[GroSid86] Grosz, B.; Sidner, C. Attention, Intentions, and the Structure of Discourse. *Computational Linguistics*, Vol. 12, No. 3 (1986).

[HovyLin97] Hovy, E.; Lin, C-Y. Automated Text Summarization in SUMMARIST. Proc. of the Intelligent Scalable Text Summarization Workshop, ACL/EACL'97 Joint Conference. (Madrid, Spain, 1997). pp. 18-24.

[JingEtAl98] Jing, H.; Barzilay, R. McKeown, K.; Elhadad, M. Summarization evaluation methods: Experiments and analysis. In the *Working Notes of the AAAI Spring Symposium on Intelligent Text Summarization*. (1998).

[KupiEtAl95] Kupiec, J.; Petersen, J.; Chen, F. (1995). A trainable document summarizer. In E. Fox, P. Ingwersen, & R. Fidel (eds.), *Proc. of the 18th ACM-SIGIR Conference on Research & Development in Information Retrieval* (Seattle, WA, EUA, 1995). ACM-SIGIR. 1995. pp. 68-73.

[Lin01] Lin, C-Y. (2001). *Summary Evaluation Environment*. http://www.isi.edu/~cyl/SEE [14 de junho de 2004].

[LinHovy02] Lin, C-Y. and Hovy, E. Manual and Automatic Evaluation of Summaries. In the *Proceedings of the Workshop on Automatic Summarization* (2002). pp. 45-51.

[LinHovy03] Lin, C-Y. and Hovy, E. Automatic Evaluation of Summaries Using N-gram Co-Occurrence Statistics. In the *Proceedings of HLT-NAACL* (2003). pp. 71-78.

[Mani01] Mani, I. *Automatic Summarization*. John Benjamins Publishing Co. 2001. Amsterdam.

[ManMay99] Mani, I.; Maybury, M.T. (eds.) *Advances in automatic text summarization*. MIT Press, Cambridge, MA. 1999.

[ManBloe98] Mani, I. and Bloedorn, E. Machine Learning of Generic and User-Focused Summarization. In the *Proc. of the 15th National Conf. on Artificial Intelligence* (AAAI 1998). pp. 821-826.

[ManiEtAl98] Mani, I.; Firmin, T.; House, D.; Chrzanowski, M.; Klein, G.; Hirschman, L.; Sundheim, B.; Obrst, L. *The TIPSTER Text Summarization Evaluation*. Final Report. 1998.

[MannTho88] Mann, W.C.; Thompson, S.A. Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. *Text*, 8 (3). (1988): pp. 243-281.

[Mar97a] Marcu, D. (1997a). From Discourse Structures to Text Summaries. In I. Mani and M. Maybury (eds.), *Proc. of the Intelligent Scalable Text Summarization Workshop* (Madrid, Spain, 1997). ACL/EACL'97. 1997. pp. 82-88.

[Mar97b] Marcu, D. The Rhetorical Parsing of Natural Language Texts. In the *Proc. of the ACL/EACL'97 Joint Conference*, (Madrid, Spain, 1997b). pp. 96-103.

[Mar00] Marcu, D. *The Theory and Practice of Discourse Parsing and Summarization*. The MIT Press. 2000. Cambridge, Massachusetts.

[MinelEtAl97] Minel, J.L.; Nugier, S.; Piat, G. How to appreciate the quality of automatic text summarization? In the *Proceedings of the ACL/EACL Workshop on Intelligent Scalable Text Summarization* (1997). pp. 25-30.

[Modo03] Módolo, M. *SuPor: an Environment for Exploration of Extractive Methods for Automatic Text Summarization for Portuguese* (in Portuguese). MSc. Dissertation. Departamento de Computação, UFSCar. 2003.

[PapiEtAl01] Papineni, K.; Roukos, S.; Ward, T. Zhu, W-J. *BLEU: a Method for Automatic Evaluation of Machine Translation*. IBM Research Report RC22176 (W0109-022). 2001.

[Par02a] Pardo, T.A.S. *DMSumm: Um Gerador Automático de Sumários*. Dissertação de Mestrado. Departamento de Computação. Universidade Federal de São Carlos. São Carlos - SP. 2002a.

[Par02b] Pardo, T.A.S. *GistSumm: Um Sumarizador Automático Baseado na Idéia Principal de Textos*. Série de Relatórios do NILC. NILC-TR-02-13. 2002b.

[ParRin02] Pardo, T.A.S.; Rino, L.H.M. (2002). DMSumm: Review and Assessment. In E. Ranchod and N. J. Mamede (eds.), *Advances in Natural Language Processing* (Faro, Portugal, 2002). Springer-Verlag, Germany. 2002. pp. 263-273 (Lecture Notes in Artificial Intelligence 2389).

[ParRin03] Pardo, T.A.S. e Rino, L.H.M. *TeMário: A Corpus for Automatic Text Summarization* (in Portuguese). NILC Tech. Report. NILC-TR-03-09. São Carlos, Outubro, 12p. 2003.

[ParEtAl03] Pardo, T.A.S.; Rino, L.H.M.; Nunes, M.G.V. GistSumm: A Summarization Tool Based on a New Extractive Method. In N.J. Mamede, J. Baptista; I. Trancoso; M.G.V. Nunes (eds.), *Proceedings of the 6th Workshop on Computational Processing of the Portuguese Language - Written and Spoken*. (Faro, Portugal, 2003): pp. 210-218. (Lecture Notes in Artificial Intelligence 2721). Springer-Verlag, Germany.

[ParEtAl04] Pardo, T.A.S; Nunes, M.G.V.; Rino, L.H.M. (2004). *DiZer: An Automatic Discourse Analyzer for Brazilian Portuguese*. XVII Brazilian Symposium on Artificial Intelligence - SBIA'04. (São Luís, Maranhão, Brazil, September 29-October 1 2004).

[RadEtAl00] Radev, D.R.; Jing, H.; Budzikowska, M. Summarization of multiple documents: clustering, sentence extraction, and evaluation. In the *Proceedings of the Workshop on Automatic Summarization*. (Seattle, WA, 2000). pp. 21-30.

[RathEtEl61] Rath, G.J.; Resnick, A.; Savvage, R. The formation of abstracts by the selection of sentences. *American Documentation*, Vol. 12, N. 2. (1961): pp. 139-141.

[Rino96] Rino, L.H.M. *Modelagem de Discurso para o Tratamento da Concisão e Preservação da Idéia Central na Geração de Textos*. Tese de Doutorado. IFSC-USP. São Carlos - SP. 1996.

[RinMod04] Módolo, M. *SuPor: An environment for AS of texts in Brazilian Portuguese.* España for Natural Language Processing – EsTal04 (2003).

[RinoEtAl04] Rino, L.H.M.; Pardo, T.A.S; Silla Jr., C.N.; Kaestner, C.A.A.; Pombo, M. *A Comparison of Automatic Summarization Systems for Brazilian Portuguese Texts*. XVII Brazilian Symposium on Artificial Intelligence - SBIA'04. (São Luís, Maranhão, Brazil, September 29-October 1 2004).

[SagLapal00] Saggion, H.; Lapalme, G. Concept identification and presentation in the context of technical text summarization. In the *Proceedings of the NAACL-ANLP Workshop on Automatic Summarization*. (Seattle, WA 2000). pp. 1-10.

[Salt89] Salton, G. *Automatic Text Processing. The Transformation, Analysis and Retrieval of Information by Computer*. Addison-Wesley. 1989.

[SalMcGill83] Salton, G.; McGill, M.J. *Introduction to Modern Information Retrieval*. McGraw-Hill. New York. 1983.

[SalBuc88] Salton, G. and Buckley, C. Term-weighting approaches in automatic text retrieval. *Information Processing and Management 24* (1988): pp. 513-523. Reprinted in: K. Sparck-Jones; P. Willet (eds.), *Readings in Information Retrieval (*1997): pp. 323-328. Morgan Kaufmann.

[SaltEtAl97] Salton, G.; Singhal, A.; Mitra, M.; Buckley, C. Automatic Text Structuring and Summarization. *Information Processing & Management* 33(2) (1997): pp. 193-207.

[Schil02] Schilder, F. Robust discourse parsing via discourse markers, topicality and position. In J. Tait, B.K. Boguraev and C. Jacquemin (eds.), *Natural Language Engineering*, Vol. 8. Cambridge University Press. 2002.

[SiCas88] Siegel, S.; Castellan, N.J. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill. 1988.

[SpaJo97] Sparck Jones, K. Summarising: Where are we now? Where should we go? *Proc. of the Intelligent Scalable Text Summarization Workshop, ACL/EACL'97 Joint Conference*. (Madrid, Spain, 1997).

[TomSan98] Tombros, A.; Sanderson, M. Advantages of query biased summaries in information retrieval. In the *Proceedings of the 21st ACM SIGIR Conference* (1998): pp. 2-10.

[WhiCar02] White, M.; Cardie, C. Selecting Sentences for Multidocument Summaries using Randomized Local Search. *Proc. of the Workshop on Automatic Summarization* (including DUC 2002). (Philadelphia, July 2002): pp. 9-18. Association for Computational Linguistics.

[WhiteEtAl00] White, J. S.; Doyon, J. B.; Talbott, S. W. Task Tolerance of MT Output in Integrated Text Processes. In *ANLP/NAACL 2000: Embedded Machine Translation Systems*. (Seattle, WA, 2000). pp. 9-16.