A Sumarização Automática de Textos: Principais Características e Metodologias*

Lucia Helena Machado Rino[1], Thiago Alexandre Salgueiro Pardo[2]

[1]NILC/Departamento de Computação – Universidade Federal de São Carlos (UFSCar)
Caixa Postal 676 – 13565-905 – São Carlos – SP – Brasil

[2] NILC/Instituto de Ciências Matemáticas e Computação – USP
Caixa Postal 668 – 13560-970 – São Carlos – SP – Brasil

lucia@dc.ufscar.br, thiago@nilc.icmc.usp.br

Abstract. *Automatic Summarization aims at simulating the main features of human summarizing, to know: to identify relevant text segments and put them together into the corresponding summaries. Summaries, in this context, are simply condensed texts of a source text. There are diverse Automatic Summarization models, which use either linguistic knowledge or statistical or empirical information. The composition of relevant information depends on the modeling process: it may be corresponding to a fully rewriting of the summary, similarly to what humans do, or it may be just a simple selection of segments and their literal reproduction as juxtaposed summary units. In this chapter, we present such diversity, illustrating both approaches by describing some automatic summarizers that have been developed at NILC.*

Resumo. *A sumarização automática pretende simular as principais características da tarefa humana: identificar segmentos relevantes de um texto e compô-los, para produzir os sumários correspondentes. Sumários, neste contexto, são sempre textos resumidos. Existem diversos modelos de sumarização automática, que envolvem o conhecimento lingüístico ou a manipulação estatística ou empírica das informações textuais. A composição das informações relevantes depende do tipo de modelagem: podem ser simuladas tanto a reescrita integral do sumário, como é, normalmente, a própria tarefa manual, quanto a simples justaposição de segmentos extraídos literalmente do texto original. Neste capítulo, apresentamos essa diversidade, ilustrando algumas propostas de ambas as abordagens com a descrição de alguns sistemas de sumarização automática desenvolvidos no NILC.*

1. A Sumarização de Textos

A sumarização de textos, de um modo geral, é uma atividade comum na vida de qualquer pessoa de nível de escolaridade médio ou superior. Textos são, se não um objeto principal de trabalho, um instrumento auxiliar para atualização ou comunicação em qualquer esfera profissional ou social. Os sumários de textos, aqui tomados em sua acepção de *resumos*, são, por sua vez, também textos. Por essa razão, constituem igualmente objetos de comunicação. Com o aumento desmedido dos meios de comunicação e distribuição, observamos hoje um acúmulo excessivo de textos de diversas naturezas, o qual nos leva, via de regra, à incapacidade de consumi-los em sua íntegra. Em geral, a razão dessa incapacidade está em nossa falta de tempo, dada a diversidade de tarefas que abarcamos na sociedade moderna. Não é à toa que, freqüentemente, recorremos às principais manchetes dos jornais diários ou às

* Sob o apoio financeiro da FAPESP (Proc. Nro. 01/08849-8), projeto em desenvolvimento no NILC – Núcleo Interinstitucional de Lingüística Computacional.

sinopses das principais notícias, elaboradas por autores de nossa preferência. Freqüentemente também assinamos semanários, como a Veja, Isto É ou, ainda, revistas mais especializadas, como a Exame ou *The Economist*, as quais nos trazem temas atuais, porém resumidos, dos principais acontecimentos ou avanços econômicos.

Considerando o meio acadêmico, vemos novos campos de uso intenso de sumários: vestibulandos brasileiros, hoje, não se dão ao trabalho de recorrer a obras completas da literatura recomendada. Recorrem, ao contrário, aos resumos tão difundidos na última década, supondo que absorverão os principais aspectos da obra correspondente sem o "pênalti" da leitura da obra completa, permitindo-lhes o sucesso no exame visado. Cientistas comumente selecionam o vasto material de atualização científica primeiramente perscrutando seu título e seu sumário, pois estes são itens obrigatórios de complementação da divulgação científica.

Com o crescente uso da Internet, essa situação só foi evidenciada: *viajar* pelas páginas de notícias a fim de apreender *o que é essencial* exige tempo, capacidade de *identificar o que é relevante*, no grande volume de informações disponível, e capacidade de mentalizar, de forma coerente, o conteúdo essencial. Todos esses fatores remetem à capacidade de *sumarizar*.

Passamos a ver, portanto, a grande difusão da tarefa de sumarização, cujos objetivos podem ser classificados de duas formas: do ponto de vista do leitor – e, portanto, do usuário de um sumário – e do ponto de vista do produtor – e, portanto, de seu escritor. Este último cenário nos dá o foco da sumarização textual, como a tarefa de *escrita de um texto condensado*, o sumário, com o objetivo de *transmitir ou comunicar somente o que é importante* de uma fonte textual de informação. Como veremos adiante, essa perspectiva é importante, pois ela direcionará a modelagem automática, para a especificação dos principais processos de tomada de decisão.

As principais premissas da sumarização podem, assim, ser enumeradas como segue:

- Está disponível um texto, aqui denominado *texto-fonte*, que deve ser condensado.
- A afirmação de que o objeto a ser sumarizado constitui um texto implica, adicionalmente, a existência de
    a) uma idéia central – o tópico principal do texto – sobre a qual se constrói a trama textual (no ensino fundamental, aprendemos que o texto deve ser desenvolvido a partir de uma idéia – nossa idéia central);
    b) um conjunto de unidades de informação que, reconhecidamente, têm relação com a idéia central em desenvolvimento;
    c) um objetivo comunicativo central que, implícita ou explicitamente, direciona tanto a seleção das unidades de informação quanto a seleção da forma como a informação será estruturada, para estabelecer a idéia pretendida.
    d) um enredo, tecido em função das escolhas antes citadas, visando transmitir a idéia central de forma coerente, a fim de atingir o objetivo comunicativo pretendido.
- Tomando por base essa relação de conceitos, a principal premissa da sumarização de textos pode ser, assim, expressa como a tarefa de *identificar o que é relevante* no texto e, então, *traçar o novo enredo*, a partir do conteúdo disponível, *preservando sua idéia central*, sem transgredir o significado original pretendido.

A não transgressão do original constitui a principal restrição da sumarização:

Em que medida e com que parâmetros ela se impõe na tarefa de sumarização?

Apesar de serem vagos os conceitos acima e, logo, insuficientemente formais para estabelecer um modelo de produção de sumários, como falantes graduados em nível médio ou superior conseguimos identificar, mesmo que intuitivamente, as premissas acima como determinantes

de dois parâmetros essenciais de julgamento da relação entre esses objetos textuais: (a) a relação do sumário com um texto-fonte (sendo este mais estendido e detalhado do que aquele) e (b) a qualidade do sumário. Como leitores, conseguimos reconhecer, e bem, o grau de proximidade entre ambos, a ponto de qualificar sumários como bons ou ruins dependendo de sua fidelidade ao que é essencial em sua fonte textual.

Como escritores, dificilmente parecemos guiados por um vínculo estrito e explícito, previamente estabelecido, para garantirmos a proximidade de um sumário com seu texto-fonte, muito embora sejamos capazes de estabelecê-lo, garantindo que o leitor possa reconhecer um sumário como o veículo de comunicação sucinta do que antes era expresso mais detalhadamente. Freqüentemente violamos, por exemplo, o objetivo comunicativo de um texto, ao produzir vários sumários, com diversas conotações. Isso é particularmente evidente em nosso meio acadêmico: ao submetermos um artigo a uma revista, primeiramente temos por objetivo convencer o editor de que o artigo deve ser publicado; ao termos o trabalho aceito, freqüentemente alteramos o sumário, agora visando o leitor da comunidade mais abrangente, de interessados no assunto em si. Assim, cada objetivo é determinante da forma como o enredo do sumário é construído. No primeiro caso, a ênfase no convencimento do leitor imediato pode fazer com que o conteúdo do mesmo seja menos técnico e, portanto, mais independente do próprio teor técnico do artigo. No segundo caso, a ênfase passa a ser na divulgação científica de seu conteúdo e, assim, convencer o leitor de que ele é válido deixa de ser relevante, passando a ser prioritário motivá-lo para a leitura do artigo completo. Essa conotação pode se relacionar, ainda, ao objetivo de atrair a atenção do maior número de leitores, implicando uma mudança radical da conotação anterior.

Outro exemplo mais comum de variadas formas de sumarização, porém, todas elas remetendo a um mesmo texto-fonte, é o de manchetes de jornais, que fazem parte de nossa vida diária. O Texto 1 da Figura 1, por exemplo, extraído de um jornal[1], pode ter as manchetes M1-M4, as quais ressaltam informações diferentes, atribuindo diversos graus de relevância ao conteúdo originário do mesmo texto-fonte. Vale notar, também, que o foco das manchetes M1-M3 é explícito no Texto 1: M1 e M3 são derivadas da sentença [1] e M2, da sentença [3]. Porém, o foco de M4 está implícito no texto, tendo sido inferido do conjunto de sentenças [8]-[10]. A inferência, neste caso, pode ser uma tentativa de recuperação do objetivo do escritor do texto (justificar o gasto com pesquisas; ressaltar a importância da pesquisa genética, etc.).

Texto 1: [1] Mosquitos alterados geneticamente em laboratório podem ajudar a combater a transmissão de doenças como a dengue. [2] A dengue é uma infecção por vírus, transmitida pela picada de mosquitos como o Aedes aegypt. [3] Em estudo publicado na edição de hoje da revista científica "Science", pesquisadores da Universidade Estadual do Colorado (EUA) criaram em laboratório um mosquito cujo organismo não aceita carregar o vírus. [4] O objetivo dos cientistas agora é fazer com que essa alteração do organismo dos mosquitos seja transmitida hereditariamente. [5] Assim, aumentaria a população de insetos refratários ao vírus. [6] A dengue provoca náuseas e dores de cabeça, articulações e músculos. [7] O tipo mais grave da doença, o hemorrágico, pode matar. [8] Em 1995, foram registrados 120 mil casos da doença no Brasil. [9] Em abril, o Ministério da Saúde anunciou um programa de combate à doença, que vai durar quatro anos e custar cerca de R$ 5 bilhões. [10] Cerca de 1250 municípios brasileiros, aproximadamente um em cada quatro, registraram casos de dengue.

Figura 1: Texto ilustrativo

[1] Corpus jornalístico do NILC (Pinheiro e Aluísio, 2003).

M1: A contribuição da pesquisa genética ao combate à dengue
M2: A criação de um mosquito resistente ao vírus da dengue
M3: O combate à transmissão da dengue com a ajuda de mosquitos alterados geneticamente
M4: A pesquisa genética pode ajudar a minimizar os custos de combate à dengue

Assim como manchetes podem ser uma forma de sumário de um texto, várias outras formas, além de textos condensados, podem ser reconhecidas como sumários, cada uma delas envolvendo pressuposições, conteúdos e características diversos, prevalecendo, contudo, sua correspondência com as respectivas fontes.

Sumários são, assim, entendidos (e usados), hoje, como objetos autônomos de comunicação. A sumarização humana, por sua vez, pode ser definida como "a tarefa de redução do tamanho de um texto-fonte, com preservação de seu conteúdo mais relevante".

Identificadas as principais características da tarefa humana de sumarização, resta saber como a Sumarização Automática de textos (doravante, referenciada por sua sigla, SA) pode incorporá-las, para simular a produção automática de sumários textuais e garantir que a correspondência entre os resultados automáticos e os textos-fonte seja, de fato, consistente. Distinguiremos, aqui, duas áreas igualmente importantes, para o projeto e desenvolvimento (P&D) dos sistemas computacionais dessa natureza, i.e., nossos *sumarizadores automáticos*[2]: a de modelagem de procedimentos para escolha e estruturação dos sumários a serem gerados automaticamente e a de avaliação dos mesmos, visando à avaliação do desempenho computacional. Nesse contexto, um sumarizador automático pode ser definido como

> Um sistema computacional cujo objetivo é produzir uma representação condensada do conteúdo mais importante de sua entrada, para consumo por usuários humanos

Para isso, ele deve ser capaz de identificar, em um texto ou em uma representação conceitual do mesmo, o que é relevante, estruturando as unidades informativas correspondentes de modo a assegurar que o sumário será coerente e consistente. Segundo Mani (2001), essa caracterização distingue a SA de outras áreas correlatas, dentre as quais destacamos:

- Recuperação de Documentos, que, para uma certa "chave de busca", visa produzir uma coleção de documentos relevantes, sem necessariamente condensá-los;
- Indexação, que visa identificar termos convenientes para a recuperação de informação;
- Extração de informação, que não necessariamente tem a condensação de informação como restrição fundamental;
- Mineração de textos, cuja principal função é identificar, nos mesmos, informações singulares, e não necessariamente informações principais, como é o caso da recuperação e preservação da idéia central, na SA.

Assim, a modelagem de um sumarizador automático terá como principais restrições as mesmas da tarefa humana: os sumários devem ser textos condensados de uma fonte (em nosso caso, um texto ou sua representação conceitual) e, como tais, devem ter um enredo claro e progressivo, desenvolvido em torno de uma idéia central, a qual deve coincidir com a idéia central da fonte. Essas restrições imporão critérios claros para a averiguação do desempenho dos sumarizadores automáticos, como veremos na Seção 4. Entretanto, a automação da tarefa não se restringe à mimetização do processo de escrita que normalmente é familiar aos falantes de uma língua, devido à sua complexidade: simular a reescrita de um texto, como faz o

[2] 'Sumarizador automático' será o termo usado para expressar, simplesmente, os sistemas computacionais que têm por objetivo sumarizar textos em língua natural.

escritor humano, é um grande problema, pois envolve processos igualmente complexos, de interpretação do texto e representação (obrigatória) de somente uma parcela dele – aquela relacionada à sua idéia central. Essa complexidade será evidenciada na Seção 2, ao descrevermos a abordagem fundamental de SA. Além desta, há um grande interesse, atualmente, pela abordagem empírica, que, diferentemente de incorporar modelos vinculados aos de comportamento humano na tarefa de sumarização, baseia-se em modelos matemáticos ou estatísticos para produzir resultados análogos.

O P&D em função dessas abordagens é, portanto, distinto: na primeira, o sumarizador automático deve incorporar modelos lingüísticos e/ou discursivos de interpretação e reescrita textual; na segunda, ele se baseia em modelos exatos de manipulação do conteúdo textual. Devido a essa diversidade, o processamento resultante caracteriza-se da seguinte forma:

- Abordagem fundamental: alto grau de representação simbólica do conhecimento lingüístico e textual e raciocínio lógico baseado em técnicas simbólicas, para estruturação e reescrita do sumário;
- Abordagem empírica: processamento prioritariamente baseado em reconhecimento de padrões derivados de informações ou distribuições numéricas. São usadas técnicas empíricas e/ou estatísticas para a extração dos segmentos textuais relevantes.

Em ambos os casos, consideram-se as principais premissas da sumarização textual, antes delineadas. Entretanto, é preciso trabalhar com as informações textuais, distinguindo-as e delimitando-as, para reconhecer tanto seu grau de relevância quanto seu inter-relacionamento, características que irão subsidiar as escolhas automáticas. Em geral, as informações textuais são associadas a unidades de conteúdo (ou unidades informativas), identificadas como unidades mínimas de significado no texto. Essas unidades podem expressar diversos níveis de detalhe e remeter a diversos contextos textuais (por sua localização), que sugiram diferentes mecanismos de compreensão e apreensão da mensagem contida no texto. Assim, a delimitação de unidades informativas simples contribui para a delimitação de contextos variados, tornando imprescindível distinguirem-se os limites textuais. É comum associar-se a uma sentença simples, por exemplo, um único significado, ou a um parágrafo, um único tópico do discurso. No Texto 1, por exemplo, distinguimos segmentos relacionados a diferentes tópicos, como demonstram as manchetes ilustrativas. Temos, nesse caso, três tópicos distintos, sendo o terceiro construído pela composição das sentenças [8]-[10].

As estruturas textuais irão contribuir igualmente para estabelecer o enredo textual e, assim, identificar as unidades que devem compor o sumário. Por isso, é importante definir a granularidade das unidades informativas, para sua delimitação. Há diferentes abordagens nesse sentido, como veremos a seguir.

2. A Sumarização Automática de Textos

As possibilidades de Sumarização Automática delineadas na seção anterior indicam o grande problema da Sumarização Automática: produzir sumários que, mesmo diversos de seus textos-fonte, reflitam sua interdependência. Duas características são essenciais nesse contexto: (a) sumários remetem, necessariamente, a seus textos-fonte; (b) sumários devem ser construídos de modo a não haver perda considerável do significado original, apesar de conterem menos informações e poderem apresentar diferentes estruturas, em relação a suas fontes. Assim, sumários são textos produzidos a partir de textos ou de suas correspondentes representações, podendo servir de indexadores ou substitutos dos mesmos. Essa distinção leva à sua classificação como sumários *indicativos* ou *informativos*, respectivamente (Sparck

Jones, 1993)[3]. Sumários indicativos não podem substituir os textos-fonte, pois não necessariamente preservam o que aqueles têm de mais importante, em termos de conteúdo e estrutura, transmitindo somente uma vaga idéia daqueles. Sumários informativos, ao contrário, contêm todos os seus aspectos principais, dispensando, por isso, sua leitura (são chamados autocontidos, nesse caso).

Em função dessa classificação, distinguem-se também sua funcionalidade e a forma de avaliar sua qualidade: sumários indicativos podem ser utilizados na classificação de documentos bibliográficos, de um modo geral, indicando seu conteúdo e agilizando o acesso às informações relevantes. Nesse caso, eles servem de *indexadores*. Sumários informativos, por serem autocontidos, servem de meios de informação, porém, apresentam uma relação mais complicada com seu texto-fonte, pois de seu objetivo dependerá bastante a avaliação sobre o quanto ele atende às necessidades do usuário. A utilidade dos primeiros sumários é mais clara e sua função mais limitada do que a dos últimos. Essas características permitem uma avaliação mais robusta de sua funcionalidade e qualidade. De um modo geral, também é mais fácil produzir automaticamente sumários indicativos do que informativos. Entretanto, ambos podem servir a diversas aplicações, ressaltando-se, especialmente, a área de Recuperação de Informação, muito importante nos dias de hoje.

Além da diferença funcional, os sumários também são comumente classificados pelo modo como são obtidos. Sparck Jones (1993a) classifica-os como *extracts* ou *abstracts,* fazendo a correspondência com o que ela chama, respectivamente, de *extração textual* e *condensação de conteúdo* (Sparck Jones, 1997), sendo este o processo correspondente à reescrita textual, antes citada. Essas formas remetem às abordagens descritas como empíricas e fundamentais, respectivamente, também denominadas abordagens baseadas em corpus e abordagens baseadas em conhecimento profundo (*knowledge-rich approaches*). Ambas as abordagens podem, ainda, explorar a estruturação do discurso, porém, distinguem-se na forma como a estrutura do mesmo é manipulada. Elas ainda delineiam arquiteturas típicas, que serão descritas abaixo.

Nesse texto, usamos o termo geral 'sumário' para nos referir tanto a *extracts* quanto a *abstracts.* Quando o sumário for derivado da metodologia fundamental, especificamente, associaremos esse termo ao próprio termo em inglês, *abstract*; quando ele for derivado da metodologia empírica, usaremos, simplesmente, a tradução literal para *extract*, i.e., 'extrato'.

De um modo geral, a SA pode ser expressa por três processos (Mani e Maybury, 1999), como ilustra a Figura 2: análise, transformação e síntese. A análise consiste em extrair uma representação computacional do texto-fonte; a transformação consiste em manipular essa representação a fim de produzir a representação do sumário. Finalmente, a síntese consiste em realizar linguisticamente essa última estrutura, produzindo o sumário, propriamente dito. Veremos que esses processos tomam diferentes formas, dependendo da abordagem considerada.

2.1. A SA Extrativa: Métodos Estatísticos e/ou Empíricos

A SA começou a ser explorada no final da década de 50, com a utilização expressiva de técnicas estatísticas de extração de conhecimento lingüístico dos textos-fonte. Luhn (1958), por exemplo, sugeriu o uso de informações estatísticas derivadas do cálculo da freqüência das palavras e de sua distribuição no texto para calcular uma *"medida relativa de significância"*[4]. Utilizou, para isso, tanto a granularidade individual (palavras), quanto a sentencial: sentenças

[3] Nessa classificação inclui-se ainda um terceiro tipo – o de sumário crítico (que seria uma resenha do texto) – não considerado nesse capítulo.

[4] Frases em itálico e entre aspas são traduções dos termos literais encontrados nas obras citadas.

mais significativas teriam *pesos maiores* e, assim, seriam escolhidas para compor o que ele chamou de *auto-abstract*. Adicionalmente, palavras mais significativas (e, portanto, de maior freqüência) correspondiam às atuais palavras-chave, tão conhecidas como representantes do conteúdo textual.

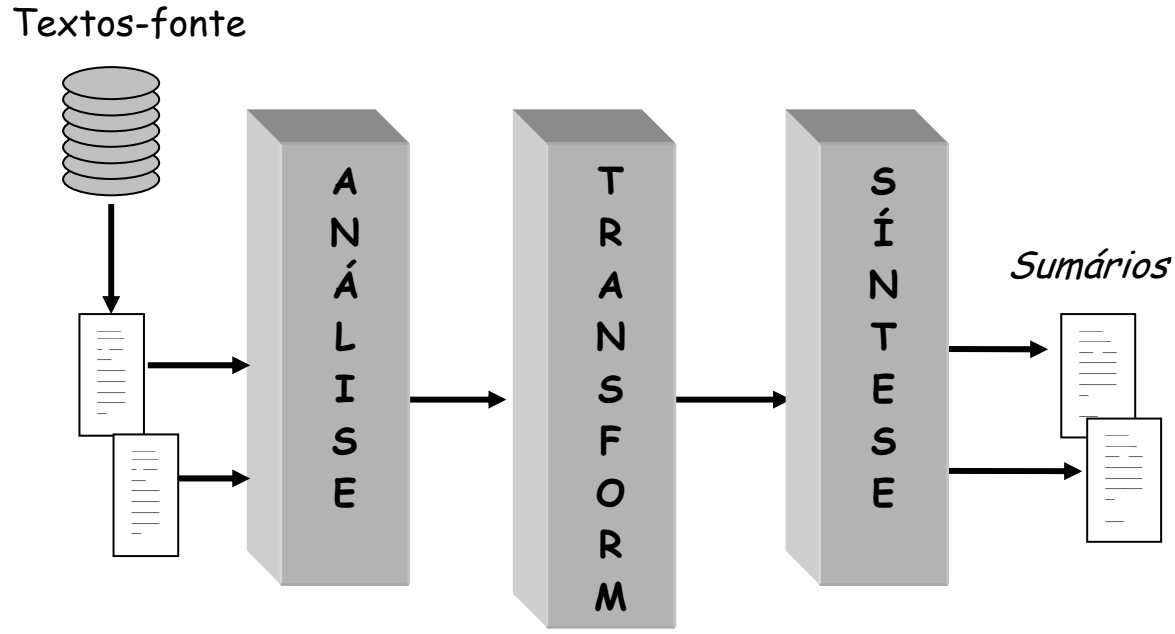


Figura 2: Arquitetura geral de um sumarizador automático

Dez anos depois, Edmundson (1969) propôs avanços sobre a metodologia de Luhn, denominando seu método de *"seleção computacional de sentenças com maior potencial de transmitir ao leitor a substância do documento"*. Além das distribuições sugeridas por Luhn, ele considerou também *"palavras pragmáticas"* (*cue words*), as quais hoje denominamos, simplesmente, *palavras sinalizadoras* (da importância do conteúdo textual, como *muito importante, significativa*, etc.), títulos e cabeçalhos, além de indicadores da localização do segmento na estrutura textual[5]. Seu sistema extrativo foi parametrizado de acordo com a influência de todos esses componentes, em uma combinação ponderada de características (normalmente, termo usado em inglês: *features*). Para essa ponderação, ele já fazia uso de dicionários eletrônicos, para reconhecer os segmentos textuais relevantes para compor os extratos[6]. Também significativa foi sua avaliação, comparando os extratos automáticos com extratos manuais e sugerindo desempenho melhor do que aquele baseado, simplesmente, na distribuição de freqüência. Entretanto, o foco desse trabalho não estava, propriamente, na garantia de progressão textual, mas sim na reprodução automática da tarefa de indicação de seu conteúdo (*document screening*), mesmo que não coeso ou coerente.

Significativo foi também o trabalho de Pollock e Zamora (1975), que sugeriu a necessidade de se restringir domínios (ou assuntos) para melhorar os resultados de métodos

[5] Sentenças mais importantes estariam, por exemplo, no começo e no fim de um documento ou seriam as primeiras e últimas de um parágrafo (Baxendale, 1958) ou, ainda, estariam logo abaixo de um título de seção.

[6] Aparentemente, nesse trabalho temos a primeira referência a *extratos* como sumários produzidos automaticamente pela metodologia de *extração de segmentos textuais*. O mesmo termo faz alusão às "*porções de um documento selecionadas para representar seu todo",* segundo Weil: Weil, B.H. (1970), Standards for writing abstracts. *Journal of the American Society for Information Science* 22(4): 351-357. (apud Pollock e Zamora, 1999, p. 43).

extrativos de SA, propondo, em adição aos trabalhos anteriores, o cruzamento de sentenças com o título da obra, para determinar aquelas significativas para um extrato. Vale notar que, neste caso, era necessário haver um título associado ao texto-fonte, para implementar o método.

De um modo geral, essas foram as obras clássicas de SA que deram origem ao que temos, hoje, de mais moderno em SA extrativa. Por longo tempo, entretanto, a exploração de métodos nessa linha ficou estagnada, devido à impossibilidade técnica para implementá-los (limitações de *hardware* e *software*, mas também de disponibilidade de recursos eletrônicos, como dicionários ou repositórios lingüísticos de grande porte). Na década de 90, vemos o ressurgimento do interesse por essa abordagem: os computadores passaram a ser de uso geral, suas memórias baratearam e recursos lingüísticos expressivos, como etiquetadores morfossintáticos e *stemmers*, tornaram-se disponíveis para o processamento textual. O conhecimento sobre manipulações estatísticas mais elaboradas pôde, assim, ser explorado para a SA de textos de domínios e gêneros variados, dando origem à metodologia baseada em corpus e caracterizando mais propriamente as diversas formas de transformação de um dado de entrada, para a o produção dos extratos. Assim, a partir da arquitetura geral (Figura 2), caracterizou-se a abordagem empírica como um processo de manipulação numérica/estatística de informações, ilustrado na Figura 3.

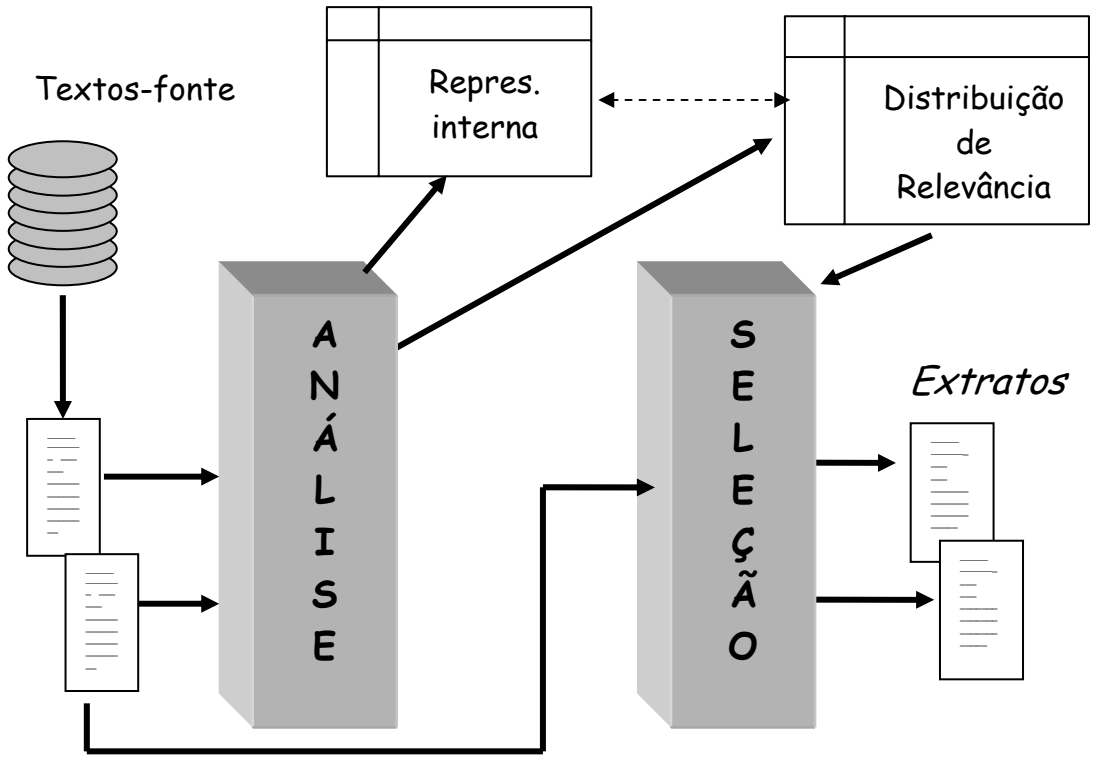


Figura 3: Sumarizador automático empírico

É na fase de análise textual que se exprime o avanço da SA extrativa: os procedimentos fundamentam-se nas distribuições clássicas de Luhn, Edmundson e Baxendale, porém, tornaram-se bastante sofisticados mais recentemente. A análise corresponde a algum tipo de esquadrinhamento e preparação do texto, para a manipulação das informações na fase de transformação que, agora, resume-se a uma simples tarefa de seleção a partir da distribuição de relevância. É comum termos, como sua representação interna, um vetor de características, sendo estas correspondentes aos aspectos significativos do texto-fonte. Por

exemplo, podem ser consideradas somente palavras de significado, ou palavras de classe aberta (substantivos, adjetivos, verbos e advérbios), as quais podem dar origem aos radicais (*stems*) correspondentes, anotados com sua categoria sintática ou morfológica. São removidas dos textos-fonte, portanto, as palavras de classe fechada (pronomes, conjunções, artigos, preposições), assim como palavras cujo significado seja irrelevante para o contexto (em geral, estas são dependentes do domínio em foco). É criada, uma lista com essas palavras, chamada lista de *stopwords*. Esta etapa da análise da entrada do sumarizador é chamada, assim, remoção de *stopwords*. As informações restantes darão origem à distribuição de relevância, que indicará os segmentos textuais selecionados para compor o extrato. Hoje também é comum utilizarem-se grandes volumes de dados textuais, para treinar o sumarizador a reconhecer informações significativas de textos em domínios específicos. As informações de entrada, neste caso, são classificadas segundo sua significância no contexto. Essa classificação serve de parametrização do sistema, para a escolha de segmentos durante a SA de textos-fonte de mesmo domínio e gênero que os utilizados no treino.

Os métodos empíricos adotam, em geral, a extração como processo fundamental: uma vez identificados como relevantes, segmentos inteiros de texto são extraídos do texto-fonte e integralmente incorporados ao extrato em construção, na mesma ordem em que eles se apresentam no texto-fonte. Assim, a síntese consiste simplesmente da justaposição dos segmentos considerados. Como é comum usar-se a medida sentencial de delimitação dos segmentos textuais (sob a suposição de que sentenças são a unidade mínima de significado), a escolha das sentenças visa à garantia de preservação do enredo (e, logo, da coerência textual).

Tipicamente, os métodos de identificação de segmentos relevantes calculam a significância de cada sentença em um texto-fonte por seu peso e, então, selecionam aquelas com maior peso (acima de um limite mínimo) para compor o extrato, incorporando os parâmetros clássicos para sua identificação e seleção (Luhn, 1958; Baxendale, 1958; Edmundson, 1969). Porém, eles manipulam grandes volumes de dados textuais para extrair as principais *features* para a identificação e seleção das sentenças, considerando domínios e gêneros particulares para melhorar o desempenho dos sumarizadores. As propostas mais recentes exploram, ainda, os papéis semânticos de cada unidade informativa (p.ex., Paice e Jones, 1993), as relações retóricas delineadas pelo inter-relacionamento de diversas unidades informativas (p.ex., Marcu, 1997a, 2000; Miike et al., 1994) e a similaridade estrutural entre sentenças (p.ex., Salton et al., 1997).

Um dos trabalhos mais importantes nessa linha é o de Kupiec et al. (1995), que propõe um sumarizador extrativo que, primeiramente, deve ser treinado para reconhecer as características textuais que parametrizam o sistema, i.e., aquelas que servem de base para a determinação da significância dos segmentos textuais, para inclusão no extrato. A hipótese principal dessa proposta é que a formulação de regras de ponderação da significância de sentenças de um texto é heurística e empírica por natureza. Assim, ela depende do treinamento do sistema sobre um corpus específico, o corpus de treino. O sistema é baseado em um classificador estatístico (bayesiano) (Mitchell, 1997) que agrupa as potenciais *features* a partir da comparação do conteúdo de textos-fonte com o conteúdo de seus respectivos extratos, construídos manualmente. Como resultado, é possível elencar um grupo de *features* ou de critérios de ponderação, para manipular as sentenças dos textos a sumarizar. A restrição, aqui, é que estes sejam similares àqueles do corpus de treino, devido à dependência de gênero e domínio, na fase de treinamento. Após vários experimentos, Kupiec et al. fixaram um conjunto de *features* consideradas significativas para seus dados (um corpus de textos sobre engenharia), que compreende o comprimento da sentença, a existência de sintagmas sinalizadores, a localização de sentenças no texto-fonte (início, meio ou fim de parágrafos), um conjunto de palavras "temáticas" (as mais freqüentes, neste trabalho, são consideradas

temáticas) e a ocorrência de substantivos próprios. Embora essas *features* tenham sido derivadas de um corpus específico, não é difícil constatar que a maioria delas é comum a outros domínios e, logo, a proposta é bastante abrangente.

Uma vez construídas as classes de *features,* são calculadas as probabilidades de todas as sentenças de um texto-fonte, de serem incluídas em um extrato. Essas probabilidades irão indicar, portanto, sua seleção (ou exclusão) do texto final. Adicionalmente, a decisão de inclusão depende, também, da taxa de compressão desejada pelo usuário: essa taxa corresponde ao *volume de redução* do texto-fonte. Considerando-se que seu tamanho possa ser medido por número de sentenças, por exemplo, ela corresponde ao número de sentenças que deverão ser *excluídas* do sumário final. Assim, ela pode ser definida pela fórmula geral (TC = Taxa de Compressão):

TC = 1 - (tamanho do extrato / tamanho do texto-fonte)

Atualmente, é comum se estabelecer a TC, principalmente na abordagem empírica, para evitar que sentenças muito longas resultem em textos pouco condensados ou, ao contrário, para estabelecer a diversidade pretendida pelo usuário, na produção de sumários. Em geral, na SA consideram-se sumários que correspondam a 10-20% dos textos-fonte e, portanto, que tenham uma taxa de compressão de 80 a 90% desses.

O trabalho de Kupiec et al. foi o marco responsável pelo *boom* da exploração de técnicas extrativas ainda mais robustas, estabelecendo a área hoje conhecida como SA baseada em corpus: métodos estatísticos de extração operam sobre sumarizadores treináveis a partir de corpora robustos de textos. A SA passou a ser, assim, um problema de classificação estatística: o objetivo é buscar uma função que calcule a probabilidade de uma sentença ser incluída no extrato (e, logo, que expresse sua significância), combinando características diversas. O problema, aqui, é determinar a contribuição relativa de diferentes *features,* condição altamente dependente do gênero textual (Mani e Maybury, 1999): textos científicos, por exemplo, podem concentrar informações relevantes no *abstract* e nas conclusões. Heurísticas distintas, derivadas da classificação das informações nos corpora, podem resolver esse problema. Essa possibilidade de treinar um sumarizador com base em textos-fonte específicos, para melhorar seu desempenho, trouxe também novas perspectivas para a avaliação dos resultados automáticos, como veremos na Seção 4.

Teufel e Moens (1999) estendem o método de Kupiec et al., adicionando à classificação probabilística a função retórica de cada sentença, associada à estrutura do discurso. Seu trabalho diverge, assim, na forma de análise: ainda é preciso esquadrinhar o texto-fonte, para produzir uma distribuição de seus segmentos. Porém, também são identificados os papéis retóricos de cada sentença no texto. A extração da distribuição retórica de um texto-fonte é baseada em sua macro-estrutura: as categorias distintas de informação que caracterizam os segmentos mais genéricos do texto são responsáveis por indicar a funcionalidade de cada segmento. Por exemplo, para textos científicos, os macro-componentes podem incluir *problema, propósito, metodologia, resultados, conclusões, trabalho futuro*, etc. Também nesta fase Teufel e Moens usam o classificador bayesiano.

Seguindo essa abordagem híbrida, foram desenvolvidos métodos mais sofisticados, que incorporam ao tratamento numérico das informações textuais também a identificação e o processamento empírico de informações lingüísticas e discursivas dos textos-fonte. Esses métodos são diferenciados por sua abordagem baseada na estruturação do discurso. A proposta de Barzilay e Elhadad (1997) é um exemplar disso: seu sumarizador explora a *coesão lexical* (i.e., o encadeamento de itens lexicais no texto), identificando nos textos-fonte as possíveis *cadeias lexicais*. Aquelas cadeias mais fortemente conectadas indicam as sentenças significativas para compor o extrato.

O trabalho fundamental dessa proposta de automação é puramente lingüístico, remetendo à coesão textual (Halliday e Hasan, 1976) e ao uso da repetição lexical para determinar os graus de coesão (Hoey, 1991). É baseada na proposta manual de cômputo das cadeias lexicais de Morris e Hirst (1991), havendo se tornado factível automaticamente pela disponibilidade de fontes robustas de conhecimento, tais como (a) uma ontologia que pudesse indicar os elos entre diversas palavras – a WordNet (Miller, 1995); (b) um etiquetador morfológico – que associa etiquetas a cada palavra, indicando sua categoria morfológica; (c) um *parser,* para identificar grupos nominais (envolvendo substantivos e adjetivos) e (d) um algoritmo de segmentação textual, responsável por delimitar, no texto-fonte, os segmentos que indicam as cadeias léxicas mais fortes.

Barzilay e Elhadad consideram somente substantivos e compostos nominais, para compor cadeias lexicais. O cômputo do relacionamento semântico das cadeias de palavras é feito de diversas formas: pela identificação de palavras idênticas ou palavras com mesmo significado; por sinonímia; por relações ontológicas de herança ou "parentesco". No primeiro caso, incluem-se a hiperonímia ou hiponímia (relações de superclasses ou subclasses, respectivamente). Por exemplo, *carro* e *Toyota* têm seus significados ontologicamente relacionados. No segundo caso, incluem-se relações de mesmo nível (também chamadas paratáticas). Por exemplo, a existente entre *caminhão* e *carro*.

A fase de análise dos textos-fonte compreende seu pré-processamento, pela seleção das palavras candidatas e segmentação do texto-fonte em tópicos, e a construção das cadeias lexicais, propriamente dita, que envolve a identificação das relações ontológicas entre as palavras. Finalmente, a síntese para a produção dos extratos baseia-se em três heurísticas distintas, para identificar as sentenças que contêm as cadeias lexicais fortes: a que focaliza a primeira ocorrência das sentenças no texto-fonte, a que identifica as sentenças que possuem os membros mais representativos e a que se concentra na significância do tópico indicado pelas sentenças. Esta última heurística sugere que um leitor pode identificar melhor o tópico de um texto simplesmente identificando suas cadeias lexicais mais representativas.

O maior problema dessa abordagem é identificar as palavras polissêmicas da língua natural: não há repositório eletrônico capaz de definir as acepções mais prováveis para casos ambíguos, pois elas são dependentes do contexto, o qual é variável. Assim, várias cadeias lexicais podem ser derivadas de uma única construção, dificultando a tarefa de identificação das informações relevantes. Outro problema, extensivo às demais abordagens empíricas, é introduzido pela impossibilidade de resolver anáforas ou de controlar o nível de detalhe dos extratos resultantes, pois não é feito o tratamento interpretativo do material indicado pelas cadeias lexicais. Essa questão, via de regra, só poderá ser adequadamente tratada pela abordagem fundamental.

Propostas como as descritas evidenciam a grande variedade de abordagens extrativas, várias delas recorrendo a técnicas de aprendizado e treinamento automáticos com base em grandes corpora de textos, resultando em técnicas mais robustas, porque mais informadas, quando comparadas aos métodos extrativos mais simples. É importante notar que elas sugerem a manipulação numérica, em geral estatística, de componentes textuais, considerando medidas que, *implicitamente*, incorporam características lingüísticas e a experiência de sumarizadores humanos. De fato, na tarefa de identificação e cópia de material dos textos-fonte para produzir os extratos, as métricas da SA extrativa modelam, sobretudo, as adotadas por sumarizadores profissionais (p.ex., Borko e Bernier, 1975; Cremmins, 1996), e, logo, estão próximas da tarefa de sumarização profissional, área clássica, anterior ao uso do computador e, portanto, de sua simulação. A manipulação numérica inicial foi, assim, incrementada com a

disponibilidade de recursos lingüísticos mais abrangentes e, conseqüentemente, de sumarizadores automáticos mais sofisticados e robustos. Por esse motivo, hoje a adoção de métodos empíricos é muito promissora, principalmente ante a urgência de se processar grandes volumes de informações textuais disponíveis eletronicamente. Assim é que podemos encontrar ferramentas de SA na Internet (como a do AltaVista) ou em ambientes de edição de textos (como o AutoResumo do MS Word™).

Os avanços da SA extrativa evidenciam ainda uma área inovadora e igualmente importante: a de avaliação. Em geral, usam-se "*gold standards*" (Kupiec et al., 1995) – padrões de referência definidos por especialistas humanos em sumarização textual – tanto para o treinamento dos sistemas quanto para sua avaliação.

Durante a fase de estagnação da SA extrativa, observamos a instalação da abordagem fundamental para a SA de textos, sobretudo a partir das idéias de Chomsky (1965): a modelagem computacional dos processos de compreensão e apreensão da estrutura textual, a fim de reescrever o texto-fonte de forma condensada, pôde ser formalizada a partir de gramáticas livres de contexto, responsáveis por analisar sintaticamente (*parsing*) os textos-fonte (de um domínio particular), para produzir sua representação conceitual.

2.2. A SA baseadas em conhecimento profundo: métodos fundamentais

Os principais problemas da abordagem fundamental (ou analítica) estão na forma como é identificada e sintetizada a informação relevante: a Figura 4 sugere a simulação do próprio processo humano de sumarização, composto da compreensão do enunciado do texto-fonte, com posteriores condensação de conteúdo e reescrita textual, conceitos já introduzidos no início desta seção, denominados por Hovy e Lin (1997) de *reescrita* e *fusão* de vários conceitos em um número menor de conceitos. Portanto, são três as etapas de SA (Sparck Jones, 1993): a construção de uma representação do significado a partir do texto-fonte (Repres. Conceitual I), a geração da representação do sumário correspondente (Repres. Conceitual II) e a sua síntese, ou realização lingüística, resultando no *abstract*, propriamente dito[7]. Essa última etapa é responsável pelas escolhas morfossintáticas da língua natural em foco, as quais não necessariamente coincidem com as apresentadas no texto-fonte.

Segundo essa arquitetura, um sumarizador automático contempla três tipos de informação: o *lingüístico*, o *informativo* (ou de domínio) e o *comunicativo*, remetendo a questões semânticas e pragmáticas que aumentam a complexidade dos sistemas, devido à necessidade de modelagem do conhecimento necessário para manipulá-la. É necessário haver uma linguagem de representação que possibilite o inter-relacionamento entre as unidades de significado (aqui chamadas de *proposições*) e engenhos de inferência capazes de interpretar o texto-fonte e gerar sua forma condensada correspondente. Como são usados métodos simbólicos e modelos computacionais de geração textual bastante complexos para a manipulação de conhecimento profundo, é possível mesmo que *abstracts* contenham informações que não se encontram no texto-fonte, decorrentes de processos inferenciais sobre seu próprio conhecimento explícito. Modelos para distinguir os diferentes graus de importância das informações dependem da caracterização dos interesses do escritor, os quais são regidos por objetivos comunicativos, e de modelos de estruturação do discurso, razão pela qual alguns métodos são também conhecidos como *métodos baseados em estruturação de discurso*.

[7] Lembrando que '*abstract'* é o termo que adotamos para diferenciar a metodologia empírica da fundamental.

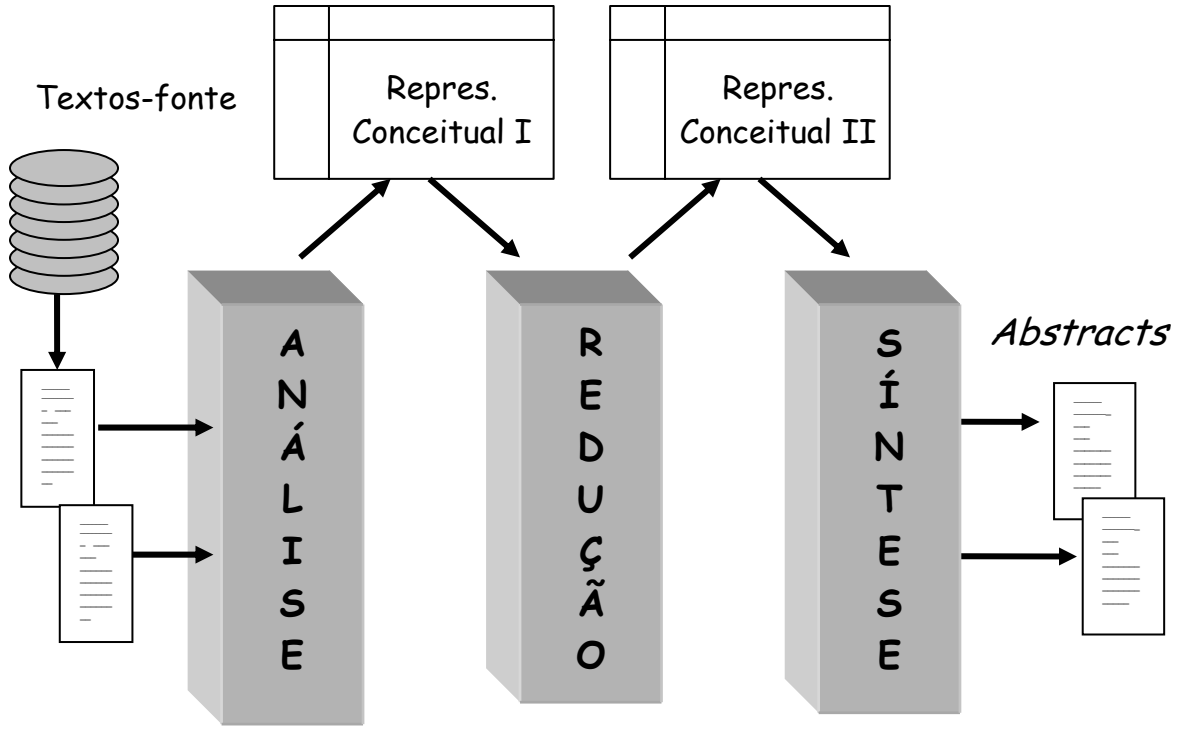


Figura 4: Sumarizador automático fundamental

O processo de análise, nessa abordagem, geralmente incorpora a um *parser* também um analisador discursivo, para produzir uma representação conceitual discursiva (e não sintática), pela qual se possam recuperar as relações entre os diversos segmentos textuais, assim como sua significância para a SA. Devido à natureza da estrutura global, são as proposições (e não suas expressões superficiais na forma de sentenças) que regem a análise. A premissa fundamental, neste caso, é que a idéia representada por um conjunto de proposições é estruturada de forma coerente antes mesmo das escolhas de vocabulário ou estrutura sintática, características intrínsecas da língua natural e não da linguagem do pensamento. Essa premissa justifica a coerência e coesão entre as unidades informativas, assim como a existência de diversos textos para uma mesma mensagem (diversas realizações lingüísticas para uma mesma estrutura conceitual).

Várias são as perspectivas dessa abordagem, para determinar as informações relevantes a partir da modelagem discursiva. A *saliência* das informações de um texto-fonte é uma propriedade importante, definida como a "*medida de proeminência relativa dos objetos ou conceitos textuais*" (Boguraev e Kennedy, 1997): as unidades informativas com grande saliência são o foco de atenção no discurso e, logo, devem ser consideradas nos *abstracts*; as com baixa saliência são periféricas e, logo, são passíveis de exclusão dos mesmos. Essa noção equivale á noção de significância ou relevância amplamente explorada na SA, que rege os critérios de escolha e agregação de segmentos textuais ou proposicionais, para a produção de sumários, de um modo geral.

O trabalho mais importante, hoje, nessa linha, é o de Marcu (1997a, 1997b, 2000), que propõe técnicas de segmentação do discurso para identificar o tópico e, a partir deste, estabelecer a saliência das informações relacionadas. A determinação das informações salientes é feita com base na estrutura retórica do texto, formalizada segundo a Teoria RST – *Rhetorical Structure Theory* (Mann e Thompson, 1988). Assim, é preciso primeiro construir a estrutura retórica do texto-fonte (tarefa de análise discursiva), para, então, determinar o conteúdo e a forma de seus possíveis sumários (tarefa de redução), ou seja, produzir a estrutura retórica do sumário correspondente. A vantagem dessa abordagem está na própria

definição das relações retóricas: elas indicam a assimetria do relacionamento proposicional, pela identificação de funções discursivas distintas. Estas, por sua vez, são construídas pela agregação de informações nucleares (os *núcleos*) e complementares (os *satélites*). Assim, Marcu explora a própria *nuclearidade* da Teoria RST, para identificar informações extraídas dos textos-fonte com diferentes graus de saliência.

O cômputo da saliência dos componentes do discurso se baseia tanto na nuclearidade quanto em sua profundidade na estrutura RST: núcleos mais próximos da raiz são considerados mais importantes do que seus satélites ou outros núcleos mais distantes da mesma. Uma possível estrutura RST para o Texto 1 é ilustrada na Figura 5 (N indicando núcelo e S o satélite). Cada proposição, neste caso, é delimitada por um segmento textual (numerado no Texto 1), sob a hipótese de que ele é a expressão superficial da proposição subjacente. As folhas da estrutura indicam, assim, as proposições, enquanto seus nós intermediários remetem às relações RST. Na Figura 5 são usadas somente as seguintes relações, com suas respectivas funções retóricas: ELABORATION (S elabora ou apresenta detalhes sobre N), LIST (as proposições fazem parte de uma lista de itens comparáveis, segundo algum critério de similaridade), JUSTIFY (S justifica N), PURPOSE (N é iniciado para a realização de S).

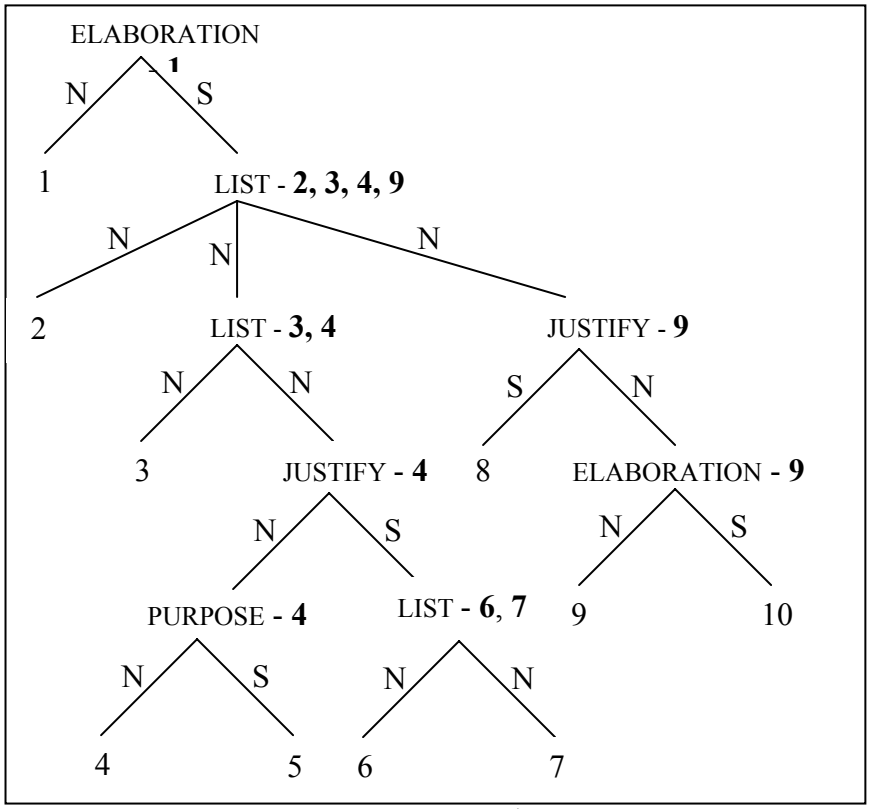


Figura 5: Estrutura RST do Texto 1

Nessa estrutura RST, as unidades mais salientes de cada segmento discursivo são indicadas junto aos nós intermediários. A ordem de precedência entre todas as proposições desse discurso é dada por 1>2>3>9>4>8>10>6=7>5 ('p1>p2' indica que p1 é mais importante que p2, assim como 'p1=p2' indica que p1 tem a mesma importância que p2). Sumários do Texto 1 podem, agora, ser construídos respeitando-se essa ordem. Variando-se o número de segmentos a incluir, podemos ter as estruturas RST 1 e 2 da Figura 6, as quais podem ser expressas superficialmente, por exemplo, pelas manchetes M1 e M2, apresentados na Seção 1, como sumários do Texto 1.

Vemos, assim, que a mensagem M1 envolve somente a relação ELABORATION entre 1 e 2; a mensagem M2 envolve também a relação LIST entre 2, 3, 4 e 9. A estrutura RST 1 tem a proposição 1 como mais saliente, enquanto a estrutura RST 2 tem a 3. Vale notar, também, que, por serem representações conceituais *profundas* da mensagem, essas mesmas estruturas poderiam derivar outras escolhas superficiais, produzindo textos diversos.

Essa proposta parece ser a mais consistente e efetiva atualmente, sendo independente de gênero e correlacionando-se à percepção que leitores têm sobre a importância de unidades textuais (Marcu, 1999). Entretanto, ela pressupõe a disponibilidade de estruturas RST para cada texto-fonte a sumarizar e, logo, requer um bom interpretador de língua natural, que gere suas estruturas retóricas. Atualmente, existem alguns interpretadores dessa natureza, para a língua inglesa (Marcu, 2000; Corston-Oliver, 1998; Schilder, 2002). Para o português, há uma proposta de análise discursiva em estágio inicial[8], associada ao modelo discursivo do sistema DMSumm, ilustrado na próxima seção.

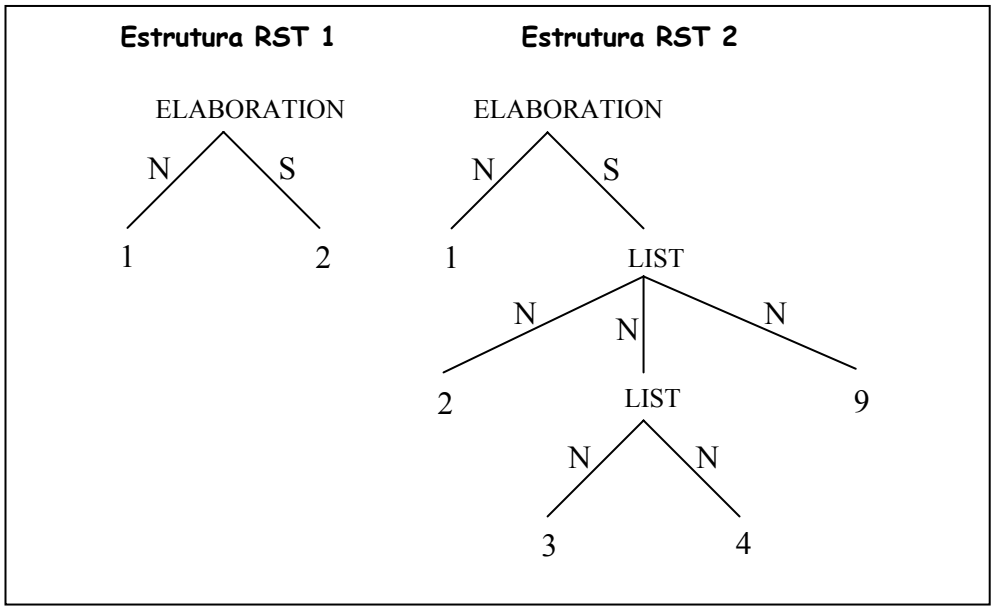


Figura 6: Possíveis estruturas RST das mensagens M1 e M2

3. Ilustrações: A experiência do NILC em Sumarização Automática

No NILC, exploramos tanto a abordagem empírica quanto a fundamental, sob o Projeto EXPLOSA[9]. Apresentamos, aqui, dois sistemas de cada abordagem (empírica e fundamental).

3.1. O GistSumm

O GistSumm, sigla para **GIST SUMM**arizer, é um sumarizador extrativo que faz uso de técnicas estatísticas simples para determinar a idéia central (o *gist*) dos textos a sumarizar. Ele se baseia na simulação da forma de sumarização humana, inicialmente identificando a idéia principal do texto e, então, acrescentando informações complementares. Assim, primeiramente ele procura a sentença que melhor expressa a idéia principal, chamada aqui de

[8] Pardo, T.A.S. (2003). Análise discursiva automática de textos em português do Brasil. Proposta de Doutorado. ICMC/USP, Junho.

[9] Sigla para **EXPLO**ração de métodos de **S**umarização **A**utomática, Proc. FAPESP Nro. 01/08849-8 (http://www.dc.ufscar.br/~lucia/PROJECTS/EXPLOSA.htm) (último acesso em maio/2003).

gist sentence, e, baseando-se nela, seleciona as demais sentenças para compor o extrato. Além das premissas básicas da SA, ele considera, portanto, que é sempre possível identificar, no texto-fonte, essa sentença. Com base nessas premissas, as hipóteses do GistSumm são as seguintes: (a) a identificação da *gist sentence* é possível com o uso de métodos estatísticos simples; (b) conhecendo-se a *gist sentence*, é possível produzir extratos coerentes por meio da justaposição de sentenças do texto-fonte relacionadas a ela. Consideramos que (a) pode ser confirmada também quando a sentença escolhida não for a *gist sentence*, mas uma aproximação significativa da mesma.

A exemplo dos métodos empíricos sem treinamento, o GistSumm compreende três processos: segmentação textual, ranqueamento e seleção de sentenças. A segmentação textual simplesmente delimita as sentenças do texto-fonte, procurando pelos sinais tradicionais de pontuação. Para o português, por exemplo, esses sinais incluem o ponto final e os sinais de exclamação e interrogação. O Texto 1 da Figura 1 (Seção 1) é segmentado automaticamente pelo GistSumm (sentenças numeradas). O ranqueamento consiste de sua ordenação, a partir de seus pesos, obtidos pela aplicação de métodos estatísticos. Ele ocorre em várias etapas, sendo que várias fases se aplicam a cada sentença. Para a sentença [7] do Texto 1, por exemplo – "O tipo mais grave da doença, o hemorrágico, pode matar." – os seguintes dados são manipulados em cada fase:

1) Vetorização das sentenças: Cada sentença é representada como um vetor (segundo Salton, 1988) cujas posições armazenam suas palavras.

O	tipo	mais	grave	da	doença	o	hemorrágico	pode	matar

2) *Case folding*, troca por canônicas e remoção de *stopwords*

Os processos de *case folding*, troca por canônicas e remoção de *stopwords* (sugeridos por Witten et al., 1994) são aplicados ao vetor de palavras. O *case folding* consiste em deixar todas as letras das palavras na mesma caixa (maiúscula ou minúscula) (por exemplo, a palavra "O" é trocada por "o"); a troca por canônicas simplesmente recupera do léxico do sistema (Nunes et al., 1996) a forma básica das palavras (por exemplo, a palavra "da" é trocada por "do"); a remoção de *stopwords* consiste em ignorar as palavras consideradas irrelevantes, cuja composição foi descrita na Seção 2.1. Na sentença do exemplo, as canônicas das palavras "mais", "da" e "o" serão eliminadas. Sua remoção é realizada em três passos: (a) as palavras iguais são unificadas em uma única posição do vetor – a posição da primeira ocorrência da palavra; (b) a freqüência de cada palavra no vetor é armazenada junto às próprias palavras; (c) a freqüência das *stopwords* é zerada. Todos esses processos, além de facilitar o processamento computacional posterior, aprimoram os resultados da sumarização. Os vetores atualizados por cada processo, a partir do vetor inicial, são mostrados abaixo:

Case folding:

o	tipo	mais	grave	da	doença	o	hemorrágico	pode	matar

Troca por canônicas:

o	tipo	mais	grave	do	doença	o	hemorrágico	poder	matar

Remoção de *stopwords*:

o	tipo	mais	grave	do	doença	hemorrágico	poder	matar
0	1	0	1	0	1	1	1	1

3) Pontuação das sentenças

No GistSumm, a pontuação das sentenças pode ocorrer pelo uso de um dos seguintes métodos: palavras-chave (Black e Johnson, 1988) ou TF-ISF (*Term Frequency-Inverse Sentence Frequency*) (Larocca Neto et al., 2000). O método das palavras-chave segue a proposta de Luhn (1958), partindo do pressuposto de que a idéia principal de um texto pode ser expressa por um conjunto de palavras, chamadas chave. O método TF-ISF[10], por sua vez, determina a importância das sentenças de um texto, para escolher aquela que melhor o represente (a mais importante, no caso).

Em geral, os extratos produzidos por esses métodos são diferentes porque eles ponderam as sentenças de forma diversa. Pelo método das palavras-chave, cada vetor recebe como pontuação a soma do número de ocorrências de cada uma de suas palavras no texto inteiro (ou seja, em todos os vetores). No vetor anterior, há somente 1 palavra, no texto todo, com as seguintes canônicas: 'tipo', 'grave', 'hemorrágico', 'poder' e 'matar'. Há também 4 palavras com a canônica 'doença'. Assim, a pontuação total da sentença [7] é 5*1+ 1*4 = 9 (X*Y: X = número de canônicas; Y = número de palavras que remetem a uma única canônica).

Diferentemente desse cálculo, pelo método TF-ISF a pontuação do vetor corresponde à média da pontuação de cada uma de suas palavras. Sendo w uma palavra, essa pontuação é calculada da seguinte forma:

$$Pontuação\,de\,w = Fw \times \log\left(\frac{nro.\,palavras\,da\,sentença}{nro.\,sentenças\,em\,que\,w\,ocorreu}\right)$$

em que *Fw* é a freqüência da palavra na sentença. Para a sentença do exemplo, a pontuação obtida é de 0,689.

Por qualquer um dos métodos, a *gist sentence* do Texto 1 será a sentença com maior pontuação, como já mencionado. Por isso, no GistSumm os métodos de pontuação das sentenças são, na realidade, métodos de *determinação da idéia principal*. Para o Texto 1, coincidentemente a sentença [3] é escolhida como *gist sentence* por ambos os métodos (neste caso, poderíamos considerar que a manchete M2 seria a mais adequada ou representativa). Essa sentença será sempre incluída no extrato, juntamente com as sentenças selecionadas por critérios de relevância e de taxa de compressão, as quais irão complementar a idéia principal. Esse processo de seleção é regido pelos seguintes passos:

1) calcula-se a média da pontuação das sentenças do texto-fonte e assume-se essa média como sendo um *cutoff*, nota de corte para eliminar sentenças irrelevantes do texto-fonte;
2) identificação das sentenças que contenham pelo menos uma palavra cuja canônica coincida com uma das canônicas da *gist sentence*;
3) dentre essas, seleção das que possuam uma pontuação maior que o *cutoff*.

Além disso, para respeitar a taxa de compressão especificada pelo usuário do sistema, o GistSumm pode eliminar desse conjunto as sentenças com menor pontuação. As Figuras 7 e 8 apresentam os extratos produzidos pelo GistSumm pelo método das palavras-chave e pelo método TF-ISF, respectivamente, com uma taxa de compressão de 60%.

[10] O método TF-ISF é uma variação do método TF-IDF (*Text Frequency-Inverse Document Frequency*) (Salton, 1988) usado na área de Recuperação da Informação.

Mosquitos alterados geneticamente em laboratório podem ajudar a combater a transmissão de doenças como a dengue. A dengue é uma infecção por vírus, transmitida pela picada de mosquitos como o Aedes aegypt. Em estudo publicado na edição de hoje da revista científica "Science", pesquisadores da Universidade Estadual do Colorado (EUA) criaram em laboratório um mosquito cujo organismo não aceita carregar o vírus.

Figura 7: Extrato produzido pelo GistSumm para o Texto 1 utilizando palavras-chave

Em estudo publicado na edição de hoje da revista científica "Science", pesquisadores da Universidade Estadual do Colorado (EUA) criaram em laboratório um mosquito cujo organismo não aceita carregar o vírus. O objetivo dos cientistas agora é fazer com que essa alteração do organismo dos mosquitos seja transmitida hereditariamente. Assim, aumentaria a população de insetos refratários ao vírus.

Figura 8: Extrato produzido pelo GistSumm para o Texto 1 utilizando TF-ISF

Por fazer uso de métodos estatísticos, o GistSumm pode ser aplicado praticamente para textos de qualquer gênero, domínio ou língua ocidental, desde que se personalizem para a língua desejada seus repositórios lingüísticos, ou seja, o léxico e o repositório de *stopwords*. Mais detalhes sobre o sistema podem ser encontrados em (Pardo, 2002a; Pardo et al., 2003).

3.2. O NeuralSumm

O NeuralSumm, sigla para ***NEURAL*** *network for* ***SUMM****arization,* é um sumarizador extrativo que utiliza uma técnica de Aprendizado de Máquina – uma rede neural do tipo SOM (*self-organizing map*) (Braga et al., 2000) – para identificar as sentenças importantes de um texto-fonte. A classificação das sentenças em graus de importância é feita pela rede neural com base em *features* extraídas das sentenças durante o processo de sumarização.

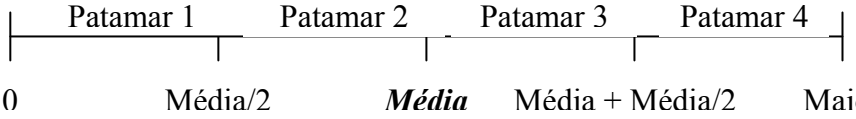
Diferentemente das outras redes neurais, uma rede do tipo SOM organiza as informações aprendidas em *clusters* (grupos) de similaridade. Justamente por isso, diz-se que esse tipo de rede é o que mais se aproxima da forma de funcionamento do cérebro humano. No NeuralSumm, as sentenças de um texto-fonte recebem sua classificação (o grau de importância) de acordo com os *clusters* da rede em que se enquadram. Em linhas gerais, o NeuralSumm extrai as *features* (descritas abaixo) de cada sentença do texto-fonte e as apresenta à rede neural, que as enquadrará em um dos *clusters* de similaridade, resultando na classificação da sentença com o valor associado a esse *cluster*.

É importante dizer que, por trabalhar com uma técnica de Aprendizado de Máquina, o NeuralSumm consiste em um sumarizador altamente experimental, pois se pode aumentar ou diminuir a rede (aumentando ou diminuindo o número de neurônios), alterar a arquitetura da rede, variar o conjunto de *features* utilizado, aumentar ou diminuir o número de *clusters* da rede, aumentar ou diminuir o tamanho do conjunto de treinamento e aumentar ou diminuir sua taxa de aprendizado e sua precisão à vontade para que se atinja a "melhor configuração" da rede, visando obter os melhores extratos.

A rede do NeuralSumm foi configurada para três *clusters*, representando as classes de sentenças *essenciais*, *complementares* e *supérfluas*, segundo as premissas básicas da SA antes delineadas. No estágio atual, ela é resultante de um treino com sentenças de um corpus de 10 textos científicos (introduções de teses e dissertações com aproximadamente 530 palavras e

19 sentenças cada, sobre Computação) em português, chamado *CorpusDT*[11] (Feltrim et al., 2001). Primeiramente, as sentenças desses textos foram classificadas de acordo com sua importância (valores *essencial*, *complementar* ou *supérflua)* por 10 juízes lingüistas computacionais e falantes nativos do português. Para cada uma delas, extraiu-se um conjunto de 8 *features*, associando a ele a classificação indicada pelos juízes. Este procedimento replica vários dos descritos na Seção 2, para a abordagem extrativa. As *features*, assim como suas premissas e valores são reproduzidos abaixo:

1. Tamanho da sentença: sentenças longas normalmente apresentam maior conteúdo informativo, sendo, portanto, relevantes para o texto (Kupiec et al., 1995). No NeuralSumm, as sentenças de um texto são enquadradas em uma escala de 4 patamares. Considerando a variável *Média* como a média do tamanho das sentenças do texto em um intervalo de 0 a *Maior* (esta representando o máximo comprimento de sentenças no texto), essa escala é definida como segue:



2. Posição da sentença: a posição da sentença no texto pode indicar sua relevância (Baxendale, 1958)[12]. Seguindo Kupiec et al. (1995), no NeuralSumm uma sentença pode estar no *início* (primeiro parágrafo), no *fim* (último parágrafo) ou no *meio* (parágrafos restantes) do texto.
3. Posição da sentença no parágrafo a que pertence: a posição da sentença no parágrafo também pode indicar sua relevância (Baxendale, 1958). Da mesma forma que anteriormente, no NeuralSumm uma sentença pode estar no *início* (primeira sentença), no *fim* (última sentença) ou no *meio* (posições restantes) do parágrafo.
4. Presença de palavras-chave na sentença: as palavras-chave são comumente utilizadas para expressar a idéia principal do texto, tendendo a se repetir no decorrer do texto (Luhn, 1958). Assim, uma sentença pode conter (*True*) ou não (*False*) palavras-chave do texto. No NeuralSumm, elas são as palavras significativas (de classe aberta) de mais alta freqüência.
5. Presença de palavras da *gist sentence* na sentença: sentenças que possuem palavras da *gist sentence* tendem a ser mais relevantes do que outras, pois fazem alusão explícita à idéia principal do texto (Pardo, 2002b). Uma sentença pode conter (*True*) ou não (*False*) palavras da *gist sentence*.
6. Pontuação da sentença com base na distribuição das palavras do texto: sentenças com alta pontuação normalmente são relevantes para o texto (Black e Johnson, 1988). A pontuação de uma sentença, neste trabalho, é resultante da distribuição de suas palavras, isto é, da divisão da soma das freqüências de suas palavras por seu comprimento (número de palavras)[13]. Essa pontuação também é enquadrada em uma escala similar à da *feature* 1, cujos patamares (*1*, *2*, *3* ou *4*) indicam a representatividade da sentença.
7. TF-ISF da sentença: sentenças com alto valor de TF-ISF (*Term Frequency-Inverse Sentence Frequency*) são sentenças singulares de um texto e, assim, podem representá-lo bem (Larocca Neto et al., 2000). O valor TF-ISF de cada sentença também é enquadrado em uma escala similar à da *feature* 1, nos patamares *1*, *2*, *3* ou *4*.

[11] Descrição disponível em http://www.nilc.icmc.usp.br/nilc/tools/corpora.htm (último acesso em maio/2003).
[12] Também confirmada por Aretoulaki (1996).
[13] Vale notar que, neste sistema, essa pontuação é distinta daquela do GistSumm.

8. Presença de palavras indicativas na sentença: palavras sinalizadoras (*cue words*) normalmente indicam a importância do conteúdo das sentenças (Edmundson, 1969; Paice, 1981). Uma sentença pode conter (*True*) ou não (*False*) palavras sinalizadoras. Essa *feature* é a única dependente de língua, gênero e domínio textuais. No NeuralSumm, atualmente customizado para textos científicos em português, as palavras sinalizadoras consideradas são *avaliação*, *conclusão*, *método*, *objetivo*, *problema*, *propósito*, *resultado*, *situação* e *solução*.

Após o treinamento da rede para esses conjuntos de *features* extraídos das sentenças do corpus de treino, os *clusters* delineados são usados para produzir extratos de textos-fonte, segundo os seguintes passos:

1. segmentação do texto-fonte em sentenças;
2. tratamento dos segmentos (remoção de *stopwords*; troca por canônicas; *case folding*);
3. extração das *features* de cada sentenças (a partir da representação interna produzida pelos passos anteriores);
4. classificação do conjunto de *features* de cada sentença segundo os *clusters* relativos aos valores *essencial*, *complementar* ou *supérfluo*;
5. seleção das sentenças com maior classificação e produção do extrato.

A seleção de sentenças para a produção do extrato (passo 5) acontece da seguinte forma:

- são selecionadas somente sentenças classificadas como *essenciais* e *complementares*;
- caso todas as sentenças do texto-fonte sejam classificadas como *supérfluas*, elas são ranqueadas pela pontuação obtida por sua distribuição de palavras (*feature* 6), selecionando-se, então, aquelas com pontuação mais alta.

Ambos os casos são ainda condicionados à taxa de compressão especificada pelo usuário do sistema. No primeiro caso, quando a taxa de compressão restringe o número de sentenças selecionadas, as sentenças *essenciais* têm prioridade sobre as *complementares*, sendo que sempre têm prioridade as de maior pontuação (pela *feature* 6).

As *features* extraídas da sentença [6] do Texto 1 ("A dengue provoca náuseas e dores de cabeça, articulações e músculos"), após a classificação de suas sentenças, são as seguintes:

1. tamanho da sentença: patamar 2 (11 palavras)
2. posição da sentença no texto: fim
3. posição da sentença no parágrafo a que pertence: meio
4. presença de palavras-chave: false
5. presença de palavras da *gist sentence*: true
6. pontuação da sentença com base na distribuição de palavras: patamar 2 (pontuação=1,428)
7. TF-ISF da sentença: patamar 3 (TF-ISF=0,731)
8. presença de palavras indicativas: false

Ao ser apresentado à rede neural, esse conjunto de *features* é enquadrado no *cluster* das sentenças supérfluas, determinando a classificação da sentença [6] como supérflua. O extrato correspondente, com taxa de compressão de 60%, é mostrado na Figura 9. Como podemos notar, a sentença [6] não foi incluída no extrato, já que é supérflua.

O objetivo dos cientistas agora é fazer com que essa alteração do organismo dos mosquitos seja transmitida hereditariamente. O tipo mais grave da doença, o hemorrágico, pode matar. Cerca de 1250 municípios brasileiros, aproximadamente um em cada quatro, registraram casos de dengue.

Figura 9: Extrato produzido pelo NeuralSumm para o Texto 1

Podemos notar, ainda, que esse extrato é ruim, pois não preserva a idéia principal do texto (conforme exposto na Seção 3.1, esta é identificada pela sentença [3]). Esse extrato sequer menciona a palavra "dengue", crucial nesse texto, conforme evidenciam as manchetes de exemplo (Seção 1). Atribuímos a esse desempenho ruim o fato de o NeuralSumm ter sido treinado com um corpus de textos científicos da Computação. Logo, ele não é adequado para sumarizar o Texto 1, de gênero jornalístico e domínio muito distinto do domínio de Computação.

Esse exemplo mostra que, apesar de o NeuralSumm incorporar um método de SA genérico o suficiente para ser aplicado a qualquer texto de qualquer gênero e domínio, é preciso treiná-lo com corpora específicos a cada alteração de gênero ou domínio. Similarmente ao GistSumm, ele também pode ser aplicado para qualquer língua ocidental, bastando que se personalizem seus repositórios lingüísticos, que incluem agora o léxico, o repositório de *stopwords* e o repositório de palavras indicativas.

3.3. O DMSumm

DMSumm é a sigla para ***D****iscourse* ***M****odeling* ***SUMM****arizer*, um gerador automático de sumários (Pardo, 2002b; 2002c) baseado em modelagem discursiva (Rino, 1996). Embora vise à SA de textos, ele não tem como entrada o próprio texto, mas sim o suposto resultado de sua interpretação. Essa delimitação foi adotada devido à inexistência de um interpretador adequado para a modelagem do discurso considerada, à complexidade de se construir um e ao interesse mais imediato de se explorar as questões peculiares da SA. Assim, no momento, a entrada para o DMSumm é produzida manualmente.

A mensagem (ou representação conceitual, cf. nomenclatura da Figura 4) de um texto-fonte é composta por três componentes: uma base de conhecimento, uma proposição central e um objetivo comunicativo. A base de conhecimento é uma estrutura semântica que contém o conhecimento expresso no texto-fonte; como comentado na Seção 1, a proposição central é a informação mais importante do texto, aquela que se quer comunicar, e o objetivo comunicativo é o objetivo que se quer atingir ao comunicar o conteúdo do texto-fonte. Com essa caracterização, o DMSumm observa as três premissas fundamentais da SA descritas na Seção 1: todo texto tem um objetivo comunicativo e uma proposição central, devendo esta ser preservada na sumarização. A hipótese principal do DMSumm está na garantia de coerência dos sumários pela interação de conhecimento de diferentes naturezas – semântica, intencional e retórica – presentes na modelagem discursiva utilizada.

A base de conhecimento é uma árvore binária cujos nós internos são rotulados por relações semânticas (representadas em itálico na Figura 10), baseadas nas relações clausais de Jordan (1992), e cujas folhas são proposições correspondentes às unidades informativas do texto-fonte. Além das relações semânticas, também é registrado o papel funcional dos segmentos no texto-fonte. Para o Texto 1, por exemplo, cuja base de conhecimento é ilustrada na Figura 10, a relação *rationale* entre os segmentos 3 e 4 indica que, com a realização de sol(3), tem-se o objetivo de realizar prop(4). Os papéis funcionais dos segmentos são expressos de duas maneiras: como etiquetas associadas a cada segmento (sit=situação, probl=problema, sol=solução, res=resultado e prop=proposição genérica) e como blocos

semânticos (Situação, Problema, Solução e Resultados) indicando o papel que um conjunto de segmentos desempenha no texto. Assim, podemos dizer que as primeiras etiquetas contemplam o nível micro-estrutural, informativo, do texto-fonte, enquanto que as últimas contemplam seu nível macro-estrutural[14]. sol(3), por exemplo, indica que o segmento 3 é uma solução para algum problema apontado no texto (que, no caso, é probl(2)) e que os segmentos 2, 6, 7, 8, 9 e 10 fazem parte da descrição do problema apresentado no texto. Essa base de conhecimento expressa todo o conteúdo disponível para a produção de um sumário, no DMSumm.

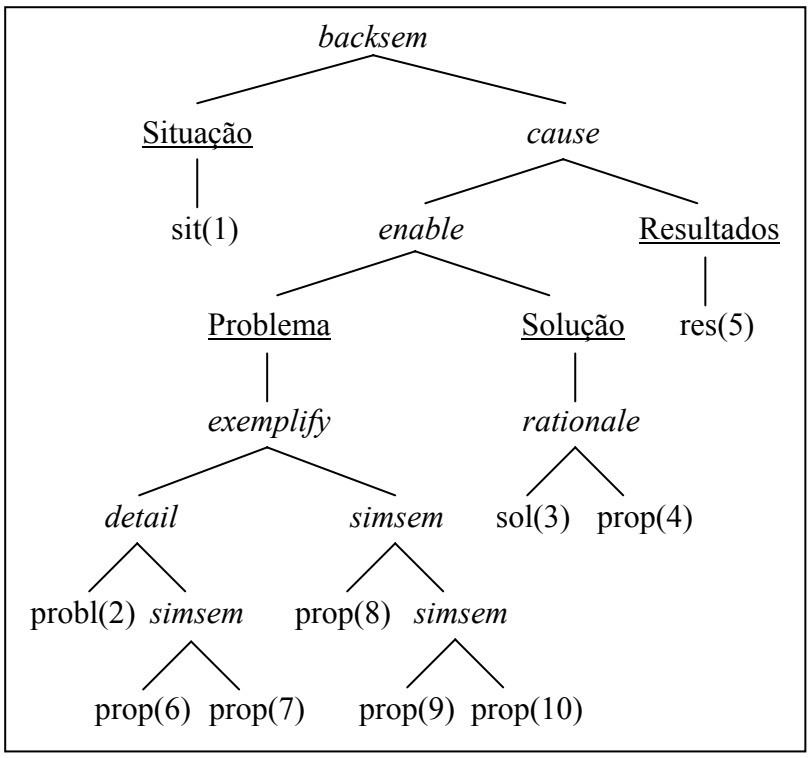


Figura 10: Base de conhecimento do Texto 1

Um possível objetivo comunicativo para o Texto 1 é *relatar a solução encontrada para o problema*. Desse objetivo, é possível, portanto, derivar a proposição central do possível sumário: a solução indicada pelo segmento 3. Ele também é responsável por delimitar a estratégia comunicativa para a construção do sumário. A proposição central, por sua vez, será a informação principal que restringirá a escolha de outros segmentos do texto-fonte, os quais deverão complementá-la, visando maior informatividade ou coerência do sumário. No DMSumm, consideram-se somente os objetivos comunicativos *descrever*, *relatar* e *discutir*.

Tendo os três componentes da mensagem do texto-fonte disponíveis como entrada (a base de conhecimento, o objetivo comunicativo e a proposição central), o DMSumm pode aplicar sua estratégia fundamental, de transformação e síntese dos possíveis sumários. São três os processos correspondentes a essas etapas de SA: a seleção de conteúdo, o planejamento textual e a realização lingüística. O processo de seleção de conteúdo simplesmente elimina da base de conhecimento suas proposições supérfluas, identificadas pela falta de relação expressiva com o objetivo comunicativo e a proposição central. São

[14] Muito embora alguns autores associem a esse nível o conhecimento retórico do texto (como o fazem Teufel e Moens, 1999), nós seguimos sobretudo a linha de Winter (1977; 1979), em que as funções dos segmentos textuais não expressam, necessariamente, qualquer omposição de objetivos retóricos ou conotação.

usadas heurísticas para essa redução de conteúdo (Rino e Scott, 1994). Por exemplo, ao excluir exemplos e detalhes da base de conhecimento do Texto 1 (relações semânticas *exemplify* e *detail*), a base de conhecimento reduzida resulta na ilustrada na Figura 11, cujas proposições devem, agora, ser reproduzidas no sumário final.

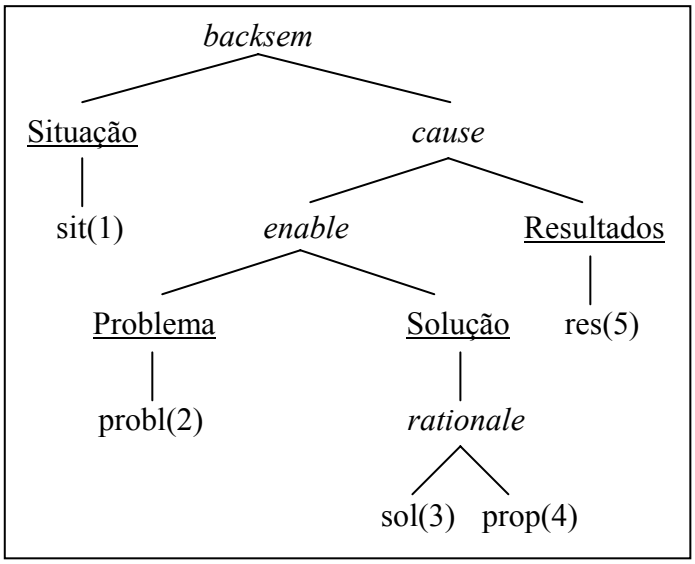


Figura 11: Base de conhecimento reduzida do Texto 1

O processo de planejamento textual organiza o conteúdo informativo restante, produzindo *planos de texto*, i.e., estruturas retóricas, por meio de um modelo de discurso (Rino, 1996) multi-nível: ele fundamenta a escolha de relações retóricas (Mann e Thompson, 1988) pelo mapeamento das relações semânticas indicadas na base de conhecimento e das relações intencionais (Grosz e Sidner, 1986) delineadas pelo objetivo comunicativo. As relações intencionais irão determinar a força retórica das unidades informativas, enquanto as semânticas irão delinear a forma como elas serão relacionados na estrutura final, isto é, se serão componentes de um núcleo ou de um satélite da estrutura retórica do sumário. Esse mapeamento é modelado computacionalmente por *operadores de plano* (Moore e Paris, 1993), artifícios que permitem identificar restrições e buscar a satisfação de condições para a determinação da estrutura e do conteúdo textual, garantindo a construção do plano de texto. Em seu estágio atual, o DMSumm incorpora 89 operadores de plano, responsáveis por gerar todos os mapeamentos possíveis entre as relações do modelo de discurso implementado. A Figura 12 apresenta um plano de texto para a mensagem antes ilustrada (gerado pelo DMSumm) de relatar o problema relacionado ao conteúdo da base de conhecimento (Figura 10) tendo como proposição central sua solução (*sol*(*3*)).

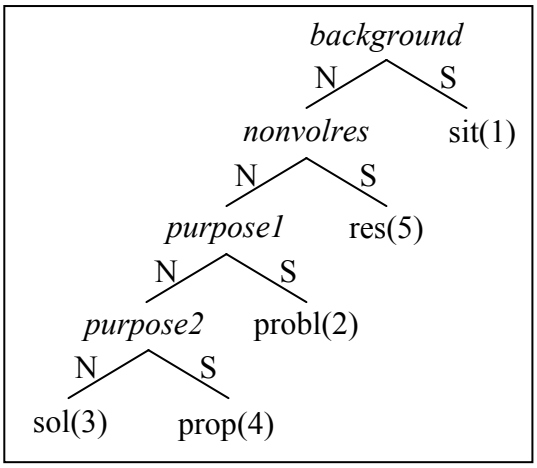


Figura 12: Plano de texto de um sumário do Texto 1

O último processo, de realização lingüística, é responsável por expressar em língua natural os planos de texto, produzindo, enfim, os sumários. Este processo é ainda simplificado no DMSumm, pois envolve somente o uso de *templates* para as escolhas de vocabulário e sintaxe, definidos por um conjunto de regras fixas de expressão das relações retóricas entre as unidades informativas. Quando necessário, são incluídos marcadores discursivos e sinais de pontuação para garantir a formação adequada do sumário. O sumário apresentado na Figura 13 é um exemplo de realização lingüística do plano de texto da Figura 12.

Sumário 1: Mosquitos alterados geneticamente em laboratório podem ajudar a combater a transmissão de doenças como a dengue. A dengue é uma infecção por vírus, transmitida pela picada de mosquitos como o Aedes aegypt. Em face disso, em estudo publicado na edição de hoje da revista científica "Science", pesquisadores da Universidade Estadual do Colorado (EUA) criaram em laboratório um mosquito cujo organismo não aceita carregar o vírus. O objetivo dos cientistas agora é fazer com que essa alteração do organismo dos mosquitos seja transmitida hereditariamente. Como resultado, aumentaria a população de insetos refratários ao vírus.

Figura 13: Sumário produzido pelo DMSumm para o Texto 1

De forma similar à ilustrada, é possível gerar vários sumários para o Texto 1, variando-se o objetivo comunicativo e/ou a proposição central (premissas apresentadas na Seção 1).

Por ser baseado em um modelo de discurso independente de língua, o DMSumm pode ser aplicado praticamente a qualquer língua, bastando que se personalize seu módulo de realização lingüística. É importante notar, também, que os textos a serem sumarizados pelo DMSumm devem apresentar uma estrutura a partir da qual possam ser derivados os componentes e as relações pertinentes ao modelo de representação da base de conhecimento, como os ilustrados na Figura 10. Ou seja, devem apresentar um problema e sua solução, assim como os resultados obtidos, etc. Esse tipo de estruturação segue, na realidade, o modelo Problema-Solução (Winter, 1976; 1977; Jordan, 1980, 1984; Hoey, 1983), que é bastante discutido na literatura e freqüentemente encontrado em textos de diversos gêneros e domínios, escritos em quaisquer línguas naturais.

Para mais detalhes sobre o DMSumm, incluindo a especificação de suas relações semânticas, intencionais e retóricas e a definição dos operadores de plano e *templates* utilizados, vide (Pardo, 2002b; 2002c) e (Pardo e Rino, 2002).

3.4. O UNLSumm

Diferentemente do DMSumm, o UNLSumm, ou ***UNL SUMM****arizer* (Martins, 2002), é um sumarizador sentencial implementado em plataforma Windows que, embora baseado em metodologia fundamental, não faz uso profundo de características discursivas ou retóricas, mas sim das características conceituais da Linguagem UNL (Uchida, 2000), a interlíngua adotada no Projeto UNL-Brasil, de tradução multilingual.

Sua entrada é uma estrutura UNL, representação conceitual, na Linguagem UNL, de cada sentença de um texto-fonte, o qual, em princípio, pode ser escrito em qualquer língua natural. Portanto, a fase de análise, nesse sistema, também é considerada independente do UNLSumm e, por essa razão, as representações conceituais são construídas ainda manualmente. A redução de uma estrutura UNL é fundamentada, assim, no inter-relacionamento semântico entre unidades de informação, formalizado pela Linguagem UNL. Supõe-se que a estrutura UNL reduzida poderá ser sintetizada com o auxílio do DeCo (Uchida, 1997), ferramenta de decodificação de UNL para qualquer língua natural desejada. Particularmente no contexto do UNLSumm, contemplamos somente o português, já que é para essa língua que o DeCo foi personalizado no Projeto UNL-Brasil. Assim, a única etapa da SA fundamental que o UNLSumm realmente contempla, no momento, é a de redução da representação conceitual original, para a produção de uma representação conceitual do sumário a gerar. Esta última estrutura, ainda em linguagem UNL, é chamada de *sumário UNL*.

A linguagem UNL é utilizada, assim, como linguagem de representação do conhecimento e fonte para o modelo de SA: o UNLSumm é baseado em conjuntos de heurísticas que, pela identificação de relações UNL, indicam os componentes menos relevantes da estrutura UNL, que serão, portanto, excluídos do sumário UNL. Diferentemente das demais propostas já apresentadas neste capítulo, este sistema concentra suas decisões em mecanismos de *exclusão*: são identificadas e extraídas as informações irrelevantes da estrutura UNL de entrada.

Para a especificação das heurísticas de identificação dos componentes supérfluos, a modelagem profunda consistiu da identificação das principais características conceituais de textos a sumarizar: foram considerados vários corpora de textos em português, para cujas sentenças foram produzidas manualmente suas estruturas UNL. Comparando a forma superficial com a conceitual, identificaram-se as correspondências lingüístico-conceituais entre construções superficiais do português e os componentes da Linguagem UNL, sobretudo aqueles remetendo aos rótulos de relação, ou *Relation Labels* (RLs), pois são estes que indicam o relacionamento conceitual entre diferentes componentes sentenciais. Estes, por sua vez, são representados em UNL pelas *Universal Words* (UWs).

O dado de entrada do UNLSumm, i.e., uma estrutura UNL, é expresso por um conjunto de relações binárias, semânticas, entre os componentes sentenciais, cujo formato é *RL(UW1,UW2)* – RL é uma relação conceitual entre dois conceitos distintos, sendo que estes podem ser simples ou compostos. Os RLs são expressos por uma cadeia de três caracteres. Por exemplo, *agt* é um RL que indica o agente de uma ação, como em "João quebra a janela da sala.", cuja relação binária é *agt(break, Joao)*.

As premissas básicas do sumarizador sentencial são as seguintes:

1. Sentenças em *pipeline* podem ser consideradas, quando conjugadas, um texto completo.
2. É possível produzir, para um texto-fonte, um sumário com mesmo número de sentenças, porém, com estruturas correspondentes condensadas, também considerando a produção de estruturas superficiais em *pipeline*, sentença por sentença. Assim, sucessivas sumarizações intra-sentenciais resultarão em representações UNL bem formadas.

3. Ao decodificar uma estrutura UNL em textos em certa língua natural, o processamento seqüencial de suas sentenças UNL ainda garantirá textos bem formados.

Por tratar somente da SA intra-sentencial, o UNLSumm impede que sejam usadas taxas de compressão variadas para a produção dos sumários. Embora não haja um consenso sobre o tamanho ideal de um sumário, normalmente considera-se que bons sumários mantenham de 5 a 30% do conteúdo do texto-fonte (Mani, 2001), ou seja, suas taxas de compressão variam de 70% a 95%. Para o UNLSumm, essas taxas são muito altas, devido ao fato de não se excluirem sentenças quaisquer da estrutura de entrada. Por essa razão, são considerados úteis também os sumários com baixa taxa de compressão.

Um exemplo do UNLSumm em operação é ilustrado na SA da sentença [3] do Texto 1 – "Em estudo publicado na edição de hoje da revista científica Science, pesquisadores da Universidade Estadual do Colorado (EUA) criaram em laboratório um mosquito cujo organismo não aceita carregar o vírus.". Sua estrutura UNL completa indica várias relações binárias, como ilustra a Figura 14.

```
[S]
obj(published,study.@indef)                          [1]
plc(published,edition.@def)                          [2]
tim(edition.@def,today)                              [3]
obj(edition.@def,magazine.@def)                      [4]
nam(magazine.@def,Science)                           [5]
mod(magazine.@def,scientific)                        [6]
plc(researcher.@pl,study.@indef)                     [7]
src(researcher.@pl,State University of Colorado)     [8]
plc(State University of Colorado,USA.@parenthesis)   [9]
agt(researcher.@pl,create.@entry.@past)              [10]
obj(create.@entry.@past,mosquito.@indef)             [11]
pos(organism,mosquito.@indef)                        [12]
aoj(accept.@not,organism)                            [13]
obj(accept.@not,carry)                               [14]
obj(carry,virus.@def)                                [15]
[/S]
```

Figura 14: Estrutura UNL da sentença [3]

O UNLSumm utiliza dois grupos de heurísticas para identificar as relações binárias indicativas de informações irrelevantes, em um total de 84 heurísticas. As heurísticas do primeiro grupo (Grupo A) identificam relações binárias que possam ser individualmente removidas da sentença original, enquanto as do segundo grupo (Grupo B) identificam grupos de relações binárias que devem ser removidas simultaneamente, não só por sua irrelevância, mas, principalmente, para garantir a coerência e coesão da sentença sumarizada. Como já mencionamos, as heurísticas se baseiam nos RLs. Por exemplo, para a sentença [3], são aplicadas duas heurísticas do grupo B, gerando o sumário UNL da Figura 15. Se este fosse decodificado novamente para o português, equivaleria à sentença "Em estudo publicado, pesquisadores criaram um mosquito cujo organismo não aceita carregar o vírus.".

```
[S]
obj(published,study.@indef)                    [1]
plc(researcher.@pl,study.@indef)               [7]
agt(researcher.@pl,create.@entry.@past)        [10]
obj(create.@entry.@past,mosquito.@indef)       [11]
pos(organism,mosquito.@indef)                  [12]
aoj(accept.@not,organism)                      [13]
obj(accept.@not,carry)                         [14]
obj(carry,virus.@def)                          [15]
[/S]
```

Figura 15: Estrutura UNL de um sumário da sentença [3]

As duas heurísticas aplicadas, nesse caso, são as seguintes:

Heurística HB.3.5

Excluir plc(a,b) + {RBs $\in$ subgrupo S1} se UWs $\in$ S1 $\notin$ RBs fora do subgrupo.

Heurística HB.7.1

Excluir src(a,b) + {RBs $\in$ subgrupo S1} se UWs $\in$ S1 $\notin$ RBs fora do subgrupo.

A heurística HB.3.5 parte do princípio de que a informação sobre o local onde uma ação ocorre não é essencial e, portanto, pode ser omitida em um sumário. Como em UNL esse tipo de informação é representada por meio do RL plc (*place*, ou lugar), então a relação binária que contém esse RL deve ser excluída. Assim, a relação binária [2], correspondendo à informação "a edição", deve ser excluída da estrutura UNL. Sua exclusão caracteriza o subgrupo composto pelas relações binárias [2], [3], [4], [5] e [6], as quais envolvem as UWs *edition* e *magazine* (mais refinadas na estrutura ilustrada): essas UWs não aparecem no restante da estrutura UNL. Desse modo, a aplicação integral da heurística implica excluir também tais relações.

Já a heurística HB.7.1 parte do princípio de que a informação sobre a origem de um objeto conceitual (no exemplo, "Universidade Estadual do Colorado") não é essencial e, portanto, também deve ser omitida. Como em UNL esse tipo de informação é representada por meio do RL src (*source*, ou fonte), essa heurística exclui a relação binária envolvendo esse RL, [8], assim como aquela caracterizada no mesmo subgrupo – a de número [9].

4. A avaliação de sumários produzidos automaticamente

Esta seção apresenta uma visão geral sobre a questão da avaliação de sistemas de SA. A subseção 4.1 introduz o tema, mostrando sua necessidade e as dificuldades associadas. A subeção 4.2 relata definições e princípios gerais de avaliação adotados em pesquisas recentes, enquanto as subseções 4.3 e 4.4 descrevem métricas e técnicas de avaliação comumente utilizadas. A subseção 4.5 mostra um estudo de caso, descrevendo a avaliação do GistSumm.

4.1. A necessidade e as dificuldades da avaliação

A avaliação de sistemas de PLN, em especial de sistemas de SA, foi bastante negligenciada no passado, mas muito importante ultimamente. É por meio da avaliação que se torna possível verificar o avanço do "estado da arte" em AS. É possível medir-se o grau de utilidade de um sistema de SA, sua adequação a determinadas tarefas, a validade de sua metodologia, etc. Em SA, pode-se dizer que a avaliação é a responsável por direcionar as pesquisas, pois ela pode indicar meios de validação e mesmo de desconsideração de orientações antes delineadas.

Esse tema é muito amplo e abrangente. Quando se fala em avaliação de um sistema, pode-se ter em mente várias facetas: (a) pode-se avaliar o desempenho computacional do sistema, isto é, o uso que ele faz de memória, seu tempo de execução, a complexidade de seu algoritmo principal, etc.; (b) pode-se considerar a usabilidade do sistema, ou seja, a crítica da clareza de sua interface ou o grau de intuição (dos possíveis usuários) necessário para seu uso ou, ainda, sua consistência e sua flexibilidade para possíveis customizações; (c) pode-se avaliar se os resultados produzidos automaticamente são satisfatórios, isto é, se são os resultados esperados, se são "corretos" ou "adequados". Na SA, contempla-se, em geral, somente esta última forma de avaliação, pois não há métodos de SA suficientemente robustos que justifiquem a análise de questões de desempenho ou interface.

Na literatura básica de SA, há muitas referências sobre métricas e métodos de avaliação. Porém, ainda não há consenso sobre a melhor forma de se avaliar um sistema dessa natureza. Há, ainda, diversos desafios nessa área, dentre os quais destacam-se (Mani e Maybury, 1999):

- A identificação do que seria um resultado "correto" para um sumarizador automático, já que, para um único texto-fonte, pode haver uma infinidade de sumários sugerindo diversas perspectivas, em função de todos os possíveis usuários e das tarefas a que podem se destinar. Por exemplo, para um leitor leigo no assunto de um certo texto-fonte, toda informação contextual pode ser importante, enquanto que, para um leitor especialista, somente a informação nova poderia lhe interessar e, assim, ser incluída no sumário. Para um leitor decidir se lerá ou não o texto-fonte correspondente a partir da leitura de um sumário, este deve ser suficientemente informativo. Entretanto, se ele servir somente para indexar documentos, ele pode ser simplesmente uma lista de palavras. Mesmo para leitores de mesmo perfil e com tarefas comuns, um sumário pode ser julgado adequado por alguns e inadequado por outros.
- A identificação de uma taxa de compressão ideal para avaliar adequadamente os sumários automáticos. Em geral, quanto mais alta a taxa de compressão, menos informativo será o sumário e vice-versa. Entretanto, essa relação de dependência não pode ser explicitamente associada a um modelo fixo, pois a informatividade depende do nível de conhecimento do usuário ao qual o sumário se destina, do tempo que ele dispõe para lê-lo e da tarefa especificada, dentre outras coisas.
- A forma como a qualidade e a informatividade de sumários automáticos podem ser avaliadas automaticamente. Nesses casos, costuma-se fazer uso do julgamento humano: leitores falantes da língua natural considerada devem dizer se os sumários automáticos são bons sumários, quando comparados a seus correspondentes textos-fonte. Essa tática torna o processo de avaliação bastante custoso, tornando preferível uma avaliação automática. Entretanto, não há nada até o momento que seja capaz de substituir de forma satisfatória o juízo humano na avaliação.
- A identificação da situação e da forma de se utilizar o julgamento humano. Apesar de a resposta para essa questão parecer simples, inferindo-se que o julgamento humano deve ser utilizado em todas as fases possíveis de uma avaliação, há o problema de se identificar o perfil adequado do juiz humano, além do custo envolvido, já mencionado. A avaliação se torna custosa, pois: (a) é difícil dispor de juízes humanos (e, quando necessário, especialistas em técnicas de sumarização) em número suficiente para a avaliação; (b) para uma avaliação robusta e abrangente, esse tipo de julgamento se torna lento e complexo; (c) há um alto grau de subjetividade no julgamento humano, sendo difícil tirar conclusões definitivas com esse tipo de avaliação.

O surgimento de conferências internacionais dedicadas somente à avaliação de sumarizadores automáticos, como a SUMMAC[15] (*Text Summarization Evaluation Conference*) (Mani et al., 1998) e a DUC[16] (*Document Understanding Conference*), evidencia a importância e necessidade da avaliação e das dificuldades inerentes à AS. A SUMMAC foi realizada em 1998 e financiada pelo governo dos EUA, sendo a primeira avaliação independente, em larga escala, de sistemas de SA. Seu principal objetivo foi tentar estabelecer padrões de avaliação e entender melhor as questões envolvidas no processo integral de construção e avaliação de sistemas de SA. A DUC, por sua vez, nasceu de um programa chamado TIDES (*Translingual Information Detection, Extraction, and Summarization*) financiado pela DARPA (*Defense Advanced Research Projects Agency*), uma agência do Departamento de Defesa dos EUA, com os mesmos objetivos da SUMMAC. Entretanto, ela tem sido realizada periodicamente e é considerada a iniciativa atual mais importante de avaliação de sistemas de SA.

Apesar da problemática envolvida na avaliação de sistemas de SA, alguns princípios gerais e definições comuns têm sido adotados, conforme ilustra a próxima seção.

4.2. Avaliação: definições e princípios gerais

Sparck Jones e Galliers (1996) foram os primeiros autores a esclarecer as possíveis diretrizes gerais para a avaliação de sistemas de PLN, as quais têm sido amplamente adotadas.

A primeira grande distinção que se faz diz respeito à forma de avaliação: ela pode ser *intrínseca* ou *extrínseca*. Uma avaliação intrínseca avalia o próprio desempenho do sistema, pela verificação da qualidade e informatividade dos sumários produzidos. São usadas métricas calculadas automaticamente ou julgamentos subjetivos, realizados por leitores humanos. A avaliação extrínseca verifica a adequação do sistema ao seu uso em tarefas específicas, distintas da SA. Por essa razão, ela é comumente chamada de *validação*. Tarefas em que a validação de sumarizadores automáticos tem se aplicado envolvem as de perguntas e respostas, de categorização de documentos e de recuperação de informação.

Quando a avaliação faz uso do julgamento humano, diz-se que ela é uma avaliação *on-line*. Caso contrário, ela é chamada de *off-line*. Dada a complexidade de se projetar uma avaliação por seres humanos, a avaliação *off-line* é, normalmente, preferível. Entretanto, métodos automáticos de avaliação que sejam tão satisfatórios quanto o julgamento humano são ainda inexistentes.

Pode-se classificar a avaliação de acordo com o que se avalia: se forem avaliados somente os resultados finais do sistema, a avaliação é chamada *avaliação black-box*. Neste caso, o sistema é visto como uma "caixa-preta", à qual não se tem acesso. Ou seja, não se avaliam os processos intermediários da sumarização. Exemplo típico desse tipo de avaliação é a comparação entre um sumário produzido automaticamente e seu correspondente texto-fonte, para verificar se ele é bom. Se forem avaliados resultados intermediários, isto é, aqueles resultantes da execução de cada processo intermediário do sistema, a avaliação é chamada *avaliação glass-box*. Caso um sistema de SA siga a arquitetura padrão composta pelos processos de análise, transformação e síntese, uma avaliação glass-box verificaria os resultados de cada um desses processos.

Uma última distinção se faz em relação à forma de comparação entre vários sistemas. Se os resultados de um sistema de SA são comparados com os resultados de outro sistema, diz-se que a avaliação é *comparativa*; caso contrário, diz-se que ela é *autônoma*. A avaliação comparativa é, normalmente, o foco das grandes conferências internacionais (SUMMAC e

[15] http://www.itl.nist.gov/iaui/894.02/related_projects/tipster_summac/ (último acesso em maio/2003).
[16] http://duc.nist.gov/ (último acesso em maio/2003).

DUC inclusas): os sistemas participantes do concurso são pontuados pelo seu desempenho e, então, são comparados pelos seus pesos.

Sparck Jones e Galliers afirmam que o mais importante na avaliação de um sistema de SA é estabelecer claramente o que se quer avaliar. Tendo isso como meta, é fácil determinar quais dos tipos anteriores de avaliação aplicar, isto é, se ela será intrínseca ou extrínseca, on-line ou off-line, black-box ou glass-box e comparativa ou autônoma. É importante esclarecer, entretanto, que estes tipos de avaliação não são exclusivos. Por exemplo, caso se queira proceder a uma avaliação intrínseca e a uma extrínseca, isso é totalmente possível e viável, dependendo somente dos objetivos ao se realizar a avaliação.

A próxima seção apresenta os métodos e métricas usuais para a avaliação intrínseca de sistemas de SA, seguindo-se aqueles da avaliação extrínseca.

4.3. Avaliação intrínseca

Em geral, a avaliação intrínseca pode envolver as medidas de *qualidade* e *informatividade* dos sumários produzidos automaticamente (Mani, 2001).

4.3.1. Qualidade dos sumários automáticos

Nos termos de Mani, medir a qualidade de sumários é verificar a sua *fluência*, ou seja, a facilidade em sua leitura e a sua clareza. Assim, é necessário o julgamento humano e, por essa razão, esse tipo de avaliação é geralmente caracterizado como avaliação *on-line.*

Os critérios para se julgar a qualidade dos sumários variam muito. Minel et al. (1997), 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 (que é um vício de linguagem que consiste em repetir o mesmo pensamento com palavras diferentes), etc. Saggion e Lapalme (2000), utilizando os critérios sugeridos por Rowley (1982), também pediram a juízes que dessem notas aos sumários, observando agora a ortografia e gramática, a indicação clara do tópico do texto-fonte, o estilo impessoal, a concisão, legibilidade e facilidade de compreensão do sumário, a presença de siglas seguidas de suas expansões, etc. Pardo e Rino (2002), seguindo os critérios de avaliação de White et al. (2000), também pediram a juízes que dessem notas a sumários de acordo com sua *textualidade*, isto é, sua coerência e coesão (Rino, 1996). Além disso, usaram também outra sugestão de Mani, de avaliação da legibilidade dos sumários, ante a legibilidade dos textos-fonte correspondentes. Para verificar se os sumários preservavam a legibilidade dos textos-fonte, foi utilizado o índice de legibilidade de Flesch (1948) adaptado ao português (Martins et al., 1996) e calculado automaticamente (*off-line*). Esse índice baseia-se no número médio de sílabas por palavra e de palavras por sentenças, para expressar o grau de dificuldade de leitura de um texto. É importante ressaltar que a legibilidade não é um critério decisivo, nem suficiente, para se afirmar que um sumário é bom. De fato, como discutido por Mani, essa medida é muito rústica, dada sua "ingenuidade" em assumir que o tamanho de palavras e de sentenças é o único fator que pode influenciar a legibilidade de um texto.

Como a avaliação da qualidade de sumários necessita de juízes humanos, tem-se procurado formas automáticas de realizar tal avaliação. A verificação da legibilidade dos sumários em relação aos textos-fonte é um exemplo de automação factível. Outras opções incluem o uso de corretores ortográficos, gramaticais ou estilísticos automáticos.

A qualidade é um bom parâmetro para se centralizar avaliações de sistemas extrativos, pois estes produzem, em geral, textos com fluência ruim. Entretanto, mesmo com fluência prejudicada, ainda é possível obter sumários úteis. Devido à complexidade de modelagem de

tantas nuances distintas sobre a qualidade e/ou utilidade de sumários automáticos, tornou-se comum avaliá-los somente pela verificação do conteúdo que eles preservam, em relação a seus textos-fonte.

4.3.2. Informatividade dos sumários automáticos

A informatividade de um sumário expressa o quanto, do conteúdo informativo original, ele contém. Na maioria das aplicações, senão todas, esse é um quesito essencial para julgar a qualidade de um sumário.

A informatividade está diretamente ligada à taxa de compressão: é comum considerar-se que, quanto maior a compressão, menos informativo será o sumário e vice-versa, muito embora isto nem sempre seja verdadeiro, principalmente se levarmos em conta a competência de domínio do leitor, a qual poderá modificar significativamente os critérios de informatividade. Assim, considerando o caso de leitores de proficiência média, pode-se dizer que a redução expressiva do conteúdo do texto-fonte pode prejudicar sensivelmente a informatividade do sumário. Desta forma, torna-se necessário determinar a taxa de compressão adequada, para não haver prejuízos consideráveis sobre a informatividade dos sumários. Em geral, essa taxa pode ser inferida por observações: analisando tarefa(s) que usuário(s) executam com o uso de sumários, é possível caracterizá-los mais apropriadamente.

Para se verificar a informatividade de um sumário, além de comparar seu conteúdo com o conteúdo de seu texto-fonte, pode-se compará-lo também com o conteúdo de um sumário de referência, normalmente denominado *sumário ideal*. Essa forma de comparação tem sido a mais utilizada atualmente, pois pode ser mais facilmente automatizada. Uma vez tendo o sumário de referência, a avaliação da informatividade do sumário automático pode basear-se em várias medidas.

Um sumário de referência para um texto-fonte pode ser conseguido de várias formas:

- ele pode ser o sumário autêntico, isto é, o sumário produzido pelo próprio autor do texto-fonte;
- ele pode ser um sumário profissional, isto é, um sumário produzido a partir do texto-fonte por um escritor especialista em técnicas de sumarização;
- ele pode ser o extrato ideal, composto somente por sentenças mais representativas do texto-fonte.

É importante deixar claro que, apesar de usualmente se selecionar sentenças completas de um texto-fonte para formar seu sumário de referência, pode-se selecionar unidades com outros critérios, p.ex., trechos de sentenças, segmentos frasais, parágrafos, etc. A granularidade desejada depende do que se tem em mente ao avaliar um sistema de SA. Os sumários autêntico e profissional, segundo a definição acima, são os únicos resultantes da reescrita do texto-fonte (e, assim, ambos seriam *abstracts*, em nossa convença terminológica).

O extrato ideal é o melhor tipo de sumário de referência para a avaliação de sistemas de SA, pois, por conter somente sentenças do texto-fonte, pode ser comparado mais facilmente com um sumário automático, já que este também se origina do mesmo texto-fonte. No caso de extratos, a comparação com o extrato ideal pode ser automatizada; no caso de *abstracts*, podem ser necessárias etapas de revisão (humana) após o processamento.

O extrato ideal também pode ser produzido de várias formas:

- ele pode ser composto pelas sentenças do texto-fonte que mais se assemelhem às sentenças do sumário autêntico;

- ele pode ser composto pelas sentenças do texto-fonte que mais se assemelhem às sentenças do sumário profissional;
- ele pode ser composto pelas sentenças do texto-fonte julgadas por humanos como essenciais para compor um sumário do texto.

Na busca pelas sentenças do texto-fonte que mais se assemelham às sentenças do sumário autêntico ou profissional (*abstracts*), pode-se fazer uso de várias medidas. A mais utilizada, sugerida por Salton (1989), é a medida do co-seno, baseada puramente na co-ocorrência de palavras. Dessa forma, para cada sentença dos *abstracts* procura-se pela sentença do texto-fonte que tenha mais palavras em comum com aquela. Ao final, a justaposição das sentenças selecionadas do texto-fonte forma o extrato ideal. Outra medida, sugerida por Teufel e Moens (2002), utiliza, além da co-ocorrência de palavras, a ordenação das palavras nas sentenças. É importante ressaltar, entretanto, que essas medidas não são perfeitas, visto que não se realiza nenhum tipo de análise semântica das sentenças. A simples verificação de co-ocorrência e ordenação de palavras não garante que duas sentenças tenham o mesmo conteúdo informativo, podendo, eventualmente, introduzir erros na produção do extrato ideal. É por isso que se aconselha, quando possível, uma revisão humana dos extratos produzidos.

Com relação à construção do sumário de referência a partir das sentenças julgadas por humanos como essenciais, deve-se ressaltar a questão da baixa concordância entre os julgamentos humanos. Como vários experimentos têm mostrado (por exemplo, Mitra et al., 1997; Rath et al., 1961), juízes humanos, em geral, concordam muito pouco sobre as sentenças que devem fazer parte de um sumário. O que se costuma fazer é selecionar somente aquelas sobre as quais há maior concordância dos juízes. Por outro lado, Marcu (1999) ressalta que, apesar de ser possível uma baixa concordância *geral* entre juízes, é possível que, pelo menos na escolha das sentenças mais importantes, a taxa de concordância seja maior. Assim, seria preciso distinguir a forma de avaliar a validade dos resultados considerando também possíveis variações de julgamento dos próprios juizes.

Tendo em mãos o sumário de referência para um texto-fonte, as formas possíveis de se verificar a informatividade de um sumário automático são:

- Cálculo automático da precisão e cobertura do sumário automático em relação ao sumário de referência. Aplicável, preferencialmente, a *extratos ideais* como sumários de referência e a sistema de SA extrativos. A precisão (P) e a cobertura (do inglês, *recall*) (R) são dadas pelas seguintes fórmulas:

$$P = \frac{número\ de\ sentenças\ do\ sumário\ automático\ presentes\ no\ sumário\ de\ referência}{número\ de\ sentenças\ do\ sumário\ automático}$$

$$R = \frac{número\ de\ sentenças\ do\ sumário\ automático\ presentes\ no\ sumário\ de\ referência}{número\ de\ sentenças\ do\ sumário\ de\ referência}$$

A precisão indica quantas sentenças do sumário de referência o sumário automático possui em relação a todas as sentenças que ele contém; a cobertura indica quantas sentenças do sumário de referência o sumário automático possui em relação a todas as sentenças que ele deveria possuir. Uma outra medida, a *f-measure*, combina as medidas de precisão e cobertura, resultando em uma medida única de eficiência do sistema: quanto mais próxima essa medida for de 1, maior a capacidade do sistema em produzir sumários ideais. A fórmula da *f-measure* é a seguinte:

$$f - measure = \frac{2 \times P \times R}{P + R}$$

- Preferencialmente para um sistema extrativo, em vez de precisão e cobertura, pode-se utilizar a medida de *utilidade* de Radev et al. (2000). Para seu cálculo, pede-se a juízes humanos que dêem notas variando em uma determinada escala (de 1 a 9, por exemplo) para todas as sentenças do texto-fonte, que expressem sua importância para compor um sumário (uma nota é chamada de *ponto de utilidade*). Calcula-se, então, a nota geral do sumário de referência pela soma das notas de suas sentenças. Deste modo, a nota geral do sumário automático deve ser próxima (ou mesmo maior) do que a nota do sumário de referência para que ele seja considerado um sumário suficientemente informativo.

  A vantagem dessa medida é que ela é mais flexível do que as medidas de precisão e cobertura, pois não penaliza tanto um sumário automático quando este não possui alguma(s) das sentenças do sumário de referência. Assim, o julgamento não se dá em relação à presença ou ausência das sentenças do sumário de referência no sumário automático, mas em relação à importância (numérica) das sentenças. É importante ressaltar, entretanto, que também pode haver uma baixa concordância entre os juízes humanos em sua atribuição de notas às sentenças do texto-fonte.

- Outra medida, mais genérica e, portanto, aplicável a outros métodos de SA além dos extrativos (i.e., àqueles que produzem *abstracts*), é a medida de *conteúdo* (Salton e McGill, 1983). Por ela, verifica-se a parcela do conteúdo do sumário de referência que é transmitida pelo sumário automático, não levando em consideração valores numéricos (quantidade de sentenças ou tamanho dos sumários, por exemplo). Essa verificação pode ser manual, subjetiva, ou auxiliada por processos automáticos. O cálculo da medida do co-seno é um exemplo de processo automático que subsidia a identificação de passagens com mesmo conteúdo.

  Pela medida de conteúdo pode-se verificar, também, quanto do conteúdo do próprio texto-fonte é preservado no sumário automático, como fizeram Brandow et al. (1994).

É importante notar que as medidas de precisão e cobertura e de utilidade também podem ser aplicadas para outros métodos além dos extrativos, necessitando, somente, de algum esforço humano para definir a forma de cômputo das medidas.

O problema em se usar sumários de referência para a avaliação da informatividade de sumários automáticos é que o sumário de referência pode ser inadequado ou até mesmo ruim. Os sumários autênticos, por exemplo, podem conter informação não apresentada no texto-fonte ou mesmo ser pouco informativos. Nesses casos, a comparação fica prejudicada, já que não há um mecanisms de compreensão para concluir por um fator comum entre variações desse tipo. É importante, portanto, selecionar as fontes utilizadas para a avaliação. Kupiec et al. (1995), por exemplo, retiraram dos sumários de referência as informações que não estavam nos textos-fonte correspondentes, resultando nos chamados sumários *gold-standard*, considerados os sumários ideais.

Além da comparação do sumário automático com o sumário de referência ou com o texto-fonte, há outras formas de se verificar a informatividade dos sumários automáticos. Mani (2001) sugere que, se um sumário for informativo, ele deve preservar os mesmos conceitos-chave de seu texto-fonte, os quais podem ser expressos por suas próprias palavras-chave. Assim, pode-se verificar, por exemplo, se as palavras-chave fornecidas pelo autor do

texto-fonte ou aquelas calculadas por algum concordanceador[17] também são as palavras-chave do sumário automático ou se, pelo menos, estão presentes nele. Pardo e Rino (2002), em um outro tipo de avaliação, pedem a juízes humanos que dêem notas a sumários automáticos de acordo com a preservação da idéia principal dos textos-fonte correspondentes. As notas, nesse caso, indicam se o sumário preserva, preserva parcialmente ou mesmo não preserva a idéia principal. Essa proposta de avaliação se baseia na hipótese de que um sumário minimamente informativo deve transmitir, pelo menos, a idéia principal do texto-fonte.

Como se pode perceber, a subjetividade e a concordância dos julgamentos humanos é um grande desafio na avaliação de sistemas de SA, tanto na qualidade como na informatividade. O problema da subjetividade pode ser amenizado pela especificação clara de critérios de avaliação, pelo estabelecimento de escalas numéricas objetivas, em vez de conceitos abstratos, e pelo treinamento dos juízes, que, apesar de custoso, normalmente produz bons resultados. Em relação à concordância entre os julgamentos, sejam quais forem os objetivos dos mesmos, a avaliação pode não ser válida ou estar comprometida se houver uma baixa concordância entre os juízes. É por esse motivo que alguns métodos para se medir a concordância entre julgamentos foram propostos. A medida Kappa (Siegel e Castellan, 1988) é a mais conhecida, bastante utilizada pelos trabalhos atuais na Lingüística Computacional.

4.4. Avaliação extrínseca

Como já mencionado, a avaliação extrínseca visa avaliar um sistema em uso, para a realização de alguma tarefa específica. Para a SA, a avaliação extrínseca pode envolver, por exemplo, os contextos de categorização de textos, recuperação de informação ou perguntas e respostas. Algumas dessas avaliações são discutidas nesta seção. É importante lembrar a avaliação extrínseca é uma forma de validação do sistema em uso: pode-se validar, por exemplo, sua metodologia e/ou seu modelo lingüístico-computacional.

4.4.1. Categorização de documentos

Em uma tarefa de categorização de documentos, muito realizada em sites de notícias e necessária para catalogação de documentos em bibliotecas, por exemplo, o objetivo é atribuir uma categoria/classe a um dado documento. Normalmene, essa atribuição é feita por juízes humanos. Em sites de notícias, por exemplo, devem-se classificar as notícias para enquadrá-las em suas devidas seções, como “ciência”, “economia”, “informática”, etc.

Em uma avaliação extrínseca de um sistema de SA para a tarefa de categorização de documentos, o que se costuma fazer é pedir aos juízes que categorizem os documentos lendo somente os sumários correspondentes. A seguir, verifica-se 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. Em relação à SA, o objetivo deste tipo de avaliação é verificar se os sumários apresentam informação suficiente para a correta classificação dos textos-fonte, a partir de sua categorização. Esse tipo de avaliação foi proposto na conferência SUMMAC, como relatam Mani et al. (1998).

[17] Um *concordanceador* é um programa que calcula dados estatísticos para um texto de entrada, por exemplo, a lista de palavras-chave, a freqüência de cada palavra do texto, etc.

4.4.2. Recuperação de informação

Em uma tarefa de recuperação de informação, o objetivo é recuperar documentos que abordem um determinado tópico, como se costuma fazer em sites de busca de informação na web. O que se faz, neste caso, é pesquisar os documentos de uma base de dados em busca daqueles cujo tópico coincida com o tópico indicado pelo usuário. Similarmente à categorização de documentos, essa tarefa também é, muito freqüentemente, realizada por seres humanos, os quais desejam selecionar, entre vários documentos, aqueles que abordam algum assunto de seu interesse.

Na avaliação extrínseca, um sistema de SA é utilizado para gerar os sumários dos documentos da base que será pesquisada. A busca de documentos é feita, então, com base nos sumários dos documentos. Ela pode ser automática ou mesmo manual. O sucesso da busca é, então, avaliado por juízes humanos. Depois, verifica-se a taxa de acerto na recuperação e o tempo necessário para a busca, como no caso da categorização. Novamente, espera-se que a taxa de acerto se mantenha e que o tempo de busca diminua. Nesses casos, mede-se se os sumários realmente preservam todos os tópicos relevantes dos documentos para que a busca possa ser feita. Além da conferência SUMMAC (Mani et al., 1998), vários outros trabalhos abordaram esta avaliação (por exemplo, Tombros e Sanderson, 1998; Jing et al., 1998; Brandow et al., 1994; etc.).

4.4.3. Perguntas e respostas

Em uma avaliação extrínseca de perguntas e respostas, tem-se por objetivo verificar se o sistema de SA produz sumários informativos ou não. Nessa avaliação, dada uma base de documentos, elaboram-se algumas perguntas de múltipla escolha para cada texto. A seguir, aplica-se o sistema de SA aos documentos para produzir os sumários correspondentes. Por fim, procede-se então à avaliação propriamente dita. Primeiro, pede-se a juízes humanos que respondam às questões sem ler os textos-fonte nem os sumários; a seguir, pede-se aos juízes que leiam os sumários e respondam as questões; em um último passo, pede-se aos juízes que leiam os textos-fonte e respondam as mesmas perguntas.

A hipótese principal, neste caso, é que, se os sumários forem devidamente informativos, os juízes conseguirão responder as perguntas satisfatoriamente lendo somente os sumários. Costuma-se pedir aos juízes que repitam o procedimento lendo os textos-fonte e não lendo nada pelas seguintes razões: se, sem ler nada, os juízes conseguem responder corretamente algumas perguntas, isso indica que as mesmas são de senso comum e, portanto, devem ser excluídas da avaliação; se, mesmo lendo o texto-fonte completo, os juízes não conseguem responder algumas perguntas, isso indica que, muito provavelmente, os sumários também não irão respondê-las satisfatoriamente. Neste caso, elas não servem para avaliá-los, novamente, e, assim, devem ser excluídas da avaliação. Ao final, restarão as perguntas e respostas que realmente servirão para avaliar a informatividade dos sumários. Dentre os trabalhos que aplicaram esta avaliação, destacam-se os trabalhos de Morris et al. (1992) e Hovy e Lin (2000).

Em geral, as avaliações extrínsecas, assim como as intrínsecas, também apresentam diversos desafios para sua realização, por exemplo:

- as avaliações extrínsecas normalmente são on-line, isto é, precisam de juízes humanos, e, como já discutido, a avaliação on-line é custosa;

- por normalmente serem on-line, as avaliações extrínsecas pedem por documentos relativamente curtos para facilitar o trabalho dos juízes humanos. Entretanto, se forem muito curtos, sequer há necessidade de sumários;
- as avaliações extrínsecas, diferentemente das intrínsecas, não indicam pontos específicos em que os sistemas de SA utilizados podem ser aprimorados. Isso ocorre porque elas medem o desempenho das tarefas nas quais os sistemas de SA estão inseridos, e não os sistemas propriamente ditos;
- às vezes, é difícil criar tarefas extrínsecas que modelem adequadamente situações do mundo real e, ao mesmo tempo, sejam passíveis de medição e possíveis de serem realizadas por juízes humanos.

4.5. Estudo de caso: avaliação do GistSumm

Como estudo de caso, será apresentada a avaliação do GistSumm – ***GIST SUMM****arizer* – já descrito na Seção 3. Será mostrado o raciocínio por trás da definição da forma de avaliação e as conclusões inferidas com base nos resultados da avaliação.

Como já explicado anteriormente, as hipóteses relativas ao processo de sumarização do GistSumm são: (I) é possível determinar a *gist sentence* de um texto por meio de métodos estatísticos simples ou, pelo menos, se aproximar dela e (II) com base na *gist sentence*, é possível construir bons extratos. Portanto, a avaliação do GistSumm deve buscar a validação dessas hipóteses.

Para avaliar a hipótese I, deve-se especificar algum procedimento em que seja possível mostrar se a *gist sentence* pode ou não ser determinada pelos métodos estatísticos simples do GistSumm. Como relatam Pardo et al. (2003), a idéia foi, então, pedir a juízes humanos que determinassem as *gist sentences* de alguns textos (10, no total) e, então, verificar se o GistSumm identificava ou não estas *gist sentences* e, caso elas não fossem identificadas como tais, se elas eram incluídas ou não nos extratos produzidos automaticamente. Dada a subjetividade do julgamento humano, a *gist sentence* escolhida para cada texto foi aquela que recebeu mais votos dos juízes. A taxa de compressão utilizada foi de 60%.

É importante ressaltar que utilizar mais de 10 textos para essa avaliação a tornaria muito custosa, pois, quanto mais textos, mais tempo os juízes levariam para ler e identificar as *gist sentences* dos textos. Sempre que se utilizam juízes humanos, é importante balancear a quantidade de esforço necessário para a realização da tarefa, a capacidade dos juízes e o tempo disponível para a realização da tarefa.

Os resultados obtidos com a avaliação acima foram os seguintes:

- Utilizando o método das palavras-chave: as *gist sentences* escolhidas pelos juízes foram identificadas em 20% dos casos (ou seja, 2 textos); em 50% dos casos (5 textos), o gistsumm escolheu *gist sentences* muito próximas das *gist sentences* indicadas pelos juízes; para os 30% restantes (3 textos), o gistsumm não conseguiu sequer uma aproximação das *gist sentences* indicadas pelos juízes, ou seja, falhou. No total, as *gist sentences* indicadas pelos juízes foram incluídas nos extratos em 70% dos casos (7 textos).
- Utilizando o método TF-ISF: as *gist sentences* escolhidas pelos juízes foram identificadas em 20% dos casos (ou seja, 2 textos); em 10% dos casos (1 texto), o gistsumm escolheu *gist sentences* vagamente próximas das *gist sentences* indicadas pelos juízes; para os 70% restantes (7 textos), o GistSumm não conseguiu sequer uma aproximação das *gist sentences* indicadas pelos juízes, ou seja, falhou. No total, as *gist sentences* indicadas pelos juízes foram incluídas nos extratos em 30% dos casos (3 textos).

Como resultado, tem-se que o método das palavras-chave é satisfatório para a determinação das *gist sentences*, validando a hipótese I, enquanto o método TF-ISF não. Por esse motivo, pode-se descartar o método TF-ISF e futuros investimentos nele, pois, se ele sequer consegue identificar as *gist sentences*, ele não gerará bons extratos.

Essa avaliação pode ser classificada como **intrínseca**, pois analisa a qualidade do sistema em si, **glass-box**, pelo fato de analisar um dos componentes do GistSumm (o "módulo" que determina a *gist sentence*), **comparativa**, por se comparar dois métodos de determinação de *gist sentences*, e **off-line**, pelo fato de não utilizar o julgamento humano na avaliação. É importante ressaltar que, apesar dos juízes terem sido utilizados para detectar as *gist sentences* dos textos, eles não foram utilizados para julgar os extratos (etapa esta que, de fato, caracteriza a avaliação como off-line ou não).

Para avaliar a hipótese II, se os extratos produzidos pelo GistSumm são bons ou não (usando somente o método das palavras-chave para determinar as *gist sentences*), Pardo et al. tiveram que recorrer ao julgamento humano. Eles estabeleceram uma escala de pontuação para medir duas características dos extratos, a preservação da idéia principal dos textos-fonte e a textualidade (vide Quadro 1). Dessa forma, por exemplo, caso um juiz achasse que um extrato preservou a idéia principal do texto-fonte, mas não apresentou textualidade, então ele deveria dar nota 7 ao extrato. A idéia principal tem a ver com a informatividade dos extratos (vide subseção 4.3.2), ou seja, para que estes sejam minimamente informativos, eles devem preservar a idéia principal dos textos-fonte, pelo menos. A textualidade, por sua vez, engloba também a qualidade dos extratos (vide subseção 4.3.1), pois verifica a coesão deste, sua "fluência" durante a leitura.

Os resultados obtidos foram: 55% dos extratos gerados pelo GistSumm estavam acima da média e 14% dos extratos estavam na média; 50% dos extratos preservaram totalmente a idéia principal e 40% preservaram parcialmente; 50% dos extratos apresentaram textualidade total e 35% apresentaram textualidade parcial. Dessa forma, 90% dos extratos preservaram totalmente ou parcialmente a idéia principal e 85% dos extratos apresentaram textualidade total ou parcial. Portanto, pode-se considerar que a hipótese II foi validada. Essa avaliação pode ser classificada como **intrínseca**, **black-box**, **autônoma** e **on-line**.

Quadro 1: Escala de pontuação dos extratos

Idéia principal	Textualidade	Nota
Preservada	Ok	9
Preservada	±	8
Preservada	Sem	7
Parcialmente preservada	Ok	6
Parcialmente preservada	±	5
Parcialmente preservada	Sem	4
Não preservada	Ok	3
Não preservada	±	2
Não preservada	Sem	1

Em uma última avaliação, o GistSumm participou da conferência DUC realizada no início de 2003, já que, por ser uma conferência de avaliação de caráter internacional, ela é confiável e, portanto, dá mais validade aos resultados obtidos com a avaliação do sistema. Na primeira etapa da avaliação, de natureza **extrínseca**, verificou-se se os extratos produzidos pelo GistSumm eram "úteis" ou não para a tarefa em que um usuário tem que selecionar que documentos ler com base nos seus sumários (vide subseção 4.4.2). Dados 624 textos-fonte, para cada extrato produzido pelo GistSumm (com uma média de 38 palavras), juízes

humanos, especialistas em técnicas de recuperação de informação, deram notas de 0 a 4 aos extratos, onde 0 indicava um extrato inútil e 4 um extrato perfeito que poderia até mesmo substituir o texto-fonte. Nesta avaliação, o GistSumm conseguiu uma nota média de 3.12, o que caracteriza seus extratos como sendo muito bons. Em uma outra etapa, agora **intrínseca**, para verificar a informatividade dos extratos, juízes humanos calcularam a cobertura (*recall*, vide subseção 4.3.2) dos extratos em relação a sumários profissionais. O GistSumm atingiu uma cobertura média de 51%. Ambas as avaliações anteriores são consideradas **black-box**, **autônomas** e **on-line**. Vale citar que, em uma última etapa da DUC, foi realizada uma avaliação **comparativa** entre os sistemas de SA participantes da conferência, porém, o GistSumm não participou desta etapa.

4.6. Considerações finais sobre avaliação de sistemas de SA

A avaliação de sistemas de SA é um assunto muito amplo. Há diversas formas de se avaliar um sistema de SA, podendo-se focalizar suas características computacionais, o design de sua interface e, mais importante, os resultados produzidos, que, neste caso, são os próprios sumários.

Avaliar, em geral, é um processo custoso, ainda mais pelo fato de precisar, com freqüência, do julgamento humano. Métodos automáticos de avaliação existem, como discutido nesta seção, mas apresentam diversos problemas e não são tão satisfatórios como o julgamento humano. Mesmo o julgamento humano pode ser problemático, dada a subjetividade desta tarefa e a baixa concordância entre juízes. Apesar das dificuldades, padrões e métricas de avaliação de sistemas de SA têm surgido na literatura, assim como as conferências internacionais de avaliação têm se tornado cada vez mais importante.

Mani (2001) afirma que a avaliação de sistemas de SA tem que nortear o desenvolvimento de tecnologia na área e ser norteada por esse mesmo processo. Entretanto, avaliar não é simplesmente seguir um "livro de receitas", pois depende das necessidades e características de cada sistema de SA e dos objetivos dos desenvolvedores do sistema, que pode ser desde melhorar o "estado da arte", em termos de modelos e métodos de sumarização, até adequar os sistemas a tarefas específicas do mundo real.

A avaliação em SA é um tema desafiador e necessário para o desenvolvimento da pesquisa, que ainda precisa ser bastante trabalhado em busca de padrões e metodologias adequadas. Citando o próprio Mani (2001, p. 224): "*if all we do is evaluation, evaluation is all we will do!*"

Agradecimentos
Agradecemos à Prof. Camilla Brandel Martins, por sua contribuição à Seção 3.4, cujo trabalho é de sua autoria.

Referências bibliográficas[18]

Aretoulaki, M. (1996). *COSY-MATS: A Hybrid Connectionist-Symbolic Approach To The Pragmatic Analysis of Texts For Their Automatic Summarisation*. PhD. Thesis. University of Manchester.

Barzilay, R.; Elhadad, M. (1997). Using Lexical Chains for Text Summarization. In the *Proc. of the Intelligent Scalable Text Summarization Workshop*, Madri, Spain. Also In I. Mani and M.T. Maybury (eds.), *Advances in Automatic Text Summarization*. MIT Press, pp. 111-121.

Baxendale, P.B. (1958). Machine-made index for technical literature – an experiment. *IBM Journal of Research and Development*, Vol. 2, pp. 354-365.

Black, W.J.; Johnson, F.C. (1988). A Practical Evaluation of Two Rule-Based Automatic Abstraction Techniques. *Expert Systems for Information Management*, Vol. 1, No. 3. Department of Computation. University of Manchester Institute of Science and Technology.

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

Borko, H.; Bernier, C.L. (1975). *Abstracting Concepts and Methods*. Academic Press. San Diego, CA.

Braga, A.P.; Ludermir, T.B.; Carvalho, A.C.P.L.F. (2000). *Redes Neurais Artificiais: Teoria e aplicações*. LTC - Livros Técnicos e Científicos Editora S.A, Rio de Janeiro.

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

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA.

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

Cremmins, E.T. (1996). *The Art of Abstracting*. Information Resource Press. Arlington, Virginia.

Edmundson, H.P. (1969). New Methods in Automatic Extracting. *Journal of the ACM*, 16, pp. 264-285.

*Feltrim, V.D.; Nunes, M.G.V.; Aluísio, S.M. (2001). *Um corpus de textos científicos em Português para a análise da Estrutura Esquemática*. Série de Relatórios do NILC. NILC-TR-01-4.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, Vol. 32, pp. 221-233.

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

Halliday, M.A.K.; Hasan, R. (1976). *Cohesion in English*. Longman.

Hoey, M. (1983). *On the Surface of Discourse*. George Allen & Unwin Ltd.

Hoey, M. (1991). *Patterns of Lexis in Text*. Oxford University Press.

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

Hovy, E.H.; C-Y. Lin. (2000). Automated Text Summarization and the SUMMARIST System. In the *Proceedings of the TIPSTER Text Program, Phase III*, pp. 197-214.

[18] Itens marcados com "*" podem ser encontrados no site do NILC (http://nilc.icmc.sc.usp.br – *Publication*s ou *Projects/EXPLOSA*).

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

Jordan, M.P. (1980). Short Texts to Explain Problem-Solution Structures – and Vice Versa. *Instructional Science*, Vol. 9, pp. 221-252

Jordan, M. P. (1992). An Integrated Three-Pronged Analysis of a Fund-Raising Letter. In W. C. Mann and S. A. Thompson (eds), *Discourse Description: Diverse Linguistic Analyses of a Fund-Raising Text*, pp. 171-226.

Kupiec, J.; Petersen, J.; Chen, F. (1995). A trainable document summarizer. In Edward Fox, Peter Ingwersen, & Raya Fidel (eds.), *Proceedings of the 18th Annual International ACM-SIGIR Conference on Research & Development in Information Retrieval*, pp. 68-73, Seattle, WA, EUA. July.

Larocca Neto, J.; Santos, A.D.; Kaestner, A.A.; Freitas, A.A. (2000). Generating Text Summaries through the Relative Importance of Topics. In the *Proceedings of the International Joint Conference IBERAMIA/SBIA*, Atibaia, SP.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development*, Vol. 2, pp. 159-165.

Mani, I. (2001). Summarization Evaluation: An Overview. In the *Proceedings of the Workshop on Automatic Summarization*. Pittsburgh, Pennsylvania.

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

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

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

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

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

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

Marcu, D. (1999). Discourse trees are good indicators of importance in text. In I. Mani and M. Maybury (eds.), *Advances in Automatic Text Summarization,* pp. 123-136. The MIT Press.

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

*Martins, C.B. (2002). *UNLSumm: Um Sumarizador Automático de Textos UNL*. Dissertação de Mestrado, DC/UFSCar, São Carlos.

*Martins, T.B.F.; Ghiraldelo, C.M.; Nunes, M.G.V.; Oliveira Jr., O.N. (1996). *Readability Formulas Applied to Textbooks in Brazilian Portuguese*. Notas do ICMSC-USP, Série Computação.

Miike, S.; Itoh, E.; Ono, K.; Sumita, K. (1994). A full text-retrieval system with a dynamic abstract generation function. In W. Bruce Croft and C.J. van Rijsbergen (eds.), *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research & Development in Information Retrieval*, pp. 152-161, July.

Miller, G. (1995). WordNet: A Lexical Database for English. *Communication of the Association for Computing Machinery* 38 (11), pp. 39-41.

Minel, J.L.; Nugier, S.; Piat, G. (1997). How to appreciate the quality of automatic text summarization? Examples of FAN and MLUCE Protocols and their Results on

SERAPHIN. In I. Mani and M. Maybury (eds.), *Proceedings of the ACL/EACL Workshop on Intelligent Scalable Text Summarization*.

Mitchell, T.M. (1997). *Machine Learning*. McGraw Hill, New York.

Mitra, M.; Singhal, A.; Buckley, C. (1997). Automatic text summarization by paragraph extraction. In the *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*.

Moore, J.D.; Paris, C. (1993). Plannig Text for Advisory Dialogues: Capturing Intentional and Rhetorical Information. *Computational Linguistics*, Vol. 19, No. 4, pp. 651-694.

Morris, J.; Hirst, G. (1991). Lexical cohesion, the thesaurus, and the structure of text. *Computational Linguistics,* 17(1): 21-48.

Morris A.; Kasper, G.; Adams, D. (1992). The Effects and Limitations of Automatic Text Condensing on Reading Comprehension Performance. *Information Systems Research*, Vol. 3, N. 1, pp. 17-35.

Nunes, M.G.V.; Vieira, F.M.V; Zavaglia, C.; Sossolete, C.R.C.; Hernandez, J. (1996). *A Construção de um Léxico da Língua Portuguesa do Brasil para suporte à Correção Automática de Textos*. Série de Relatórios Técnicos do ICMC-USP, no. 42.

Paice, C. D. (1981). The automatic generation of literature abstracts: an approach based on the identification of self-indicating phrases. *Information Retrieval Research*. Butterworth & Co. (Publishers).

Paice, C.D.; Jones, P.A. (1993). The identification of important concepts in highly structure technical papers. In R. Korfaghe, E. Rasmussen, and P. Willett (eds.), Proc. of the 16th ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 69-78. ACM Press, June.

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

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

*Pardo, T.A.S. (2002c). *Descrição do DMSumm: um Sumarizador Automático Baseado em um Modelo Discursivo*. Série de Relatórios do NILC (DC-UFSCar). NILC-TR-02-02.

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

*Pardo, T.A.S.; Rino, L.H.M.; Nunes, M.G.V. (2003). GistSumm: A Summarization Tool Based on a New Extractive Method. To appear in the *Proceedings of the 6th Workshop on Computational Processing of the Portuguese Language - Written and Spoken*. Faro, Portugal.

*Pinheiro, G.M. e Aluísio, S.M. (2003). *Corpus NILC: Descrição e Análise Crítica com Vistas ao Projeto Lacio-Web*. Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação – ICMC, Universidade de São Paulo, N. 190.

Pollock, J.J.; Zamora, A. (1975). Automatic Abstracting Research at Chemical Abstracts Service. *Journal of Chemical Information and Compute Sciences* 15(4): 226-232.

Pollock, J.J.; Zamora, A. (1999). Reedição de (Pollock e Zamora, 1975).

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

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

*Rino, L.H.M. (1996). *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.

*Rino, L.H.M.; Scott, D. (1994). *Automatic generation of draft summaries: heuristics for content selection.* ITRI Techn. Report ITRI-94-8. University of Brighton, England.

Rowley, J. (1982). *Abstracting and Indexing*. Clive Bingley, London.

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

Salton, G. (1988). *Automatic Text Processing*. Reading, MA: Addison-Wesley.

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

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

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

Schilder, F. (2002). 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.

Sparck Jones, K. (1993). What might be in a summary? In Krause Knorz and Womser-Hacker (eds.), *Information Retrieval 93*, pp. 9-26. Universitatsverlag Konstanz. June.

Sparck Jones, K.; Galliers, J.R. (1996). Evaluating Natural Language Processing Systems. *Lecture Notes in Artificial Intelligence*, Vol. 1083.

Sparck Jones, K. (1997). "Summarising: Where are we now? Where should we go?" *Proc. of the Intelligent Scalable Text Summarization Workshop, ACL/EACL'97 Joint Conference*. Madrid, Spain, p. 1.

Siegel, S.; Castellan, N.J. (1988). *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill.

Teufel, S.; Moens, M. (1999). Argumentative Classification of Extracted Sentences as a First Step Towards Flexible Abstracting. In Inderjeet Mani and Mark T. Maybury (Eds.), *Advances in Automatic Text Summarization*. Massachusetts Institute of Technology Press.

Teufel, S.; Moens, M. (2002). Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status. *Computational Linguistics*, Vol. 28, N. 4, pp. 409-445.

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

Uchida, H. (1997). *DeConverter Specification, Version 1.0*. Tech. Rep. UNL-TR1997-010, UNU/IAS/UNL Center, Tokyo, Japan.

Uchida, H. (2000). *Universal Networking Language: An Electronic Language for Communication, Understanding and Collaboration*. UNL Center, IAS/UNU, Tokyo (também disponível no site www.unl.ias.unu.edu).

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

Winter, E.O. (1976). *Fundamentals of Information Structure*. Hatfield Polytechnic, Hertfodshire, England.

Winter, E.O. (1977). A Clause-Relational Approach to English Texts. A Study of Some Predictive Lexical Items in Written Discourse. *Structional Science*, Vol. 6, N. 1, pp. 1-92.

Winter, E.O. (1979). Replacement as a Fundamental Function of the Sentence in Context. In *Forum Linguitics*, Vol. 4, N. 2, pp. 95-133.

Witten, I.H.; Moffat, A.; Bell, T.C. (1994). Managing Gigabytes. *Van Nostrand Reinhold*. New York.