INTRODUÇÃO À SUMARIZAÇÃO AUTOMÁTICA

Camilla Brandel Martins
Thiago Alexandre Salgueiro Pardo
Alice Picon Espina
Lucia Helena Machado Rino

Sumário

A sumarização, em geral, é uma atividade bastante comum. Quando se narra um evento a uma pessoa, costuma-se fazer um resumo do que aconteceu e não uma narração completa e detalhada. Inconscientemente, as pessoas estão sempre sumarizando, quer oral, quer textualmente. Exemplos de sumários escritos incluem notícias de jornais, artigos de revistas, resumo de textos científicos, entre muitos outros. Por sua utilidade e freqüência, há um grande interesse em automatizar esse processo. Neste relatório, veremos as principais características a serem consideradas para a sumarização automática.

Índice

1. Introdução

A sumarização, em geral, é uma atividade bastante comum. Quando se narra um evento a uma pessoa, costuma-se fazer um resumo do que aconteceu, e não uma narração completa e detalhada. Inconscientemente, as pessoas estão sempre sumarizando. Muito freqüentes também são os sumários escritos, como, por exemplo, notícias em jornais, artigos de revistas, resumo de textos científicos, entre muitos outros.

Os sumários produzidos a partir de textos são particularmente úteis, podendo servir como indexadores ou ser autocontidos. No primeiro caso, os sumários são lidos para descobrir qual é o assunto do texto-fonte correspondente e, caso seja de interesse, o leitor remete ao texto completo, para informar-se. No segundo caso, os sumários já são considerados informativos o suficiente e, portanto, o leitor pode dispensar o texto de origem e, ainda assim, apreender suas informações principais.

Devido a sua utilidade e freqüência e aos avanços na área de Processamento de Línguas Naturais (PLN), é de grande interesse a automação do processo de sumarização. A sumarização automática vem sendo objeto de estudo desde os primórdios da computação. Já no final da década de 50 começaram a surgir alguns métodos estatísticos para extrair as sentenças principais de um texto. As pesquisas continuaram nas décadas seguintes, trazendo vários avanços à área, conforme descreveremos neste relatório, sob a ótica de duas abordagens principais do PLN: a superficial e a profunda, as quais caracterizam métodos distintos de sumarização automática. A abordagem superficial utiliza, sobretudo, métodos experimentais e estatísticos, enquanto a profunda está relacionada a teorias formais e lingüísticas.

Na Seção 2 são apresentados os conceitos principais da sumarização, assim como as classificações mais comuns de sumários, de acordo com vários autores. A Seção 3 dá uma visão geral dos problemas e abordagens da sumarização automática, sobretudo em relação ao processamento superficial e profundo, enquanto que as Seções 4 e 5 exploram, respectivamente, cada um desses modelos de processamento. A Seção 6 apresenta conclusões e comentários gerais.

2. Conceitos de Sumarização

Quando se fala de sumarização, é necessário referir-se ao que se entende por um sumário. No campo da sumarização humana, por exemplo, encontramos vários tipos de sumários: resenhas de notícias jornalísticas, sinopses do movimento da bolsa de valores, sumários de textos novelísticos, extratos de livros científicos, resumos de previsões meteorológicas, etc. Cada um desses tipos envolve pressuposições e características diversas, assim como conteúdos e correspondência com suas fontes de teores variados. Por exemplo, um determinado autor de um sumário jornalístico pode considerar que um título espetacular para um texto que descreve um acidente automobilístico envolvendo um ator de proeminência nacional (Paulo Zulu, por exemplo) deve mencionar sua morte. Assim, seu sumário pode ser “Acidente mata Paulo Zulu”. Para o mesmo evento, outro sumarizador humano pode priorizar o desfecho trágico do evento, ignorando a vítima. Neste caso, seu sumário pode ser “Acidente termina em tragédia”.

Variações dessa natureza são comuns na produção humana de sumários. Podemos argumentar que, no primeiro exemplo, o autor pressupôs que, ao mencionar a vítima, sua mensagem surtiria um efeito espetacular sobre os leitores, compelindo-os a ler o texto correspondente. No segundo exemplo, podemos supor que o autor ignora a vítima proposital ou casualmente, ou por resolver não chocar explicitamente seus leitores ou por desconhecer a proeminência da vítima. Vale notar, no entanto, que

ambas as formas podem ser associadas – e serão – ao mesmo texto descritivo do acidente envolvendo 'Paulo Zulu'. Ambos os exemplos de sumários, neste caso, apresentam ainda uma característica bastante peculiar: a de se considerar que títulos – e, portanto, formas não textuais – podem também sumarizar seus correspondentes textos.

Com esse exemplo, referimo-nos a três características possíveis na sumarização humana: a variação de conteúdo informativo e a multiplicidade sentencial ou estrutural dos sumários e, portanto, a possibilidade de se produzir mais de um sumário para um mesmo texto-fonte.

Para efeito de sumarização automática, essas possibilidades levam a um grande problema: o de modelá-las de modo adequado, para que os resultados automáticos reflitam a diversidade de sumários sem que estes percam sua interdependência com os textos-fonte correspondentes.

Além dessas características, várias outras, também apreendidas mediante análise do processo humano de sumarização, irão interferir no projeto e desenvolvimento de sumarizadores automáticos. O importante é notar que a) sumários remetem, necessariamente, a eventos ou textos originários dos mesmos e b) sumários devem ser construídos tendo em mente que não pode haver perda do significado original, muito embora contenham menos informações e possam apresentar diferentes estruturas, em relação a suas fontes.

Por essas razões, é necessário, primeiramente, que se definam claramente os conceitos básicos relativos ao que se consideram sumários para, em seguida, analisar os fundamentos que permitirão a modelagem automática. É visando esclarecer essas noções que apresentamos, nesta seção, algumas classes de sumários segundo a visão de alguns autores em particular para, a seguir, descrever o processo, propriamente dito.

2.1. Sumários, Índices e Extratos

Hutchins (1987) classifica sumários científicos em três tipos: indicativos, informativos e sumários de crítica (*evaluative*). Para ele, sumários indicativos contêm apenas os tópicos essenciais de um texto, não necessariamente contendo detalhes de resultados, argumentos e conclusões. Sumários informativos, por sua vez, são considerados substitutos do texto, devendo conter todos os seus aspectos principais. Assim, se o texto-fonte correspondente for organizado em função de, por exemplo, dados, métodos, hipóteses e conclusões, um sumário informativo deveria conter as informações principais de cada um desses tópicos. Sumários de crítica assumem a função de avaliadores que, por exemplo, apresentam uma análise comparativa do conteúdo do texto-fonte com o contexto de outros trabalhos relacionados a esse conteúdo, na área específica em foco. Segundo Hutchins, seria mais simples produzir automaticamente sumários indicativos, devido à complexidade de se modelar adequadamente a sumarização humana para os demais tipos de sumários.

Mais recentemente, Sparck Jones (1993a) procurou esclarecer a distinção entre sumários e índices de modo a permitir uma melhor forma de proceder à modelagem lingüístico-computacional. Segundo ela, sumários são também textos, podendo ser substitutos para o documento original. Seriam equivalentes, portanto, aos sumários informativos de Hutchins. No entanto, ela considera que um sumário também pode ser utilizado como índice. Sumários indexadores, no entanto, não poderiam ser utilizados como substitutos para seus textos-fonte, pois não necessariamente estariam preservando o que os originais têm de mais importante, em termos de conteúdo e estrutura. Ao contrário, estariam transmitindo somente uma vaga idéia do texto correspondente, podendo, inclusive, apresentar uma forma não textual (uma lista de itens, por exemplo).

Para índices em formato textual, os sumários de Sparck Jones poderiam ser comparáveis aos sumários indicativos de Hutchins.

Outra diferença apontada por Sparck Jones (1993b), entre sumários e índices, reside na avaliação de sua funcionalidade e qualidade. Índices, por exemplo, são utilizados na classificação de documentos bibliográficos, de um modo geral, indicando seu conteúdo e agilizando o acesso às suas informações relevantes. Através deles, o leitor de uma enciclopédia, por exemplo, pode ir direto ao volume ou página que trata do assunto que lhe interessa. Dessa forma, a utilidade de índices é mais clara e sua função mais limitada do que as de sumários, permitindo uma avaliação mais robusta de sua funcionalidade e qualidade. Por outro lado, sumários apresentam uma relação mais complicada com relação a seu texto-fonte. Do seu objetivo dependerá bastante a avaliação sobre o quanto ele atende às necessidades do usuário.

Extratos são outra forma de sumários, podendo ser entendidos como uma composição de segmentos relevantes de um texto. Na sumarização automática, sentenças inteiras podem ser extraídas de um texto e justapostas, sem qualquer modificação, para compor um sumário (Paice, 1981). O problema, neste caso, é identificar, no texto-fonte, aquelas sentenças que sejam relevantes para transmitir a idéia principal do texto. Métodos estatísticos (p.ex., o método das palavras-chave) são explorados com esse fim, como veremos adiante.

As variações conceituais ilustradas acima, sobre o que se consideram sumários, embora compreensíveis do ponto de vista de um falante, são noções ainda obscuras para o projeto e desenvolvimento de sumarizadores automáticos. Além de métodos estatísticos, são explorados também métodos simbólicos (ou profundos), para se tentar obter modelos computacionais que reflitam as principais características da sumarização humana. Apresentaremos, mais adiante, os aspectos relevantes de ambas as abordagens, para a sumarização automática.

2.2. A Sumarização Humana

O ponto central da sumarização é reconhecer, em um texto, o que é relevante e o que pode ser descartado, para compor um sumário. Esse é também um dos pontos mais problemáticos e sujeito a controvérsias. A importância de uma sentença ou trecho de um texto pode depender de vários fatores: a) dos objetivos do autor do sumário, b) dos objetivos ou interesse de seus possíveis leitores e c) da importância relativa (e subjetiva) que o próprio autor (ou leitor) atribui às informações textuais, fazendo com que mesmo a estruturação do sumário esteja sujeita à complexidade de se determinar forma e conteúdo a partir do fonte correspondente.

Segundo Hutchins, a análise do conteúdo de documentos é uma das atividades mais importantes de um sistema de informação e entender o modo como profissionais (indexadores ou sumarizadores humanos) produzem seus textos pode levar a avanços consideráveis para a automação do processo. No entanto, há uma grande dificuldade em saber exatamente como esses profissionais trabalham, devido à subjetividade da tarefa. Em alguns casos, em que os autores seguem uma seqüência direta de raciocínio e construção, torna-se possível modelá-la. Por exemplo, quando eles esquadrinham, selecionam, constróem o sumário e o revisam a partir de um texto-fonte, pode-se obter uma caracterização modular de cada uma dessas tarefas. Contudo, Endres-Niggemeyer (1990) observa que o modo como eles realizam cada um desses passos ainda pode variar de acordo com o indivíduo, dificultando a modelagem necessária.

Hutchins argumenta ainda que, em textos expositivos, o leitor se lembra somente da idéia central – ou argumento principal (*gist*) – do texto-fonte, classificando as demais

características como positivas ou negativas, em função dessa idéia. O problema é, novamente, identificar qual é a idéia central do texto a ser sumarizado, a fim de preservá-la no(s) sumário(s) correspondente(s). Esse problema está relacionado tanto à forma (ou estrutura), quanto ao conteúdo do texto, como veremos a seguir.

2.3. A Estrutura Textual

Para conseguir produzir um bom sumário, é necessário antes que o sumarizador consiga reconhecer e reproduzir as idéias principais do texto. De forma semelhante, um sistema de sumarização automática precisa, de alguma forma, entender o que está sendo apresentado. A compreensão do discurso, por sua vez, envolve o reconhecimento das estruturas do texto. Assim, antes da sumarização propriamente dita, faz–se necessário investigar a estrutura do discurso.

Seguindo a idéia de micro e super-estrutura de van Dijk (1979), Hutchins (1987) estabelece uma distinção entre a micro-estrutura e a macro-estrutura de textos. A micro-estrutura representa as relações entre seqüências de sentenças no texto; a macro-estrutura representa relações entre blocos de sentenças e a organização global dos textos.

No nível micro-estrutural, o entendimento de textos requer a recuperação da progressão temática das sentenças e cláusulas, para a recuperação adequada do significado textual. Desse modo, esse nível é associado ao inter-relacionamento semântico entre os componentes textuais (palavras, frases ou blocos textuais). No nível macro-estrutural, a recuperação do significado requer a construção de padrões organizacionais. Segundo Hutchins, evidências sugerem que o que o leitor de um texto lembra do conteúdo se assemelha à rede semântica de sua macro-estrutura. Ou seja, ele se lembra da seqüência de episódios e dos participantes principais, em textos narrativos, e da idéia principal de um argumento ou de suas características positivas e negativas, em textos expositivos. Os detalhes micro-estruturais, neste caso, dificilmente seriam retidos. Portanto, sumários expressariam a macro-estrutura de um texto segundo a interpretação de um indivíduo, com base no seu conhecimento previamente adquirido (Hutchins, 1987). Para recuperar os padrões organizacionais, Hutchins sugere quatro macro-regras que podem ser utilizadas para generalizar e condensar um texto: Delição, Generalização, Construção (todas originalmente propostas por van Dijk), além de uma regra de composição, conforme segue:

1) Delição: exclusão de propriedades e atributos, ou até de cláusulas completas, considerados irrelevantes.
   Exemplo: “A menina ganhou um gatinho branco” pode ser sumarizada pela delição do qualificativo “branco”, resultando em “A menina ganhou um gatinho”.

2) Generalização: substituição de hipônimos por seus hiperônimos ou de detalhes por uma idéia ou frase mais abstrata, que envolva os detalhes, deixando-os implícitos.
   Exemplo (Hutchins, 1987): “Pedro viu aves.” pode ser a generalização da seqüência “Pedro viu um falcão. Pedro viu um urubu.” ou, ainda, da sentença “Pedro viu um falcão e um urubu.”.

3) Construção: esta regra, semelhante à de generalização, age como seu complemento sintagmático: enquanto a generalização é basicamente pragmática, a construção depende da micro-estrutura textual. Neste caso, a macro-proposição construída é resultante de uma seqüência de micro-proposições.

Exemplo (Hutchins, 1987): a seqüência "Pedro comprou tijolos, areia, cimento, colocou os alicerces, ergueu paredes..." permite construir a sentença "Pedro construiu uma casa".

4) Composição delição-construção: Consiste em delitar sentenças que expressam pré-condições ou motivações para a descrição de ações. Como resultado da delição, faz-se necessário rever – ou reconstruir – o texto. Exemplo: para a seqüência de sentenças "Pedro queria ir ao cinema". "Pedro foi ao cinema", a primeira pode ser omitida, sem perda de significado.

Na década de 80, as macro-regras de delição e construção foram bastante investigadas e utilizadas em implementações de sistemas de sumarização automática (Correira, 1980; Fum et al., 1982), enquanto a de generalização foi pouco explorada, devido à necessidade de se recorrer a um *thesaurus* interno para se fazer referência a possíveis generalizações. Hutchins observa, contudo, que esse processo envolve mais do que a consulta a informações de *thesaurus*, sendo necessário investigar mais profundamente os métodos de análise semântica, não apenas para a sumarização automática, mas para o PLN em geral.

2.4. O Conteúdo Textual

Para se determinar o conteúdo relevante de um texto-fonte, para compor um sumário, além dos sinais superficiais (sejam eles estruturais ou lingüísticos) aos quais recorre o autor do texto em sua escrita, outros fatores também devem ser considerados pelo sumarizador humano. Destacam-se, principalmente, a) o domínio do assunto específico, que o sumarizador detém para entender e, portanto, abstrair ou generalizar as informações que ele lê no texto-fonte e b) o conhecimento prévio que ele possa ter, como experienciador nesse domínio.

São relatadas, por exemplo, várias experiências no campo da psicolingüística, sobre o fato de sumários apresentarem informações que não necessariamente se encontram aparentes nos respectivos textos-fonte. Essa observação remete, muitas vezes, a evidências de que, ao contrário do que se espera normalmente, sumários são construídos de forma independente, ou até mesmo antes, de seus textos correspondentes, apresentando, inclusive, objetivos diversos dos mesmos (Mushakoji, 1993). Isto ocorre, por exemplo, quando se deseja submeter sumários de trabalhos científicos à aprovação de um comitê de programa de uma conferência internacional: na maioria das vezes, o artigo completo sequer foi idealizado, mas seu sumário já é produzido e enviado para avaliação. Caso o trabalho seja aceito para apresentação, constrói-se, então, o texto correspondente que, por sua vez, pode conter um sumário (*abstract*) muito diverso do primeiro, submetido para aceitação. Muito embora de caráter diverso, vale notar, entretanto, que esses sumários remetem ao mesmo texto-fonte, mesmo quando contêm informações diversas ou adicionais, em relação ao fonte.

A razão atribuída para essa caracterização está no fato do que se denominou *authoring summarizing*, em oposição à sumarização profissional (Mushakoji and Nozoe, 1992). No primeiro caso, supõe-se que o escritor é o próprio autor do trabalho que está sendo relatado: por conhecer o assunto, ele recorre a informações variadas, em relação ao fonte; no segundo caso, o texto é enviado a um escritor independente, que não necessariamente domine sequer o assunto em foco, razão que o levaria a se limitar ao conteúdo aparente no texto-fonte correspondente. Essa distinção justifica o fato do

conteúdo textual de um sumário diferir do conteúdo de seu fonte, ainda que remeta claramente ao mesmo.

Muito embora essas considerações sobre o conteúdo textual sejam altamente relevantes para o entendimento do processo de sumarização, elas são de difícil modelagem computacional, pois apresentam um alto grau de subjetividade e requerem uma representação bastante complexa do conhecimento de mundo/domínio, além de um modelo do escritor/leitor também elaborado, para dar conta de decisões variáveis, em relação ao conteúdo textual. Em geral, os modelos explorados para a sumarização automática se baseiam no conteúdo explícito dos textos-fonte e em suas características estruturais, quando se baseiam em metodologias profundas. Quando contemplam técnicas superficiais, baseiam-se também em suas características estruturais. Veremos, a seguir, alguns exemplos de sistemas de ambas as naturezas.

3. A Sumarização Automática

A sumarização automática vem sendo explorada desde a década de 50, quando começaram a surgir os primeiros métodos para a produção de extratos, sendo o método das palavras-chave (Luhn, 1958) o mais significativo então. Entretanto, como métodos "cegos", fazendo uso de técnicas superficiais, os resultados apresentavam inúmeros problemas, quer de coesão, quer de coerência (Hutchins, 1987), razão pela qual a área ficou praticamente estagnada nas décadas seguintes, voltando a ser objeto de interesse com o advento da Internet e, portanto, com o aumento considerável de documentos disponíveis *on-line* e com a necessidade de se "digerir" informações em larga escala (isto é, em grande quantidade e no menor tempo possível).

Sparck Jones (1993b) atribui a ausência de progresso significativo na área até meados da década de 90 à dificuldade de se modelar adequadamente o processo, resultando na impossibilidade de se obter sumários automáticos de qualidade. Contudo, ela observa que, para aplicações restritas, é possível explorar critérios estruturais ou composicionais em certa profundidade, baseados nos aspectos lingüísticos do texto-fonte.

Como já mencionamos, a abordagem profunda toma do processamento humano da linguagem (e, portanto, também da psicologia cognitiva) a base interpretativa para a compreensão do texto-fonte e posterior extração/produção do(s) sumário(s) correspondente(s). A base fundamental dessa metodologia é que um bom sistema automático de PLN deve fazer uso intensivo de regras gramaticais e de habilidades de inferência lógica, além de manipular grandes bases de conhecimento de mundo, além do próprio conhecimento lingüístico (Sampson, 1987). Segundo essa abordagem, a sumarização automática é baseada em teorias lingüísticas formais. Entendido dessa forma, o sistema computacional deveria simular a inteligência humana, para obter um processamento eficiente da língua. Porém, a grande maioria de pesquisadores entende que a tarefa de simular a inteligência humana para um domínio aberto, no PLN, ainda está fora do alcance.

Baseados em tal argumento, alguns pesquisadores utilizam as técnicas estatísticas como métodos superficiais, que seriam de mais simples aplicação, não necessitando de algoritmos complexos, podendo ser aplicados para quaisquer conjuntos de entrada, sejam estruturas lingüísticas bem formadas ou não. Por outro lado, tais métodos, ditos "cegos", não levam em consideração todo o conhecimento potencial da língua, tampouco os critérios de compreensão ou critérios lingüísticos necessários a essa tarefa.

Não se pode afirmar que uma das abordagens seja mais eficiente que a outra, pois cada uma apresenta fatores positivos e negativos, devendo, antes, ser complementares. Segundo Klavans e Resnick (1996), embora no passado parecesse necessário escolher entre essas duas abordagens, os pesquisadores passaram a buscar uma combinação coerente das mesmas, investigando amplamente tecnologias híbridas, ou seja, fazendo uso de metodologias simbólicas e técnicas estatísticas.

O sumarizador de eventos de Maybury (1993), por exemplo, utiliza uma abordagem híbrida para lidar com mensagens de um simulador de batalha. As mensagens de eventos passam inicialmente por uma triagem, a qual seleciona as mais relevantes utilizando uma métrica baseada tanto na freqüência (ou seja, técnica estatística) quanto na singularidade e importância específica do evento para o domínio (técnica simbólica). Os agentes das mensagens resultantes são missões de guerra (tais como destruir ou abastecer), os quais são então ordenados de acordo com a freqüência na lista de eventos. Para freqüências iguais, utiliza-se uma ordem de importância pré-estabelecida. Por exemplo, ataque aéreo tem maior importância que abastecimento.

Outro exemplo é o sistema SUMMARIST (Hovy and Lin, 1997), que combina conhecimento simbólico com técnicas estatísticas para sumarizar um documento ou conjunto de documentos. O sistema utiliza diferentes técnicas de ambas as abordagens para identificar o tópico de um documento, interpretá-lo e gerar o sumário.

Quaisquer que sejam a metodologia e o foco da sumarização automática, passaram a ocupar papel de destaque, ainda, os critérios de avaliação, tanto de desempenho dos sumarizadores automáticos, quanto da qualidade dos próprios sumários produzidos. Tornaram-se freqüentes, ainda, as pesquisas sobre sumarização multi-documentos, cujo objeto de estudo é a produção automática de sumários de um assunto, explorando-se vários textos-fonte que discorrem sobre o mesmo.

Nos capítulos seguintes, descrevemos algumas técnicas e métodos bastante conhecidos de sumarização automática pelas abordagens superficial e profunda.

4. Abordagem Superficial

Nesta seção, daremos exemplos de alguns métodos ou sistemas associados à abordagem superficial, sem, contudo, pretender sermos exaustivos. Antes, nossa intenção é indicar as possibilidades de leitura adicionais.

4.1. Principais Métodos Explorados nas Décadas de 70 e 80

❖ Método das Palavras-Chave

Este método parte do pressuposto que as idéias principais de um texto podem ser expressas por algumas palavras-chave. Segundo Black e Johnson (1988), conforme as idéias vão sendo desenvolvidas no texto, os termos-chave aparecem com maior freqüência. A idéia é, então, determinar a distribuição estatística das palavras-chave do texto e, a partir de sua freqüência, extrair as sentenças que as contenham, agrupando-as de forma a constituir um sumário, na ordem em que aparecem originalmente.

O programa escrito por Luhn (1958), por exemplo, cria sumários utilizando este método. Dentre as palavras-chave encontradas, o programa considera somente aquelas de classe aberta, ou seja, as que carregam significado[1].

[1] Aquelas cuja categoria lexical se encontra no conjunto de substantivos, verbos, advérbios e adjetivos licenciados na língua natural em foco.

Earl (1970) apresentou uma variação deste método, tomando como hipótese a observação de que os substantivos mais freqüentes de um texto são responsáveis pela maioria de suas palavras-chave, sendo, portanto, mais representativos do conteúdo textual. Ele privilegiou substantivos em relação a verbos, por exemplo, porque eles poderiam indicar a progressão temática do texto. Por exemplo, na sentença "A casa desmoronou", a palavra "casa" seria mais representativa que o verbo "desmoronou", porque indicaria o tópico sobre o qual se fala. Caso o autor decidisse mudar o foco para o evento, propriamente dito, introduzido nesse exemplo pelo verbo "desmoronar", seria necessário nominalizá-lo e, portanto, o substantivo continuaria sendo a forma-chave mais representativa do conteúdo textual. Assim, uma possível progressão temática mudando o foco de "casa" para "desmoronar" seria construída, por exemplo, pela sentença "O desmoronamento ocorreu devido aos ventos fortes".

- **Método das Palavras-Chave do Título**

Este método é uma variação do anterior, diferindo daquele no sentido de que somente o título do texto a ser sumarizado é considerado para se buscar as palavras-chave que irão nortear a sumarização. Caso o texto contenha subtítulos ou, mesmo, um sumário, também estes poderiam ser considerados para a extração das palavras-chave. Para aplicá-lo, parte-se do pressuposto de que o título do texto é bem formado e, por si só, já expressa o conteúdo textual (o que, normalmente, é o caso, quando se tem bons títulos). A hipótese principal, neste método, é que as palavras componentes de um título têm maior probabilidade de serem representativas do tópico principal do texto correspondente.

A diferença básica desse método em relação ao anterior é que a lista de palavras-chave que indicarão as sentenças relevantes para um sumário incluirá as palavras-chave do título. Ainda, a distribuição de sentenças selecionadas será normalizada, para que aquelas cujas palavras-chave também ocorram no título tenham maior peso e, portanto, sejam escolhidas incondicionalmente para compor o sumário (Edmundson, 1969).

- **Método da Localização**

Baxendale (1958) também verificou que a posição de uma sentença em um texto poderia ser associada à sua importância no contexto textual. Por exemplo, a primeira e a última sentença de um parágrafo podem conter suas idéias principais e, portanto, estas seriam as sentenças consideradas para a produção de um sumário. Baxendale mostrou que, em 85% de uma amostra de 200 parágrafos, a sentença topical era a primeira e, em 7%, a última. Nos 8% restantes, as informações relevantes encontravam-se entre a primeira e a última de um mesmo parágrafo. Apesar do resultado aparentemente pouco significativo de 7% associado à última sentença, o autor considerou que esta também é importante para o sumário, pois constitui o elo de ligação com o parágrafo seguinte e, portanto, com a primeira sentença deste parágrafo, de alta topicalidade no contexto textual. Assim, para que a coesão textual fosse mantida, Baxendale sugeriu que deveriam ser incluídas em um sumário tanto a primeira quanto a última sentença de cada parágrafo.

❖ Método das Palavras Sinalizadoras (*Cue Phrases*) ou dos Marcadores Lingüísticos

Similarmente aos dois métodos anteriores, este método também consiste em atribuir valores a sentenças de um texto, selecionando as de maiores pesos. O método faz uso de um dicionário composto por palavras consideradas relevantes no domínio do texto, previamente construído, a partir do qual as sentenças consideradas importantes têm pesos associados positivos. Sentenças cujas palavras não constem do dicionário têm pesos negativos (Paice, 1981). Assim, o peso de uma sentença passa a ser uma média ponderada entre valores negativos e positivos, conforme a especificação dicionarizada. Em um texto científico, por exemplo, as palavras *conclusões* e *resultados* serão, muito provavelmente, altamente significativas, estando presentes no dicionário. Sua ocorrência em alguma sentença implicará, conseqüentemente, o aumento da relevância da mesma, em relação à sua escolha para compor um sumário.

Textos de gêneros distintos teriam, correspondentemente, outros dicionários e, portanto, seus próprios marcadores da importância do conteúdo textual. Cada gênero, neste caso, delineia uma distribuição de marcadores distinta, a partir da qual a modelagem computacional pode resultar razoavelmente satisfatória para a sumarização automática. No entanto, a correspondência entre os dicionários de marcadores e os diferentes gêneros textuais não é biunívoca e, assim, as intersecções podem levar a imprecisões na seleção dos componentes de um sumário (por exemplo, quando os marcadores dicionarizados sequer aparecem no texto-fonte).

Vale notar que, muito embora este método se assemelhe ao método das palavras-chave, eles diferem porque o dicionário, neste método, não é composto por palavras-chave ocorrentes no texto-fonte, mas por palavras consideradas significativas como marcas da relevância de outros componentes textuais.

❖ Método Relacional

Também chamado de método adaptativo, o método descrito por Skorokhod'ko (1972) é também considerado um método baseado em conhecimento (Black & Johnson, 1988). Ele utiliza uma representação gráfica do texto, na qual relações entre sentenças são criadas dependendo da relação semântica entre suas palavras. Sentenças semanticamente relacionadas a um grande número de outras sentenças, cuja delição levaria a uma maior dificuldade no entendimento do texto, recebem um peso alto e, portanto, são mais prováveis de serem escolhidas para extração e composição do sumário. A relação semântica entre as sentenças é identificada, por exemplo, pela ocorrência de substantivos comuns. Outras heurísticas de natureza similar também são definidas para se identificar as sentenças a serem selecionadas.

❖ Método da Frase Auto-Indicativa

Este método, descrito por Paice (1981), faz uso de um indicativo explícito de quão significativa é a sentença a ser selecionada para compor o sumário. Uma frase auto-indicativa apresenta uma estrutura cuja ocorrência é freqüente no texto e indica explicitamente que a sentença se refere a algo importante sobre o assunto do texto. Exemplos de frases auto-indicativas são *O objetivo deste artigo é investigar...*; *Neste artigo, é descrito um método para…*, etc. De acordo com esse método, as frases auto-indicativas permitiriam somente a produção de sumários indicativos, ou seja, aqueles que ajudam a identificar o tópico de um texto, sem, no entanto, discorrer sobre o mesmo. Este método é similar ao método dos marcadores lingüísticos, pois também

depende, de certa forma, do gênero textual. No entanto, não trata de palavras (unitárias) de conteúdo do domínio do texto-fonte, propriamente, mas de frases relativamente genéricas entre diversos gêneros textuais que irão indicar o conteúdo relevante.

Os métodos superficiais descritos acima, muito embora tenham levado a resultados interessantes para alguns tipos ou gêneros textuais, muito freqüentemente levavam a sumários mal construídos (desconexos e/ou incoerentes) devido à justaposição de sentenças extraídas de seu contexto original. Um dos maiores problemas encontrados, por exemplo, era o da resolução anafórica: por serem métodos "cegos", sem qualquer resolução analítica, os processos automáticos não distinguiam quando era necessário recuperar o contexto e os possíveis pares referentes/referenciados de sentenças inter-relacionadas anaforicamente, antes de isolar algumas delas de seu contexto. Desse modo, tais referências se perdiam, com reflexos altamente prejudiciais para os sumários resultantes.

Os pesquisadores tentaram resolver tais problemas sem muito sucesso. Por exemplo, Black & Johnson (1988), ao utilizar o método da frase indicativa, selecionavam também algumas sentenças anteriores à de interesse, quando esta contivesse pronomes anafóricos (considerando, obviamente, que o referenciado teria ocorrido previamente ao referente). O escopo de investigação, neste caso, remetia somente ao início do mesmo parágrafo da sentença de interesse, isto é, eles não consideravam que a anáfora poderia se estender além de um parágrafo.

Já Paice controlava a escolha de sentenças de um sumário usando três valores limitantes: os dois primeiros correspondendo, respectivamente, ao tamanho mínimo e máximo estimados para o sumário; o terceiro sendo o tamanho desejado, computado a partir dos outros dois. Esses parâmetros podiam ser ajustáveis no sistema. Por exemplo, inicialmente cada extrato podia consistir de uma única sentença indicativa. Caso houvesse referências anafóricas, as sentenças relacionadas à anáfora eram adicionadas ao sumário. Esse processo era repetido para cada sentença que contivesse uma anáfora, até que cada referência fosse resolvida ou até que o limite superior para o tamanho do sumário fosse alcançado. Neste caso, o algoritmo tentava "neutralizar" as referências restantes delitando-as ou modificando-as.

As regras de Paice para a escolha de sentenças com referências anafóricas envolviam o reconhecimento e o tratamento das anáforas, que são indiscutivelmente problemas complexos do PLN, extrapolando o uso de técnicas superficiais para a sumarização automática, pois dependem da interpretação textual. Para tentar contornar essa complexidade, Paice ainda detalhou quando seria possível reconhecer, por exemplo, um pronome anafórico. Assim, em "Nossas investigações mostraram que isto é verdade.", o pronome demonstrativo "isto" é anafórico, cujo referenciado deve ser identificado no próprio texto. No entanto, pronomes de mesma categoria, em outros contextos, não o seriam. Este é o caso de "Este artigo apresenta o método...", onde "este" não indica uma referência anafórica. Paice determinou que nenhuma palavra componente de frases indicativas poderia ser anafórica. Outra delimitação que ele impôs ao seu sistema, específica para a língua inglesa, diz respeito à ocorrência da palavra *this*: esta só pode ser anafórica se o referenciado ocorrer em uma janela de tamanho oito, isto é, entre as oito primeiras palavras da mesma sentença.

Decisões dessa natureza, no entanto, continuaram remetendo à necessidade de se considerar métodos mais informados para a sumarização automática, tópico este que será abordado mais adiante. Antes disso, porém, vamos apresentar variações de métodos empíricos, exploradas na última década, que podem resultar em avanços significativos em relação aos descritos acima.

4.2. A Sumarização Automática e a Mineração de Textos (*Text Mining*)

Na década de 90, com a disponibilidade de grandes quantidades de dados e os avanços tecnológicos para sua manipulação eletrônica, houve novamente um interesse bastante grande pela aplicação de métodos estatísticos em larga escala. Esses métodos são comumente chamados métodos baseados em *corpora* e têm demonstrado que é possível, para alguns gêneros textuais ou domínios particulares do conhecimento, obter-se sumários bons e úteis para algum objetivo previamente estabelecido. *Corpora on-line* impulsionaram novos métodos para estudos em uma variedade de áreas, além da sumarização automática, como aquisição de conhecimento lexical, revisão gramatical, tradução automática e *data mining*.

Corpora considerados relevantes para a investigação empírica em diversas áreas incluíram, inicialmente, o *Brown Corpus* (Francis e Kucera, 1982) e o *Birmingham Corpus* (Sinclair, 1987), ambos com mais de um milhão de palavras. No entanto, atualmente, esses *corpora* já foram superados por outros *corpora* textuais. O *University Centre for Computer Corpus Research on Language* (UCREL), por exemplo, um centro de pesquisas da Universidade de Lancaster, reúne 21 *corpora*, totalizando cerca de 26 milhões de palavras, além de inúmeros textos em CD-ROM (http://www.comp.lancs.ac.uk/computing/research/ucrel/). Outros *corpora* atualmente bastante explorados são aqueles desenvolvidos especialmente para tarefas dedicadas do PLN, como os utilizados nas TREC (*Text REtrieval Conference*) ou nas MUC (*Message Understanding Conference*).

Esses *corpora*, também utilizados na área de *Data Mining*, motivaram o surgimento da área de *Text Mining*, ou *Text Data Mining* (TDM – Hearst, 1999), área de bastante interesse para as áreas de Recuperação da Informação e Sumarização Automática. Enquanto em *Data Mining* contempla-se dados estruturados, em *Text Mining* busca-se o estudo das relações existentes entre componentes de textos não estruturados. Esse inter-relacionamento pode ser interno, isto é, relativo a apenas um texto, ou externo, abrangendo vários textos, dependendo do objetivo da aplicação. Para a recuperação da informação, por exemplo, é interessante buscar informações representativas e, portanto, distintivas de algum texto em particular. Portanto, para distingui-lo, é necessário investigar vários textos. Para a sumarização automática, no entanto, é importante identificar informações relevantes em um dado contexto textual e, portanto, basta contemplar um texto por vez – aquele que será sumarizado.

Em *Text Mining*, utiliza-se também grande volume de dados, agora textuais, a fim de se buscar padrões relevantes por comparação. A dificuldade, neste caso, é lidar com a rica gama de informações, em geral desestruturadas e, assim, de difícil recuperação automática.

Estamos interessados, particularmente, em como *Text Mining* pode ser utilizada na sumarização automática. Assim, uma das primeiras abordagens remete, naturalmente, à obtenção de palavras ou sentenças-chave de um texto para a composição do sumário correspondente. Veremos, a seguir, como essa questão é tratada no âmbito da mineração de dados textuais, descrevendo alguns sistemas em particular.

❖ O Sistema ANES

O sistema ANES (*Automatic News Extraction System*) (Rau e Brandow, 1993) é um aplicativo para sumarização automática de textos jornalísticos, produzindo sumários por meio da extração de sentenças. Ele utiliza a freqüência de palavras para identificar,

em um texto, palavras relativamente únicas ao documento e que, portanto, tendem a indicar os tópicos ou carregar informações importantes.

A partir de um conjunto de documentos, o sistema compara a freqüência das palavras de cada documento à freqüência de palavras em um conjunto de treinamento. O sistema guarda seu número de ocorrências, a partir da qual a palavra recebe um peso, relativo também ao número de ocorrências no *corpus* de treinamento. Palavras que ocorrem mais freqüentemente no documento do que no conjunto de treinamento recebem um peso maior. Isso equivale à medida chamada TF-IDF (*Text Frequency-Inverse Document Frequency*), utilizada na Recuperação de Informações.

A medida TF-IDF é derivada da estatística e se baseia na freqüência de termos. No método das palavras-chave, as palavras mais freqüentes de um texto são consideradas representativas. Há, no entanto, o cuidado de não se considerar as palavras de domínio fechado, como artigos ou pronomes, que não carregam significado. Assim como essas palavras são muito freqüentes sem, no entanto, serem relevantes ou expressarem informações topicais, há palavras que aparecem muito constantemente em diversos textos, não sendo, portanto, úteis para expressar a individualidade do texto. A medida parte do princípio de que uma palavra será representativa em um texto se ocorrer diversas vezes no texto em questão e for pouco freqüente em outros textos.

Usando essa medida, palavras cujo peso indicam que carregam informações importantes são separadas em uma lista, chamada lista de *signature words*. São também adicionadas a esta lista palavras do título, ainda que pouco freqüentes. Esta lista contém, assim, as palavras que são relativamente únicas ao documento. Em seguida, o peso das sentenças é calculado a partir do peso das palavras que a compõem. Ou seja, o peso resultante será a soma dos pesos das palavras da sentença que também estiverem presentes na lista de *signature words*.

O sistema ANES utiliza também algumas métricas básicas para a seleção das sentenças que vão compor o extrato, tais como a localização da sentença no documento, a presença de palavras anafóricas, o tamanho desejado para o extrato e o tipo de extrato a ser gerado. Para aumentar a legibilidade e tentar assegurar a coesão, eventualmente a primeira ou a primeira e segunda sentenças de um parágrafo são adicionadas, caso a segunda ou terceira, respectivamente, contenham *signature words*.

Na fase formal de testes, foram obtidos resultados que indicaram um desempenho de 74,4%, o que foi considerado satisfatório. No entanto, em comparação com um método que consistia em extrair apenas as primeiras sentenças de notícias o sistema ANES apresentou resultados inferiores. Os autores consideram, então, que, apesar dos resultados indicando métricas promissoras, eles podem melhorar se forem utilizadas outras técnicas, por exemplo, atribuir pesos maiores a sentenças pertencentes ao começo do texto.

❖ Ferramenta de *Text Mining* para Sumarização Automática e *Clustering*

O trabalho desenvolvido por Larocca Neto et. al. (2000) consiste de uma ferramenta para *text mining* que realiza as tarefas de *clustering* de documentos e sumarização textual. Conforme o sistema ANES que descrevemos, o algoritmo de sumarização é também baseado na freqüência inversa de palavras, com algumas modificações. A principal peculiaridade do método é que a sumarização pode ser aplicada a um texto específico, sem a necessidade de um grande *corpus* de textos.

Nos casos cuja freqüência das palavras é primordial, é necessário formatar o texto para que palavras iguais com formatação diferente não sejam erroneamente consideradas palavras distintas. Do mesmo modo, várias designações de um mesmo

verbo devem ser consideradas em conjunto. Portanto, inicialmente o texto de entrada sofre um pré-processamento, consistindo de três passos:

1) *Case Folding*

Consiste em converter todos os caracteres do documento para um mesmo formato, ou seja, letras maiúsculas ou minúsculas.

2) *Stemming*

Consiste em converter cada palavra à sua forma radical, ou seja, sua forma neutra sem inflexões verbais ou de número.

3) Remoção de *Stop Words*

Stop Words são palavras de classe fechada, ou seja, que não carregam significado, tal como artigos e pronomes. A remoção é feita verificando a presença em uma lista de 524 *stop words*.

Os procedimentos 2) e 3) se aplicam para a língua inglesa. No caso de algoritmos específicos para línguas diferentes do inglês, é utilizado um processo baseado em n-grams. Assim, alternativamente, o sistema aplica a representação *n-gram* em vez de executar esses passos. Um *n-gram* é uma parte de uma palavra consistindo de n caracteres. Por exemplo, a palavra DATA pode ser representada pelos trigramas *_DA, DAT, ATA, TA_* ou pelos bigramas _D, DA, TA, entre outros.

O algoritmo de sumarização representa cada sentença como um vetor de pesos TF-ISF (*Term Frequency – Inverse Sentence Frequency*), métrica similar à TF-IDF. A diferença é que a noção de documento é substituída pela noção de sentenças. Assim, a importância de uma palavra w em uma sentença s, denotada por TF-ISF(w,s), é calculada através da fórmula

$$\text{TF-ISF}(w,s) = \text{TF}(w,s)*\text{ISF}(w)$$

onde TF(w,s) é o número de vezes que a palavra w ocorre na sentença s, e a freqüência inversa da sentença é obtida através da fórmula

$$\text{ISF}(w) = \log(|S|/\text{SF}(w))$$

onde a freqüência sentencial SF(w) é o número de sentenças nas quais a palavra w ocorre.

Em seguida, para cada sentença s, o peso médio de cada sentença, denominado Avg-TF-ISF(s) é calculado através da média aritmética dos pesos TF-ISF(w,s) de todas as palavras w pertencentes à sentença, ou seja:

$$\text{Avg-TF-ISF}(s) = \sum_{i=1}^{W(s)} \frac{\text{TF-ISF}(i,s)}{W(s)}$$

onde W(s) é o número de palavras na sentença s.

Finalmente, são selecionadas as sentenças com os maiores valores de Avg-TF-ISF(s), baseadas em um valor mínimo (*threshold*) definido pelo usuário.

Em testes, a ferramenta apresentou boas taxas de compressão. De um texto-fonte com 75 sentenças, por exemplo, ela produziu um sumário com 10 sentenças. Os autores também consideraram que os sumários apresentam alta qualidade, capturando as idéias

principais do texto original. No entanto, o método também apresenta problemas, quando a extração de sentenças leva a sumários não coesos e, muitas vezes, incoerentes.

4.3. Uma Ilustração da Aplicação do Método de Palavras-Chave para a Sumarização Automática

O exemplo apresentado aqui foi extraído de um trabalho de Iniciação Científica cujo principal objetivo é investigar a aplicabilidade do processamento estatístico de dados textuais à sumarização automática. Foi usado o *WordSmith* (http://www.liv.ac.uk/~ms2928/wordsmit.html) para o processamento textual, cujas ferramentas estatísticas permitem obter a distribuição de freqüência de componentes textuais, listas de palavras-chave de determinado dado ou, mesmo, o contexto de ocorrência de palavras especificadas pelo usuário. A partir dos resultados gerados automaticamente, foram construídos sumários (manualmente) seguindo-se algum dos métodos superficiais apresentados na seção anterior. A metodologia adotada, assim como um exemplo que evidencia o objetivo de tal trabalho, são ilustrados nesta seção.

- **Seleção de Amostra Textual**

O *WordSmith* requer uma amostra de dados sobre o qual se possa proceder ao tratamento estatístico. Foi utilizada, nesta etapa, uma amostra de textos do *corpus* do NILC[2], atualmente o maior *corpus* disponível da língua portuguesa (com, aproximadamente, 37 milhões de palavras). Por questão de praticidade, optou-se por textos corrigidos de tamanho médio (cerca de uma página) no domínio de *esportes*.

O *corpus* simplificado consiste de 20 textos, dentre os quais se encontra o TEXTO 1 ilustrado abaixo.

TEXTO1:

Clair está pronto para ir a Atlanta
Corredor catarinense se preparou com afinco para ganhar corrida de domingo

Do clima ameno de Santa Catarina para o calor do Rio. Esta foi a trajetória de Clair Wathier, um dos 12 maratonistas brasileiros com índice olímpico que participarão domingo da Maratona do Rio - uma realização da Confederação Brasileira de Atletismo, Comitê Olímpico Brasil e Prefeitura do Rio, com patrocínio da Antarctica e promoção do JORNAL DO BRASIL. A mudança de temperatura, no entanto, não é um obstáculo intransponível para o atleta, de 29 anos. "Estou muito bem preparado. Respeito meus adversários, mas confio muito em mim. Espero estar em Atlanta".

Clair está se preparando desde novembro de 95 para essa maratona e acredita que se o domingo for muito úmido e quente será ruim para todos. "É uma incógnita, mas será uma prova totalmente tática. A decisão virá após os 30 quilômetros", opina o atleta de São Miguel do Oeste, que está treinando em Petropólis para se acostumar aos poucos

[2] NILC – Núcleo Interinstitucional de Lingüística Computacional de São Carlos.

com o clima do Rio. "Estou tomando todos os cuidados necessários e fazendo uma alimentação rica em carboidratos e proteínas".

Dono do sexto melhor tempo (2h14min10) entre os 12 atletas com índice, Clair acredita que seus principais adversários serão Valdenor Pereira dos Santos, Daniel Ferreira e Luís Carlos da Silva. Há 13 anos participando de corridas de rua, o corredor catarinense confia em sua experiência internacional: foi campeão nas maratonas de Buenos Aires, Karpi (Itália), Florença (Itália) e Vigarano (Itália). Na Maratona do Rio, Clair tem uma preocupação extra. "Não quero me desgastar demais, porque, caso me classifique, tenho de estar em forma para os Jogos Olímpicos".

Kits - Para retirarem os kits que usarão no dia da corrida, os corredores deverão comparecer ao Hotel Luxor Continental, na Rua Gustavo Sampaio, no Leme, de hoje até sexta-feira, das 12h às 21h. No sábado, o horário de atendimento ao público é das 10h às 20h. O kit inclui a camiseta da prova, o número do corredor e as duas senhas para entrada obrigatória no dia da corrida.

No sábado, às 16h, somente para os atletas de elite e seus técnicos, haverá o congresso técnico, em que serão dadas as informações quanto aos procedimentos a serem tomados na prova.

❖ Extração de Informações Relevantes do TEXTO1

A partir do software *WordSmith* gerou-se uma lista de todas as palavras do TEXTO1, com suas respectivas freqüências, bem como uma lista de palavras-chave para esse texto. São utilizadas duas ferramentas distintas do *WordSmith* com esse fim: para a geração das palavras-chave, é usado um *corpus* de referência que, a partir da distribuição de freqüência simples das palavras componentes do texto, indica quais as mais relevantes para serem consideradas representativas do texto, propriamente dito e, portanto, figurarem na lista de palavras-chave.

A partir desses dados estatísticos, utilizou-se o método das palavras-chave para a geração manual de sumários do TEXTO1.

❖ Sumário do TEXTO1

O Sumário1 ilustra a aplicação do método das palavras-chave ao TEXTO1. A lista de palavras-chave apontada pelo *WordSmith* indicava, com alto grau de freqüência, a palavra CLAIR como representativa do texto original, dentre outras. Esta foi a palavra escolhida para a elaboração do Sumário1, que foi realizada simplesmente extraindo-se, do texto original, as sentenças que continham tal palavra, mantendo-se a mesma ordem textual.

SUMÁRIO1:

Clair está pronto para ir a Atlanta

Esta foi a trajetória de Clair Wathier, um dos 12 maratonistas brasileiros com índice olímpico que participarão domingo da Maratona do Rio - uma realização da Confederação Brasileira de Atletismo, Comitê Olímpico

Brasil e Prefeitura do Rio, com patrocínio da Antarctica e promoção do JORNAL DO BRASIL.

Clair está se preparando desde novembro de 95 para essa maratona e acredita que se o domingo for muito úmido e quente será ruim para todos.
Dono do sexto melhor tempo (2h14min10) entre os 12 atletas com índice, Clair acredita que seus principais adversários serão Valdenor Pereira dos Santos, Daniel Ferreira e Luís Carlos da Silva.
Na Maratona do Rio, Clair tem uma preocupação extra. "Não quero me desgastar demais, porque, caso me classifique, tenho de estar em forma para os Jogos Olímpicos".

- **Análise do SUMÁRIO1**

A partir da leitura do SUMÁRIO1, pode-se notar, inicialmente, que logo no primeiro parágrafo ocorre um segmento de texto *non-sequitur*[3], devido à existência do pronome demonstrativo 'Esta', cujo termo referenciado, existente originalmente, encontra-se ausente no sumário. De fato, este termo correspondia à primeira sentença do texto original, que foi omitida do sumário por não conter a palavra-chave CLAIR. Apesar de haver outros problemas, que se referem sobremaneira a escolhas lexicais que interferem no estilo do sumário e poderiam ser mais bem elaboradas, o restante do mesmo apresenta uma boa progressão temática. A ausência de progressão temática também pode ser notada em alguns segmentos textuais, sendo este um problema que também pode implicar a ocorrência de textos *non-sequitur*, conforme apontam pesquisas anteriores (refira-se à Seção 3).

Comparando o SUMÁRIO1 com o texto original, nota-se que foram retiradas deste (a) a segunda sentença do primeiro parágrafo; (b) a primeira sentença do segundo parágrafo; (c) a primeira e a última sentença do terceiro. Com exceção de (a), que caracteriza o problema de *non-sequitur* já descrito, os demais parágrafos são distribuídos da seguinte forma:

1. Segundo parágrafo: da segunda sentença em diante, temos detalhes da primeira. Logo, os detalhes foram compreensivelmente extraídos do sumário.
2. Terceiro parágrafo: De modo análogo ao anterior, a segunda sentença é detalhamento da primeira e, logo, também é justificável sua exclusão do sumário. A última sentença, por sua vez, repete o nome do esportista CLAIR, e, assim, pode refletir a importância da informação textual. De fato, introduz-se nessa sentença uma nova informação (e, portanto, a mudança de tópico): a preocupação extra de CLAIR. Justifica-se, assim, sua inclusão no sumário.
3. Quarto e quinto parágrafos: Totalmente excluídos do sumário, a justificativa se encontra no fato destes apresentarem informações complementares para a participação na maratona, sendo que elas não indicam a idéia central do sumário.

Vale notar que essa análise é, de certa forma, subjetiva, podendo haver outros sumários a partir da escolha de outras palavras-chave igualmente relevantes ou de outras interpretações para o mesmo texto. Entretanto, ela corrobora critérios de sumarização

[3] Textos caracterizados pela ocorrência de "lacunas" de raciocínio, sem garantia de progressão temática

apresentados por outros autores (por exemplo, Hoey, 1983, Hutchins, 1987 ou Sparck Jones, 1993b).

Do ponto de vista de aproveitamento estatístico na sumarização automática, esse exemplo mostra que é possível produzir-se um sumário bastante razoável a partir das sentenças apontadas como chaves no texto. Ele tem tamanho razoável e contém, de uma maneira geral, apenas informações relevantes e representativas, quando comparado ao original. Alguns problemas, no entanto, foram encontrados.

Uma análise mais profunda do texto ilustrado indica, por sua vez, que o casamento direto do padrão *CLAIR* com os componentes textuais exclui os casos em que outros itens lexicais se referem ao próprio esportista de nome *CLAIR*. Seria de se supor que, para a análise da relevância das informações a compor o sumário, está em consideração a entidade denotada pelo item lexical e não o item lexical, propriamente dito. Neste caso, denotações distintas para a mesma entidade *CLAIR* deveriam também ser consideradas no sumário, o que não ocorre quando se utiliza o método ilustrado, que segue única e exclusivamente o padrão determinado pela palavra-chave escolhida. Para que qualquer sumário fosse construído levando-se em conta uma análise de relevância mais ampla, seria necessário que se considerasse um processo de interpretação textual e, portanto, uma metodologia mais "informada" que a adotada nesse experimento e mesmo nas abordagens superficiais, que são cegas, como processos computacionais.

Assim como ilustra o exemplo acima, procedeu-se à geração (manual) de outros sumários para esse mesmo texto, assim como de sumários para outros textos selecionados no domínio de esportes, a partir de palavras-chave apontadas pelo WordSmith como relevantes para cada caso. O mesmo tipo de análise lingüística e de conteúdo ilustrado acima foi realizado para cada um dos sumários obtidos, comparando-os com seus textos-fonte correspondentes. O objetivo principal dessa tarefa foi verificar se as idéias principais dos textos originais poderiam ser reproduzidas nos sumários construídos com base na distribuição estatística dos componentes textuais, assegurando-se que os resultados automáticos seriam de boa qualidade. No entanto, alguns problemas de natureza lingüística, por exemplo, a ocorrência de textos *non-sequitur*, notadamente podem prejudicar a compreensão, apontando para a necessidade de se adotar técnicas que dêem conta de características mais profundas, quer relacionadas à coerência, quer relacionadas à coesão textual.

4.4. Os Métodos Superficiais: Sumário

Apesar de tentarem resolver o problema da falta de coesão textual, que implicará, quase que certamente, na falta de coerência, as alternativas apontadas acima eventualmente levam a sumários mais longos do que o desejado. Além disso, ainda não garantem a coesão, tampouco a coerência, caracterizando os textos *non-sequitur*). Este problema, crucial na produção textual, não é resolvido até agora por qualquer um dos métodos superficiais de sumarização automática apresentados, uma vez que eles se baseiam única e exclusivamente na noção de relevância, em função da posição e do contexto textual, que, por sua vez, são, muitas vezes, função do gênero textual. Assim, mapear as informações que, em determinado contexto, possam ser mais relevantes do que outras, ainda constitui o maior problema para a construção de uma ferramenta computacional que distinga o conteúdo a compor sumários de textos-fonte. Mesmo na sumarização humana, tais noções são, muito freqüentemente, dificilmente exprimíveis, resultando em uma grande dificuldade para a automação do processo. É por essa razão que, em geral, faz-se necessário considerar também a abordagem profunda do PLN, lançando mão da interpretação textual, para obter-se uma distribuição estrutural que

indique, por outros mecanismos, quando alguma informação é relevante para a preservação do significado textual no sumário correspondente. Este caso será discutido a seguir.

5. Abordagem profunda

Dentre as limitações de alguns dos métodos superficiais abordados na seção anterior, pode-se citar, por exemplo, as seguintes:

- O método do título não garante um bom sumário, se considerarmos que muitos textos não possuem títulos significativos;
- O método das palavras-chave pode gerar um sumário cujo tamanho não seja significativo em relação ao texto-fonte, se usado sozinho;
- O método da localização não garante um sumário representativo, devido à impossibilidade de se prever com clareza a posição em que uma informação relevante irá se encontrar, no texto-fonte.

Esses métodos tomam decisões de condensação normalmente limitadas apenas à identificação, seleção e exclusão/extração de segmentos textuais. Em oposição a eles, a abordagem profunda contempla o conhecimento lingüístico e/ou extralingüístico associado ao texto de origem, a fim de compor seu(s) possível(is) sumário(s). Esse conhecimento envolve, por exemplo, as relações semânticas e retóricas, no nível lingüístico, ou as relações intencionais, no nível extralingüístico, as quais serão mapeadas no modelo lingüístico, computacional, na maioria das vezes envolvendo a manipulação simbólica.

Trabalhando com base nas relações discursivas subjacentes ao texto-fonte, modelos de sumarização automática podem lançar mão de informações em nível profundo para garantir a produção de sumários coerentes e coesos. Sob essa perspectiva, a sumarização automática passa a ser um processo de geração textual com restrições de condensação e preservação do significado apontado pela fonte textual, ou seja, da mensagem do texto-fonte.

Considerando-se que todo texto produzido por um escritor tem uma proposição principal, que compõe a mensagem que o texto deve transmitir ao leitor, a idéia na abordagem profunda é (Rino, 1996): a) identificar, no texto-fonte, qual sua proposição central; b) identificar quais as informações complementares à proposição central que, no sumário, poderão contribuir para a transmissão da mensagem e para a preservação da idéia principal do texto original e c) permitir a estruturação do sumário com base em (a) e (b) e em restrições e normas de coerência e coesão, para que o resultado final seja, também, um texto, no sentido lingüístico (Couture, 1985). Entretanto, identificar e manipular computacionalmente os aspectos e recursos necessários para a realização de (a-c) implicam processos complicados, já que mesmo a mensagem é uma entidade complexa, composta por vários conceitos, além da proposição central de um texto.

O segmento de texto abaixo, por exemplo, pode ter como proposição central a informação de que o livro *Viagem ao Centro da Terra*, de Júlio Verne, é um bom livro (na linguagem simbólica de Primeira Ordem, algo como PC1 = livro(X) & bom(X)). No entanto, diferentes leitores podem apreender, a partir desse mesmo texto, diferentes proposições centrais, implicando diferentes acepções e, portanto, mensagens diversas transmitidas pelo mesmo texto. Por exemplo, uma segunda proposição para esse texto poderia, muito bem, ser que o livro de Júlio Verne merece ser comprado (PC2 =

livro(X) & autoria(X,"Júlio Verne") & indicacao_compra(X,alta)), estando esta proposição implícita no texto ilustrado.

> "O livro Viagem ao Centro da Terra é um dos grandes best-sellers em todo o mundo, permitindo que o leitor entre em um mundo nunca imaginado e se delicie com as aventuras fictícias e cheias de surpresas pelas quais algumas pessoas passam ao tentarem explorar um novo mundo localizado na própria Terra. Envolvente e encantador, este livro pode ser encontrado em qualquer livraria perto da sua casa..."

Além da decisão dicotômica de expressar claramente ou deixar implícita a proposição central do texto, o escritor pode também optar por sugeri-la de forma mais rebuscada, utilizando a própria trama do discurso. Neste caso, a proposição central pode estar "diluída" no texto – ou distribuída em vários de seus segmentos – tornando, assim, mais difícil seu reconhecimento. Qualquer um desses recursos é utilizado pelo escritor com base no efeito que ele deseja produzir no leitor e, portanto, em função dos objetivos comunicativos que ele pretende alcançar com seu texto. No caso do exemplo acima, dois são os possíveis objetivos comunicativos relacionados às proposições centrais PC1 e PC2: OC1 = informar [que a obra Viagem ao Centro da Terra é boa] ou OC2 = persuadir [o leitor a comprar o livro], com as correspondentes implicações na linguagem simbólica:

$$\text{livro}(X) \;\&\; \text{bom}(X) \rightarrow \text{informar}(\text{bom}(X)).$$

$$\text{livro}(X) \;\&\; \text{autoria}(X,\text{"Júlio Verne"}) \;\&\; \text{indicacao\_compra}(X,\text{alta}) \rightarrow \text{persuadir}(\text{leitor},\text{comprar}(X)).$$

Além desses objetivos comunicativos, outros podem ser derivados a partir do conteúdo informativo do texto ou, como ilustra esse exemplo, a partir, unicamente, das possíveis proposições centrais pretendidas pelo autor. Por exemplo, PC1 pode implicar OC2 ou, ainda, PC2 pode implicar OC1 e, portanto, poderemos ter também

$$\text{livro}(X) \;\&\; \text{bom}(X) \rightarrow \text{persuadir}(\text{leitor},\text{comprar}(X)).$$

ou

$$\text{livro}(X) \;\&\; \text{autoria}(X,\text{"Júlio Verne"}) \;\&\; \text{indicacao\_compra}(X,\text{alta}) \rightarrow \text{informar}(\text{bom}(X)).$$

Sob essa perspectiva, para a produção textual passa a ser necessário considerar tanto a proposição central a ser transmitida, quanto o objetivo comunicativo a ser atingido com o texto final. Neste sentido, fica evidente que a mensagem a ser transmitida é composta por, no mínimo, dois conceitos distintos: a proposição central que o autor pretende veicular e seu objetivo comunicativo, com o texto produzido. Ainda, ambos os conceitos podem variar, sendo que a diversidade da proposição central do texto pode implicar a diversidade de objetivo comunicativo. Contudo, o inverso também pode ser verdadeiro, isto é, a diversidade de objetivos comunicativos pode levar o autor a estabelecer diferentes proposições centrais, escolhendo-as no elenco de possíveis informações que contribuirão para ressaltar seus objetivos.

Desse modo, uma mensagem associada a um texto-fonte contém, ainda, um terceiro componente, indicado no elenco de informações disponíveis para a composição textual. Consideraremos, assim, que uma mensagem-fonte de um sumário é expressa a)

pela proposição central do texto-fonte; b) por seu objetivo comunicativo e c) por seu conteúdo informativo, ao qual denominaremos base de conhecimento do sumário (Rino, 1996). Sob essa perspectiva, construir um sumário consiste em preservar a mensagem-fonte [do texto original], a qual deve ser reconhecida por meio da análise lingüística e estrutural do texto-fonte[4].

Nas seções seguintes são apresentados os principais problemas da abordagem profunda, em função de sua dependência de uma mensagem-fonte a ser preservada, assim como uma arquitetura clássica de sumarização textual. São descritos, ainda, alguns modelos lingüísticos para a sumarização automática.

5.1. A Sumarização Automática na Abordagem Profunda

O problema de se construir sumários automaticamente remete às máximas de Grice (1975), a saber:

- Qualidade: informar precisamente somente o que pode ser evidenciado no texto-fonte;
- Quantidade: dizer exatamente o que se requer, nem mais, nem menos;
- Relevância: transmitir somente o que é relevante, dependendo da meta comunicativa e do conhecimento do leitor;
- Modo: evitar obscuridade e ambigüidade e escrever de forma ordenada e breve.

Na sumarização automática, tais máximas relacionam-se, ainda, a outras medidas, tais como:

- Grau de abstração: grau com que o escritor ignora detalhes;
- Clareza: relacionada ao esforço de compreensão do leitor;
- Grau de detalhe: relacionado à quantidade e qualidade da informação transmitida;
- Grau de explicitação: grau de evidência das informações e de seu inter-relacionamento;
- Nível de informatividade: relacionado à referência a unidades significativas de conteúdo.

Essas características, por sua vez, estão relacionadas à observação (humana) de que não necessariamente um sumário mais conciso é obscuro, tampouco um mais explicitamente marcado[5] garante clareza ou é de melhor qualidade. Um sumário menos marcado, por exemplo, pode exigir maior esforço de compreensão, pela necessidade de inferência do que ficou implícito. Essas questões envolvem tanto o conhecimento do domínio quanto a competência lingüística do leitor e estão vinculadas, ainda, a considerações de gênero e tipo de texto em foco, os quais, por sua vez, caracterizam eventos comunicativos particulares, que podem ser evidenciados por recursos de linguagem ou recursos estruturais especiais, além de formas lingüísticas características do gênero ou da tipologia sob enfoque.

A distinção entre gênero e tipo de um texto é delicada, sendo o entendimento de ambos polêmico até mesmo no meio literário. Entretanto, é consenso que o tipo textual

[4] Como observamos na Seção 2.2., sumários podem ser construídos, que apresentem objetivos comunicativos diversos de seus textos-fonte correspondentes. Ainda assim, eles devem ter um forte elo com os originais, que permita reconhecê-los como formas condensadas daqueles.

[5] Isto é, com marcadores superficiais das relações de discurso subjacentes.

pode ser caracterizado como uma subclasse do gênero, pelo fato dele envolver aspectos comunicativos, além de estruturais e informacionais. Exemplos de gênero e tipo de textos incluem (Rino, 1996):

- Gênero: poesia, prosa, ficção, novelas, estórias, instruções, artigos científicos, relatórios, anúncios, piadas, seminários, informativos, receitas, etc;
- Tipo: persuasivo, expositivo, analítico, descritivo, dissertativo, narrativo, argumentativo, etc.

A consideração do gênero e tipo de textos na sumarização automática pode ser de grande valia para ajudar a reconhecer a estrutura textual e a especificar um modelo de processamento automático correspondente. Tal metodologia é utilizada, por exemplo, para a especificação de esquemas (McKeown, 1985) e heurísticas de condensação que podem ser tipificadas para um determinado domínio de textos (por exemplo, no discurso científico no domínio da Física, como sugerem Rino e Scott (1994)).

Essas características devem, portanto, ser consideradas para a geração de sumários bons, coerentes e coesos, que reflitam adequadamente o texto original.

Uma das abordagens profundas, para a sumarização automática, está em distinguir o que é relativo a decisões de ordem estrutural – e, portanto, de coerência – e o que é relativo a decisões de ordem superficial – e, portanto, de coesão. Sob esse ponto de vista, coerência e coesão são propriedades complementares, cada uma delas remetendo a um nível distinto de representação textual (ou discursiva) e ambos contribuindo para a transmissão da mensagem que o autor pretende veicular com seu texto.

Propriedades coesivas servem, por exemplo, para evidenciar relações estruturais, definidas no nível profundo, no nível superficial. Sua forma de expressão, freqüentemente, é refletida por marcadores do discurso. Por exemplo, para a relação de Causa-Efeito entre as proposições relativas a *'Choveu forte'* (causa) e *'eu fechei a janela'* (efeito), a realização superficial pode ser dada pela junção de duas sentenças, como segue:

"*Choveu forte. Portanto, eu fechei a janela.*"

ou

"*Fechei a janela porque choveu forte.*"

cujos marcadores discursivos são, respectivamente, portanto e porque.

Essa mesma relação pode ser expressa sem os marcadores explícitos, como em

"*Choveu forte. Eu fechei a janela.*"

Ainda assim, o sentido próprio da relação de Causa-Efeito é recuperável, desde que leitor e escritor partilhem o mesmo conhecimento de mundo, que permita ao leitor inferir a relação discursiva pretendida pelo escritor. O marcador da relação causal, neste caso, serve como elemento coesivo fatual explícito.

Vê-se, pela discussão acima, que são vários os fatores que interferem nas escolhas estruturais e superficiais durante a produção textual, tanto humana quanto automática. Ao se considerar a abordagem profunda, o inter-relacionamento entre medidas nebulosas como clareza, abstração, informatividade, coerência e coesão, conceitos intuitivos para os seres humanos, é de difícil incorporação e manipulação em um sistema computacional, pois remetem ao nível profundo, de estruturação e garantia

da coerência textual, assim como ao nível superficial, de expressão adequada do nível profundo e, portanto, da correta manipulação das informações lingüísticas disponíveis para se produzir o texto final, propriamente dito, em função da estruturação imposta pela necessidade de comunicação da mensagem pretendida.

Uma possível arquitetura para essa abordagem de dois níveis, na sumarização automática, é ilustrada na próxima seção.

5.2. Uma Arquitetura de um Sistema de Sumarização Automática

Segundo Rino (1996), a sumarização automática pode ser caracterizada como um sistema de geração de textos tradicional, no qual são consideradas, prioritariamente, restrições de sumarização. A diferença entre a geração e a sumarização, neste caso, está no fato de, na geração, considerar-se todo o conteúdo informativo, para dizer tudo da melhor forma possível, enquanto que, na sumarização, deve-se dizer o mínimo em relação ao texto-fonte, cuidando para que não haja perda de seu significado essencial (Hovy, 1988).

Como um processo de geração textual, a sumarização automática pode ser caracterizada também classicamente, pelos três processos principais abaixo:

- Seleção do conteúdo: escolha das informações a serem expressas no texto final;
- Organização (ou planejamento textual) do conteúdo: estruturação das informações a serem expressas;
- Realização gramatical: expressão lingüística das informações selecionadas.

A Figura 1 ilustra essa arquitetura[6]. Tal sistema pode ser implementado de três formas:

- Em *pipeline*: o início de um processo se dá após o término do outro. Assim, seleção, planejamento e realização são realizados em seqüência;
- Intercalado: há uma certa interação entre planejamento e realização, sendo que somente um dos dois processos é ativado por vez;
- Integrado: o planejamento e a realização são interdependentes e são continuamente ativados.

[6] As siglas denotam o seguinte: ‘M’ - Mensagem; ‘ED’ – Estrutura de Discurso; ‘S’ – Sentença [final].

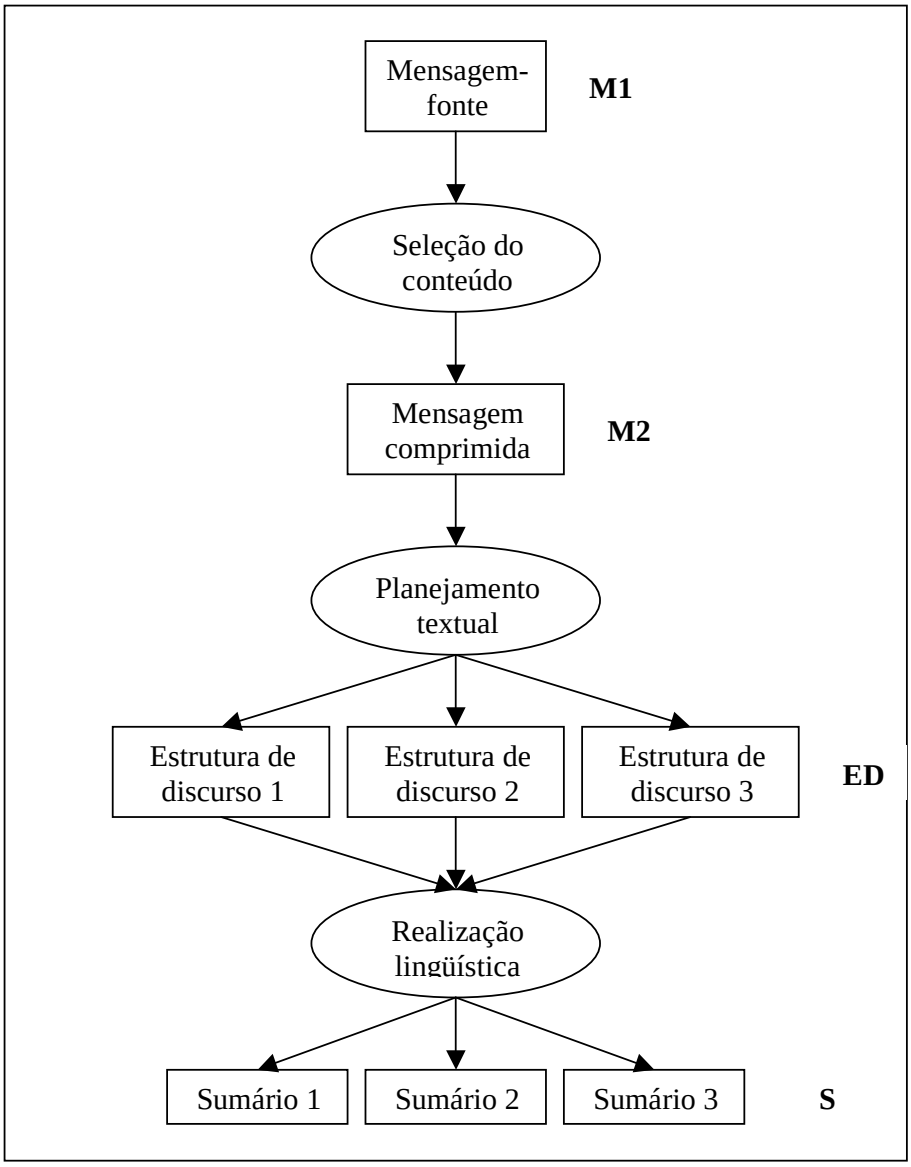


Figura 1 – Arquitetura de um sumarizador automático de três passos

O seguinte exemplo ilustra os dados representados nesse sistema, após cada um dos processos assinalados. Por motivos de simplificação, excluímos dessa arquitetura o processo de interpretação do texto-fonte, considerando que a entrada do sistema já é sua mensagem-fonte correspondente (M1).

1. Texto-fonte: "*O menino beijou a menina bonita. O menino ficou feliz.*"
2. Resultado do processo de interpretação:

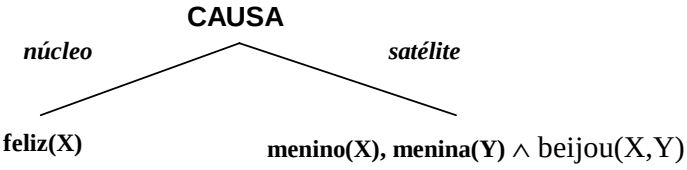
M1: ɜ (X,Y) ((menino(X) & menina(Y) & bonita(Y) & beijou(X,Y)) → feliz(X))

3. Resultado do processo de seleção do conteúdo:

M2: ɜ (X,Y) ((menino(X) & menina(Y) & beijou(X,Y)) → feliz(X))

4. Estrutura de discurso possível:

ED:



5. Resultado final:

S: "*O menino beijou a menina e ficou feliz.*"

Nesse exemplo, o formalismo escolhido para representar as mensagens (M) resultantes da interpretação do texto-fonte é o Cálculo de Predicados de Primeira Ordem (Clocksin and Mellish, 1981). A estrutura de discurso será explicada mais adiante, bastando entender agora que ela indica a relação causal entre o fato do menino beijar a menina e ficar feliz por esse motivo.

Como mostra a mensagem comprimida (M2), foi eliminado, de M1, o qualificativo 'bonita(Y)', relativo a 'menina(Y)'. Uma possível heurística de seleção de conteúdo pode ser, assim, *'elimine qualquer qualificativo de uma entidade do discurso'*. Outras heurísticas similares a essa podem ser definidas para a seleção do conteúdo. Certamente, é necessário haver um modelo de domínio no sistema, juntamente com seu léxico, para que se distingam qualificativos ou propriedades de entidades conceituais que possam servir a heurísticas dessa natureza. Esse modelo, assim como o léxico, deverão ser consultados durante a seleção do conteúdo, mediante a base de heurísticas de seleção. Nos demais níveis de processamento, podem existir outros procedimentos, determinísticos ou não, algoritmizáveis ou baseados em heurísticas, que sirvam também aos propósitos da sumarização automática. A diferença se dará pelo tipo de informação contemplada em cada nível que, certamente, implicará conjuntos de procedimentos distintos.

A sumarização de M1 se baseou tanto no número de palavras quanto no número de sentenças. Entretanto, de modo geral, a sumarização pode se refletir também no número de segmentos de discurso (que são unidades de significado mínimo em um texto[7]) e no número de relações retóricas (p.ex., *causa-efeito*, no exemplo acima), as quais indicam as relações existentes entre os segmentos de discurso. O uso de relações de tal natureza é necessário para que se permita a estruturação coerente de textos em segmentos de discurso.

O processo de realização lingüística, como já mencionamos, é responsável por expressar a estrutura de discurso resultante do processo de planejamento textual (ED) em sua forma lingüística (S). Esse processo pode se dar de várias formas, a saber (Reiter and Dale, 2000; Scott and Souza, 1990):

- Pela utilização de sistemas computacionais já disponíveis, por exemplo, FUF (Elhadad, 1991), que utiliza *f-structures* (Kaplan e Bresnan, 1982) como

[7] Segmento difere de sentença no sentido de que uma sentença pode conter vários segmentos de discurso.

representação de entrada do realizador, ou Nigel (Mann and Matthiessen, 1985), que utiliza um modelo sistêmico de representação discursiva (Halliday, 1985);

- Utilizando-se *case frames*: normalmente, dependentes do mapeamento semântico possível, para a produção de estruturas morfossintáticas licenciadas pela língua natural em foco. Os *case frames* são, portanto, função das relações retóricas do discurso, mapeando em funções semântico-sintáticas da língua;
- Por *templates* ou *canned texts*: estruturas rígidas (mais rígidas que as anteriores), especificadas em função das possíveis formas gramaticais da língua em foco e preenchidas a partir das propriedades semânticas, estruturais, indicadas na estrutura de discurso a ser realizada lingüisticamente.

Como exemplo da utilização de um modelo de estruturação de discurso para a sumarização automática, apresentamos, a seguir, a Teoria de Estruturação Retórica, ou *Rhetorical Structure Theory* (RST).

5.3. A Teoria RST

❖ Principais Características da RST

A Teoria RST (Mann and Thompson, 1987), inicialmente formalizada para a análise lingüística de textos e, portanto, para a interpretação, é atualmente a teoria de coerência mais influente nos modelos de PLN para a geração textual (Reiter and Dale, 2000), servindo, conseqüentemente, também para a sumarização automática. Pelo menos teoricamente, ela define um conjunto de relações de coerência – chamadas relações retóricas ou relações RST – capaz de representar todas as possíveis relações existentes entre os segmentos de um texto. Há controvérsias a respeito da completude do conjunto de relações RST originalmente definido por Mann e Thompson. Vários outros autores modificaram, complementaram ou discutiram sua especificação (por exemplo, Marcu (1996) e Hobbs (1985)), buscando maior clareza para fins computacionais. A dificuldade de se definir relações dessa natureza reside no caráter interpretativo e subjetivo da língua, posto que remete à apreensão do significado. Além disso, tais relações supostamente devem cobrir todas as possíveis relações estruturais na produção textual.

Uma relação RST possui um núcleo (N) e um satélite (S), sendo N composto pelo segmento conceitual que indica a informação principal da relação e S, a informação secundária, que influencia de alguma forma a interpretação que o leitor faz da informação contida no núcleo. A definição de cada relação pressupõe a existência de quatro tipos de informação necessários, quer para determiná-la durante um processo de interpretação textual, quer para decidir pelo uso de uma relação particular durante a geração textual. Essas informações são definidas na forma de *templates*, cujos campos são os seguintes:

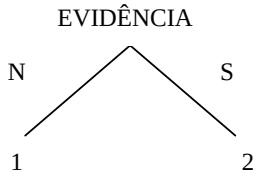
- Restrições sobre o núcleo (N);
- Restrições sobre o satélite (S);
- Restrições sobre a combinação do núcleo e do satélite;
- Efeito: especificação do efeito que a relação em questão causa no leitor quando este interpreta um texto ou efeito pretendido pelo escritor, ao selecionar tal relação para compor seu texto, relacionando N e S.

Algumas das relações RST são ilustradas a seguir. Detalhes sobre o conjunto completo, segundo a Teoria RST, podem ser encontrados na obra de referência.

Seguem, ainda, textos ilustrativos de cada uma dessas relações, subdivididos por proposições, isto é, cada segmento numerado corresponde a uma proposição no nível conceitual (e, portanto, profundo) de representação do discurso.

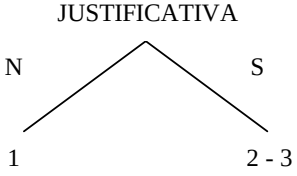
Nome da relação: EVIDÊNCIA
Restrições sobre N: o leitor pode não acreditar em N com um grau de satisfação qualquer. *Restrições sobre S*: o leitor acredita em S ou o acha válido. *Restrições sobre N + S*: a comparação do leitor sobre S aumenta a crença do leitor em N.
Efeito: a crença do leitor em N aumenta.

1. O programa realmente funciona.
2. Em somente alguns minutos, eu consegui acessar todos os meus dados.



Nome da relação: JUSTIFICATIVA
Restrições sobre N: nenhuma. *Restrições sobre S*: nenhuma. *Restrições sobre N + S*: a compreensão do leitor sobre S aumenta a prontidão do leitor para aceitar o direito do escritor de apresentar N.
Efeito**:** a prontidão do leitor para aceitar o direito do escritor de apresentar N aumenta.

1. O próximo espetáculo musical será agendado para o dia 21 de julho.
2. Darei mais detalhes posteriormente,
3. mas esta é uma boa hora para reservar um lugar em seu calendário.



Nome da relação: CONCESSÃO
Restrições sobre N: o escritor tem uma visão positiva sobre a situação apresentada em N. *Restrições sobre S*: o escritor afirma que a situação apresentada em S não se verifica. *Restrições sobre N + S*: o escritor mostra uma incompatibilidade aparente ou potencial entre as situações apresentadas em N e em S; o escritor estima as situações apresentadas em N e em S como compatíveis; o reconhecimento da compatibilidade entre as situações apresentadas em N e em S aumenta a estima do leitor em relação à situação apresentada em N.
Efeito: a estima do leitor em relação à situação apresentada em N aumenta.

1. Darei mais detalhes posteriormente,
2. mas esta é uma boa hora para reservar um lugar em seu calendário.

Como se pode notar pelos exemplos acima, núcleos e satélites podem corresponder a proposições elementares ou a outros segmentos textuais mais complexos, envolvendo proposições elementares relacionadas gradativamente por meio de outras relações RST. Desse modo, um texto estruturado segundo o modelo RST passa a ser uma árvore hierárquica com qualquer nível de profundidade e complexidade. Quando se tratar de um processo interpretativo, pode haver várias árvores (ou planos) RST resultantes, dependendo da interpretação[8].

❖ O Uso da RST na Sumarização Automática

Entre as relações RST, algumas são consideradas mais fracas do que outras, no sentido de contribuírem pouco para o enunciado ao qual se relacionam, servindo apenas para fornecer-lhe mais detalhes. Por exemplo, as relações CIRCUNSTÂNCIA, que envolve as noções de tempo e lugar, e ELABORAÇÃO, que fornece detalhes adicionais sobre seu núcleo informativo. Devido a essa observação, tais relações passaram a ser consideradas como uma forma de se sumarizar: uma vez identificadas em uma estrutura de discurso, as relações fracas podem indicar os segmentos de discurso supérfluos e, portanto, passíveis de exclusão do texto original. Hobbs (1985), por exemplo, sugere o corte integral de tais relações, assim como de seus segmentos subordinados, da estrutura de discurso, enquanto Sparck Jones (1993) sugere a seleção das relações SUMÁRIO e, talvez, INTERPRETAÇÃO e REFORMULAÇÃO, para formar o conteúdo do sumário. Sugestões dessa natureza, comuns no início da década de 90, não levaram a bons

[8] Devido à possibilidade de cada leitor ter a sua própria interpretação textual, diferente da interpretação de outros leitores.

resultados, pois segmentos significativos do texto-fonte (estruturado como uma árvore RST) eram desconsiderados na composição do sumário.

O corte de relações RST inteiras e/ou a integração de núcleos das relações, somente, pode implicar a perda da coerência do texto, com a geração de parágrafos muito curtos, a perda ou o excesso de pontuação e a existência de expressões referenciais ou marcadores de discurso cujos referentes se perdem no texto resultante. Este problema, de certa forma, já havia sido observado com o uso de metodologias superficiais. Na abordagem profunda baseada na representação RST, a grande deficiência para a produção de bons sumários se encontra em sua formalização incompleta, que se deve a vários fatores, dentre os quais vale a pena ressaltar o fato de algumas relações (por exemplo, LISTA e JUNÇÃO[9]) não serem claramente especificadas para que possam servir aos propósitos de sumarização automática e o fato da RST ter sido projetada para a interpretação e não para a geração, o que acarreta problemas adicionais para a produção textual automática, cujas decisões estruturais precisam ser especificadas a *priori*, para que possibilitem escolhas de expansão da estrutura de discurso. Isto já não ocorre na interpretação, cuja derivação da árvore RST é dependente da estrutura do texto-fonte, já disponível.

Além desses problemas, existe um outro, inerente da própria formulação RST, apontado por Scott e Souza (1990), relacionado à realização superficial. Mais precisamente, às possibilidades de se marcar as relações de discurso na superfície textual, como no exemplo que elas fornecem, para os segmentos de discurso A: "*my car is not British*" e B: "*my car is a Renault*", que podem estar relacionados pela relação RST EVIDÊNCIA, tomando-se A como núcleo e B como satélite: EVIDÊNCIA(A,B). De acordo com a definição desta relação, as seguintes construções lingüísticas se tornam possíveis: "*Since my car is not British, it is a Renault.*" ou "*My car is not British, therefore it is a Renault.*" Claramente, tais expressões, embora licenciadas por definição, são inadequadas, porque contradizem o sentido lógico.

Tentando sanar tais dificuldades, mais recentemente as propostas de O'Donnel (1997) e Marcu (1996, 1997) têm demonstrado alguns avanços na abordagem profunda baseada na RST. O'Donnel sugere somente a exclusão dos satélites de uma estrutura de discurso, em função de uma lista de relações de discurso que indicam detalhes (e, portanto, informações que podem ser excluídas de um sumário). Marcu, por outro lado, propõe um modelo mais elaborado de manipulação das relações RST, utilizando a nuclearidade das relações juntamente com seu efeito prescrito (pela teoria), procurando, assim, determinar melhor as informações relevantes para o sumário.

Rino (1996) também propõe um modelo de sumarização automática com base na teoria RST, porém, ela sugere um refinamento da representação discursiva com base nas intenções subjacentes ao discurso e na caracterização semântica da mensagem-fonte, considerando que um sumário deve manter, no mínimo, a proposição central e o objetivo comunicativo inalterados em relação a seu fonte. A caracterização intencional desse modelo é baseada na Teoria GSDT, que é descrita a seguir.

❖ Um exemplo do Uso de Intenções na Sumarização Automática: A Teoria GSDT

Considerando-se que a estruturação de um discurso consiste na etapa crucial da geração textual, posto que é nesta fase que se deve garantir o inter-relacionamento adequado e consistente entre unidades proposicionais (Hobbs, 1985; Mann and Thompson, 1987), Rino propõe um modelo de sumarização automática no qual a

[9] Relações inseridas em versões mais recentes da RST.

estruturação do discurso é baseada na determinação de relações de coerência (ou relações RST) a partir das intenções subjacentes ao discurso, sendo que estas são definidas em função dos segmentos informacionais em foco, para a produção do sumário correspondente.

O modelo de intenções adotado é o de Grosz e Sidner (1986) – a Teoria GSDT (*Grosz and Sidner Discourse Theory*) – que prescreve o relacionamento entre unidades básicas ou segmentos do discurso com base nas intenções subjacentes a cada um dos segmentos. Neste modelo, são essas intenções que individualizam o discurso e o tornam coerente, sendo importantes, assim, para justificar a estrutura do discurso, assim como sua própria origem.

Segundo a GSDT, qualquer discurso é uma composição de três componentes distintos, que interagem entre si, a saber:

- A estrutura lingüística: seqüência de termos e cláusulas do discurso;
- A estrutura de intenções: intenções dos segmentos do discurso e suas relações;
- O foco de atenção: informação sobre os objetos, propriedades, relações e intenções do discurso que é mais saliente em um determinado momento.

Esses três componentes juntos fornecem a informação necessária para que qualquer escritor/leitor determine como um segmento de discurso se encaixa no resto do mesmo, identificando a razão de ter sido expresso, assim como seu significado.

A estrutura lingüística é composta pelos segmentos do discurso que preenchem alguma função em relação ao discurso como um todo. Apesar da segmentação do discurso ser auxiliada por marcadores lingüísticos, não se tem uma metodologia exata para identificar as funções discursivas, pois a segmentação depende da interpretação que, por sua vez, é subjetiva.

A estrutura de intenções contém as intenções de cada segmento do discurso, assim como a especificação do modo como ele pode se relacionar com as intenções dos demais segmentos. Apesar do leitor não conseguir controlar todo o conjunto de intenções subjacentes a um texto (que, teoricamente, é infinito), ele deve reconhecer seu inter-relacionamento por meio de algumas relações características, definidas na GSDT como segue:

- Relação de dominância (DOM): indica qual a intenção que contribui para o entendimento de outra.
  Forma de expressão: I1 *DOM* I2
  Leitura: I2 contribui para o entendimento de I1; I1 domina I2

- Relação de satisfação/precedência (SAT-PREC): indica qual a intenção que satisfaz outra e, portanto, que deve precedê-la.
  Forma de expressão: I1 *satisfaction-precedes* I2
  Leitura: I1 deve ser satisfeita antes de I2

O foco de atenção, diferentemente das duas relações acima, é uma estrutura que contém a informação em foco em cada momento do discurso, sendo esta o repositório central da informação contextual, necessário para processar os segmentos do discurso a cada instante de desenvolvimento e, assim, garantir a progressão temática.

Considerando-se uma abordagem profunda na qual o que importa é a) identificar os segmentos relevantes para a composição de um sumário e b) identificar como cada segmento pode contribuir para a estruturação do sumário, (a) dependerá, substancialmente, do relacionamento possível entre cada segmento e o segmento

indicador da proposição central do discurso, de modo a expressar a escala de relevância em relação a essa proposição central e (b) consistirá em b.1) identificar a intenção subjacente a cada segmento e b.2) o modo como cada intenção de cada segmento pode contribuir para a satisfação do objetivo comunicativo central do discurso em elaboração. É possível, assim, identificar as partes relevantes para compor o sumário, levando em consideração as intenções subjacentes ao discurso em relação à proposição central e ao seu objetivo comunicativo. Desse modo, em relação à arquitetura de um sumarizador automático conforme descrita pela Figura 1, as informações essenciais e complementares da mensagem-fonte que serão consideradas na mensagem-fonte do sumário, assim como as informações supérfluas, que deverão ser descartadas, serão identificadas com base em regras intencionais, definidas a partir da GSDT.

No modelo de Rino, as regras intencionais são especificadas com base nas relações de dominância (DOM) ou precedência (SAT-PREC) existentes entre segmentos informativos, a partir do domínio em foco, isto é, da base de conhecimento que compõe a mensagem-fonte do sumarizador. Para uma certa mensagem-fonte (lembrando que esta é resultado de um processo interpretativo, fora do foco do sumarizador ilustrado), uma vez observada a distribuição intencional entre seus constituintes, faz-se possível um mapeamento entre essa estrutura intencional e a estrutura de discurso, que nada mais é do que uma árvore RST. Por exemplo, caso se verifique a relação Y DOM X e X SAT-PREC Y entre dois segmentos quaisquer X e Y de um texto, pode-se gerar a relação retórica MEANS na estrutura de discurso resultante. Neste caso, o segmento X poderia estar descrevendo um problema no texto-fonte, enquanto que Y uma possível solução para esse problema.

5.4. A Pragmática e a Sumarização Automática

Na metodologia profunda, vários aspectos podem ainda influenciar o tamanho e o conteúdo do sumário. Algumas características pragmáticas usadas na geração automática de textos podem ser aplicadas também para a sumarização, acarretando mudanças no sumário produzido de acordo com o modelo do usuário, a relação entre o falante e ouvinte e os objetivos do falante. Entre as características de estilo, Hovy (1988) explora, por exemplo, o tempo despendido para a preparação do discurso (*haste* – que pode ser classificado como *escasso*, *pouco*, *suficiente* ou *ilimitado*) e formalidade (*formality – coloquial*, *normal* ou *formal*), que são particularmente interessantes para se traçar a estrutura e conteúdo de um sumário.

Segundo Hovy, enquanto textos formais normalmente possuem sentenças longas, aqueles com baixa formalidade, ou coloquiais, apresentam sentenças mais curtas. Dessa forma, na sumarização, poderiam ser utilizadas algumas de suas regras para diminuir a formalidade e, conseqüentemente, o tamanho do sumário, controlando o seu grau de compressão.

Já as regras de tempo despendido para a preparação textual podem determinar quantos tópicos serão incluídos no sumário. Quanto menor o tempo, menos se deve incluir no sumário, seja devido à limitação de espaço ou devido ao nível de conhecimento do usuário ao qual o sumário é destinado. Por exemplo, se o usuário for especialista no assunto, pode ser atribuído o valor mínimo ao tempo (*escasso*) de forma a forçar a exclusão de detalhes no sumário a ser gerado. Essas regras envolvem ainda conceitos como distância e subordinação social entre falante e ouvinte, que influenciam também o número de detalhes e o nível de parcialidade do texto produzido, interferindo, conseqüentemente, no critério de sumarização também.

Como exemplo da interação entre as métricas de Hovy para a geração de sumários, caso se tenha um nível coloquial de formalidade e tempo escasso, o sumário será altamente condensado, contendo somente o tópico principal do discurso. A utilização do método inverso também é válida, por exemplo, se for necessário gerar um sumário com condensação alta, pode-se optar por uma formalidade coloquial e tempo escasso. A Tabela 1 sintetiza a interação entre as regras de estilo de Hovy e as implicações no conteúdo do sumário.

TEMPO / *FORMALIDADE*	***Escasso***	***Pouco***	***Suficiente***	***Ilimitado***
Coloquial	Sumarização alta; apenas tópico principal	Sumarização média	Sumarização média; tópico principal, detalhes desconhecidos	Sumarização baixa; tópico principal, detalhes relevantes
Normal	Sumarização média; tópico principal, poucos detalhes	Sumarização média; tópico principal	Sumarização média; tópico principal, detalhes importantes	Sumarização baixa; tópico principal, detalhes relevantes
Formal	Sumarização média; tópico principal.	Sumarização média; tópico principal	Sumarização baixa; tópico principal, detalhes importantes	Sumarização baixa; tópico principal e correlatos, detalhes relevantes

Tabela 1 – Interação entre regras de estilo e conteúdo do sumário

Com relação à realização lingüística dessas características textuais, Hovy também sugere regras. Por exemplo, caso se queira um texto formal com sentenças pesadas, deve-se utilizar cláusulas adverbiais, voz passiva e tempos complexos. As regras sugeridas para a característica formalidade são mostradas na Tabela 2.

Objetivo	**Estratégia**	**Regra**
Texto mais formal	Sentenças longas	Selecionar opções relacionadas com outros tópicos de sentenças.
	Sentenças complexas	Selecionar opções subordinadas em cláusulas relativas, que reúnam dois ou mais tópicos, justapostas em relações e frases multi-predicativas.
	Sentenças pesadas	Uso de cláusulas adverbiais, as quais são colocadas no início de sentenças, uso de voz passiva, tempos complexos, evitar elipse.
	Cláusulas pesadas/formais	Uso de adjetivos e cláusulas adverbiais, substantivos duplos, advérbios, *stress words*. Nominalizar verbos e advérbios. Pronominalizar. Não se referir diretamente ao interlocutor.

	Frases/palavras formais	Evitar expressões populares, contrações e gramática “duvidosa”.
Texto menos formal	Sentenças simples	Não reunir tópicos de sentenças, não justapor tópicos e evitar subordinação em cláusulas relativas.
	Sentenças curtas	Incluir no máximo uma cláusula adverbial por sentença (no fim do predicado), voz ativa, evitar tempos complexos, usar elipse.
	Cláusulas simples	Incluir no máximo um adjetivo, frases pequenas e simples, pronominalizar, usar verbos e advérbios em vez de suas formas nominais, referência direta a interlocutores.
	Frases/palavras informais	Selecionar palavras comuns, usar expressões populares e contrações.

Tabela 2 – Regras para formalidade

5.5. A Abordagem Profunda: Sumário

O tratamento profundo da sumarização traz grandes vantagens para o processamento computacional, permitindo capturar relações e conhecimentos até então normalmente ignorados pelas técnicas estatísticas. Tais conhecimentos, lingüísticos ou não, permitem determinar melhor o conteúdo informativo do sumário, fornecendo os recursos necessários para se levar em consideração o usuário e o objetivo a que o sumário se destina, além de ajudar a resolver os problemas discutidos na Seção 3. Entretanto, como já foi mencionado, a modelagem profunda encontra problemas para a definição, estruturação e manipulação do conhecimento profundo, dificuldades que podem motivar a adoção de técnicas superficiais, de mais fácil aplicação, apesar de resultados não necessariamente confiáveis.

6. Conclusões e Comentários

A sumarização automática tem se tornado uma área promissora de investigação, projeto e desenvolvimento pelas mais diversas razões, sendo uma das principais o fato de haver, atualmente, volumes imensos de dados disponíveis, principalmente *on-line*, que requerem do usuário uma capacidade de absorção “infinita” do conhecimento, além de uma enorme disponibilidade de tempo.

O grande volume de dados estimula as pesquisas tanto em virtude da necessidade de ferramentas que auxiliem a busca de informações, como também porque permite a análise estatística de grandes *corpora* de texto em busca de padronizações que levem a novos modelos para as áreas da Lingüística Computacional.

As pesquisas têm como objetivo final o de oferecer disponibilidade de recursos automáticos que permitam ao usuário a) acessar informações on-line; b) apreender a mensagem principal – ou o que for mais relevante – em relação ao fonte, de modo a manter-se atualizado e c) selecionar material relevante para uma leitura ou utilização adequada a seus objetivos rápida e eficientemente.

Nesse sentido, houve uma grande evolução no campo da sumarização automática e áreas relacionadas, tais como interpretação ou geração textual. Entretanto, há ainda muitos problemas que precisam ser solucionados para que a sumarização

automática de textos seja plenamente realizada, desde a extração de textos até a condensação de conteúdo. Para isso têm contribuído pesquisas recentes sobre identificação conceitual (Saggion and Lapalme, 2000), identificação de marcadores de discurso (Chan et. al., 2000), sumarização multi-documentos (Radev et. al., 2000; Goldstein et. al., 2000; Ando et. al., 2000), visualização de conteúdo (Ando et. al., 2000) e avaliação (Minel et. al, 1997; Oka e Ueda, 2000) e uso (Radev et. al.), bem como sobre técnicas e sistemas para sumarização automática propriamente ditos (Hovy and Lin, 1997; Mitra et. al, 1997, Aone et. al, 1997).

Analisando o estado da arte da área de sumarização automática, verificamos que, apesar dos avanços, não existe ainda uma única. Como vimos neste trabalho, a sumarização automática pode ser tratada em nível superficial e/ou profundo, mas cada qual apresenta tanto méritos quanto dificuldades. As últimas tendências estão em adotar metodologias híbridas, enfatizando a construção de sistemas voltados para a funcionalidade e para o usuário e aprimorando os métodos de avaliação.

Referências Bibliográficas

Ando, R. K., Boguraev, B. K. , Byrd, R. J. and Neff, M. S. (2000). Multi-Document Summarization by Visualizing Topical Content. In: *Proceedings of the ANLP/NAACL Automatic Summarization Workshop*, pp. 79-88. Seattle, Washington.

Aone, C., Okurowski, M. E., Gorlinsky, J. and Larsen, B. (1997). A Scalable Summarization System Using Robust NLP. In: I. Mani and M. Maybury (eds.) *Intelligent Scalable Text Summarization ACL 1997 Workshop*, pp. 66-73. Madrid, Spain.

Baxendale, P. B. (1958). Machine-made index for technical literature – an experiment. *IBM Journal of Research and Development*, 2, pp. 354-361.

Black, W. J. and Johnson, F. C. (1988). *EXPERT SYSTEMS FOR INFORMATION MANAGEMENT. A Practical Evaluation of Two Rule-Based Automatic Abstraction Techniques.* Department of Computation. University of Manchester Institute of Science and Technology. Volume 1. Number 3.

Chan, S. W. K., Lai, T. B. Y., Gao, W. J. and T'sou, B. K. (2000). Mining Discourse Markers for Chinese Textual Summarization. In: *Proceedings of the ANLP/NAACL Automatic Summarization Workshop*, pp. 11-20. Seattle, Washington.

Clocksin, W. and Mellish, C. (1981). *Programming in Prolog*. Springer-Verlag, Germany.

Correira, A. (1980). Computing Story Trees. *American Journal of Computational Linguistics* 8 (3/4), pp. 135-149.

Couture, B. (1985). Systemic Network for Analysing Writing Quality. In J.D. Benson and W.S. Greaves (eds.), *Systemic Perspectives on Discourse*, Vol. 2, pp. 67-78. Ablex Publishing Corporation.

Earl, L. L. (1970). Experiments in automatic abstracting and indexing. *Information Storage and Retrieval*, 6 (4), pp. 313-334.

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

Elhadad, M. (1991). *FUF: the Universal Unifier User Manual, Version 5.0*. Department of Computer Science, Columbia University, New York.

Endres-Niggemeyer, Brigitte (1990). A procedural model of abstracting and some ideas for its implementation. In *Second International Congress on Terminology and Knowledge Engineering*, pp. 230-243, vol.1. Indeks Verlag: Frankfurt, Germany.

Francis, W. e Kucera, H. (1982). *Frequency Analysis of English Usage*, Houghton Mifflin.

Fum, D. et al. (1982). Forward and Backward Reasoning in Automatic Abstracting. In: J. Horecky (ed.) *COLING 82*, pp. 83-88. North Holland: Amsterdam.

Goldstein, J., Mittal, V., Carbonelle, J. and Kantrowitz, M. (2000). Multi-Document Summarization By Sentence Extraction. In: *Proceedings of the ANLP/NAACL Automatic Summarization Workshop*, pp. 40-48. Seattle, Washington.

Grice, H. P. (1975). Logic and Conversation. In P. Cole and J. L. Morgan (eds.), *Syntax and Semantics. Volume 3: Speech Acts*, pp. 41-58. Academic Press, New York.

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

Halliday, M.A.K. (1985). *An Introduction to Functional Grammar*. Edward Arnold, London.

Hearst, M. A. (1999). Untangling Text Data Mining. In: *Proceedings of ACL´99, the 37th Annual Meeting of the ACL*, University of Maryland, USA.

Hobbs, J.R. (1985). *On the Coherence and Structure of Discourse*. Technical Report CSLI-85-37, Center for Study of Language and Information, Stanford University.

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

Hovy, E. (1988). *Generating Natural Language under Pragmatic Constraints*. Lawrence Erlbaum Associates Publishers, Hillsdale, New Jersey.

Hovy, E. and Lin, C. (1997). *Automated Text Summarization in SUMMARIST*. In: I. Mani and M. Maybury (eds.) *Intelligent Scalable Text Summarization ACL 1997 Workshop*, pp. 39-46. Madrid, Spain.

Hutchins, John. (1987). *Summarization: Some problems and Methods*. In: Jones. Meaning: The frontier of informatics. Cambridge. London, pp. 151-173.

Kaplan, Ronald M. and Bresnan, Joan (1982). Lexical-Functional Grammar: A Formal System for Grammatical Representation. In Joan Bresnan (ed.), *The Mental Representation of Grammatical Relations.* Cambridge, MA, MIT Press.

J. Klavans and P. Resnik (eds.) (1996). *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*. (Selected Papers of the 32nd Meeting of the ACL). Las Cruces, New Mexico.

Larocca Neto, J.; Santos. A.D.; Kaestner, C.A.A., and Freitas, A.A. (2000). Generating Text Sumaries through the Relative Importance of Topics. In: M. C. Monard and J. S. Sichman (eds.) *Lecture Notes in Artificial Intelligence*, (International Joint Conference 7th Ibero-American Conference on AI, 15th Brazilian Symposium on AI IBERAMIA-SBIA 2000), vol. 1952, pp. 300-309. Atibaia, São Paulo, Brazil.

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

Mann, W.C. and Matthiessen, C.M. (1985). Nigel: A Systemic Grammar for Text Generation. In R. Benson and J Greaves (eds), *Systemic Perspectives on Discourse: Selected Papers from the Ninth International Systemic Workshop*. Ablex. London, England.

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

Marcu, D. (1996). Building up rhetorical structure tress. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, volume 2, pp. 1069-1074, Portland, Oregon.

Marcu, D. (1997). *The rhetorical parsing, summarization, and generation of natural language texts*. Ph.D. Thesis, Department of Computer Science, University of Toronto.

Maybury, M. (1993). Automated Event Summarization Techniques. In: *Seminar Report of Summarizing Text for Intelligent Communication Seminar*. Dagstuhl, Germany.

McKeown, K.R. (1985). *Text Generation: Using Discourse Strategies and Focus Constraints to Generate Natural Language Text*. Cambridge University Press.

Minel, J., Nugier, S. and Piat, G. (1997). How to Appreciate the Quality of Automatic Text Summarization. In: I. Mani and M. Maybury (eds.) *Intelligent Scalable Text Summarization ACL 1997 Workshop*, pp. 25-30. Madrid, Spain.

Mitra, M., Singhal, A. and Buckley, C. (1997). Automatic Text Summarization by Paragraph Extraction. In: I. Mani and M. Maybury (eds.) *Intelligent Scalable Text Summarization ACL 1997 Workshop*, pp. 39-46. Madrid, Spain.

Mushakoji, S. (1993). Constructing 'Identity' and 'Differences' in Original Scientific Texts and Their Summaries: Its Problems and Solutions. *Seminar Report of Summarizing Text for Intelligent Communication Seminar*. Dagstuhl, Germany.

Mushakoji, S. and Nozoe, A. (1992). Toward qualified medical abstracts: Rethinking the process of producing author abstracts. In K.C. Lun et. al. (eds), *Medinfo'92*, Elsevier.

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

Oka, M. and Ueda, Y. (2000). Evaluation of Phrase-Representation Summarization based on Information Retrieval Task. In: *Proceedings of the ANLP/NAACL Automatic Summarization Workshop*, pp. 59-68. Seattle, Washington.

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

Radev, D. R., Jing, H. and Budzikowska, M. (2000). Centroid-based Summarization of Multiple Documents: Sentence Extraction, Utility-based Evaluation, and User Studies. In: *Proceedings of the ANLP/NAACL Automatic Summarization Workshop*, pp. 21-30. Seattle, Washington.

Rau, L. F. and Brandow, R. (1993). Domain-Independent Summarization of News. In: B. Endres-Niggemeyer, J. Hobbs e K. Sparck Jones (eds.), *Seminar Report of Summarizing Text for Intelligent Communication Seminar*. Dagstuhl, Germany.

Reiter, E. and Dale, R. (2000). *Building Natural Language Generation Systems*. Cambridge University Press.

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. and Scott, D. (1994). Automatic Generation of Draft Summaries: Heuristics for Content Selection. In A.I.C. Monaghan (ed.), *Proceedings of the Third International Conference on the Cognitive Science of Natural Language Processing*. Dublin City University, Ireland.

Sagion, H. and Lapalme, G. (2000). Concept Identification and Presentation in the Context of Text Summarization. In: *Proceedings of the ANLP/NAACL Automatic Summarization Workshop*, pp. 1-10. Seattle, Washington.

Sampson, G. (1987). Probabilistic models of analysis. In G. Leech, R. Garsude and G. Sampson (eds.), *The computational analysis of English*, pp. 16-29. Longman. Harrow.

Scott, D.R. and Souza, C.S. (1990). Getting the Message Across in RST-Based Text Generation. In R. Dale; C. Mellish and M.Zock (eds), *Current Research in Natural Language Generation*, pp. 47-73. London, Academic Press.

Sinclair, J., ed. (1987). *Looking Up: An Account of the COBUILD Project in Lexical Computing.* Collins.

Skorokhod´ko, E.F. (1972). Adaptive method of automatic abstracting and indexing. In: *IFIP Congress*, Ljubjana, Yugoslavia, vol. 71, pp. 1179-1182. Amsterdam. North Holland.

Sparck Jones, K. (1993a). *Discourse Modelling for Automatic Summarising*. Technical Report. No. 290, University of Cambridge.

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

van DIJK, T.A. (1979). Recalling And Summarizing Complex Discourse. In: Burghart, W. and Hölker, K. *Text Processing Textverarbeitung.* Berlin, Walter de Gruyter.