4
A sumarização automática

Para nos referirmos à sumarização automática, devemos situar a área da Linguística Computacional (LC) e seu campo de estudo. A LC vem desenvolvendo ou discutindo teorias sobre as línguas naturais humanas que possam ser testadas através de métodos computacionais e, por outro lado, provendo ou aproveitando-se de descrições destas línguas que sejam suficientemente detalhadas e objetivas de modo a que possam ser formalizadas para uso computacional. Nesse sentido, por um lado contribui para o conhecimento linguístico, e, por outro, beneficia-se do conhecimento gerado pela Linguística para suas aplicações. Trata-se de área interdisciplinar, que reune pesquisadores em linguística e em ciência da computação. Segundo Vieira & Lima (2001), além da pesquisa teórica, muitas pesquisas desenvolvidas na área de LC ou de Processamento de Linguagem Natural (PLN) tem o objetivo prático de "interpretar e/ou gerar informação fornecida em linguagem natural".

As aplicações de PLN vão de reconhecedores e sintetizadores da fala, corretores ortográficos e gramaticais, tradutores automáticos, sistemas de diálogo em linguagem natural, sistemas de recuperação de informação, até sistemas de geração de textos, dentre estes últimos sistemas de produção automática de resumos. Para nossa pesquisa nos interessam especificamente os geradores de resumo ou sumarizadores.

4.1
A geração de sumários automáticos

A necessidade de se gerar sumários automaticamente cresce com a propagação de informação via rede mundial de computadores. A sumarização automática é um processo que possibilita a apresentação das informações mais

importantes advindas de diversas fontes de informação e possibilita que o leitor descubra qual é o assunto do texto fonte correspondente e decida se quer aprofundar ou não sua leitura.

A meta mais ambiciosa no desenvolvimento de programas que gerem sumários automaticamente é a de chegar o mais próximo possível de um sumário humano e, para chegar a esse ponto teríamos de saber descrever exatamente como um leitor constrói um modelo mental do texto lido e como transforma esse modelo em um novo texto. Podemos com certeza afirmar que essa é uma ambição exagerada. Um objetivo mais factível seria o de gerar textos-resumo que contenham informações relevantes do texto fonte.

Conforme Martins et al. (2001), o interesse pela sumarização automática inicia-se já nos anos cinqüenta, quando utilizaram-se métodos estatísticos para extração das frases principais de um texto. Com a rápida revolução digital e a consequente explosão da quantidade de informação textual disponível na internet, que gera a necessidade de filtrar o que é informação relevante para um determinado leitor ou objetivo, os anos noventa viram crescer o interesse pela sumarização automática. A partir dessa necessidade sempre crescente, as tentativas de aprimoramento de programas sumarizadores resultaram em avanços na área de estudo. Para os autores, os sumários também são ferramentas especialmente úteis para a divulgação de textos e/ou conhecimento codificado em outras linguagens que podem, se apresentados resumidamente, ser acessados de forma rápida e eficiente. Mani & Maybury (2001) ressaltam ainda que o estudo da sumarização automática vem se diversificando em diversos campos, com o desenvolvimento de ferramentas que produzem resumos multilíngues, multimídia e multi-documentais.

A sumarização automática de textos é um processo que visa produzir uma versão mais curta de um texto fonte, que contenha suas idéias principais (MANY, I.; MAYBURI, M.T., 2001). Isto é, os sumários devem ter o máximo de informações relevantes no mínimo de espaço. Na concepção de Hutchins (1987, apud MARTINS et al., 2001), o sumário que mais se adapta à condição de um sumarizador automático

seria o do tipo indicativo ou descritivo, um sumário que seleciona informações (sentenças ou sintagmas) relevantes do texto fonte, mas não tem compromisso com a textualidade. Servem basicamente de indicadores ou indexadores de conteúdos presentes no texto fonte. Para o autor, os demais tipos de sumário são muito complexos para se adequarem à sumarização automática.

O resultado do processo de sumarização pode ser um extrato (do termo *extract*, em inglês) ou um sumário (do termo *abstract*, em inglês): o primeiro corresponde à simples justaposição de sentenças do texto fonte consideradas importantes, preservando a ordem original do mesmo. O extrato é um texto mais curto e que possui as informações mais importantes do original. O segundo tipo, por sua vez, altera a estrutura e/ou o conteúdo das sentenças originais, fundindo-as e/ou reescrevendo-as, para generalizar ou especificar as informações do texto fonte. O resultado desse processo é uma estrutura linguístico/textual diferenciada daquela do texto fonte (LEITE et al., 2001).

Os autores também afirmam que, para a produção de sumários automáticos de qualidade, há necessidade de compreensão textual. Apesar dos avanços na área, os sumarizadores atuais não são capazes de representar a textualidade com a eficiência do sumário humano. O uso de um modelo de compreensão e de discurso poderiam ajudar a melhorar a textualidade dos sumários obtidos automaticamente. Abre-se aí, ao nosso ver, uma área de pesquisa que precisa ser explorada. Acreditamos, com Mani & Maybury (2001, p.7), que qualquer alternativa de pesquisa que possa trazer contribuição pertinente ao estudo da sumarização automática é válida.

Martins et al (2001) destacam dois métodos de abordagem na estruturação de programas para a elaboração de sumários automáticos: o que contempla a estrutura profunda do texto e o que leva em consideração apenas sua estrutura superficial.

Os autores dizem que na abordagem profunda é levado em consideração todo o processamento da linguagem, incluindo-se os processos cognitivos, como base para a compreensão de um texto fonte e a geração de seu sumário. Nessa perspectiva, um sistema de PLN seria fundamentado em conhecimento linguístico, habilidades de

inferência lógica e conhecimento de mundo. Isto é, um sistema computacional desse nível implica um estágio de desenvolvimento distante daquele em que a pesquisa se encontra, mesmo para a descrição desses fenômenos e de sua integração na mente humana.

Na prática, tem-se desenvolvidos programas que se baseiam mais em métodos superficiais do que em profundos, devido à dificuldade inerente a esses métodos. Os métodos superficiais, destacam os autores, são mais simples e não necessitam de algorítmos complexos. Chamados de "cegos", eles não consideram todo o arsenal linguístico disponível na mente humana, e se utilizam de técnicas estatísticas para sumarizar.

Há programas que contemplam ambas as metodologias, usando um método híbrido. Como exemplo do uso do método híbrido, Martins et al. destacam o sumarizador de eventos desenvolvido por Maybury (1993), que seleciona, dentre as mensagens de um programa simulador de batalhas, as mais importantes, baseado na freqüência, singularidade e importância do evento. Por exemplo, numa batalha, a uma mensagem que reporte um ataque inimigo será atribuído peso maior do que a outra que relate evento não tão crucial. O SUMMARIST (HOVY e LIN, 1997) é outro sistema híbrido que utiliza técnicas das duas abordagens.

Dentre os principais métodos desenvolvidos entre as décadas de 70 e 80 que usam a abordagem superficial são destacados pelos autores o método das Palavras-Chave, que parte do pressuposto de que algumas palavras indicam as idéias principais de um texto. O programa determina a distribuição estatística das palavras-chave e, de acordo com sua freqüência, extrai as sentenças que as contenham, juntando-as ordenadamente em um sumário. Além da frequência no texto, dá-se mais relevância às palavras que carregam significado, como os substantivos, descartando-se outras, por exemplo, palavras gramaticais, como não indicativas de conteúdo do texto.

O método das Palavras-Chave do Título é uma variação do anterior, que busca as palavras mais importantes no título do texto, partindo da hipótese de que no título está expresso o conteúdo textual fundamental para a compreensão de um texto. A

lista de palavras-chave que orienta o sumarizador incluirá então também as palavras-chave do título, às quais se dará peso maior. Assim, a distribuição das sentenças selecionadas será normalizada em função das palavras-chave que aparecem no título.

O método da Localização, de Baxendale (1958), parte do pressuposto de que as informações principais se encontram na primeira e na última sentença de um texto e que estas devem ser impreterivelmente anexadas a um sumário.

O método das Palavras Sinalizadoras, ou "Cue Phrases", ou ainda dos Marcadores Linguísticos atribui valores às sentenças e seleciona as que têm maior peso. Ele lança mão de um dicionário construído previamente de forma manual sobre um domínio de conhecimento (por exemplo, num texto científico, as palavras "teoria", "métodos", "resultados") e valora positivamente as sentenças que possuem as palavras incluídas nesse dicionário.

Martins et al. (2001) citam ainda o método da Frase Auto-indicativa, descrito por Paice (1981). A sentença a ser selecionada para fazer parte do sumário é escolhida pela presença de palavras ou expressões que indicam partes importantes no texto. Por exemplo, uma sentença como "A abordagem teórica que adotamos é..." seria selecionada para um resumo de texto científico. A diferença desse método para o de palavras-chave está na unidade considerada para a seleção. Aqui acredita-se que a estrutura da frase denota aspectos importantes do texto e leva-se em consideração o gênero textual para a escolha das sentenças indicativas. Num texto narrativo, por exemplo, poderia ser selecionada uma sentença como "Era uma vez..." Vimos que os gêneros, de acordo com Bakthin (2003), possuem uma estrutura "relativamente estável". Essa estabilidade se revela no uso de certos elementos linguísticos que indicam a relevância ou não da informação no texto.

Os autores afirmam que os métodos que se apoiam na estrutura superficial não geram sumários bem construídos. Esses falham principalmente no que se refere à resolução anafórica, porque são métodos que não levam em conta a necessidade de recuperar o contexto e as referências no inter-relacionamento entre sentenças, gerando-se sumários desconexos e incoerentes. Foram realizadas diversas tentativas

com o objetivo de sanar estas dificuldades, mas ainda existe a real necessidade de pesquisa sobre métodos que usem mais informações de natureza linguístico-textual-pragmática.

Por outro lado, a década de 90 vê o crescente interesse no desenvolvimento de métodos estatísticos, que têm sido utilizados com grande sucesso. Eles são fundamentados em corpora e, segundo os pesquisadores, demonstram a possibilidade de gerar bons sumários, dependendo do gênero textual. O uso de corpora na área de Mineração de Dados (Data Mining) possibilitou o aparecimento da área de Mineração de Textos (Text Mining ou Text Data Mining) nas áreas de Recuperação da Informação e Sumarização Automática. Em aplicações de "Data Mining", busca-se informação em dados estruturados, organizados em bases ou bancos de dados dos quais se conhece – ou pode-se prever – a forma de organização. Já em Text Mining, "busca-se o estudo das relações existentes entre componentes de textos não estruturados" (MARTINS et al., 2001).

Para os pesquisadores, no campo da sumarização automática, interessa especificamente a identificação das informações importantes do texto. O objetivo de usar técnicas de Text Mining é justificado pela "obtenção de palavras ou sentenças-chave de um texto para a composição do sumário correspondente". Usando o enfoque de mineração de textos, foram desenvolvidos vários sistemas. O sistema ANES (Automatic News Extraction System), de Raul & Brandon (1993), extrai sentenças para a sumarização automática de textos jornalísticos. Ele identifica, por meio de métodos estatísticos por frequência inversa, palavras que provavelmente destacam os tópicos e outras informações importantes por aparecerem apenas em um dado texto e não aparecerem em um conjunto de outros textos. Essas palavras, hipoteticamente, caracterizam esse determinado texto em oposição ao conjunto de textos. Elas recebem então um peso maior e são separadas em uma lista de "signature words". As sentenças que farão parte do sumário são escolhidas dependendo do peso dado às palavras da lista. O sistema também leva em consideração a localização das sentenças no texto, as anáforas, o tamanho e o tipo do extrato desejado.

Além do sistema ANES, os autores destacam o trabalho desenvolvido por Larocca Neto et al. (2000), que realiza tarefas de "clustering", que identifica relações de co-ocorrência entre palavras do texto, agrupando-as. O conjunto dessas palavras, os *clusters*, seriam indicadores do conteúdo do texto.

As conclusões de pesquisas indicam que o maior problema na geração de sumários continua a ser a distinção entre o que possa ser relevante ou não para a composição de um sumário. Nota-se que essa distinção também é imprecisa em relação aos sumários gerados por humanos. Por esse motivo, os autores consideram que a abordagem profunda é necessária para a obtenção de uma distribuição estrutural que privilegie outros mecanismos além da "identificação, seleção e exclusão/extração de segmentos textuais".

> "A abordagem profunda contempla o conhecimento linguístico e/ou extralinguístico 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 linguístico, ou as relações intencionais, no nível extralinguístico, as quais serão mapeadas no modelo linguístico, computacional, na maioria das vezes envolvendo a manipulação simbólica" (MARTINS et al. 2001).

Assim, uma metodologia de abordagem profunda deve levar em consideração o texto fonte como um todo para que o sumário resultante possa ser fidedígno ao texto. Ainda que informações oriundas de conhecimentos intuitivos dos seres humanos sejam relevantes para a sumarização, mas de difícil acesso ao desenvolvimento de métodos de sumarização automática, é possível obterem-se pistas de distinção de ordem estrutural. Por exemplo, os autores citam os elementos coesivos que podem ser recuperados por meio da análise das estruturas do discurso.

Uma das teorias computacionais que tratam da estruturação do discurso é a Teoria de Estruturação Retórica (RST). Segundo Mann e Thompson (1987), a RST é uma abordagem descritiva que identifica a estrutura hierárquica de um texto, descreve as relações lógico-semânticas entre suas partes, apontando o ponto de transição de uma relação para outra e os termos a ela relacionados.

A RST propõe uma forma geral de descrever as relações lógicas, ou de coerência, entre sentenças num texto, indicando quais delas são mais relevantes para sua compreensão. A teoria tem sido usada como ferramenta analítica para uma grande variedade de tipos de textos e como ferramenta para a representação de conteúdo a ser gerado em sistemas de geração de textos, e, portanto, também de resumos.

São quatro os objetos de estudo da RST: as relações de coerência, os esquemas, a aplicação dos esquemas e as estruturas. A definição de relações identifica relações lógico-semânticas que podem ocorrer entre duas porções de um texto, por exemplo, a relação causal entre proposições ou a relação de evidência entre uma proposição e outra. Os esquemas definem os padrões através dos quais uma determinada porção do texto pode ser analisada em relação a outras porções do mesmo texto. A aplicação das convenções dos esquemas define as formas como um esquema pode ser instanciado com maior flexibilidade. A estrutura de um texto é definida em termos de composição da aplicação dos esquemas.

No que se refere ao uso da RST para a sumarização automática, Martins et al. (2001) ressaltam que ela foi projetada para interpretação e não para a geração automática de textos, que depende de decisões estruturais mais específicas do que as fornecidas pela RST, o que pode levar a uma indefinição no momento de lexicalizar um plano de conteúdo do texto a ser gerado. Apesar disso, a RST e seus desdobramentos são um dos modelos teóricos mais utilizados em abordagens que pretendam dar conta da semântica do texto.

4.2
O sumarizador automático GistSumm

Dentre os diversos programas desenvolvidos para a sumarização automática, vamos nos deter especificamente naquele que utilizamos em uma das etapas da pesquisa: o GistSumm (PARDO, 2002). Ele foi escolhido para esta pesquisa, principalmente, por estar disponível e ser de fácil acesso e uso. O GistSumm é um

sumarizador de abordagem superficial, que se baseia na idéia principal do texto a ser sumarizado e pode ser usado para processar textos em várias línguas e gêneros e revertê-los em sumários. Para a geração de sumários pelo programa GistSumm, Pardo considera as premissas de que todo texto possui uma idéia principal, e de que "é possível identificar em um texto uma sentença que melhor represente sua idéia principal, isto é, a sentença-gist" (2002).

O desenvolvimento desse programa sumarizador levou à pesquisa de duas hipóteses: uma que comprova a possibilidade da identificação da sentença principal do texto fonte, ou de uma que se aproxime disso; e outra, ainda em fase de verificação, da possibilidade da produção de sumários coerentes por meio da justaposição de sentenças relacionadas à sentença-gist.

O GistSumm, segundo Pardo (2002) é um sumarizador que usa os métodos genérico e extrativo. Genérico porque produz sumários de uma forma geral, sem especificação do gênero, e extrativo porque seleciona e justapõe sentenças inteiras de um texto fonte, revertendo-as num sumário, sem modificações estruturais. O programa é dotado de três sequências: segmentação sentencial, ranqueamento e seleção de sentenças.

Na primeira fase, ele segmenta o texto fonte identificando as sentenças pelos sinais usuais de pontuação, tais como: ponto final, ponto de exclamação e de interrogação. Nessa fase também é verificada a presença de abreviaturas para diferenciação do sinal de ponto final.

Na segunda etapa, ou ranqueamento, o programa atribui uma pontuação às sentenças identificadas na fase anterior. A sentença que obtiver maior pontuação é escolhida como a sentença-gist e, a partir desta, serão então escolhidas todas as outras que farão parte do sumário. A seleção das sentenças que comporão o sumário ocorre com o seguinte processo, de acordo com Pardo (2005):

a) case folding: todas as letras das sentenças são transformadas em letras minúsculas, para uniformização;

b) stemming: as palavras do texto são substituídas por suas raízes;

c) remoção de stopwords: o programa seleciona todas as palavras, remove as palavras de classe fechada (palavras gramaticais, que não possuem significado por si mesmas, ou palavras que são demasiado comuns ou julgadas irrelevantes);

d) pontuação das sentenças: essa pontuação pode ocorrer por meio de três métodos estatísticos: o *Keywords*, no qual o programa pontua cada sentença de acordo com as palavras-chave, pois parte do pressuposto que a idéia principal de um texto pode ser expressa por palavras-chave; o *Average Keywords* ou a média de palavras-chave, em que o programa avalia cada sentença no texto de acordo com o número de palavras que contém e a presença de palavras-chave; e o método TF-ISF (*Term Frequency- Inverse Sentence Frequency*) que determina quais orações são importantes em um texto para escolher a que melhor o represente.

e) Ranqueamento das sentenças em função da pontuação obtida no passo anterior: o programa examina a pontuação de cada sentença. A sentença que tiver maior pontuação é escolhida para ser a "gist", a que melhor expressa a idéia principal.

Para uma oração ser escolhida para compor o sumário, ela precisa estar em conformidade com os critérios de correlação com a idéia principal (deve ter pelo menos uma palavra da sentença "gist") e de relevância, computada pelo escore maior da média de todas as sentenças no texto. O número de sentenças selecionadas para fazer parte do sumário é ainda limitado por uma taxa de compressão, isto é, a porcentagem, computada em número de palavras, que especifica o tamanho do sumário em relação ao texto fonte.

Segundo Balage Filho et al. (2007), o GistSumm passou por várias avaliações, obtendo resultados de desempenho muito favoráveis, alcançando reconhecimento e uso na área de pesquisa. Esses resultados indicam que o sumarizador escolhe a sentença "gist" com grande confiabilidade. Os pesquisadores indicam o método Keywords como o mais adequado e concordam que a taxa de compressão em 40% é a melhor a ser empregada para sumários de textos. Isso significa que o sumário

resultante apresenta 60% do tamanho, em palavras, do texto correspondente. Em nossa pesquisa, também utilizamos o GistSumm Keywords em 40% para comparação com os resumos elaborados por um grupo de alunos universitários, por ter sido a taxa de compressão que, em nossa avaliação, alcançou melhores resultados.