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

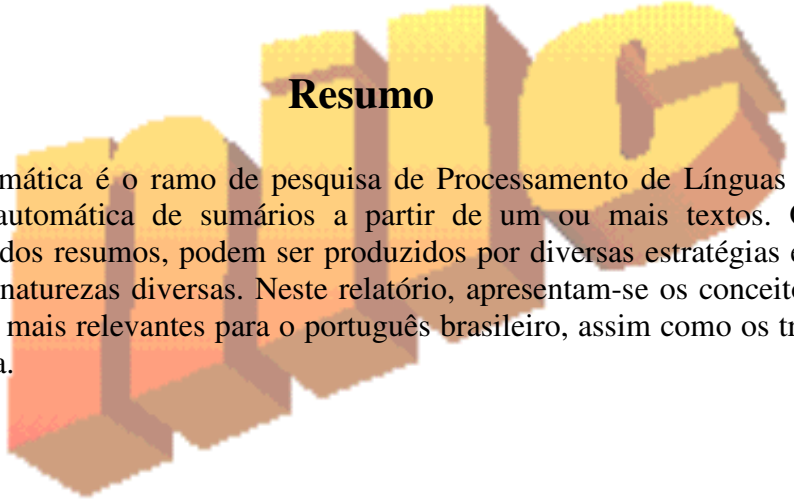
Sumarização Automática: Principais Conceitos e Sistemas para o Português Brasileiro

Thiago Alexandre Salgueiro Pardo

NILC-TR-08-04

Maio, 2008

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



Resumo

Sumarização Automática é o ramo de pesquisa de Processamento de Línguas Naturais que visa à produção automática de sumários a partir de um ou mais textos. Os sumários, comumente chamados resumos, podem ser produzidos por diversas estratégias e pelo uso de conhecimentos de naturezas diversas. Neste relatório, apresentam-se os conceitos básicos da área e os trabalhos mais relevantes para o português brasileiro, assim como os trabalhos mais importantes da área.

ÍNDICE

1. Introdução

A Sumarização Automática (SA) trata da produção automática de sumários a partir de um ou mais textos-fonte e é uma subárea de pesquisa de Processamento de Línguas Naturais (PLN).

Segundo Mani (2001), um sumário é a versão mais curta de um texto. A ABNT (Associação Brasileira de Normas Técnicas) define o termo sumário como se referindo a índice, ou seja, a enumeração das partes de um trabalho. Em SA, sumário pode tanto se referir a índice quanto a resumo propriamente dito. Em português, o termo sumário foi importado da língua inglesa e, por isso, foi adotado como padrão nos trabalhos da área.

Sumários são objetos que nós, humanos, utilizamos nas mais diversas atividades. Usamos sumários para pautar nossas decisões sobre comprar um livro, ler um artigo científico ou alugar um filme para assistir em casa, para nos informarmos da estrutura de um documento ou para acompanhar os dados da previsão meteorológica, para escolher uma página retornada por um buscador na Web para acessar, etc. Muitas vezes, sumários também são úteis para outras tarefas automatizadas em PLN. Por exemplo, o uso de sumários pode melhorar certos aspectos da recuperação de informação e da categorização de textos. Sumários também podem ajudar nas questões de usabilidade de interfaces e de acessibilidade e inclusão digital, por exemplo, na apresentação de dados em aparelhos celulares (cujos visores pequenos restringem a quantidade de informação que pode ser mostrada) e em simplificação de textos para leitores com pouco domínio da língua, respectivamente.

Como ilustração, a Figura 1 mostra um texto completo (chamado texto-fonte) e a Figura exibe 2 seu sumário gerado automaticamente pelo sistema GistSumm (Pardo et al., 2003a) disponível na Web.

Os apaixonados receberam hoje uma boa notícia dos cardiologistas: o amor faz bem ao coração.

Enquanto os apaixonados enviavam cartas e rosas vermelhas em comemoração ao Dia dos Namorados (dia de São Valentim, comemorado em muitos países), a Federação Mundial do Coração (WHF, na sigla em inglês) divulgou um comunicado pedindo aos casais de todo mundo que demonstrem suas emoções com liberdade.

"Os namorados têm outra razão para comemorar porque estudos mostram que estar apaixonado e ser correspondido nos ajuda a manter a saúde e é particularmente bom para nossos corações", afirma o comunicado do WHF, que tem sua sede em Genebra, na Suíça.

A federação, cujo objetivo é combater doenças cardíacas e reúne 166 sociedades de cardiologia de 97 países, também acrescentou que o amor reduz o estresse, a depressão e a ansiedade – três fatores de risco associados às doenças do coração.

"Uma em cada três mortes no mundo ocorrem devido a problemas no coração e derrame, seis vezes superior do que as mortes associadas à Aids", afirmou o professor Philip Poole-Wilson, cardiologista do Imperial College, em Londres, e presidente da federação.

"É por essa razão que estamos ressaltando a importância de adotar um estilo de vida saudável e o impacto positivo que o amor pode ter para a saúde".

De acordo com a WHF, muitos estudos publicados demonstraram que fatores psicológicos, assim como os físicos, estão envolvidos com a doença cardíaca. Em uma pesquisa de cinco anos, 10 mil homens com risco elevado de desenvolver angina (dor no peito) foram questionados se a mulher com quem estavam demonstrava seu amor por eles. Aqueles que responderam "sim" tinham a metade do risco de apresentar a condição.

Figura 1 – Ilustração de texto-fonte

Enquanto os apaixonados enviavam cartas e rosas vermelhas em comemoração ao Dia dos Namorados (dia de São Valentim, comemorado em muitos países), a Federação Mundial do Coração (WHF, na sigla em inglês) divulgou um comunicado pedindo aos casais de todo mundo que demonstrem suas emoções com liberdade.

A federação, cujo objetivo é combater doenças cardíacas e reúne 166 sociedades de cardiologia de 97 países, também acrescentou que o amor reduz o estresse, a depressão e a ansiedade – três fatores de risco associados às doenças do coração.

Figura 2 – Ilustração de sumário correspondente ao texto da Figura 1

Humanos são capazes de produzir tipos variados de sumários e até mesmo sumários diferentes para um mesmo texto. Na realidade, é pouco provável que um mesmo texto tenha sumários idênticos produzidos por diferentes pessoas (Rath et al., 1961), pois o julgamento do que é importante em um texto para ser preservado no sumário, a forma de escrita do sumário e o próprio estilo de escrita de cada pessoa são fatores que introduzem uma gama de variedades nos sumários. E, em geral, é difícil afirmar qual é o melhor sumário. Pode haver vários sumários igualmente bons.

A SA tenta simular a produção humana de sumários e sua riqueza, apesar dos sumários produzidos na área atualmente ainda serem inferiores aos dos humanos. De qualquer forma, é inegável a necessidade da SA nos dias de hoje, em que há uma quantidade imensa de informação disponível, principalmente on-line. Para se ter uma idéia, uma pesquisa conduzida pelo IDC (*International Digital Center*) calculou que havia 281 exabytes (281 bilhões de gigabytes) de dados digitais em 2007, superando em mais de 10% as previsões. Ainda pior: cada vez mais as pessoas têm menos tempo para processar a informação de que necessitam. Nesse cenário, a SA tem se tornado uma ferramenta indispensável, sendo considerada um dos maiores desafios atuais do PLN.

Apresentam-se, neste relatório, os conceitos básicos de SA e os trabalhos mais relevantes na área. Para o leitor interessado em mais detalhes, sugere-se a leitura das obras de Rino e Pardo (2003, 2007) e de Martins et al. (2001).

2. Conceitos Básicos de Sumarização

Sumários podem ser classificados como informativos, indicativos ou críticos (Mani e Maybury, 1999) quanto à função que exercem. Sumários informativos, ou autocontidos, contêm as informações principais do texto organizadas de forma coerente e coesa. Além disso, estes sumários têm que apresentar boa progressão temática, serem gramaticais e legíveis, ou seja, as características que atribuem "textualidade" aos textos. Diz-se que esses sumários podem dispensar a leitura do texto-fonte. Sumários indicativos, por sua vez, não substituem o texto-fonte, mas apenas dizem do que ele trata. Índices, se considerados sumários, são classificados como indicativos. Por fim, sumários críticos, ou avaliativos, apresentam opiniões além do conteúdo esperado. Resenhas de livros são exemplos de sumários críticos.

Em relação à audiência a que se destinam, sumários podem ser classificados como genéricos ou focados nos interesses dos leitores. Sumários genéricos trazem as informações mais importantes dos textos-fontes correspondentes, sem se preocupar com os leitores. Sumários focados nos interesses dos leitores, por outro lado, customizam as informações que trazem em função do conhecimento destes. Por exemplo, se o leitor é leigo no assunto do texto-fonte, um sumário com mais informações contextuais faz-se útil; para um leitor especialista no assunto, o sumário deve conter somente a informação nova ou essencial do texto.

Em termos de formação, sumários podem ser classificados como extratos ou *abstracts* (Sparck Jones, 1993a). Extratos são sumários compostos por trechos inalterados do texto-fonte. Eles são construídos por operações de cópia e cola de trechos do texto-fonte, literalmente. *Abstracts*, por sua vez, apresentam partes (ou mesmo tudo) reescritas, ou seja, há algum nível de modificação na estrutura e/ou significado dos trechos extraídos do texto-fonte. O termo *abstract* também foi importado da língua inglesa e assim é utilizado em português. Não é incomum encontrar o termo traduzido como "abstrato" em português, apesar de pouco aceito na área. Em geral, o termo sumário é utilizado para se referir tanto a extrato quanto a *abstract*.

Sumários podem ser construídos basicamente por 2 abordagens definidas em função da quantidade e do nível de conhecimento lingüístico que utilizam: a abordagem superficial e a profunda. Estas também podem se mesclar de variadas formas, dando origem a uma abordagem híbrida. Antes de definir tais abordagens, faz-se necessário explicitar os níveis básicos de conhecimento que podem ser utilizados em SA, assim como na área maior de PLN. A Figura 3 ilustra esses níveis e sua organização em termos de abstração lingüístico-computacional e sua complexidade para o tratamento computacional.

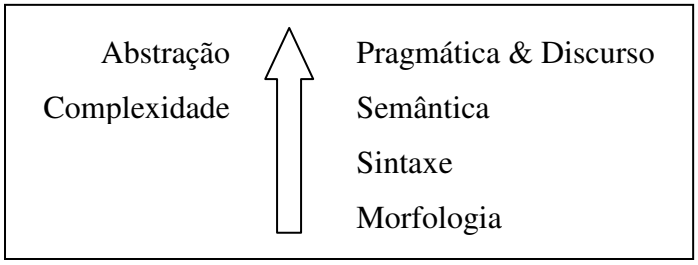


Figura 3 – Níveis de conhecimento em SA e PLN

Quanto mais se sobe da morfologia em direção à pragmática e ao discurso, maior é a abstração lingüístico-computacional e mais difícil se torna obter uma representação formal do nível de conhecimento e, por conseguinte, mais complexo é processar computacionalmente tal nível.

Cada um dos níveis de conhecimento é muito rico e conta com muitos pesquisadores na Letras e na Lingüística que os investigam. Em PLN, apenas os aspectos mais relevantes de cada nível para a área são abordados. Em PLN, no geral, tem-se que:

- na morfologia, estuda-se a formação das palavras, isoladas de seus contextos de ocorrência; trabalha-se com radicais, prefixos e sufixos, derivações e desinências, dentre outros temas;
- na morfossintaxe, nível intermediário entre a morfologia e a sintaxe, lidam-se com as classes gramaticais (ou etiquetas morfossintáticas) das palavras, ou seja, substantivos, verbos, adjetivos e advérbios, dentre outras; diz-se que este nível de análise é morfossintático (e não morfológico) porque requer conhecimento do contexto da palavra para que se determine sua classe gramatical;
- na sintaxe, trata-se da função das partes das sentenças, isto é, se elas exercem a função de sujeito, predicado, objetos, adjuntos, complementos, dentre outras; é comum utilizar as noções de sintagmas (nominais, verbais, adverbiais, preposicionais, etc.) em vez da função;
- na semântica, lida-se em geral com a representação e determinação do significado lexical, sentencial e, muitas vezes, textual;
- na pragmática e no discurso, são considerados aspectos de organização das informações do texto, intenções subjacentes ao texto, contexto em que o texto foi produzido e em que é veiculado, etc.

Como exemplificado pelo nível da morfossintaxe, os níveis do conhecimento não ocorrem isoladamente. Há uma interação entre eles. Por exemplo, para se determinar a correta estruturação sintática de uma sentença, faz-se necessário conhecer as etiquetas morfossintáticas das palavras e seus significados. Similarmente, para se fazer a análise discursiva de um texto, aspectos sintáticos e semânticos são determinantes.

A primeira abordagem para a SA, chamada "superficial", faz pouco ou nenhum uso de conhecimento lingüístico para produzir sumários. Quando usa conhecimento, este se restringe, em geral, aos níveis morfossintático e sintático. Nesta abordagem, é comum se fazer uso de dados estatísticos e empíricos. Por exemplo, um método para se construir extratos se baseia na seleção e justaposição das sentenças do texto-fonte que contêm as palavras mais freqüentes do texto. Muitas vezes, o conhecimento morfossintático e sintático é utilizado para se decidir se palavras (por exemplo, adjetivos e advérbios) e componentes sintáticos (por exemplo, adjuntos adverbiais, agentes da passiva e vocativos) podem ser omitidos ou não no sumário.

A segunda abordagem, referenciada por "profunda", faz uso massivo de conhecimento lingüístico, utilizando teorias e modelos formais da língua. Faz-se uso, em geral, de léxicos, *wordnets*, gramáticas, analisadores sintático-semânticos e discursivos, etc. Léxicos e *wordnets* são utilizados para se determinar e classificar as palavras do texto; gramáticas e analisadores sintático-semânticos são a base para se estruturar as sentenças e determinar seus significados básicos; analisadores discursivos costumam ser utilizados para se identificar as partes mais importantes dos textos. Em particular, a teoria discursiva RST (*Rhetorical Structure Theory*) (Mann e Thompson, 1987) e os analisadores automáticos correspondentes (por exemplo, Marcu, 1997; Pardo e Nunes, 2008) têm sido a base da maioria dos trabalhos na SA profunda.

Sumarizadores da abordagem superficial costumam produzir extratos, enquanto sumarizadores da abordagem profunda podem gerar *abstracts*. Apesar do desenvolvimento relativamente simples de sumarizadores pela abordagem superficial e de seu baixo custo, é consenso na área que os métodos superficiais produzem sumários de qualidade inferior aos sumários produzidos por métodos profundos, em geral (Mani, 2001; Sparck Jones, 2007). Uzêda et al. (2007) e Leite et al. (2007), em particular, demonstraram a superioridade dos métodos profundos em uma ampla avaliação comparativa que realizaram.

Em termos do número de textos processados, o processo de sumarização pode ser classificado como monodocumento ou multidocumento. A SA monodocumento tradicional produz o sumário de um único texto-fonte; a multidocumento, por sua vez, produz um sumário a partir de uma coleção de textos-fonte. A SA multidocumento tem ganhado muito destaque devido à quantidade cada vez maior de informação disponível. Mani (2001) afirma que a tarefa de sumarização multidocumento não é intuitiva para humanos, mas McKeown et al. (2005) demonstram que, apesar das dificuldades, sumários multidocumentos produzidos tanto automaticamente quanto por humanos se mostraram muito úteis em experimentos que simulavam a apreensão de informação por humanos.

Na SA multidocumento, além de se identificar o que é informação importante e irrelevante no conjunto de textos, novos desafios surgiram e desafios antigos se tornaram mais complexos, por exemplo, eliminação de informação redundante do sumário, ordenação (temporal ou não) dos segmentos textuais que comporão os sumários, fusão de segmentos textuais com informações complementares, manutenção da coerência do sumário, etc. Deve-se levar em conta também que os textos podem se originar de fontes diferentes e, em geral, são escritos por pessoas diferentes e, portanto, têm estilos diversos.

A Tabela 1 sintetiza as várias classificações possíveis dos sumários.

Tabela 1 – Classificação de sumários

Critério	**Classificação**
Função	Indicativo, informativo ou crítico
Audiência	Genérico ou focado nos interesses do leitor
Formação	Extrato ou *abstract*
Abordagem	Superficial, profunda ou híbrida
Número de textos-fonte	Monodocumento ou multidocumento

O sumário da Figura 2 produzido pelo sistema GistSumm pode ser classificado como um extrato informativo, genérico, monodocumento e da abordagem superficial.

Idealmente, as etapas de um processo de sumarização podem ser visualizadas como se mostra na Figura 4 (Mani e Maybury, 1999).

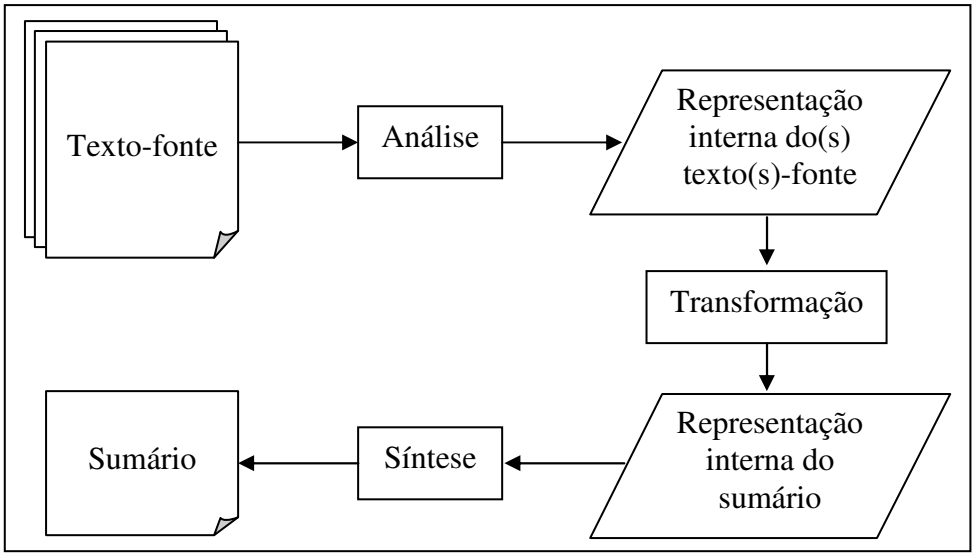


Figura 4 – Etapas da SA

Na etapa inicial de análise, um ou mais textos-fonte são processados e uma representação interna com todo o conteúdo dos textos é produzida. Essa representação deve ser formal o suficiente para ser processada automaticamente. A etapa de transformação realiza o processo de sumarização sobre a representação interna dos textos-fonte, produzindo a representação interna do sumário, a qual contém o conteúdo mais importante a ser veiculado textualmente. Por fim, a etapa de síntese expressa em língua natural a representação interna do sumário.

Todo este processo deve ser guiado pela taxa de compressão, ou seja, o tamanho desejado do sumário. Por exemplo, um sumário com taxa de compressão de 70% diz que este deve ter tamanho equivalente a 30% do tamanho do texto-fonte (medido em número de palavras, em geral). Não é incomum se encontrar esta definição invertida, em que se afirma, por exemplo, que a taxa de compressão para o exemplo anterior é de 30% ou seja, sobram 30% do texto-fonte para se formar o sumário.

Em princípio, as etapas da SA acima são independentes das classificações anteriores de sumários, mas podem ser adaptados de acordo com a metodologia de SA seguida. Em alguns casos, as etapas de análise e síntese são extremamente simples ou mesmo inexistentes, podendo consistir somente de um levantamento da freqüência das palavras do texto (na abordagem superficial, por exemplo); em outros, estas etapas e as representações internas com as quais se trabalha podem ser muito complexas (na abordagem profunda, por exemplo),

exigindo muitos recursos e ferramentas lingüístico-computacionais auxiliares. Tal raciocínio também se aplica à etapa de transformação.

É importante que se diga que as etapas de SA se assemelham muito ao processo realizado na grande subárea de Geração Automática de Textos dentro de PLN (vide, por exemplo, Reiter e Dale, 2000). Muitas vezes, aliás, o processo de SA pode ser visto e tratado como um processo de geração textual.

A avaliação de sistemas de sumarização é um tema que tem sido muito investigado. A razão é a grande dificuldade em fazê-lo sob várias perspectivas: há muitos sumários diferentes que podem ser igualmente bons; a avaliação humana, que aparentemente é mais apropriada, é cara, demorada, não reprodutível e suscetível a erros e inconsistências típicos do julgamento humano; a avaliação automática não é robusta o suficiente para discernir entre todos os quesitos que devem ser julgados em um sumário. Apesar de tudo isso, avanços significativos foram feitos na área ultimamente.

Há diversas facetas de um sistema de SA que podem ser avaliadas, por exemplo, o desempenho computacional (em termos de complexidade de tempo e espaço), a usabilidade do sistema e o quão bons são os sumários produzidos. Esta última faceta é o foco da maioria das avaliações realizadas.

Sparck Jones e Galliers (1996) apresentam detalhadamente todos os pontos que devem ser levados em consideração na avaliação de sistemas de PLN. Estes pontos são integralmente aplicáveis à SA e, por isso, são brevemente discutidos abaixo.

A avaliação de sistemas de SA pode ser classificada como intrínseca, quando se avalia a qualidade do sistema em si, ou extrínseca, quando se avalia o impacto do sistema de SA em aplicações que necessitam de tal ferramenta, por exemplo, aplicações de perguntas e respostas e de recuperação de informação. Atualmente, há uma forte tendência em se realizarem avaliações extrínsecas, pois as avaliações intrínsecas têm demonstrado que tem havido pouca melhora na qualidade dos sumários produzidos.

Em termos do que se avalia em um sistema, a avaliação pode ser do tipo caixa preta (*black-box* em inglês), em que se avalia somente a saída final do sistema, ou transparente (*glass-box* em inglês), em que a saída de cada módulo interno do sistema é avaliado. A avaliação caixa-preta tem sido dominante nas pesquisas em SA.

A avaliação é dita on-line quando faz uso de humanos para julgar os sumários. Quando não o faz e, portanto, a avaliação é feita via métricas automáticas, a avaliação é dita off-line. Diante da dificuldade de se lidar com a avaliação humana, a avaliação off-line tem sido amplamente preferida pela comunidade de pesquisa em SA. Para tanto, boas métricas automáticas de qualidade de sumários se fazem necessárias. Algumas serão discutidas a seguir.

Uma última distinção que se faz é se a avaliação é autônoma ou comparativa. Avaliação autônoma se refere ao julgamento em isolado de um sistema. Avaliação comparativa, por sua vez, refere-se ao estabelecimento de métricas de avaliação comuns a vários sistemas de SA e sua posterior comparação. Em SA e em PLN como um todo, a avaliação comparativa tem sido muito desejada, pois permite mensurar o estado da arte.

Independentemente do tipo de avaliação que se conduz, boas métricas de avaliação de sumários são necessárias. Neste ponto, é importante diferenciar os dois principais critérios que devem ser avaliados em um sumário: qualidade e informatividade. Pode-se dizer que, na maioria dos trabalhos em SA, qualidade de um sumário diz respeito ao seu grau de legibilidade, gramaticalidade e fluência, dentre outros quesitos, enquanto informatividade refere-se ao quanto de informação o sumário apresenta.

As questões de qualidade costumam ser avaliadas por humanos e por métricas de inteligibilidade de textos, por exemplo, o índice de Flesch (1948), já adaptado ao português brasileiro (Martins et al., 1996). Para informatividade, há muitas métricas.

As medidas clássicas de cobertura e precisão relacionam o conteúdo do sumário produzido automaticamente com o conteúdo de um sumário de referência correspondente. Sumário de referência refere-se a um sumário considerado ideal, normalmente produzido por um humano. Cobertura indica o quanto de informação do sumário de referência o sumário automático contém; precisão indica o quanto de informação do sumário de referência o sumário automático contém em relação a tudo que este contém. Se os sumários automático e de referência forem extratos, o cômputo destas medidas pode ser automático (comparando-se as sentenças em comum nos dois sumários); caso contrário, a avaliação tem que ser manual ou mecanismos mais sofisticados de contagem de informação em comum devem ser levados em conta. Há muitas variações de medidas de informatividade, muitas baseadas nos conceitos de cobertura e precisão.

Como discutido anteriormente, as métricas automáticas têm sido preferidas na área. A métrica mais utilizada, chamada ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) (Lin e Hovy, 2003), consiste basicamente no cômputo automático de n-gramas (conjuntos de palavras em seqüência) em comum entre um sumário automático e um ou mais sumários de referência. Quanto mais n-gramas em comum houver, maior a nota (entre 0 e 1) atribuída ao sumário automático. Os autores desta medida demonstraram que ela é tão boa quanto humanos em ranquear sumários de acordo com sua informatividade.

A ROUGE pertence a uma família de métricas de avaliação, como apresentado por Soricut e Brill (2004), aplicáveis para várias tarefas de PLN. Recentemente, medidas correlatas a ROUGE têm surgido e começam a se difundir na área. Valem destacar os métodos da pirâmide (Nenkova e Passonneau, 2004) o dos elementos básicos (*basic elements* em inglês) (Hovy et al., 2006). Em vez de comparar n-gramas, o método da pirâmide compara os conceitos expressos nos sumários. O método dos elementos básicos compara os pares de elementos (palavras ou expressões) do sumário e suas relações (sintáticas, em geral).

Para se ter idéia da importância da avaliação em SA, há conferências internacionais dedicadas ao tema. A mais importante chama-se TAC (*Text Analysis Conference*) (até recentemente era chamada *Document Understanding Conference* - DUC) e é realizada anualmente. Além do problema da sumarização, também lida com perguntas e respostas e tarefas correlatas.

3. Sistemas de Sumarização Automática para o Português Brasileiro

O GistSumm (Pardo et al., 2003a) foi o primeiro sistema de SA superficial disponível para o português. Inicialmente, o sistema identifica a sentença mais importante do texto-fonte, chamada sentença-gist. A seguir, procura por sentenças que complementam esta sentença, as quais, juntamente com a sentença-gist, constituem o extrato. A sentença-gist é aquela que apresenta as palavras mais freqüentes do texto. As sentenças que a complementam são escolhidas dentre aquelas que têm alguma palavra em comum com a sentença-gist. Recentemente, o GistSumm foi estendido para realizar sumarização multidocumento (Pardo, 2005) e para lidar com textos estruturados, como teses e artigos científicos (Balage et al., 2007).

O NeuralSumm (Pardo et al., 2003b) é um sumarizador superficial baseado em uma rede neural artificial, mais especificamente, uma rede de Kohonen. São fornecidas à rede características de cada sentença do texto-fonte. Para cada sentença, a rede produz como saída a classe "importante", "complementar" ou "supérflua". Sentenças supérfluas são descartadas, enquanto sentenças importantes e complementares são consideradas para serem incluídas no sumário. As características extraídas de cada sentença para serem fornecidas à rede são o tamanho da sentença, posição da sentença no texto, posição da sentença no parágrafo ao qual

pertence, se há ou não palavras-chave na sentença, se há ou não na sentença palavras da sentença-gist (calculada de forma similar ao que é feito no GistSumm), se há palavras indicativas ou não na sentença e pontuação da sentença com base na distribuição das palavras no texto, dentre algumas outras. Utilizando outro modelo de redes neurais e as mesmas características do NeuralSumm, Orrú et al. (2006) apresentam o sumarizador chamado SABio.

Pertencente à abordagem profunda, o DMSumm (Pardo e Rino, 2002) é um gerador automático de sumários, e não um sumarizador propriamente dito. Como entrada, em vez de receber um texto-fonte, o sistema recebe a representação interna do texto-fonte, construída manualmente segundo algumas teorias discursivas propostas por Mann e Thompson (1987), Grosz e Sidner (1986) e Jordan (1992). O sistema constrói a representação interna do sumário pela seleção e reestruturação das informações mais importantes da representação interna do texto-fonte. A representação interna do sumário é, então, sintetizada, de forma muito simples, no sumário final.

O UNLSumm (Martins e Rino, 2002), assim como o DMSumm, é um gerador de sumários. Como entrada, recebe a representação do texto-fonte codificada na interlíngua UNL (*Universal Networking Language*) (Uchida, 2000). O núcleo deste sumarizador consiste em um conjunto de heurísticas que cortam partes menos importantes da representação em UNL, produzindo uma representação interna do sumário em UNL.

O melhor sumarizador conhecido para o português é o SuPor (Leite et al., 2007; Rino et al., 2004). O sumarizador, pertencente à abordagem superficial, utiliza aprendizado de máquina bayesiano para combinar diversas características sentenciais e outros métodos completos de sumarização (também vistos como características) para decidir quais são as sentenças que devem ser selecionadas para o sumário. Dentre as características, há o tamanho e posição da sentença, presença ou não de palavras freqüentes e sinalizadoras, ocorrência de nomes próprios e encadeamento lexical, dentre outras. Na mesma linha, há o sistema ClassSumm (Larocca Neto et al., 2002) que, apesar de se sair pior do que o SuPor nas avaliações que participou, é um dos melhores para a língua portuguesa.

O sumarizador superficial TF-ISF-Summ (Larocca Neto et al., 2000) utiliza a medida TF-ISF (*Term Frequency - Inverse Sentence Frequency*), derivada da medida tradicional TF-IDF da área de recuperação de informação, para ranquear as sentenças de um texto em função da representatividade de suas palavras. As mais bem ranqueadas são selecionadas para compor o sumário.

Souza e Nunes (2001) e Pereira et al. (2002) apresentam sumarizadores superficiais com base em algoritmos de detecção de palavras-chave, baseando-se na idéia de que sentenças com mais palavras-chave são mais importantes no texto.

Antiqueira (2007) apresenta diversos métodos de sumarização superficial baseados em medidas e características de redes complexas, que são tipos especiais de grafos. Inicialmente, o texto-fonte é representado como um grafo. Em seguida, as informações provenientes do grafo fornecem dados sobre a importância das sentenças, sendo que as mais importantes são selecionadas para o sumário. Na mesma linha, Pardo et al. (2006) demonstram que é possível utilizar redes complexas para avaliar comparativamente sumários automáticos e produzidos por humanos.

Na abordagem profunda, Uzêda et al. (2007) exploram diversos métodos de sumarização baseados na teoria discursiva proposta por Mann e Thompson (1987), mostrando que os métodos mais conhecidos desta linha têm desempenho similar, além de que são melhores do que métodos superficiais clássicos da literatura. A teoria utilizada fornece uma ordenação dos segmentos textuais (sentenças e orações) em função de sua importância discursiva. Os segmentos mais importantes são selecionados para formar o sumário. Assim como o DMSumm e o UNLSumm, esses métodos necessitam da representação interna dos

texto-fonte como entrada. Carbonel et al. (2007) e Gonçalves et al. (2008) utilizam a mesma teoria discursiva como base para seus sistemas de sumarização, mas adicionam módulos de resolução de cadeias de co-referência aos sistemas, de forma que os sumários não tenham anáforas não resolvidas. O primeiro sistema, chamado VeinSum, faz uso da Teoria das Veias (Cristea et al., 1998) durante a produção do sumário; o outro sistema, chamado CorrefSum, realiza pós-edição nos sumários, utilizando heurísticas de resolução anafórica desenvolvidas com base em análise de córpus.

4. Principais Trabalhos da Área

A área de sumarização é muito vasta e conta com inúmeros trabalhos. Além dos trabalhos descritos na subseção anterior para o português brasileiro, outros também devem ser comentados, sendo que muitos destes influenciaram diretamente os trabalhos anteriores.

As obras de Mani e Maybury (1999) e de Mani (2001) são consideradas os livros-texto da área: a primeira é uma coleção dos artigos mais relevantes até a data da publicação, vários dos quais são citados abaixo; o último abrange toda a área, com suas definições, abordagens e trabalhos mais representativos. Luhn (1958) e Edmundson (1969) apresentam trabalhos pioneiros baseados na freqüência de palavras e expressões sinalizadoras, sendo que suas metodologias são citadas até hoje. Sparck Jones (1993a, 1993b, 1997) tem papel decisivo no estabelecimento da área e nos rumos que a área seguiu. Kupiec et al. (1995) introduzem com sucesso o uso de aprendizado de máquina em sumarização, inspirando diversos trabalhos posteriores. Barzilay e Elhadad (1997) apresentam o conceito de cadeias lexicais e seu uso em sumarização, conceito este muito empregado em PLN. Salton et al. (1997) introduzem o que é considerado hoje um dos principais trabalhos de sumarização com uso de grafos. Recentemente, Mihalcea (2005) explora com sucesso o uso de grafos para sumarização com inspiração em algoritmos de recuperação de informação. Hovy e Lin (1997) apresentam o sistema SUMMARIST de grande impacto na área. McKeown e Radev (1995) e Mani e Bloedorn (1999) são dois dos primeiros trabalhos na área de sumarização multidocumento, sendo ainda bastante referenciados. Knight e Marcu (2002) inovam ao utilizar modelos estatísticos sofisticados para a compressão sentencial, fomentando um ramo de pesquisa que tem ganhado cada vez mais destaque. Mann e Thompson (1987) introduzem a teoria discursiva RST e sinalizam seu uso em potencial para sumarização, sendo que o trabalho mais representativo que segue esta linha foi o de Marcu (2000). Recentemente, Wolf e Gibson (2006) revisitam a teoria discursiva proposta e seu uso em sumarização, propondo um dos trabalhos que é considerado um dos grandes avanços recentes na área. Na linha discursiva, Teufel e Moens (2002) realizam a classificação dos segmentos textuais de artigos científicos para posterior seleção para o sumário.

Referências

ANTIQUEIRA, L. **Desenvolvimento de Técnicas Baseadas em Redes Complexas para Sumarização Extrativa de Textos**. Dissertação de Mestrado. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. 2007.

BALAGE FILHO, P.P.; PARDO, T.A.S.; NUNES, M.G.V. **Summarizing Scientific Texts: Experiments with Extractive Summarizers**. In the Proceedings of the Seventh International Conference on Intelligent Systems Design and Applications – ISDA, pp. 520-524. Rio de Janeiro-RJ, Brazil. October, 22-24. 2007.

BARZILAY, R.; ELHADAD, M. **Using Lexical Chains for Text Summarization**. In the Proceedings of the Intelligent Scalable Text Summarization Workshop, Madri, Spain. 1997.

CARBONEL, T.I.; PELIZZONI, J.; RINO, L.H.M. **Validação Preliminar da Teoria das Veias para o Português e Lições Aprendidas**. In the Proceedings of the V Workshop on Information and Human Language Technology. Rio de Janeiro-RJ. 2007.

CRISTEA, D.; IDE, N.; ROMARY, L. **Veins Theory: A Model of Global Discourse Cohesion and Coherence**. In the Proceedings of the Coling-ACL, pp. 281-285. Montreal, Canadá. 1998.

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

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

GONÇALVES, P.N.; VIEIRA, R.; RINO, L.H.M. **CorrefSum: Referencial Cohesion Recovery in Extractive Summaries**. In the Proceedings of the International Conference on Computational Processing of Portuguese. 2008.

GROSZ, B. and SIDNER, C. **Attention, Intentions, and the Structure of Discourse**. Computational Linguistics, Vol. 12, No. 3. 1986.

HOVY, E. and LIN, C-Y. **Automated Text Summarization in SUMMARIST**. In the Proceedings of the Intelligent Scalable Text Summarization Workshop, pp. 18-24. Madrid, Spain. 1997.

HOVY, E.H.; LIN, C-Y.; ZHOU, L.; FUKUMOTO, J. **Automated Summarization Evaluation with Basic Elements**. In the Proceedings of the Fifth International Conference on Language Resources and Evaluation. Genoa, Italy. 2006.

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

KNIGHT, K. and MARCU, D. **Summarization beyond sentence extraction: A Probabilistic Approach to Sentence Compression**. Artificial Intelligence, Vol. 139, N. 1, pp. 91-107. 2002.

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

LAROCCA NETO, J.; SANTOS, A.D.; KAESTNER, C.A.A.; FREITAS, A.A. **Document clustering and text summarization**. In the Proceedings fo the 4th International Conference on Practical Applications of Knowledge Discovery and Data Mining, pp. 41-55. 2000.

LAROCCA NETO, J.; FREITAS, A.A.; KAESTNER, C.A.A. **Automatic text summarization using a machine learning approach**. In the Proceedings of the XVI Brazilian Symposium on Artificial Intelligence, pp. 205-215. 2002.

LEITE, D.S.; RINO, L.H.M.; PARDO, T.A.S.; NUNES, M.G.V. **Extractive Automatic Summarization: Does more linguistic knowledge make a difference?** In C. Biemann, I. Matveeva, R. Mihalcea, and D. Radev (eds.), Proceedings of the HLT/NAACL Workshop on TextGraphs-2: Graph-Based Algorithms for Natural Language Processing, pp.17-24. 26 April, Rochester, NY, USA. 2007.

LIN, C-Y. and HOVY, E.H. **Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics**. In the Proceedings of the Language Technology Conference. Edmonton, Canada. 2003.

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

MANI, I. **Automatic Summarization**. John Benjamins Publishing Co., Amsterdam. 2001.

MANI, I. and BLOEDORN, E. **Summarizing Similarities and Differences among Related Documents**. Information Retrieval, Vol. 1, N.1, pp. 35-67. 1999.

MANI, I. and MAYBURY, M.T. **Advances in automatic text summarization**. The MIT Press, Cambridge, MA. 1999.

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

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

MARTINS, T.B.F.; GHIRALDELO, C.M.; NUNES, M.G.V.; OLIVEIRA JR., O.N. **Readability Formulas Applied to Textbooks in Brazilian Portuguese**. Notas do ICMSC-USP, Série Computação. 1996.

MARTINS, C.B.; PARDO, T.A.S.; ESPINA, A.P.; RINO, L.H.M. **Introdução à Sumarização Automática**. Relatório Técnico RT-DC 002. Departamento de Computação, Universidade Federal de São Carlos. São Carlos-SP, Fevereiro, 38p. 2001.

MARTINS, C.B. and RINO, L.H.M. **Revisiting UNLSumm: Improvement through a case study**. In the Proceedings of the Workshop on Multilingual Information Access and Natural Language Processing, Vol. 1. p. 71-79. Sevilha, Espanha. 2002.

MCKEOWN, K. and RADEV, D.R. **Generating summaries of multiple news articles**. In the Proceedings of the 18th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, pp. 74-82, Seattle, WA. 1995.

MCKEOWN, K.; PASSONNEAU, R.J.; ELSON, D.K.; NENKOVA, A.; HIRSCHBERG, J. **Do summaries help? A task-based evaluation of multi-document summarization**. In the Proceedings of the 28th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, pp. 210-217. 2005.

MIHALCEA, R. **Language Independent Extractive Summarization**. In the Proceedings of the 43th Annual Meeting of the Association for Computational Linguistics. 2005.

NENKOVA, A. and PASSONNEAU, R. **Evaluating content selection in summarization: the pyramid method**. In the Proceedings of NAACL-HLT. 2004.

ORRÚ, T.; ROSA, J.L.G.; NETTO, M.L.A. **SABio: An Automatic Portuguese Text Summarizer Through Artificial Neural Networks in a More Biologically Plausible Model**. In the Proceedings of the 7th Workshop on Computational Processing of Written and Spoken Portuguese, pp. 11-20. 2006.

PARDO, T.A.S.; NUNES, M.G.V.; OLIVEIRA JR., O.N.; ANTIQUEIRA, L.; COSTA, L.F. **Using Complex Networks for Language Processing: The Case of Summary Evaluation**. In the Proceedings of the International Conference on Communications, Circuits and Systems, Vol. 4, pp. 2678-2682. IEEE Press. 2006.

PARDO, T.A.S. **GistSumm - GIST SUMMarizer: Extensões e Novas Funcionalidades**. Série de Relatórios do NILC. NILC-TR-05-05. São Carlos-SP, Fevereiro, 8p. 2005.

PARDO, T.A.S. and NUNES, M.G.V. **On the Development and Evaluation of a Brazilian Portuguese Discourse Parser**. Revista de Informática Teórica e Aplicada - RITA. 2008.

PARDO, T.A.S.; RINO, L.H.M.; NUNES, M.G.V. **GistSumm: A Summarization Tool Based on a New Extractive Method**. In the Proceedings of the 6th Workshop on Computational Processing of the Portuguese Language - Written and Spoken. Faro, Portugal. 2003a.

PARDO, T.A.S; RINO, L.H.M.; NUNES, M.G.V. **NeuralSumm: Uma Abordagem Conexionista para a Sumarização Automática de Textos**. In Anais do IV Encontro Nacional de Inteligência Artificial – ENIA, pp. 1-10. Campinas-SP, Brasil. 2 a 8 de Agosto. 2003b.

PARDO, T.A.S. and RINO, L.H.M. **DMSumm: Review and Assessment**. In E. Ranchhod and N. J. Mamede (eds.), 3rd International Conference: Portugal for Natural Language Processing – PorTAL (Lecture Notes in Artificial Intelligence 2389), pp. 263-273. Faro, Portugal. 2002.

PEREIRA, M.B.; SOUZA, C.F.R.; NUNES, M.G.V. **Implementação, Avaliação e Validação de Algoritmos de Extração de Palavras-Chave de Textos Científicos em Português**. Revista Eletrônica de Iniciação Científica. Ano II, Vol. 2, N. 1. 2002.

RATH, G.J.; RESNICK, A.; SAVVAGE, R. **The formation of abstracts by the selection of sentences**. American Documentation, Vol. 12, N. 2, pp. 139-141. 1961.

REITER, E. and DALE, R. **Building Natural Language Generation Systems**. Cambridge, University Press. 2000.

RINO, L.H.M. e PARDO, T.A.S. **A Sumarização Automática de Textos: Principais Características e Metodologias**. In Anais do XXIII Congresso da Sociedade Brasileira de Computação, Vol. VIII: III Jornada de Minicursos de Inteligência Artificial, pp. 203-245. Campinas-SP, Brasil. 2 a 8 de Agosto. 2003.

RINO, L.H.M.; PARDO, T.A.S.; SILLA JR., C.N.; KAESTNER, C.A.; POMBO, M. **A Comparison of Automatic Summarization Systems for Brazilian Portuguese Texts**. In the Proceedings of the 17th Brazilian Symposium on Artificial Intelligence – SBIA (Lecture Notes in Artificial Intelligence 3171), pp. 235-244. São Luis-MA, Bazil. September, 29 - October, 1. 2004.

RINO, L.H.M. e PARDO, T.A.S. **A coleção TeMário e a avaliação de sumarização automática**. In D. Santos (ed.), Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa, pp. 267-276. IST Press, Lisboa. 2007.

SALTON, G.; SINGHAL, A.; MITRA, M.; BUCKLEY, C. **Automatic Text Structuring and Summarization**. Information Processing & Management, Vol. 33, N. 2, pp. 193-207. 1997.

SPARCK JONES, K. **Discourse Modelling for Automatic Summarisation**. Tech. Report No. 290. University of Cambridge. UK, February. 1993a.

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

SPARCK JONES, K. and GALLIERS, J.R. **Evaluating Natural Language Processing Systems**. Lecture Notes in Artificial Intelligence, Vol. 1083. 1996.

SPARCK JONES, K. **Automatic summarising: a review and discussion of the state of the art**. Technical Report UCAM-CL-TR-679. University of Cambridge. 2007.

SOUZA, C.F.R. e NUNES, M.G.V. **Avaliação de Algoritmos de Sumarização Extrativa de Textos em Português**. Relatórios Técnicos do ICMC-USP. NILC-TR-01-09. 2001.

TEUFEL, S. and MOENS, M. **Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status**. Computational Linguistics, Vol. 28, N. 4, pp. 409-445. 2002.

UCHIDA, H. **Universal Networking Language: An Electronic Language for Communication, Understanding and Collaboration**. UNL Center, IAS/UNU, Tokyo. 2000.

UZÊDA, V.R.; PARDO, T.A.S.; NUNES, M.G.V. **Avaliação de Métodos de Sumarização Automática de Textos Baseados na Rhetorical Structure Theory**. Revista de Iniciação Científica - RIC. Centro de Tecnologia Educacional para Engenharia - CETEPE, USP/São Carlos. 2007.

WOLF, F. and GIBSON, E. **Coherence in Natural Language: Data Structures and Applications**. The MIT Press. 2006.