A sumarização automática de textos em português

Lucia Helena Machado Rino
Departamento de Computação
Centro de Ciências Exatas e de Tecnologia
UFSCar, São Carlos - SP
lucia@dc.ufscar.br

Resumo

Este artigo trata da sumarização automática sob perspectivas comunicativas e lingüísticas, associando relações estruturais em três níveis distintos de representação: o de intenções, o de coerência e o semântico. Nesse contexto, consideramos que a proposição central de um sumário deve coincidir com a proposição central de sua mensagem fonte, sendo que a preservação dessa proposição constitui um fator crítico no processo de sumarização. Ilustramos neste artigo a aplicação de um modelo de discurso dirigido por objetivos que tem a função de associar as restrições características de cada um dos níveis de representação estrutural, para garantir a preservação da proposição central do discurso, assim como para buscar a satisfação de seu objetivo comunicativo. Esse modelo especifica o mapeamento de intenções em relações de coerência ao mesmo tempo que verifica a dependência semântica entre as unidades de conteúdo de uma mensagem fonte. Na sumarização automática, o modelo se torna a base de um planejador de textos, mais especificamente, de um planejador de *mensagens fontes de sumários*, onde a proposição central do discurso é enfatizada. Por construção, qualquer mensagem fonte de um sumário satisfaz o objetivo comunicativo do discurso. O modelo proposto se aplica a diferentes gêneros, domínios ou linguagens e, em particular, a textos em português, como será ilustrado neste trabalho.

Introdução

Este artigo ilustra a aplicação de um modelo de discurso para a preservação da idéia central de um texto na sumarização automática. A modelagem proposta utiliza regras comunicativas e lingüísticas, ressaltando o processamento profundo do discurso. Logo, o sumarizador é delimitado à estratégia de *estruturação* dos sumários. Nesse contexto especial, a entrada do sumarizador é chamada *mensagem fonte primária* e sua saída é um plano de texto chamado *mensagem fonte do sumário*. Esta mensagem é derivada de um subconjunto do conhecimento da mensagem primária, estruturado de acordo com regras definidas em função do modelo de discurso proposto. Veremos aqui que esse modelo se aplica a textos escritos em português, desde que eles sigam um padrão de representação particular para a mensagem fonte primária.

O processo de sumarização

Neste trabalho, a sumarização compreende dois processos: a seleção do conteúdo relevante de uma mensagem fonte primária e sua organização coerente. Em ambos, três características principais devem ser observadas: a satisfação do objetivo comunicativo do discurso, a veiculação de sua proposição central e o inter-relacionamento coerente das unidades de conteúdo necessárias para satisfazer as restrições anteriores.

Na sumarização, a preservação da proposição central do discurso é crucial: a derivação da mensagem fonte do sumário envolve um exame minucioso da contribuição de *segmentos de discurso* ao discurso, no sentido de complementar a proposição central e, logo, torná-la proeminente, e também no sentido de contribuir para a satisfação do objetivo comunicativo do discurso. A natureza do processo de sumarização em foco é caracterizada, portanto, pelo aspecto comunicativo e lingüístico do discurso.

Em nível de seleção de conteúdo, o modelo reconhece informações essenciais de um sumário. Em nível de planejamento do discurso, ele se baseia em várias teorias, dentre as quais destacamos a GSDT (*Grosz and Sidner Discourse Theory*), de Grosz e Sidner (1986) e a RST (*Rhetorical Structure Theory*), de Mann and Thompson (1987). De acordo com a GSDT, há relações estruturais entre intenções que determinam a contribuição de segmentos de discurso entre si e, logo, sua contribuição ao discurso como um todo. Essas relações permitem determinar um "espaço de contribuição" entre sub-objetivos que, uma vez satisfeitos, contribuem para a satisfação do objetivo principal do discurso. Assim, o modelo de sumarização proposto é dirigido por objetivos e a sumarização consiste em construir os segmentos de discurso cujos sub-objetivos sejam relevantes no contexto discursivo. Como cada objetivo pode ser associado a uma intenção comunicativa, a função principal do modelo proposto é mapear as relações estruturais entre intenções em relações de coerência. Esse planejamento depende ainda da estrutura semântica das unidades informacionais, especificada na mensagem fonte primária. Portanto, o modelo agrega intencionalidade, coerência e semântica para construir o discurso de um sumário com base em restrições progressivas de planejamento, especificadas durante a construção do enredo de sub-objetivos.

Restrições do modelo de discurso

As seguintes restrições são fundamentais para a sumarização de mensagens fontes primárias com base no modelo de discurso sugerido:

1. A forma da mensagem fonte primária: Toda mensagem fonte primária do sumarizador é uma composição de três elementos: o *objetivo comunicativo*, a *proposição central* a veicular e a *base de conhecimento*, uma estrutura que indica o conhecimento específico do domínio, fornecendo o conteúdo informativo para a elaboração do sumário. Esta base de conhecimento é o único componente disponível para condensação e é considerada completa. Os demais componentes da mensagem fonte primária delineiam o contexto particular do discurso, segundo o qual as unidades de conteúdo podem ser supérfluas, complementares ou essenciais. Na seleção do conteúdo, modifica-se somente a base de conhecimento primária, extraindo-se suas proposições supérfluas. Como resultado, temos uma base de conhecimento mais concisa, cujas unidades de informação devem ser organizadas coerentemente para a produção da estrutura de discurso do sumário, chamada aqui de *mensagem fonte do sumário*.

2. A forma da base de conhecimento: Qualquer base de conhecimento no nosso modelo é uma estrutura hierárquica de unidades proposicionais inter-relacionadas semanticamente: em nível super-estrutural, ela segue o padrão *Problema-Solução* (P-S) (Hoey, 1983; Winter, 1976), caracterizado genericamente pela seqüência *Situação-Problema-Solução-Avaliação.* Esta seqüência pode ser bastante complexa, p.ex., a avaliação pode ser geral e, logo, parte da conclusão do texto, ou pode ser particular para cada um dos resultados obtidos com a resolução do problema. Casos particulares como esse geram outras formas distintas da seqüência genérica, p.ex., *Situação-Problema-Solução-Resultados-Conclusão*. O padrão P-S expressa o raciocínio lógico de resolução de problemas, que pode ser indicado na superfície textual por uma metalinguagem própria que indica o tipo de atividade do autor, refletida em suas decisões de escrita (Jordan, 1980). Em nível micro-estrutural, as relações semânticas (p.ex., causa-efeito - *cause*, instrumental - *enable* ou contextual - *background*) são comuns a várias línguas, p.ex., português, inglês ou francês. A base de conhecimento não apresenta estrutura discursiva, mas puramente semântica, seguindo sempre o mesmo padrão da seqüência Problema-Solução. No entanto, algumas de suas relações podem ser quase que diretamente mapeadas em relações de coerência (ver, p.ex., (Hobbs, 1985) e (Mann and Thompson, 1987)).

O padrão P-S é verificado também em textos em português. Por exemplo, a Figura 1 ilustra a estrutura proposicional de uma base de conhecimento da mensagem fonte primária que dá origem ao resumo "Deficientes Visuais", ilustrado abaixo. Neste resumo, orações numeradas correspondem a proposições do discurso.

Resumo "Deficientes Visuais" (Bueno, 1992):

1a. Este artigo trata de uma pesquisa realizada com alunos da disciplina de Orientação e Mobilidade - Habilitação de Deficientes Visuais,
1b. da Faculdade de Educação da Universidade de São Paulo.
2a. Os resultados mostraram que o treinamento de orientação e mobilidade para deficientes visuais foi eficiente
2b. para produzir mudanças de comportamento verbal e
2c. de ação nos sujeitos/alunos em Habilitação de Deficientes Visuais.

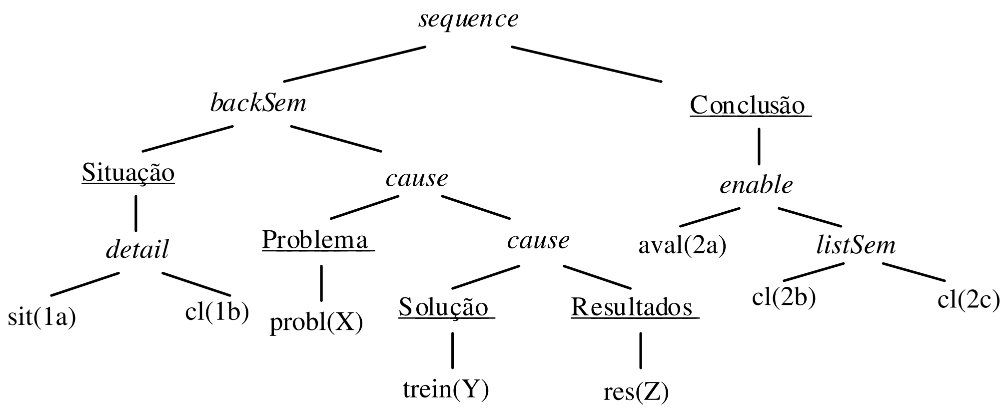

Figura 1: Uma base de conhecimento de uma mensagem fonte primária

Nessa base de conhecimento, palavras sublinhadas representam os super-componentes da estrutura informacional da mensagem fonte primária, reconhecidos como entidades conceituais do padrão P-S. Seguindo esse padrão, o grupo de todos os conceitos distintos constitui uma taxonomia conceitual por meio da qual se pode identificar cada unidade de informação de um texto, ou seja, de sua base de conhecimento correspondente. Assim, as folhas da base de conhecimento ilustrada são proposições anotadas segundo essa classificação conceitual. Por exemplo, *probl(X)* indica que X é uma unidade de conteúdo que expressa um problema.

Ocorrências de X, Y e Z na base ilustrada indicam conceitos sugeridos implicitamente no resumo, por decisões de realização lingüística. Porém, na representação semântica profunda essas informações estão presentes e podem ser expressas, p.ex., por:

• *X expressa o problema abordado, i.e., a dificuldade de orientação e mobilidade de deficientes visuais;*

• *Y indica* (mas não especifica) *que a solução do problema está no treinamento dos indivíduos, para melhorar sua orientação e mobilidade*;

• *Z indica os resultados que são observados em função do comportamento dos indivíduos após o treinamento*.

De acordo com essas proposições, podemos inferir que a avaliação do treinamento indicada por (2a) é resultado de uma análise dos resultados em Z, levando à conclusão de que a solução expressa por Y, i.e., o treinamento, é um meio válido que leva a uma mudança positiva no comportamento dos indivíduos. Esta análise é expressa pela palavra *eficiente* em (2a), que indica a avaliação dos resultados, não explicitados no texto. Dessa forma, essa avaliação "habilita" a expressão das asserções seguintes, i.e., de (2b) e (2c), ou seja, (2a) está em relação de *enablement* com (2b) e (2c), como mostra a Figura 1.

3. A forma da mensagem fonte do sumário: Diferentemente da mensagem fonte primária, esta estrutura não é uma composição de três componentes, mas sim uma única representação da estrutura profunda do discurso, que já incorpora objetivo comunicativo e proposição central no seu próprio enredo proposicional. As unidades da base de conhecimento são escolhidas e inter-relacionadas por relações de coerência que sejam apropriadas para a satisfação dos sub-objetivos comunicativos de cada segmento de discurso que contribui ao discurso como um todo. Este processo é garantido pelo planejamento dirigido por objetivos comunicativos, com a utilização de operadores de plano (Moore and Paris, 1993), recurso computacional cujo enredo constitui a própria estruturação do discurso. Esses operadores combinam relações estruturais em três níveis distintos: o intencional, o semântico e o de coerência. A proposição central é preservada garantindo-se que ela ocupe posição de destaque na mensagem fonte do sumário. Construída dessa forma, uma possível mensagem fonte do sumário "Deficientes Visuais" é obtida, p.ex., da mensagem fonte primária composta da base de conhecimento da Figura 1, do objetivo comunicativo *Relatar* e da proposição central *Enfatize a eficiência dos resultados obtidos para o treinamento de orientação e mobilidade para deficientes visuais.* Esta proposição é expressa na superfície pela oração (2a), que constitui a idéia central do resumo. Em nível profundo, ela deve ser a unidade principal a preservar na mensagem fonte do sumário. Veremos a seguir como obter esta mensagem para o resumo ilustrado, com o objetivo de preservar a proposição central durante o planejamento do mesmo.

Obtenção da mensagem fonte do sumário

No processo de sumarização apresentado, a obtenção de mensagens fontes estruturadas coerentemente é o objetivo final. Compreendendo os processos de seleção e organização do conteúdo relevante da mensagem fonte primária, a sumarização consiste em:

• condensar a base de conhecimento da mensagem primária, preservando sua estrutura hierárquica original;

• organizar coerentemente o conteúdo da base de conhecimento reduzida.

Ambos os processos são dirigidos pelas restrições fundamentais de satisfazer o objetivo comunicativo e veicular a proposição central, sendo que essas duas entidades completam a especificação da mensagem fonte primária. Em nosso modelo, esses processos são executados em *pipeline*. Para o resumo ilustrado acima, p.ex., a associação da intenção *Relatar a eficiência do treinamento*, ou *relatar(aval(2a))*, à base de conhecimento ilustrada na Figura 1 delineia as restrições necessárias para a seleção do conhecimento relevante nesse contexto específico (objetivo comunicativo: *relatar*; proposição central: *aval(2a)*), levando à base de conhecimento reduzida da Figura 2. Nesta fase, o objetivo comunicativo indica sub-estruturas da base de conhecimento diretamente relacionadas à sua satisfação, por meio de seus super-componentes (especialmente, *Situação* e *Conclusão*). Ainda,o processo de seleção verifica que a estrutura resultante "abrigue" a proposição central, para que ela seja preservada na estrutura de discurso final.

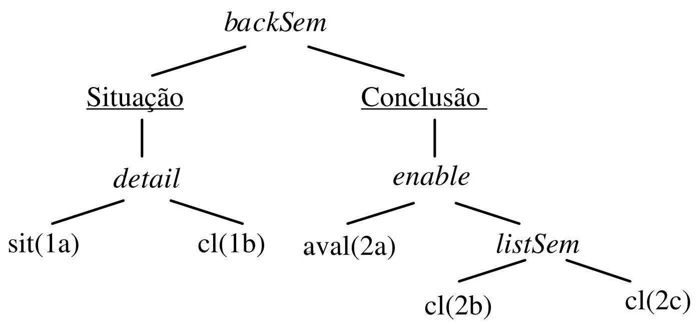

Figura 2: Base de conhecimento reduzida da mensagem fonte primária

A base de conhecimento reduzida, juntamente com o objetivo *Relatar* e a proposição central *aval(2a)*, servem de subsídios para o processo de estruturação do discurso, que finalmente produz a mensagem fonte do sumário exibida na Figura 3. Este processo é fundamentalmente dirigido pela busca da satisfação do objetivo comunicativo, com base na GSDT e, em termos computacionais, na aplicação de operadores de plano (Moore e Paris, 1993).

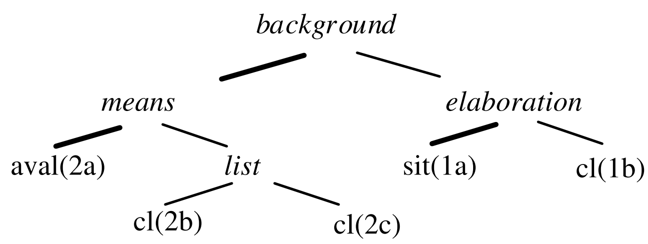

Figura 3: Mensagem fonte do resumo "Deficientes Visuais"

Logo, diferentemente da mensagem fonte primária, onde a estrutura hierárquica da base de conhecimento é semântica, esta mensagem é uma estrutura de discurso, i.e., um plano de texto onde as unidades de informação se relacionam por meio de relações de coerência (que são os nós intermediários da árvore ilustrada). Esse plano serve de base para o resumo "Deficientes Visuais": realizado lingüisticamente, ele dá origem à composição das sentenças (1) e (2), sendo que a idéia central é indicada na segunda e é correspondente à proposição central em posição "mais nuclear" na mensagem fonte do sumário (ramos nucleares são indicados em negrito na Figura 3).

Metodologia para a obtenção do modelo de discurso

O modelo de discurso que serve de base para os processos de seleção e organização do conteúdo foi obtido pela associação dos seguintes fatores:

Análise lingüística: Textos científicos escritos por nativos da língua inglesa forneceram a taxonomia conceitual das entidades significativas do discurso e o próprio padrão P-S, com o auxílio da interpretação e do reconhecimento de palavras ou orações sinalizadoras da *metalinguagem* (p.ex., *treinamento*, *produção*, *eficiente*) e de sinais lingüísticos (p.ex., as frases sublinhadas no resumo "Deficientes Visuais"), comumente usados pela comunidade de discurso em foco.

Reconhecimento de aspectos comunicativos: O discurso científico é caracterizado por textos típicos, p.ex., textos descritivos, expositivos ou argumentativos, que levam aos objetivos comunicativos *Descrever, Relatar* e *Discutir*, respectivamente. Esses objetivos são os únicos de nosso modelo que podem dar origem a um sumário e, logo, compor uma mensagem fonte primária. Eles podem ainda ser satisfeitos por outros sub-objetivos, p.ex., uma descrição de uma característica pode se dar pela descrição de um atributo diferenciável da mesma. Desse modo, especificamos uma *rede de herança de objetivos* que determina a contribuição de cada sub-objetivo à satisfação de um objetivo principal. Essa contribuição é especificada em função das relações de dominância e precedência entre intenções da GSDT (Grosz and Sidner, 1986). Apesar de altamente abstratas e independentes do comportamento lingüístico, essas relações intencionais são associadas a relações de coerência e permitem, portanto, a busca da estruturação coerente do discurso.

Especificação de relações de coerência: As relações de coerência espelham intenções comunicativas, ao mesmo tempo que expressam a dependência conceitual entre unidades de conteúdo. Dessa forma, elas agregam os aspectos informativo e intencional do discurso em uma unidade de significado. Em nosso modelo, elas são derivadas da RST (Mann and Thompson, 1987).

Esses fatores caracterizam a associação de relações estruturais do discurso em três níveis distintos de representação: o da intencionalidade, o da coerência e o da semântica. Ao buscarmos a estruturação coerente de um discurso, buscamos a determinação de relações de coerência em um contexto específico delimitado pelas intenções (ou objetivos comunicativos) e pelo relacionamento semântico das unidades informacionais. Assim, o modelo proposto complementa a teoria RST com o aspecto intencional da GSDT, garantindo a especificação de relações de coerência adequadas aos propósitos comunicativos do discurso, que não violem a dependência semântica original indicada pela mensagem fonte primária. A agregação entre intencionalidade e coerência é crítica para a sumarização, pois é ela que determina a contribuição de cada segmento à preservação da proposição central do discurso e à satisfação de seu objetivo comunicativo.

Formas de implementação do modelo de sumarização

O modelo proposto para a estruturação coerente de um sumário, cujo cenário particular é apresentado na Figura 4, não foi ainda implementado. Entretanto, exploramos alguns aspectos dessa implementação, a serem investigados com profundidade no futuro.

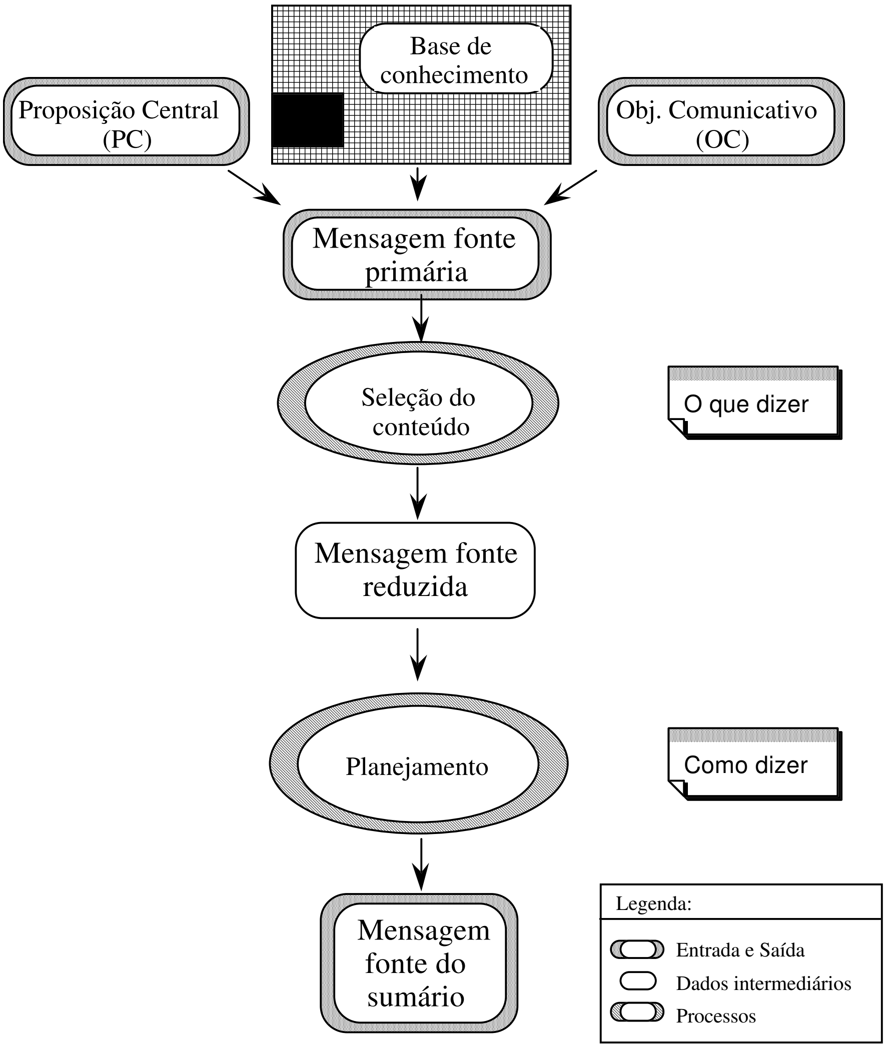

Figura 4: Um cenário particular de sumarização

Consideramos particularmente o processamento profundo do discurso devido aos seguintes motivos:

- O processamento superficial somente reflete a coerência por meio de recursos da linguagem, mas não a cria (Halliday and Hasan, 1976). Logo, o problema central da coerência do discurso deve ser tratado em nível profundo.
- Os aspectos problemáticos da geração automática são, primordialmente, os que abordam questões de discurso, e não de gramática.
- Experiências acadêmicas substanciais para a construção de linearizadores com enfoque na língua inglesa já foram realizadas, p.ex., FUF (Elhadad, 1991) ou Penman (Mann, 1983). No entanto, não se obteve até o momento avanço substancial sobre as questões discursivas, particularmente sobre a questão da coerência do discurso na geração automática de um modo geral e, tampouco, na sumarização. Resolvendo-se a questão discursiva, a realização gramatical pode se beneficiar dos subsídios fornecidos pelos linearizadores existentes. Em

particular, é possível colocar em perspectiva sua re--utilização para uma gramática da língua portuguesa.

• Por se tratar de uma arquitetura em *pipeline*, um realizador tático pode ser acoplado ao sumarizador automático proposto, uma vez solucionados os problemas básicos de processamento do discurso. Para isso, basta que as mensagens fontes de sumários sejam especificadas em uma linguagem de interface com o realizador.

• A existência de linearizadores, mesmo que voltados à língua inglesa, permite que a investigação das questões de sumarização se concentre na abordagem discursiva da geração automática, particularmente na escolha de conteúdo adequado e em sua organização em função dos objetivos comunicativos.

Desse modo, o sumarizador como planejador de textos "herda" da geração clássica diversas características, sendo que cada um de seus processos deve ser capaz de manipular a informação de modo a manter a integridade do sistema como um todo, garantindo que a proposição central da mensagem fonte primária seja preservada e ocupe posição de destaque na mensagem fonte do sumário. O subconjunto da base de conhecimento original, resultante do processo de seleção de conteúdo e relevante para um sumário, corresponde ao que McKeown (1985) chama de *relevant knowledge pool* e constitui o fundamento para a estruturação do sumário, juntamente com a especificação do objetivo comunicativo e da proposição central. Para reconhecer e utilizar esse *pool* de informações, os seguintes critérios devem ser observados:

Critério 1: A necessidade de concisão restringe tanto a seleção quanto a organização do conteúdo.

Critério 2: O objetivo comunicativo delineia não só as possíveis informações do *pool*, como também os segmentos de discurso que contribuem para sua satisfação. Como conseqüência, o espaço de (sub-)objetivos comunicativos determina a estratégia retórica do discurso.

Critério 3: A estratégia retórica (e, portanto, a própria ordem em que informações ou segmentos de discurso são relacionados) é regida também pela proposição central, por ser a informação que deve ocupar posição de destaque na estrutura de discurso do sumário.

Implementação do processo de seleção do conteúdo

A seleção se baseia tanto no contexto quanto no significado e função de cada segmento de informação, segundo as relações semânticas (nível micro-estrutural) e a estrutura típica da base de conhecimento (nível super-estrutural). Heurísticas de seleção de conteúdo atuam nos dois níveis de representação - o super-estrutural e o micro-estrutural. No nível super-estrutural, elas ressaltam os segmentos de discurso para possível exclusão; no nível micro-estrutural, elas ressaltam proposições (simples ou complexas) com o mesmo fim, utilizando seu inter-relacionamento semântico: em nível informativo, as relações semânticas delineiam unidades que podem ser consideradas supérfluas. Ao excluir uma proposição do par, a relação que as une perde sua função na estrutura da base de conhecimento e é omitida. Esta operação é correspondente a considerar somente o núcleo de uma relação RST em uma árvore hierárquica. No entanto, neste nível do processamento não estamos trabalhando com a estrutura de discurso, mas sim com a estrutura semântica típica do discurso sob investigação. Mais detalhes desse processo podem ser encontrados em (Rino and Scott, 1994).

Implementação do planejador

Para a organização do conteúdo, destacamos os modelos de processamento que utilizam esquemas e operadores de planos (Cawsey, 1993; Hovy, 1988; Maybury, 1992;

McKeown, 1985; Moore and Paris, 1993; Moore, 1995; Paris, 1988). A construção do "espaço estratégico" se inicia com a determinação de esquemas (McKeown, 1985) que podem satisfazer o objetivo comunicativo. A definição desses esquemas é possível devido ao discurso altamente estereotipado sob enfoque. Eles se distinguem dos operadores de plano pela sua função: enquanto esquemas determinam os possíveis operadores iniciais (i.e., aqueles que caracterizam a busca da satisfação do objetivo comunicativo), os operadores buscam a construção hierárquica de segmentos de discurso que contribuem para o objetivo comunicativo, com base em seus objetivos particulares.

A construção gradativa do discurso corresponde à busca do conjunto de operadores de plano que, aplicados com sucesso, satisfazem o objetivo comunicativo. Diferentes planos podem existir devido a diferentes inter-relacionamentos entre os operadores que, em última instância, determinam formas retóricas distintas para a estruturação do discurso particular. Detalhes sobre esse procedimento são encontrados em (Rino, 1996).

Contribuições e conclusão

O trabalho de pesquisa apresentado neste resumo é um trabalho original, pela combinação das questões comunicativa e lingüística. Na geração automática, tem-se perseguido uma modelagem que agregue esses componentes; na sumarização, conhecemos uma única proposta nessa mesma linha de pesquisa, de Sparck Jones (1993), mas não temos conhecimento da continuidade dessa proposta.

Nosso modelo aprofunda as questões abordadas por Sparck Jones. Embora ele tenha sido obtido originariamente de uma análise de textos científicos em inglês, mostramos aqui que ele se aplica também a textos em português, pois as relações semânticas, as relações de coerência e as relações intencionais utilizadas não são exclusivas da linguagem, tampouco do domínio particular de qualquer texto. Em (Rino, 1996) aplicamos o modelo a textos completos e variados, tanto em seu domínio quanto em seu gênero, e os resultados obtidos manualmente são satisfatórios. Logo, é possível utilizá-lo na sumarização automática de textos em português, desde que se considere mensagens fontes primárias que obedeçam o padrão Problema-Solução em sua estrutura hierárquica da base de conhecimento.

Em termos práticos, as restrições de planejamento determinadas pela associação dos três níveis distintos de representação das relações estruturais do discurso podem ser incorporadas aos operadores de plano na forma de restrições para sua aplicação ou de especificações particulares para sua expansão em outros operadores de plano. Estes, em última instância, podem ser diretamente expressos por relações de coerência e, assim, é possível construir estruturas de discurso coerentes.

Desse modo, a viabilização computacional da sumarização automática de textos em português depende da re-utilização de alguns formalismos já existentes, que deve ser enriquecida com a modelagem proposta para a preservação da idéia central de um texto.

A arquitetura do sumarizador automático sugerida na seção anterior pode acomodar a modelagem de discurso proposta, para a qual apresentamos uma aplicação específica no âmbito da língua portuguesa. Essa abordagem é independente da linguagem escolhida, por tratar somente dos aspectos profundos do processamento do discurso. Essa abrangência do modelo faz deste um trabalho proeminente e relevante para a lingüística computacional de um modo geral e, em particular, para a língua portuguesa. Sua extensão ou averiguação para esta língua constitui um objeto de estudo promissor e sugere vários aspectos a abordar, dentre os quais destacamos inicialmente a própria implementação do modelo para o planejamento dirigido por objetivos comunicativos e, posteriormente, o acoplamento de um realizador gramatical da língua portuguesa para expressar mensagens fontes de sumários na forma textual.

A questão da obtenção da mensagem fonte primária, que deveria ser resultante de um processo de interpretação automática, também deve ser alvo de estudo, embora não tenha sido abordada neste trabalho. Tanto o tópico de geração apresentado neste artigo quanto o de interpretação são, certamente, de interesse para a comunidade de lingüística computacional, extrapolando o âmbito da comunidade falante da língua portuguesa.

Referências

Bueno, G.A. (1992). Orientação e mobilidade na habilitação de deficientes visuais. *Rev. Fac. Educação*, 18 (2), pp. 205-215.

Cawsey, A. (1993) Planning Interactive Explanations. *Int. Journal of Man-Machine Studies* 38, pp. 169-199.

Elhadad, M. (1991) *FUF: the Universal Unifier User Manual, Version 5.0*. Department of Computer Science, Columbia University, 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. and Hasan, R. (1976) *Cohesion in English*. Longman Group Ltd.

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

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

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

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

Mann, W.C. (1983) *An Overview of the Penman Text Generations System*. ISI Research Report, ISI/RR-83-114, USC.

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

Maybury, M.T. (1992) Communicative Acts for Explanation Generation. *Int. Journal of Man-Machine Studies* 37, pp. 135-172.

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

Moore, J.D. (1995) *Participating in Explanatory Dialogues: Interpreting and Responding to Questions in Context.* The MIT Press.

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

Paris, C. L. (1988) Tailoring Object Descriptions to the User's Level of Expertise. *Computational Linguistics*, Vol. 14, No. 3, pp. 64-78.

Rino, L.H.M. (1996). *Modelagem de Discurso para o Tratamento da Concisão e Preservação da Idéia Central na Geração de Textos*. PhD. Thesis. Universidade de São Paulo, Brasil.

Rino, L.H.M. and Scott, D. (1994) Automatic generation of draft summaries: heuristics for content selection. In A.I.C. Monaghan (ed.), *Third International Conference on the Cognitive Science of Natural Language Processing*. Dublin City University, Ireland.

Sparck Jones, K. (1993) *Discourse Modelling for Automatic Summarising*. Tech. Rep. No. 290. University of Cambridge, UK.

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