Diferenças de conteúdo de satélites das relações Parenthetical e Purpose

Élen Cátia Tomazela (elen_tomazela@dc.ufscar.br)
Lucia Helena Machado Rino (lucia@dc.ufscar.br)

Universidade Federal de São Carlos

A fim de assegurar sumários automáticos inteligíveis, este estudo centra-se na tarefa de analisar como a exclusão indiscriminada de satélites de relações RST pode prejudicar o entendimento de uma mensagem veiculada total ou parcialmente. Utilizamos, para esse fim, os sumários automáticos produzidos pelo VeinSum [Carbonel et al., 2007] para os 50 textos jornalísticos que compõem o *Corpus* Summ-it [Collovini et al., 2007]. O VeinSum é o sumarizador proposto pelo projeto ProCaCoSA, concluído em Outubro de 2008, cujo foco foi assegurar que os sumários produzidos apresentassem *clareza referencial*, isto é, a propriedade que um texto deve ter, de facilidade de identificar a quem ou a que um determinado pronome ou sintagma nominal se refere. Analisamos, então, como a exclusão de um satélite pode prejudicar a clareza referencial de um sumário, pois, caso esse contenha uma expressão anafórica e seu antecedente não esteja explícito na versão reduzida, sua inteligibilidade faz-se prejudicada.

O VeinSum é um sumarizador de estruturas RST e, portanto, recebe as estruturas dos textos-fonte que se pretende sumarizar como entrada. A RST, no entanto, não prevê o tratamento do fenômeno correferencial. Devido a isso a Teoria das Veias, ou VT [Cristea et al., 1998] também é utilizada pelo sumarizador. Ela se baseia numa árvore RST para determinar um conjunto de unidades de discurso que possa conter possíveis antecedentes de uma expressão anafórica, denominado *domínio de acessibilidade referencial*, ou *acc,* o qual deve também ser incluído no sumário caso a unidade textual que contém uma anáfora o for. Para determinar as unidades que farão parte de um sumário, o VeinSum deve obedecer simultaneamente à VT e ao Modelo de Saliência [Marcu, 1997]. Este é, na verdade, o módulo que atribui graus de relevância, ou saliência, às unidades textuais, para que o sumarizador possa escolher as que farão parte do sumário. Ele se baseia na posição que elas ocupam na árvore RST. Esse modelo também não julga a interdependência referencial entre diversas unidades textuais para classificá-las por saliência. Por esse motivo, no VeinSum, ele é associado à VT: a cada unidade saliente incluída, acrescenta-se no sumário seu *acc* inteiro, para que não haja quebras de clareza referencial.

O problema, aqui, é que a VT determina o *acc*, sobretudo, topologicamente, isto é, ela não faz uso explícito de conhecimento semântico para descobrir quais unidades são correferentes. Assim, o cálculo do *acc* não é preciso e ele pode indicar unidades textuais que nada têm a ver com o contexto referencial de uma anáfora. Como resultado, informações de importância secundária também podem ser incluídas no sumário, comprometendo a taxa de compressão e a informatividade como um todo.

Para esses casos, nossa proposta consiste na diminuição do *acc,* através do uso de etiquetas semânticas provenientes do *parser* PALAVRAS [Bick, 2000] e da WordNet [Fellbaum, 1998]. Tal proposta consiste na comparação da etiqueta semântica do núcleo do sintagma nominal anafórico e os demais núcleos dos sintagmas nominais que fazem parte do acc. As heurísticas resultantes deverão descartar as unidades textuais que não possuam, em seus núcleos, etiquetas semânticas iguais, similares ou

hierarquicamente relacionadas à etiqueta da expressão anafórica e preservar as unidades textuais que compartilhem tais etiquetas. Com a diminuição do acc, o Modelo de Saliência será melhor respeitado e o nível de informatividade do sumário gerado tende a melhorar, o que reflete mais fielmente o conteúdo do texto-fonte.

Além dos casos em que a VT aponta unidades textuais que nada têm a ver com contexto referencial da anáfora como parte do *acc*, existem ainda os casos em o real antecedente da anáfora não faz parte desse conjunto. São os casos em que o sumário apresenta obscuridade referencial, as chamadas quebras de clareza referencial. Essas podem ser de várias naturezas, como por exemplo: pronomes sem antecedente, siglas e nomes parciais sem seus referentes completos, entre outras. Após uma análise dessas quebras, verificou-se que grande parte delas se deve à exclusão da relação retórica PARENTHETICAL. Essa é de natureza estrutural e, segundo a tradução de [Pardo, 2005], apresenta informação extra relacionada ao núcleo, mas que não está expressa no fluxo principal do texto, mas sim através de parênteses ou travessões. A princípio, a exclusão dessa relação seria natural para a modelagem de SA, pois supõe-se que, em geral, informações parentizadas constituem detalhes e, assim, poderiam ser excluídas, porém essas informações podem, muitas vezes, ser essenciais para o entendimento da mensagem, ou ainda, ajudar o leitor a compreender melhor o texto. A seguir, ilustramos alguns casos do *corpus* Summ-it.

O conteúdo da relação PARENTHETICAL

Um exemplo de quebra de clareza referencial pode ocorrer quando o referente completo de uma sigla expressa no fluxo principal do texto encontra-se parentizado, como ilustrado em (1). A relação PARENTHETICAL correspondente teria como informação nuclear (N) na árvore RST a sigla e como informação satélite (S), o conteúdo parentizado[1], o qual, neste caso, remete ao referente. Se não incluirmos esse satélite no sumário, a sigla CCNE ficará sem referente, o que poderá trazer problemas de compreensão para o leitor.

> (1) No início do mês, o CCNE (Comitê Consultivo Nacional de Ética), órgão que orienta o governo francês sobre aspectos éticos da biotecnologia, reforçou a posição da ministra, alegando que "o conhecimento da seqüência de um gene não pode ser assimilado como produto patenteado e, portanto, não é patenteável".

Além disso, essa relação pode ser utilizada para explicar um termo técnico específico utilizado no decorrer do texto, o que é incluído para esclarecer o leitor leigo sobre um termo que não lhe é comum, como em (2) e (3):

> (2) O grupo já tem em funcionamento um outro espectrômetro que também "pesa" moléculas orgânicas, mas usando uma técnica mais antiga: o bombardeamento da amostra e feito por partículas de califórnio (um elemento radioativo) em vez de laser.
>
> (3) A partir dessas observações, eles calcularam a órbita do astro e constataram que se trata de um objeto que passa a maior parte do tempo no halo galáctico - a região que circunda a Via Láctea.

[1] Na relação PARENTHETICAL a informação veiculada por S apresenta detalhes da informação expressa pelo N e pode se manifestar entre parênteses ou como um aposto. Nos segmentos textuais de exemplo ela se encontra sublinhada.

Outras vezes, porém essa relação pode trazer informações extras que, se retiradas do sumário, não prejudicarão a mensagem veiculada, como no caso de uma página de internet citada entre parênteses ou quando a informação veiculada é usada para especificar algo, como ilustrado em (4), (5) e (6):

(4) Os resultados estão na edição de hoje da revista "Science" (www.sciencemag.org).

(5) Só ontem, cinco pessoas morreram na Ásia - quatro delas em Hong Kong, de onde a doença se espalhou pelo mundo.

(6) O impacto dos destroços teria ocorrido às 18h06 (horário de Brasília), numa área remota do oceano.

Assim, é possível verificar que o conteúdo da relação PARENTHETICAL pode, muitas vezes, ser excluído em SA, porém não se pode generalizar que essa relação poderá sempre ser descartada em um sumário sem prejuízo à veiculação da mensagem. O mesmo ocorre com o conteúdo da relação PURPOSE, descrita mais detalhadamente a seguir.

O conteúdo da relação PURPOSE

Problemas de informatividade nos sumários também podem ocorrer envolvendo a relação retórica PURPOSE. Essa relação pode conter trechos que, se não incluídos no sumário, podem prejudicar a mensagem veiculada, ou ainda, os que não farão falta para o entendimento da mesma, caso não sejam incluídos. Ilustramos abaixo trechos do *corpus* Summit.

Caso os trechos sublinhados não constem no sumário, a mensagem veiculada fica incompleta em (7), (8) e (9):

(7) Um grupo do Instituto de Física da USP está ajudando a inflar ainda mais o ego da categoria, ao construir o primeiro aparelho no país que usa laser para analisar moléculas biológicas.

(8) As mudanças nas populações de pingüins também serviram como indicativo de que as coisas não andavam bem: as espécies que usavam geleiras para se abrigar e procriar começaram a diminuir na região, enquanto os pingüins acostumados a mar aberto colonizaram a península.

(9) Capitaneados por Sir David King, o principal assessor científico do governo britânico, os pesquisadores não pouparam esforços para demonstrar que o aquecimento global já está pondo em risco as vidas e a economia humanas em diversas regiões.

Já em (10) e (11), a exclusão do conteúdo dessa relação não é prejudicial ao entendimento da mensagem:

(10) “Aquela área toda do rio é, na sua maior parte, arenosa. Se você tirar a cobertura vegetal para colocar pasto ou agricultura, a primeira chuva arrasta uma enormidade de areia para o rio” diz Emiko de Resende.

(11) Segundo o Ibama, nas propriedades multadas houve desmatamento nas margens do rio, com a finalidade de abrir espaço à agricultura ou à criação de gado.

Portanto, percebemos, após esse estudo de quebras de clareza referencial e de problemas de informatividade, que somente a VT e o Modelo de Saliência não são suficientes para a identificação de trechos que não podem ser excluídos do sumário, já que não fazem processamento intraoracional e se baseiam somente na configuração da estrutura RST. Para os casos ilustrados, é preciso investigar padrões de similaridade

entre os trechos que podem ou não ser excluídos, para que sejam marcados no momento da seleção de unidades textuais a serem incluídas no sumário e, conseqüentemente, melhorar sua qualidade.

Equipe de trabalho: Élen Cátia Tomazela (aluna), Lucia Helena Machado Rino (orientadora)

LaLiC - Laboratório de Linguística Computacional
Página de Internet: lalic.dc.ufscar.br

Referências Bibliográficas

Bick, E. 2000. *The Parsing System PALAVRAS: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework.* Arhus: Arhus University.

Carbonel, T. I.; Pelizzoni, J. M.; Rino, L. H. M. 2007. VEINSUM: Um Modelo de Sumarização Automática de Textos Baseado em Estruturas Retóricas. *CoPG - Congresso de Pós-Graduação da USFCar*, São Carlos - SP.

Collovini, S.; Carbonel, T. I.; Fuchs, J. T.; Coelho, J. C.; Rino, L. H. M.; Vieira, R. 2007. Summ-it: Um corpus anotado com informações discursivas visando à sumarização automática. In V. Quental, C. Oliveira (eds.). *Proc. of the V Workshop on Information and Human Language Technology (TIL'2007, CD-ROM). XXVII Congresso da Sociedade Brasileira de Computação (SBC'2007)*, Rio de Janeiro - RJ.

Cristea, D.; Ide, N.; Romary, L. 1998. Veins Theory: A Model of Global Discourse Cohesion and Coherence. *Proc. of the Coling/ACL 1998*, pp. 281-285.

Fellbaum, C. D. 1998. *WordNet: an electronic lexical database*. Cambridge: The MIT Press.

Marcu, D. 1997. The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts. PhD Thesis. *Department of Computer Science*, Toronto, Canada: University of Toronto.

Pardo, T. A. S. 2005. Métodos para Análise Discursiva Automática. *Instituto de Ciências Matemáticas e de Computação*, pp. 211. São Carlos: Universidade de São Paulo.