A importância do tratamento co-referencial para a sumarização automática de textos

Lucia Helena Machado Rino[1], Eloize Rossi Marques Seno

Departamento de Computação – UFScar
[1]Também Departamento de Letras – UFScar
Universidade Federal de São Carlos
Caixa Postal 565 – São Carlos – SP
lucia@dc.ufscar.br, eloize@mail.fpte.br

Resumo. *Um dos principais problemas de textualidade de sumários automáticos é a ausência de resolução anafórica: quando uma anáfora é escolhida para a composição do sumário, mas seu antecedente não, sua compreensão e, logo, sua correspondência com o texto-fonte pode ser prejudicada. Na Sumarização Automática, esse problema é grave porque não existem recursos computacionais sofisticados, nem mesmo de resolução anafórica. Este artigo discute especialmente o tratamento da co-referenciação para garantir que sumários automáticos sejam coerentes. A importância desse tratamento é reforçada pelas teorias adotadas, de uso promissor na Lingüística Computacional.*

Palavras-chave. *Sumarização Automática; Lingüística Computacional; Estruturação de textos.*

Abstract. *One of the main problems concerning the textuality of automatic summaries is the lack of anaphoric resolution: when an anaphor is chosen to compose a summary but its antecedent is not, comprehension and, thus, correspondence of the summary with its source text may be jeopardized. In Automatic Summarization, this problem is serious because there are no sophisticated computer resources or tools so far, even for anaphora resolution. This article discusses especially the coreferencing problem to guarantee that automatic summaries be coherent. The importance of such a proposal is reinforced by the underlying theories, which are promisingly used in Computational Linguistics.*

Keywords. *Automatic Summarization; Computational Linguistics; Text Structuring.*

1. Introdução

A sumarização automática de textos, como uma aplicação do PLN (Processamento Automático das Línguas Naturais), restringe-se historicamente a duas metodologias básicas: a baseada em informações lingüísticas – abordagem profunda ou rica em conhecimento – e a baseada em informações estatísticas ou empíricas – abordagem superficial ou pobre em conhecimento. Ser rica ou pobre em conhecimento, nesse contexto, significa diferenciar o tipo de conhecimento incorporado ao sistema, que lhe servirá de base para decisões de processamento: na abordagem profunda adotam-se modelos lingüísticos ou discursivos que remetem à área fundamental de estudo das línguas naturais; na superficial, modelos empíricos, matemáticos ou estatísticos, são

usados visando, sobretudo, ganhos computacionais, embora visem implicitamente a modelagem lingüística. A dicotomia entre as duas abordagens é crucial para a complexidade e computabilidade dos sistemas em construção. Atualmente, é muito mais simples construir um sistema de PLN superficial, dada a disponibilidade de modelos e de recursos computacionais. No entanto, as aplicações de PLN de interesse utilizam esses recursos somente em etapas de pré-processamento, de cujos resultados partirá o processamento computacional central da aplicação real.

As condições características dos sumarizadores automáticos incluem a identificação de segmentos textuais relevantes para compor um sumário e a preservação da mensagem principal do texto a sumarizar (texto-fonte). Ambas as tarefas, quando realizadas pelo leitor humano, remetem à interpretação do texto. Se escritor competente, o leitor consegue identificar as informações mais importantes e apreender a mensagem do texto, bastando-lhe reescrever o sumário sob essas restrições, para garantir que haja uma única mensagem comum ao sumário e seu texto-fonte. Um sumarizador de três passos (Sparck Jones, 1993) – interpretação, transformação e reestruturação da informação mais relevante – decorre a modelagem dessa tarefa humana. Porém, as propostas mais significativas fazem a correspondência de unidades de informação, ou *proposições*, com segmentos textuais. Embora questionável do ponto de vista discursivo (Grimes, 1975; Grosz & Sidner, 1986; Polanyi, 1988; 1993), ela torna os sistemas computáveis, já que modelos de apreensão do discurso são de difícil computabilidade[1].

Na abordagem superficial, modelos de cômputo da concentração e representatividade de grupos textuais diversos existem que usam o reconhecimento de padrões textuais como principal método para a Sumarização Automática[2] adotam essa abordagem (p.ex., Edmundson, 1969; Salton et al., 1994), que, em geral, se originaram com a proposta de Luhn (1958), de cômputo da freqüência de meras palavras do texto: assumindo-se somente as palavras *ricas em significado* (S, ADJ, V e ADV), seria possível construir sumários de textos-fonte classificando-as descendentemente por sua freqüência textual. O problema dessa proposta é que as sentenças, extraídas de seu contexto sem qualquer processamento lingüístico ou reestruturação textual, são meramente justapostas no sumário final. Claramente, a coerência fica em risco e também a preservação do tópico ou idéia central do texto. Mais recentemente, outras propostas superficiais para contornar esse problema trouxeram avanços significativos para a área (p.ex., Hearst, 1997; Barzilay & Elhadad, 1997), várias delas fundamentadas em modelos lingüísticos de coesão lexical, destacando-se os de Halliday & Hasan (1976) e de Hoey (1991). No entanto, o problema da textualidade persiste na maioria dos sistemas propostos, razão pela qual medidas diversas de utilidade, abrangência de conteúdo, etc., são utilizadas para avaliar o desempenho dos sumarizadores automáticos[3].

Do ponto de vista prático, sumários podem ser não-estruturados em textos, como uma lista de tópicos principais de um texto, ou índice (normalmente utilizada na Ciência da Informação), ou uma lista de palavras-chave (uma das acepções da Recuperação da Informação), claramente úteis para fins específicos, mas cujo caráter de textualidade é exigido do leitor. Enquanto os métodos superficiais podem se aproximar desse tipo de resultado, a Sumarização Automática profunda busca sumários *textuais* e, portanto, *resumos*, na acepção mais comum no Brasil. Vale notar que adotamos o termo *sumário*, em vez de *resumo*, por sua correspondência explícita com o termo *sumarização automática*, processo que dá nome à área, nacional ou internacionalmente.

Caracterizadas as duas grandes abordagens da Sumarização Automática, motivamos a abordagem profunda e descrevemos as principais propostas correlatas na Seção 2. A base

teórica de um sistema profundo que contempla critérios de textualidade é descrita na Seção 3, juntamente com ilustrações do problema em foco. Avaliações preliminares realizadas em comparação com outros sistemas de Sumarização Automática são apresentadas (Seção 4), que delineiam perspectivas interessantes para a resolução co-referencial na Sumarização Automática (Seção 5).

2. Motivações para a abordagem profunda de Sumarização Automática

O P&D de um sumarizador automático profundo se apresenta como principal proposta de Sumarização Automática, quando não há recursos ou ferramentas computacionais robustos para elaborar suas decisões centrais. Mesmo quando eles estão disponíveis, a produção de textos, no sentido estrito, não é a condição básica dos métodos superficiais, embora seus resultados permitam recuperar parte da trama discursiva. Considerando o fenômeno em foco neste artigo, da co-referenciação anafórica, abordagens profundas limitadas, e mesmo as superficiais, podem levar à escolha de uma sentença anafórica referencial para compor um texto sem que sua correspondente sentença antecedente o seja. Esta já é, em si, uma boa motivação para a abordagem profunda. Entretanto, não é suficiente, se a condicionarmos à existência de resolvedores referenciais automáticos, que ainda constituem um dos tópicos mais complexos de pesquisa no PLN (Mitkov, 2002).

A co-referenciação anafórica é estabelecida entre uma expressão de referência (anáfora) e um termo que a antecede no texto (antecedente). Embora ela seja definida como o fenômeno estabelecido quando cadeias lexicais são reproduzidas ao longo do texto a partir de uma referência a uma entidade já introduzida na comunicação (Milner, 2003), aqui ela é limitada às indicações da superfície textual. Mais especificamente, contemplamos somente descrições definidas, isto é, aquelas formadas por sintagma nominais iniciados por um artigo definido, por exemplo, *o presidente, o médico do hospital,* etc. (Russell, 1905)[4]. Havendo a interpretação textual anterior à sumarização automática, um resolvedor anafórico poderia identificar o encadeamento referencial. As etapas de seleção de conteúdo e estruturação do sumário teriam, então, as unidades informativas plenamente resolvidas e a garantia de coerência ficaria inteiramente condicionada à estruturação textual do sumário.

O dilema estabelecido é que a resolução anafórica é vital para a maioria dos sistemas de PLN: embora as descrições definidas tenham um significado independente de seus antecedentes (Mitkov, 2002), a ausência de resolução anafórica pode induzir à descontinuidade referencial e, assim, a mensagens incompatíveis de um sumário, se comparado a seu texto-fonte. No caso extremo, ela impedirá a compreensão, permitindo que a quebra de cadeias de co-referência (CCRs) inviabilize a remissão textual do interlocutor (Koch, 2005). É importante notar que esse dilema é tão mais sério quanto a complexidade de se modelar os aspectos do mundo real relacionados ao fenômeno da referenciação. É por esse motivo que, no PLN, limita-se a natureza da referenciação ao fenômeno superficial, assumindo-se que o referente seja sempre o próprio antecedente no texto e, como conseqüência, que o encadeamento co-referencial se encontra auto-contido na entrada do sistema.

Na abordagem profunda, a detecção de CCRs exige recursos robustos, como analisadores morfológicos e anotadores sintáticos ou dicionários e ontologias eletrônicos. Requer-se pelo menos o *parsing* parcial para identificar uma descrição definida, restando, então, a busca pelo escopo de referenciação: segmentos textuais candidatos a antecedente devem ser automaticamente determinados. Por razões de computabilidade, esse escopo deve ser limitado. Métodos variados para sua deteminação incluem os modelos lineares (Cristea et al., 2000), que

buscam nas sentenças imediatamente precedentes à anáfora seu antecedente; os modelos hierárquicos (Cristea et al., 1998), que consideram unidades de discurso hierarquicamente precedentes. Cristea et al. (2000) mostram que os hierárquicos apresentam esforço de busca menor que os lineares, sendo menos custosos computacionalmente e, possivelmente, mais eficazes ao indicar antecedentes potenciais. No entanto, a identificação automática de unidades estruturais, baseada na *Veins Theory* (VT), não é satisfatória (Mitkov, 2002): após se determinar as unidades candidatas a antecedentes, resta ainda selecionar uma delas, o que requer ainda analisadores semânticos, meta também ambiciosa para o PLN. Várias teorias propõem o tratamento dessa questão, dentre as quais destacam-se, pelo valor computacional[5], a *Centering Theory* (Grosz et al., 1995), a *Binding Theory* (Chomsky, 1981; 1995), a Discourse Representation Theory (Kamp & Reyle, 1993) e a própria VT, a qual retomaremos na Seção 3.

Já a abordagem superficial é motivada pelo expressivo desenvolvimento de ferramentas robustas para o PLN e em corpora representativos que viabilizam procedimentos empíricos de treinamento automático dos sistemas computacionais. Padrões *colocacionais* (ou padrões de co-ocorrência), em geral, são aprendidos, alguns deles baseados em anotações sintáticas. Padrões de preferências léxicas também são usados para buscar o escopo de referenciação, com medidas estatísticas dos candidatos. É importante notar que essas propostas são pesadamente dependentes de textos pré-processados e, portanto, de ferramentas específicas, e também da construcao de corpora robustos, de preferência anotados sintaticamente. De um modo geral, as abordagens atuais privilegiam fenômenos distintos de co-referenciação. A única que tem como foco as descrições definidas é a de Vieira e Poesio (2000), que propõem um sistema superficial baseado em informações estruturais do texto, lexicais (extraídas de um dicionário/recurso eletrônico) e em informações anotadas manualmente ou adquiridas de corpora. O mesmo problema anterior se aplica também a esta proposta: embora haja avanços para o inglês, ainda não é possível se estabelecer um ambiente adequado para o processamento superficial do português. Além disso, as diversas propostas ainda não foram avaliadas consistentemente, devido aos diversos graus de pré-processamento, à adoção de diferentes amostras de teste e aos focos diversos de cada uma. Essas são as principais razões da abordagem profunda relatada aqui, restrita à busca de preservação de CCRs inter-oracionais em textos em português.

3. Base teórica do sumarizador automático profundo RHeSumaRST

Visando o tratamento automático da co-referenciação, propusemos a implementação do protótipo RHeSumaRST (**R**egras **He**urísticas de **Suma**rização de estruturas **RST**)[6], que é baseado nos modelos de estruturação retórica do discurso da *Rhetorical Structure Theory* – RST (Mann & Thompson, 1987) e de coerência global do discurso da *Veins Theory* (Cristea et al., 1998), já referida anteriormente.

O problema específico do RHeSumaRST é evitar quebras de CCRs nos sumários. Heurísticas são escolhidas que, identificando informações supérfluas em uma estrutura retórica (ou estrutura RST) de um texto-fonte, buscam garantir que sua exclusão não prejudicará a recuperação de possíveis elos co-referenciais. Ou seja, elas condicionam à inclusão de um componente anafórico em um sumário a inclusão de seu antecedente. O sistema admite como entrada somente a estrutura RST do texto-fonte a sumarizar e como saída, ou a estrutura RST do sumário ou um texto cuja realização lingüística é elementar. Assim, o problema principal das quebras de CCRs é tratado somente pelo módulo central de um sumarizador de três passos (o de seleção do conteúdo e estruturação do sumário). Essa restrição, embora severa, teve o objetivo

de viabilizar a investigação teórica proposta. Veremos, na Seção 5, que ela poderá ser resolvida parcialmente no futuro próximo.

Em relação aos modelos teóricos adotados no RHeSumaRST, dois fatores tornam a RST interessante: a nuclearidade pode ser usada para determinar segmentos relevantes, como originalmente proposto, e a escolha das unidades discursivas para compor os sumários pode basear-se na nuclearidade. Supõe-se que a proeminência dos segmentos seja indicada pelo objetivo comunicativo do escritor e, assim, que seja possível recuperar seu propósito de construção do texto e, portanto, sua idéia central. Um texto bem-formado será, então, representado por uma estrutura RST completa, isto é, por unidades elementares do discurso (ou UEDs) fortemente conectadas por relações retóricas. UEDs são unidades informativas mínimas, ou unidades mínimas de significado. Por permitir a organização das informações em núcleos (Ns) e satélites (Ss), as relações RST de interesse para a sumarização automática são as mono-nucleares, isto é, as que se manifestam entre um N e um S. Outra característica de interesse pela RST é a existência de algoritmos para segmentar e estruturar textos retoricamente (Marcu, 1997; 1999; 2000). Embora eles não tenham sido ainda incorporados ao RHeSumaRST, suas entradas, estruturadas manualmente, seguem essas propostas: as UEDs correspondem, simplesmente, a orações demarcadas por sinais de pontuação (sentenças completas) ou orações explicitamente marcadas discursivamente, como ilustra o texto abaixo.

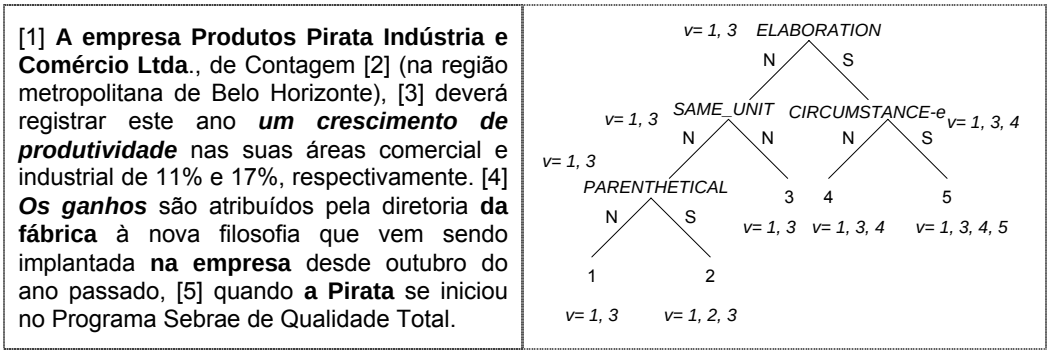

Figura 1. Estrutura RST do texto 'Empresa Pirata'

Essa é a primeira simplificação necessária para viabilizar o processo computacional: não havendo um modelo do mundo real ou a interferência humana, a simulação do processamento discursivo não encontra respaldo em modelos automáticos. Assim, as questões mais abrangentes do processamento do discurso não são consideradas. Orações restritivas, por exemplo, não são consideradas UEDs, embora elas possam corresponder a unidades mínimas de significado. Tampouco UEDs anafóricas são resolvidas, como discutimos na Seção 2, e as descrições definidas não remeterão a seus antecedentes nominais em qualquer estrutura RST, sendo representadas por UEDs dissociadas. O texto 'Empresa Pirata', que apresenta anáforas nas UEDs [4] e [5] (em negrito) e na UED [4] (em negrito e itálico), tem como antecedentes, respectivamente, as UEDs [1] e [3]. Como não existe resolução anafórica, essas UEDs segmentadas na superfície textual são as próprias folhas da estrutura RST correspondente (Figura 1).

Embora seja a nuclearidade que torna a RST interessante, como comprovam outrs trabalhos (Ono et al., 1994; O'Donnell, 1997), evitar a quebra de CCRs implica considerar a inclusão de alguns Ss em um sumário, para facilitar ou possibilitar a compreensão. Por exemplo, supondo que a idéia central a ser veiculada por um sumário

do texto exemplificado fosse a expressa pela UED [4], ou seja, a razão do progresso da empresa, o RHeSumaRST poderia considerar como supérfluos os Ss [2] e [5]. Ainda poderia tomar como sumário mínimo somente [4], já que ela é nuclear e, assim, a informação mais proeminente na relação circunstancial[7]. Entretanto, a ausência de resolução anafórica impede que a sumarização automática seja bem sucedida, devido à quebra das duas CCRs (***Os ganhos*** são atribuídos pela diretoria **da fábrica** à nova filosofia que vem sendo implantada **na empresa** desde outubro do ano passado).

Esse exemplo mostra que, para a sumarização automática, o desvinculamento de CCRs pode ser desastroso. Em síntese, decisões baseadas somente na nuclearidade indicada pela estruturação RST de textos-fonte não suprem as condições mínimas para a textualidade, em um contexto em que a estruturação textual não inclui a resolução das anáforas expressas em UEDs. A RST, por si só, não dá conta de resolver esse problema, razão pela qual o RHeSumaRST incorpora também a VT, a qual abrange a composicionalidade das unidades do discurso. Com ela, associam-se estruturas RST à delimitação do domínio de acessibilidade referencial de cada UED com base na nuclearidade: a veia de uma UED é definida como um conjunto de unidades do discurso que podem conter o antecedente de uma anáfora. A vantagem de associar ambas as teorias se deve, também, à existência de um algoritmo claro de anotação das veias sobre estruturas RST. Para a UED [4] acima, sua veia é indicada pelas UEDs [1] e [3], além de conter a própria UED (anotação *v=1,3,4* na Figura 1)[8]. Assim, os antecedentes de ambas as anáforas de [4] podem estar presentes nas UEDs [1] ou [3]. Essa veia de [4] obriga que, ao escolher [4], os demais componentes da mesma veia também sejam escolhidos, para garantir a inexistência de quebra de CCRs.

Assim, as heurísticas do RHeSumaRST se baseiam em duas hipóteses principais, associando a RST à VT, respectivamente: (a) os satélites de relações RST podem ser supérfluos e, portanto, excluídos de uma estrutura RST de um sumário; (b) os satélites que contêm antecedentes de termos anafóricos já inclusos na estrutura de um sumário não podem ser excluídos. Vale notar que, na Sumarização Automática, não há associação similar de ambas as teorias, muito embora Cristea et al. (2005) proponham o uso da VT para esse fim. No entanto, eles não utilizam estruturas RST, mas árvores binárias como representações de estruturas textuais, para a Sumarização Automática baseada em foco. No RHeSumaRST várias heurísticas podem ser aplicadas sucessivamente para diferentes relações RST. O limite de poda é dado pela taxa de compressão, expressa, nessa abordagem, pelo número aproximado de UEDs que o sumário supostamente irá conter. Genericamente, as heurísticas são representadas por uma única regra condicional:

Seja $T=\{t_1...t_N\}$ um conjunto de UEDs que compõem uma estrutura RST de um texto-fonte, $S=\{s_1...s_M\}$, $M<N$, um conjunto que contém somente as UEDs de T candidatas a compor a estrutura RST de um sumário desse texto e $V=\{v_1...v_L\}$, $L \leq N$, um conjunto de UEDs que constituem a veia de uma UED s_i, $s_i \in S$, para $1 \leq i \leq M$. A heurística genérica de poda de UEDs da estrutura RST de um texto-fonte é dada abaixo:

Se t_i $(1 \leq i \leq N)$ for satélite de uma relação RST qualquer R e $t_i \notin V$, para
V = veia de uma UED s_j, $s_j \in S$ $(1 \leq j \leq M)$,
então exclua t_i de T, reestruturando a árvore RST.

4. Avaliações preliminares do RHeSumaRST

O RHeSumaRST já foi avaliado sob duas perspectivas também sugeridas nas DUC: informatividade e coerência (Seno & Rino, 2005a). A primeira visou verificar se as heurísticas

permitiam preservar as informações mais relevantes do texto-fonte; a segunda, se elas garantiam a coerência dos sumários, isto é, a inexistência de quebra de CCRs. Para o cômputo da informatividade, foi usada a ferramenta ROUGE (Lin, 2004a; 2004b)[9], que compara a informatividade de sumários gerados por sumarizadores diversos. Foram utilizados dois outros sumarizadores automáticos: o de Marcu (1997) e um *baseline*, cujos sumários são construídos pela poda de todo satélite das estruturas RST. Para a forma mais expressiva de utilização da ROUGE – a ROUGE-1 – o RHeSumaRST foi mais informativo que o *baseline,* porém menos que o modelo de saliência de Marcu, sobre um corpus de teste de 10 textos jornalísticos que já possui seus sumários de referência, o TeMário, com um total de 5.277 palavras, aproximadamente 1,5 página cada texto (http://www.linguateca.pt/Repositorio/TeMario).

A avaliação da coerência das estruturas RST foi mais complicada, devido à necessidade de comparação manual: cada sumário produzido pelos mesmos três sistemas foi comparado com seu correspondente texto-fonte, anotado com as CCRs[10]. Desse modo, foi possível identificar as quebras de CCRs nos sumários, para os casos em que a anáfora era comprovada no texto-fonte (isto é, quando não era uma referência nova). O RHeSumaRST apresentou o menor índice de quebras de CCRs nos sumários (5%, contra 8% e 15% no *baseline* e no modelo de saliência, respetivamente). Embora esse resultado fosse esperado, pois os modelos usados para comparação não tratam explicitamente a preservação dos elos co-referenciais, seu desempenho ficou muito próximo ao dos outros, para justificar todo o esforço necessário de modelagem e processamento estrutural. Como o corpus de teste utilizado era muito pequeno, outra avaliação da coerência foi realizada (Seno & Rino, 2005b), agora considerando outro corpus jornalístico[11], composto de 20 textos anotados retoricamente por especialistas em RST, que também os anotaram com suas CCRs. Somente o *baseline* foi considerado neste segundo experimento. Embora o corpus de teste ainda fosse pequeno, o desempenho do RheSumaRST foi bem melhor (4% de quebra de CCRs, contra 18%). Essa melhora indica o potencial do RHeSumaRST para a garantia da textualidade, no que concerne a prevenção de quebras de CCRs nos sumários automáticos. Vale notar também que ela se deve à uniformidade e consistência de anotação do corpus, realizada cuidadosamente por dois analistas RST. Esses resultados, embora preliminares, indicam-nos perspectivas interessantes no processamento profundo da co-referenciação para a Sumarização Automática.

Por fim, embora a proposta de condicionamento da sumarização automática de textos à preservação das veias de UEDs já incluídas em sumários seja de autoria de Cristea et al. (1998), eles não consideram o algoritmo de determinação da saliência das UEDs em estruturas RST de Marcu, tampouco a proposta original de Mann & Thompson (1987). A diferença entre esses dois modelos está no nível de independência entre UEDs que venham a constituir uma mesma CCR: no exemplo do texto 'Empresa Pirata', um sumário mínimo, pelo modelo de saliência, levaria à escolha das UEDs [1] e [3], e, mesmo que [4] fosse considerada o foco, o cômputo da saliência permanece independente do problema da co-referenciação. No modelo da VT, [1], [3] e [4] são definitivamente interdependentes. Ao incorporar explicitamente essa interdependência à manipulação de estruturas RST, o RHeSumaRST apresenta contribuição adicional às propostas individuais.

5. Perspectivas de avanços no tratamento profundo da co-referenciação

Considerando o estado atual da Sumarização Automática no Brasil, caracterizado pela inexistência de recursos dicionarizados sofisticados (sobretudo ontológicos) para a adoção expressiva de métodos empíricos, a resolução profunda do RheSumaRST é promissora pela

simplicidade de elaboração dos algoritmos de cômputo da saliência de UEDs e de reconhecimento das veias de estruturas RST. O trabalho braçal de P&D do sistema, devido à anotação manual das CCRs e das estruturas retóricas, pode hoje ser superado com sua associação ao DiZer (Pardo, 2005), um analisador discursivo baseado no mesmo modelo de Marcu, mas voltado, agora, ao processamento de textos em português. Assim, textos reais poderão ser dados como entrada ao RheSumaRST. Com isso, a conjugação das teorias RST e VT poderá ser explorada mais rigorosamente, visando a escalabilidade e robustez do sistema.

Claramente, o tratamento computacional do fenômeno de co-referenciação é importante, por ser esse um recurso estilístico muito usado nos textos em língua natural e, assim, também nos textos em português. Um estudo realizado por Coelho (2004) mostrou que 30% dos sumários automáticos gerados pelo GistSumm (Pardo et al., 2003) apresentavam quebras de co-referência introduzidas por descrições definidas. Vieira & Salmon-Alt (2002) e Vieira et al. (2002) também apontam um alto índice de ocorrência de descrições definidas em corpus de textos jornalísticos, gênero cuja sumarização automática é de grande interesse, pela disponibilidade em larga escala desse material na Internet. Embora o P&D do RheSumaRST tenha focalizado o português, por sua abordagem profunda, ele pode ser utilizado para outras línguas naturais, já que suas heurísticas são totalmente independentes da língua-objeto do texto-fonte. Sua importância também é delineada por outras áreas do PLN que podem se beneficiar da Sumarização Automática, destacando-se a de Recuperação ou Extração da Informação, assim como a de Q&A (*Question-Answering*), que podem usar sumários como instrumentos. Ele permitirá, por exemplo, recuperar respostas a perguntas estabelecendo os elos co-referenciais entre entidades ou eventos das perguntas e as entidades ou eventos dos documentos candidatos.

Do ponto de vista lingüístico, o RheSumaRST também permitirá explorar as condições de modelagem que possam inovar sobre os modelos já existentes, visando à clareza textual. Uma das áreas que mais diretamente pode se beneficiar da iniciativa do sistema é a de processos de escrita auxiliadas por computador. O SciPo (Feltrim, 2004), por exemplo, baseia-se também em estruturas retóricas para sugerir a forma de textos científicos, podendo incorporar um módulo de sumarização. Por fim, restará comparar, no futuro, o desempenho de sistemas superficiais e profundos, ante as mesmas restrições, para verificar se todo o esforço de modelagem discursiva realmente se justifica.

---

[1] Para maior caracterização das limitações da área de Sumarização Automática, vide (Mani & Maybury, 1999).

[2] Distinguiremos, aqui, a área (Sumarização Automática) do processo (sumarização automática) adotando letras maiúsculas para a primeira.

[3] Vide o site da DUC – *Document Understanding Conference* – para maior abrangência das avaliações atuais (http://www-nlpir.nist.gov/projects/duc/data.html, Ago/2005).

[4] On Denoting. Apud (Mitkov, 2002).

[5] Todas elas referenciadas em (Mitkov, 2002).

[6] Agradecimentos especiais a Leandro Hanada, implementador do protótipo.

[7] Esta escolha dependeria de uma estratégia de aplicação de uma heurística que privilegia [4] e considera as demais UEDs supérfluas.

[8] Minimamente, a veia de cada UED de uma estrutura RST é constituída pela própria UED.

[9] Adotada nas duas últimas DUC, de 2004 e 2005 (http://www-nlpir.nist.gov/projects/duc/data.html, Ago/2005).

[10] Agradecemos César Coelho (2004) pelas anotações.

[11] Corpus Rhetalho, www.nilc.icmc.usp.br/~thiago/rhetalho.html, Ago/2005.

Referências

BARZILAY, R. & ELHADAD, M. "Using Lexical Chains for Text Summarization". In I. Mani & M.T. Maybury (eds.), *Advances in Automatic Text Summarization*. Cambridge, MIT Press, 1999. p.111-21.

COELHO, J.C.B. Cadeias de co-referência aplicadas à sumarização automática. *Mostra de Iniciação Científica* – MIC' 2004. UNISINOS, São Leopoldo – RS, 2004.

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

CRISTEA, D.; Ide, N.; Marcu, D.; Tablan, V. An empirical investigation of the relation between discourse structure and coreference. *In the Proceedings of the Coling/ACL' 2000*, 2000. pp. 208-214. Saarbrücken, Germany.

CRISTEA, D.; POSTOLACHE, O.; PISTOL, I. Summarization Through Discourse Structure. *In the Proceedings of the 6th International Conference on Computational Linguistics and Intelligence Text Processing – CICLing 2005,* Mexico, 2005.

EDMUNDSON, H.P. "New Methods in Automatic Extraction". *Journal of the Association for Computing Machinery*, v.16, n.2, 1969, p. 264-85.

FELTRIM, V. 2004. Uma abordagem baseada em córpus e em sistemas de crítica para a construção de ambientes Web de auxílio à escrita acadêmica em português. Tese de doutorado. ICMC-USP, 169 p.

GRIMES, J.E. The Thread of Discourse. The Hague: Paris. Mouton, 1975.

GROSZ, B. & SIDNER, C. "Attention, Intentions, and the Structure of Discourse". *Computational Linguistics*, v.12, n.3, 1986.

HALLIDAY, M. A.K.; HASAN, R. Cohesion in English. Longman, 1976.

HEARST, M.A. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics,* 23(1):33-64, March, 1997.

HOBBS, J. R. On the Coherence and Structure of Discourse. Tech. Report CSLI-85-37. Stanford University, Center for the Study of Language and Information, 1985.

HOEY, M. Patterns of Lexis in Text. Oxford University Press, 1991.

KOCH, I.V. Referenciação e orientação argumentativa. Em I.V. Koch, E.M. Morato e A.C. Bentes (orgs.), *Referenciação e Discurso*, , 2005. pp. 33-52. Editora Contexto. São Paulo.

LIN, C. ROUGE: a Package for Automatic Evaluation of Summaries. *In Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004),* Barcelona, Spain, 2004a.

_____ Looking for a Few Good Metrics: Automatic Summarization Evaluation - How Many Samples Are Enough?. *In Proceedings of the NTCIR Workshop 4*, Tokyo, Japan, 2004b.

LUHN, H.P. "The Automatic Creation of Literature Abstracts". In *IBM Journal of Research and Development*, v.2, n.2, 1958. pp. 159-65.

MANI, I. & MAYBURY, M. (eds). Advances in Automatic Text Summarization. Cambridge, MIT Press, 1999.

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

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

_____ Discourse trees are good indicators of importance in text. In I. Mani and M. Maybury (eds.), *Advances in Automatic Text Summarization,* 1999. pp. 123-136, The MIT Press.

_____ The Theory and Practice of Discourse Parsing and Summarization. The MIT Press. Cambridge, Massachusetts, 2000.

MILNER, J. C. Reflexões sobre a referência e a correferência. In M.M. Cavalcante, B.B. Rodrigues, A. Ciulla. (eds.), *Referenciação*. Editora Contexto, 2003.

MITKOV, R. Anaphora Resolution. Longman, U.K, 2002.

O'DONNELL, M. RST-Tool: An RST Analysis Tool. *Proceedings of the 6th European Workshop on Natural Language Generation*, March 24 – 26. Gerhard-Mercator University, Duisburg, Germany, 1997.

ONO, K.; SUMITA, K.; MIIKE, S. Abstract Generation Based on Rhetorical Structure Extraction. *In the Proceedings of the International Conference on Computational Linguistic – Coling-94*, 1994. pp 344-348, Japan.

PARDO, T.A.S.; RINO, L.H.M.; NUNES, M.G.V. GistSumm: A Summarization Tool Based on a New Extractive Method. In N.J. Mamede, J. Baptista, I. Trancoso, M.G.V. Nunes (eds.), 6th Workshop on Computational Processing of the Portuguese Language - Written and Spoken, 2003. pp. 210-218 (Lecture Notes in Artificial Intelligence 2721). Springer-Verlag, Germany.

PARDO, T.A.S. 2005. Métodos para Análise Discursiva Automática. Tese de Doutorado. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. São Carlos-SP, Junho, 211p.

POLANYI, L. A formal model of the structure of discourse. *Journal of Pragmatics,* 12:601-638, 1988.

_____ Linguistic dimensions of text summarization. In Brigitte Endres-Niggemeyer, Jerry Hobbs, Karen Sparck Jones (eds.), *Working Notes of the Dagstuhl Seminar on Summarizing Text for Intelligent Communication*. Dagstuhl, Germany, December 13-17, 1993.

SALTON, G.; ALLAN, J.; BUCKLEY, C.; SINGHAL, A. Automatic analysis, theme generation, and summarization of machine-readable texts. *Science*, 264 (3), 1994. pp. 1421-1426. June.

SENO, E.R.M. & RINO, L.H.M. Summarizing RST trees focusing on referential chains: A case study. In *III Workshop em Tecnologia da Informação e da Linguagem Humana – TIL*, São Leopoldo, Brazil, 2005a.

_____ Co-referential chaining for coherent summaries through rhetorical and linguistic modelling. In the RANLP'2005 *Workshop on Crossing Barriers in Text Summarization Research*. Borovets, Bulgaria. September, 2005b.

SPARCK JONES, K. "What might be in a summary?" KNORZ, G.; KRAUSE, J.; & WOMSER-HACKER, C. (eds). *Information Retrieval*. Universitatsverlag Konstanz, 1993. pp. 9-26.

VIEIRA, R. & POESIO, M. An empirically-based system for processing definite descriptions. *Computational Linguistics,* 26(4), 2000. pp. 525-579.

VIEIRA, R. & SALMON-ALT, S. Nominal Expression in Multilingual Corpora: Definite and Demonstratives. In *the Proc. of the Language Resources and Evaluation Conference - LREC 2002*, Las Palmas, 2002.

VIEIRA, R.; SALMON-ALT, S.; SCHANG, E. Multilingual corpora annotation for processing definite descriptions. In Elisabete Ranchhod and Nuno J. Mamede (eds.), *Recents Advances in Natural Language Processing* (Lecture Notes in Artificial Intelligence 2389). Springer-Verlag, Germany, 2002.