Sumarização Automática Multidocumento com Mapas de Relacionamento

Rafael Ribaldo[1], Thiago A. S. Pardo[1], Lucia H. M. Rino[2]

[1]Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo
[2]Departamento de Computação, Universidade Federal de São Carlos

ribaldo@grad.icmc.usp.br, taspardo@icmc.usp.br, lucia@dc.ufscar.br

Abstract. *This paper presents a proposal for multi-document summarization based on relationship maps among sentences and on information provided by the CST (Cross-document Structure Theory).*

Resumo. *Apresenta-se, neste artigo, uma proposta de se realizar sumarização multidocumento com base em um mapa de relacionamentos entre sentenças aliado a informações da teoria CST (Cross-document Structure Theory).*

1. Introdução

Com a crescente quantidade de informações à disposição do ser humano na rede e o escasso tempo para absorção destas, a análise manual de cada informação/documento torna-se quase impossível. Logo, o tratamento automático de textos passa a ser necessário, como a recuperação e extração de informação, e a sumarização automática (SA) multidocumento. Formalmente, a SA multidocumento, foco deste trabalho, é definida como a tarefa de se produzir automaticamente um único resumo (também chamado sumário) a partir de um conjunto de textos que versam sobre um mesmo assunto (Mani, 2001). Aplicações práticas incluem, por exemplo, a produção de sumários a partir de documentos retornados por um buscador *web* com base em consultas dos usuários (como normalmente se faz em buscadores como o Google, ou mesmo no aplicativo Google Notícias), produção de índices de obras complexas (como coletâneas de artigos científicos sobre um determinado tema), síntese de opiniões diversas (em comum ou contraditórias) sobre determinado assunto para análise posterior e possível tomada de decisão, etc.

Neste projeto, utilizam-se grafos e suas medidas para representar e sumarizar textos juntamente com o modelo de representação multidocumento CST (*Cross-document Structure Theory*) (Radev, 2000). Após a modelagem dos textos como um grafo, em que sentenças são representadas como nós e as arestas indicam proximidade lexical entre as sentenças, o grafo é enriquecido com as informações da CST para especificar como as partes dos textos, ou melhor, os segmentos/sentenças, se relacionam. Portanto, as conexões existentes no grafo podem ser definidas, além da proximidade lexical, pelas relações previstas pelo modelo mencionado, por exemplo, se há uma relação CST de equivalência, os segmentos apresentam o mesmo conteúdo. Acredita-se que esta união, o uso da CST aliado ao poder de representação dos grafos, permita produzir sumários de melhor qualidade.

Na seção seguinte são apresentados os conceitos básicos da área, sendo seguidos pelo detalhamento deste projeto, na Seção 3.

2. Conceitos Básicos

Sumários podem ser formados de acordo com algum propósito, podendo ser genéricos ou baseados em interesses do usuário. Com relação ao primeiro, o qual é utilizado neste projeto, dá-se importância aos segmentos textuais, de acordo com alguma medida ou método, para formar um sumário informativo que contenha as informações mais salientes dos textos-fonte. Tal processo ocorre um pouco diferente quando se trata dos interesses do usuário, pois, quando se tem alguma palavra-chave ou consulta, por exemplo, a ponderação dos segmentos pode mudar.

Diferentes métodos podem ser utilizados de forma a conseguir encontrar os segmentos textuais mais salientes de determinado conjunto de textos, como o método de Salton et al. (1997), onde há o descobrimento de segmentos textuais importantes decorrentes de suas relações com outros. Neste projeto, tal método foi adaptado para o cenário multidocumento, uma vez que, originalmente, o método de Salton et al. sumariza um único texto. Além de encontrar segmentos importantes para se compor um sumário, deve-se tomar cuidado com a redundância, e até mesmo com a replicação de textos, que pode acabar ocorrendo, pois dois ou mais segmentos que têm conteúdos parecidos são grandes candidatos à seleção e, se levados todos ao sumário, este poderá conter informações redundantes.

Com a utilização da CST, espera-se aprimorar razoavelmente o descobrimento de segmentos salientes, pois informações semânticas são introduzidas para a análise. Um bom representante dessa linha é o trabalho desenvolvido por Jorge e Pardo (2010) que, a partir das relações semânticas encontradas, monta um ranque de sentenças de acordo com suas importâncias para determinar quais delas vão compor o sumário.

Por fim, a criação de um sumário envolve outro fator relevante: seu tamanho. A quantidade de informação que se deseja ter em um resumo deve ser limitada para que este possa cumprir seu propósito, sem adicionais. Logo, o usuário deve poder delimitar um certo valor, chamando de Taxa de Compressão, a qual determina o tamanho do sumário final com relação ao documento de origem, em número de palavras, normalmente.

3. Mapas de Relacionamento Multidocumento

Como antes mencionado, propõe-se neste projeto utilizar a proposta de Salton et al. enriquecida com informações da CST. Primeiramente, o método não mais modelará um texto (cenário original monodocumento), mas sim um conjunto finito de textos, fazendo com que as sentenças possam se relacionar com outras que são fornecidas por um documento diferente do primeiro. Assim que os textos forem definidos, estes serão representados em um grafo, onde seus vértices são sentenças e as ligações são valores numéricos que indicam o quão próximas elas são lexicalmente. Tais valores podem ser calculados de diferentes formas, por exemplo, pela tradicional similaridade de cosseno (quanto menor o ângulo que separa duas sentenças, mais similares elas são), a qual é utilizada neste projeto.

Feita tal modelagem, as sentenças começam a ser selecionadas de acordo com seu grau (número de ligações) para seguir, segundo Salton et al., por dois caminhos para compor o sumário: denso e profundo. No primeiro, o sumário é construído com os nós mais densos (maior grau); já, no segundo, parte-se do melhor vértice para escolher, além deste, seus filhos mais importantes, na ordem em que aparecem nos textos. Com relação ao segundo caminho (profundo), é notável que, por exemplo, se um filho, vindo de um determinado texto, tem um pai de outro texto, o primeiro pode então concorrer à classificação de sentença mais importante; agora, caso o filho e o pai sejam do mesmo texto, este filho somente será analisado se ele vier após seu pai, na ordem do texto, caso contrário é descartado (visando-se preservar a coerência e a coesão do sumário ao respeitar a ordenação das sentenças nos textos).

Por fim, o sumário é construído com as sentenças mais salientes de todos os textos analisados. É claro que muitas arestas deverão indicar um grau de similaridade altíssimo (devido à redundância entre textos), por isso calcula-se o limite de redundância que duas sentenças podem ter entre si. Neste trabalho, utiliza-se a média dos valores de maior e menor cosseno do grafo para descartar sentenças redundantes, ou seja, caso um nó tenha um valor de similaridade maior do que o calculado como limitante em relação a uma sentença selecionada previamente para o sumário, a sentença deste nó não é levada ao sumário, pois é considerada redundante. Além disso, o usuário pode limitar o tamanho do sumário, utilizando a Taxa de Compressão anteriormente definida. Neste trabalho, usa-se uma taxa de 70%, o que faz com que o sumário tenha 30% do número de palavras do maior texto-fonte.

Com relação ao método mencionado enriquecido com a CST, as informações das relações presentes nesse modelo servirão para aprimorar ainda mais a seleção das sentenças mais relevantes e de maneira a desempatar possíveis sentenças similares. Pode-se utilizar esse

conhecimento de duas maneiras: a primeira é considerar o número total de relações CST de uma sentença, ou seja, não importando seu tipo. A segunda forma é a consideração dos tipos de relações que há entre as sentenças e ponderá-las de acordo com seu nível de redundância. Por exemplo, uma relação CST de Identidade leva peso máximo, pois as sentenças que compõem tal relação são exatamente iguais.

A base de aplicação do método e técnicas apresentadas é o córpus CSTNews (Aleixo e Pardo, 2008), que é composto por 50 grupos de textos jornalísticos, no qual cada grupo versa sobre um mesmo assunto, com os textos já devidamente anotados (manualmente) segundo a CST. Para este córpus, geraram-se resumos automáticos para todos os seus grupos de textos.

Além disso, os grupos de textos contêm, cada um, um resumo feito manualmente. Neste trabalho, os sumários extrativos manuais também foram produzidos, selecionando-se sentenças completas dos textos que correspondessem em conteúdo às sentenças dos sumários manuais. Esses extratos de referência foram produzidos para fins de avaliação.

Além da comparação com os extratos, deve-se também examinar os sumários automáticos com relação aos sumários manuais originais. Para isso, tal análise apoia-se no uso da ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) (Lin e Hovy, 2003), que é um pacote de métricas para a avaliação de sumários. Com ela, é possível comparar os resumos automáticos com os manuais e colher resultados que declarem o quão próximo eles se encontram. Os resultados foram considerados muito bons, próximos do estado da arte.

Abaixo, na Figura 1, é apresentado um exemplo de sumário gerado neste projeto, utilizando-se um grupo do córpus mencionado, cujo número de sentenças é igual a dez, e com uma taxa de compressão de 70%.

A ginasta Jade Barbosa, que obteve três medalhas nos Jogos Pan-Americanos do Rio, em julho, venceu votação na internet e será a representante brasileira no revezamento da tocha olímpica para Pequim-2008. Na América do Sul, a chama passará por Buenos Aires, onde Jade participará do revezamento, no dia 11 de abril.

Figura 1. Exemplo de sumário automático

Para trabalhos futuros, deseja-se tratar o reconhecimento/descobrimento topical de documentos, ou seja, quais são as sentenças que representam cada subtópico do texto. Feito isso, pode-se então utilizar um terceiro caminho proposto por Salton et al., chamado de Caminho Denso-Segmentado, onde os tópicos são considerados em primeiro lugar, para que as sentenças pertencentes a eles sejam analisadas. Com isso, espera-se uma maior abrangência do assunto que um texto trata.

Agradecimentos

À FAPESP, pelo apoio a este trabalho.

Referências

Aleixo, P. e Pardo, T.A.S. (2008). *CSTNews: Um Córpus de Textos Jornalísticos Anotados segundo a Teoria Discursiva Multidocumento CST (Cross-document Structure Theory)*. Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, N. 326.

Jorge, M.L.C. and Pardo, T.A.S. (2010). Experiments with CST-based Multidocument Summarization. In the *Proceedings of the ACL Workshop TextGraphs-5: Graph-based Methods for Natural Language Processing*.

Lin, C.Y. and Hovy, E. (2003). Automatic Evaluation of Summaries Using N-gram cooccurrence Statistics. In *Proceedings of the Language Technology Conference*.

Mani, I. (2001). *Automatic Summarization*. John Benjamins Publishing Co., Amsterdam.

Radev, D.R. (2000). A common theory of information fusion from multiple text sources, step one: Cross-document structure. In the *Proceedings of the 1st ACL SIGDIAL Workshop on Discourse and Dialogue*.

Salton, G.; Singhal, A.; Mitra, M.; Buckley C. (1997). Automatic Text Structuring and Summarization. *Information Processing & Management*, Vol. 33, N. 2, pp. 193-207.