Um Método de Sumarização Multidocumento Baseado em Grafos e Informação Semântico-Discursiva

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

1. Objetivos

Com a crescente quantidade de informações a disposição do ser humano, principalmente on-line, o tratamento automático de textos passa a ser necessário, como a sumarização multidocumento, a qual permite a produção de um sumário a partir de um grupo de textos sobre o mesmo assunto.

O objetivo deste trabalho é explorar o uso de estruturas do tipo grafo para representar e sumarizar textos, considerando não apenas informações de natureza superficial, mas também as relações semântico-discursivas da CST (*Cross-document Structure Theory*) (Radev, 2000). Esta teoria ajuda, por meio das relações mencionadas, a identificar igualdades e diferenças entre sentenças e tratá-las adequadamente nos sumários. Por exemplo, uma relação de Equivalência indica que sentenças têm informações similares e que somente uma delas pode ser selecionada para compor o sumário, portanto.

2. Métodos e Procedimentos

Os materiais utilizados para a realização deste projeto foram: a) um córpus de textos jornalísticos anotados segundo a CST e com sumários humanos; b) um segmentador textual; c) um sistema de radicalização de palavras; d) e, para a análise dos resultados, o pacote de métricas ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) (Lin e Hovy, 2003), que compara automaticamente sumários automáticos com humanos.

Para a produção de sumários automáticos, foi adaptado o método de sumarização de Salton et al. (1997), que foi originalmente desenvolvido para sumarizar um único texto. O método determina a importância de uma sentença pelo seu nível de similaridade com relação ao restante do texto, ou seja, quanto mais uma determinada informação é reiterada no conjunto de textos, mais relevante ela é. A inclusão da CST nesse método permite ainda um refinamento da seleção das sentenças para compor o sumário, como mencionado anteriormente.

3. Resultados e Conclusões

Foram gerados sumários (dos quais um deles encontra-se na Figura 1) para todo o córpus, sendo que os mesmos puderam ser avaliados pela ROUGE e, seus resultados, comparados com alguns dos melhores trabalhos da área. Esta comparação mostrou que a qualidade dos sumários gerados por este projeto está muito próxima aos trabalhos do estado da arte, que se baseiam fortemente em conhecimento lingüístico.

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

Agradecimentos

À FAPESP, pelo apoio financeiro.

Referências

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

Radev, D.R. (2000). A common theory of information fusion from multiple text sources, step one: Cross-documet 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, No, 2, pp. 193-207.