Ordenação de Sentenças em Sumários Multidocumento

Jader Bruno Pereira Lima ***Thiago A. S. Pardo***
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo

1. Objetivos

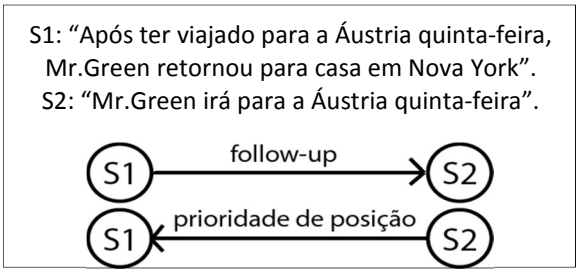
Com o grande volume de dados presentes na internet, faz-se necessário o estudo de diferentes formas de se analisar e processar essas informações, que são apresentadas predominantemente na forma textual. Neste contexto, nos deparamos com a área de sumarização automática multidocumento, que tem como objetivo extrair um sumário, ou resumo, de vários textos que versam sobre um determinado assunto.

A ordem em que são apresentadas as sentenças em um sumário extraído de mais de um texto fonte pode ter grande relevância em sua coerência e coesão textual (Barzilay et al., 2002), pois cada texto fonte possui uma ordem própria de apresentação de ideias e fatos.

Neste trabalho, além de se explorarem métodos simples, propõe-se um método de ordenação de sentenças de sumários multidocumento baseado na análise das relações CST (*Cross-document Structure Theory*) (Radev, 2000). Estas relações tem o objetivo de relacionar as sentenças de acordo com o tipo de informação que apresentam, por exemplo, relações de contradição, sequência temporal, equivalência de conteúdo, etc.

2. Métodos e Procedimentos

O método proposto tem suas operações baseadas no grafo (G) construído a partir de todas as relações CST existentes entre todas as sentenças de todos os textos fonte. Cada tipo de relação nos dá uma informação semântica sobre seu par de sentenças, e, para algumas dessas relações, é possível estabelecer uma ordenação relativa entre a posição dessas sentenças no sumário, ou seja, qual das sentenças deve aparecer primeiro no sumário, de modo que a coerência e a coesão deste seja a melhor possível. Por exemplo, a relação CST *Follow-up,* que temos entre S1 e S2 (Quadro 01), nos diz que a sentença S1 apresenta informação adicional, a qual tem acontecido desde S2. Logo, em um sumário em que estão contidas as duas sentenças, vindas de dois textos fonte diferentes, a sentença S1 deve ser apresentada antes da sentença S2.



Quadro 01: Exemplo de uma relação CST

A partir do grafo G, obtemos um segundo grafo G' que nos dá a informação de posição relativa entre as sentenças, dada a semântica da relação CST. Então, a ordenação topológica desse grafo G' nos dá as restrições de posicionamento entre todas as sentenças de todos os textos fonte.

Também se investigaram métodos de ordenação de sentenças que consideravam a penas a posição delas nos textos de origem e seus tamanhos.

3. Resultados e Conclusões

O método com base em CST está sendo avaliado nesta etapa do projeto, porém, os métodos mais simples já foram implementados e avaliados. O método que resultou em sumários de maior qualidade, até o momento, foi o que utilizou a posição da sentença em seu texto fonte.

Agradecimentos

À FAPESP, pelo apoio financeiro.

Referências Bibliográficas

Barzilay, R., McKeown, K., & Elhadad, M. (2002). Inferring Strategies for Sentence Ordering in Multidocument News Summarization. *Journal of Artificial Intelligence Research*, Vol. 17, pp. 35-55.

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.*