Tratamento de redundância e senso comum

Vinícius R. Uzêda[1], **Thiago A. S. Pardo**[1], **Sandra M. Aluísio**[1]

[1]Núcleo Interinstitucional de Linguística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação (ICMC)
Universidade de São Paulo (USP) - Campus de São Carlos
Caixa Postal 668 – 13.566-590 – São Carlos – SP – Brasil

vruzeda@grad.icmc.usp.br, {taspardo,sandra}@icmc.usp.br

Abstract. *This paper describes the methodologies and approaches in study to detect and remove redundant information from written texts, helping tasks in the textual simplification process, enabling less-capable readers to obtain knowledge through facilitated material interpretation.*

Resumo. *Este artigo descreve metodologias e abordagens em estudo para a detecção e remoção de informações redundantes em textos escritos, tarefas que auxiliam no processo de simplificação textual, habilitando leitores menos aptos a absorverem conhecimento através da interpretação do material facilitado.*

1. Introdução

Leitores com pouca formação, denominados pela UNESCO (Organização das Nações Unidas para a Educação, a Ciência e a Cultura) como Analfabetos Funcionais (AF), e pessoas com problemas de saúde como afasia e dislexia apresentam dificuldades na leitura de textos em geral. A complexidade dos termos escolhidos, o comprimento das sentenças, tudo se torna empecilho no entendimento do conteúdo passado.

Dentro desta meta, este trabalho procurou desenvolver uma ferramenta para remover informações redundantes de textos no âmbito de Simplificação Textual (ST). [Siddharthan 2003] concluiu que AF's tem dificuldade em tratar grandes quantidades de informação simultaneamente. Se esta informação for redundante, a já limitada utilização da memória de trabalho destes leitores estará sendo inutilmente ocupada, dificultando o acesso destes à captação de informações novas.

2. Proposta

O objetivo deste trabalho é, dado um texto qualquer, eliminar sentenças redundantes para facilitar sua leitura. Para isto, a proposta realizada foi a de mapear este texto em um grafo completo, cujos vértices seriam suas sentenças e as arestas representariam o nível de similaridade entre o par de sentenças que ligam.

Estas medidas de similaridade foram classificadas em quatro categorias, conforme mostra a Tabela 1. Cada medida vai ser descrita nas subseções seguintes.

2.1. Medida de Identidade

A métrica que foi utilizada é denominada *cosine* [Salton 1988], ou cosseno em português, e é dada pela fórmula da Figura 2.1.

Table 1. Medidas de similaridade

Tipo	Descrição
Identidade	Palavras idênticas entre as sentenças
Sinonímia	Palavras sinônimas entre as sentenças
Estrutural	Similaridades na estrutura sintática das sentenças
Senso Comum	(Auto-infligida) Presença de informações de senso comum

$$identidade(S_1, S_2) = cosseno(S_1, S_2) = \frac{2 \cdot \|elementos\_em\_comum(S_1, S_2)\|}{\|S_1\| \cdot \|S_2\|}$$

Figure 1. Medida de identidade via cosseno

Aonde $\|X\|$ denota o número de palavras do elemento X e $elementos\_em\_comum(S_1, S_2)$ determinará o conjunto de palavras idênticas entre as sentenças. Para todas estas operações, ignoram-se palavras que não acrescentam qualquer informação ao texto, como artigos ou preposições, conhecidas como *stopwords*, e reduzem-se as palavras a suas formas primitivas, tornando "*palavrões*" e "*palavras*" em "*palavra*", para tornar a medida mais maleável.

Com essas modificações, sentenças idênticas terão $identidade(S_1, S_2) = 1$, e as distintas, $identidade(S_1, S_2) = 0$.

2.2. Medida de Sinonímia

Primeiramente, para se determinarem os sinônimos das palavras, utilizou-se a WordNet.Br [da Silva et al. 1991], versão brasileira do projeto do Instituto Tecnológico de Massachussets (MIT), a WordNet [Fellbaum 1998]. Dada uma palavra em sua forma primitiva, uma consulta simples ao banco de dados fornecido retorna o conjunto de *synsets*, agrupamentos de sinônimos, a que a palavra pertence.

A medida implementada então segue a fórmula exibida na Figura 2.2.

$$sinonimia(S_1, S_2) = \frac{2 \cdot \sum_{W_1 \in S_1} \max_{W_2 \in S_2} (cosseno(synset(W_1), synset(W_2)))}{\|S_1\| \cdot \|S_2\|}$$

Figure 2. Medida de sinonímia

Através dessa medida, sentenças idênticas terão $sinonimia(S_1, S_2) = 1$, e sentenças totalmente distintas, $sinonimia(S_1, S_2) = 0$, usando-se as mesmas considerações utilizadas na medida de identidade.

2.3. Medida de similaridade Estrutural

Agora, introduz-se o conceito de Basic Elements (BE) [Hovy et al. 2005]. Um BE é composto de uma entidade sintática principal (substantivo, verbo, adjetivo ou conjuntos

Dois libaneses foram indiciados pela bomba de Lockerbie em 1991.
(libaneses, dois, cardinal)
(indiciados, libaneses, acusação)
(indiciados, bomba, crime)
(indiciados, 1991, tempo)

Figure 3. BE's para texto exemplo

adverbiais), um dependente simples e uma relação. A Figura 2.3 mostra um exemplo de BE's para uma frase simples.

A fórmula $estrutural(S_1, S_2) = cosseno(BE(S_1), BE(S_2))$ é então utilizada, onde $BE(X)$ constrói o conjunto de BE's para o texto em X. Isso ainda garante que a pontuação será um valor entre 0 e 1, com 0 representando sentenças distintas e 1, sentenças idênticas.

2.4. Medida de Senso Comum

Além de sentenças repetitivas no texto, é possível que uma sentença traga informações que o leitor consideraria óbvias. Essas informações podem ser tratadas como senso comum para aquele perfil de leitores.

Para implementar essa medida, busca-se cada palavra na base do OpenMind Common Sense Brasil (OMCS-Br), uma base de dados de informações de senso comum construída a partir de dados fornecidos por usuários [Carlos 2008] nos moldes do OMCS do MIT [Singh et al. 2002]. As tuplas resultantes tem um par de conceitos relacionados por senso comum e um campo que representa a frequência com que consta na base; se a sentença contiver as duas metades da tupla, será acrescentada a frequência da informação no banco à medida.

Essa medida não gera valores entre 0 e 1, mas qualquer valor real positivo. Antes de associar-se as demais medidas, faz-se uma uniformização da mesma, mas não existe o conceito de identidade das sentenças, já que esta medida é auto-infligida, ou seja, a sentença causa redundância a si mesma.

3. Conclusão

Este trabalho ainda precisa ser devidamente avaliado. Para tanto, está-se construindo um corpus de textos manualmente anotados quanto à redundância. Resultados concretos não devem tardar, mas, mesmo antes disso, acredita-se que os efeitos da aplicação de tal sistema no processo de ST deve ter resultados bastante interessantes.

Trabalhos futuros refinaram as medidas, atribuindo-lhes pesos diferenciados. Ainda pode ser o caso de que mais medidas sejam implementadas, conforme outros trabalhos interessantes forem identificados.

References

Carlos, A. J. F. (2008). Aplicando senso comum na edição de objetos de aprendizagem. Master's thesis, Universidade Federal de São Carlos (UFScar), São Carlos - SP.

da Silva, B. C. D., Moraes, H. R., Oliveira, M. F., Hasegawa, R., Amorim, D. A., Paschoalino, C., and Nascimento, A. C. (1991). 3d hierarchies for animation. In Magnenat-Thalmann, N. and Thalmann, D., editors, *New Trends in Animation and Visualization*. John Wiley & Sons ltd.

Fellbaum, C., editor (1998). *WordNet: An Electronic Lexical Database (ISBN: 0-262-06197-X)*. MIT Press, first edition.

Hovy, E., Lin, C.-Y., and Zhou, L. (2005). Evaluating duc 2005 using basic elements. Proceedings of Document Understanding Conference (DUC). Vancouver, B.C., Canada.

Salton, G., editor (1988). *Automatic text processing*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Siddharthan, A. (2003). Syntactic simplification and text cohesion. Technical report, Research on Language and Computation.

Singh, P., Lin, T., Mueller, E. T., Lim, G., Perkins, T., and Zhu, W. L. (2002). Open mind common sense: Knowledge acquisition from the general public. In *On the Move to Meaningful Internet Systems, 2002*, pages 1223–1237, London, UK. Springer-Verlag.