Ferramenta Automática de Simplificação Textual

Erick G. Maziero[1], Thiago A. S. Pardo[1], Sandra M. Aluísio[1]

[1]Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo (USP)
Caixa Postal: 668 - CEP: 13560-970 – São Carlos – SP – Brazil
erickgm@grad.icmc.usp.br, taspardo@icmc.usp.br, sodfandra@icmc.usp.br

Abstract. *This abstract discusses a automated tool to simplify text, composed of manually created rules, which are applied on the syntactic structure of texts in Brazilian Portuguese to make them as intelligible as possible, syntactically. The simplified texts, in addition to address an audience with difficulties in reading, benefit systems that use written text, since its structure is simplified.*

Resumo. *Este resumo trata sobre uma ferramenta automática de simplificação textual, composta de regras manualmente criadas, que são aplicadas sobre a estrutura sintática de textos do Português do Brasil a fim de torná-los o mais inteligíveis possível, sintaticamente. Os textos simplificados, além de atender a um público alvo com dificuldades de leitura, beneficiam sistemas que atuam sobre o texto escrito, dado sua estrutura simplificada.*

1. Introdução

No âmbito do projeto PorSimples (Simplificação Textual do Português para Inclusão e Acessibilidade Digital), foi desenvolvida uma Ferramenta Automática de Simplificação Textual que tem como objetivo tornar o texto o mais inteligível possível a fim de atender a necessidade do público com baixo nível de letramento ou problemas cognitivos (como afasia e dislexia) na leitura de textos, dado que este público tem dificuldades em ler textos com sentenças longas e de encontrar as associações entre os componentes de uma sentença. Este projeto visa suprir a ausência de sistemas de simplificação textual para o Português do Brasil.

A ferramenta de simplificação textual simbólica automática é composta de regras definidas manualmente, que, quando aplicadas a qualquer texto, tornam este mais simples. Simplicidade textual aqui se refere à estrutura sintática das sentenças, em que a ordem de seus constituintes é sujeito-verbo-objeto, e cada sentença é composta de apenas uma oração, na voz ativa.

2. Trabalhos Relacionados

Os sistemas de simplificação textual automática são encontrados mormente para a língua inglesa, sendo novidade para a língua portuguesa do Brasil.

A simplificação textual é tida como o processo de reduzir a complexidade nos diversos níveis do texto (léxico, sintático e discursivo), mantendo sua significação inicial (Max, 2006). O objetivo da simplificação é tornar o texto mais compreensível ao usuário humano (ou outro programa de tratamento de texto) através de tratamento

automático (Siddharthan, 2003), e pode ser vista como uma tarefa de tradução do texto de uma forma livre para uma forma simplificada (sintática e lexical), mantendo a semântica do texto tanto quanto possível (Nowell, 2000).

Siddharthan (2003) desenvolveu um simplificador sintático para a língua inglesa. Seu sistema utiliza um processo de três estágios, a saber: análise, transformação e regeração, dando importância ao nível discursivo do texto, considerando a interação entre as sentenças do texto, não as tratando individualmente.

Inui et al. (2003) propuseram um sistema baseado em regras para pessoas surdas; utilizaram-se de experiências de professores de surdos e criaram manualmente um conjunto de aproximadamente mil regras. Criaram diversas paráfrases para uma sentença e utilizaram um classificador para escolher a mais simples. Esta escolha se faz necessária, pois as paráfrases podem conter problemas como conjugação verbal e regência.

3. A Ferramenta Automática de Simplificação

A Figura 1 apresenta a arquitetura da ferramenta de simplificação, em que se têm produtos intermediários da simplificação tais como as sentenças de entrada analisadas sintaticamente e as sentenças simplificadas que ainda contêm fenômenos sintáticos a serem simplificados. A entrada é um arquivo de texto sem qualquer marcação e a saída é um arquivo XML com a indicação dos fenômenos simplificados.

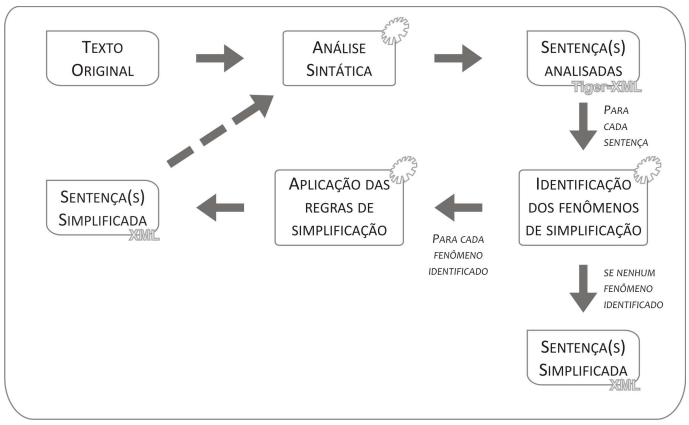

Figura 1. Arquitetura da Ferramenta de Simplificação

A simplificação é direcionada pela identificação dos fenômenos sintáticos, sentença a sentença. Assim, quando uma sentença contém mais de um fenômeno, cada um deles é simplificado por sua vez. A cada simplificação, a(s) sentença(s) resultante(s) deve(m) passar por nova análise sintática, dado que a simplificação faz a alteração da estrutura sintática original, necessitando nova análise sintática. O processo termina quando não é identificado fenômeno sintático a ser simplificado.

Esse processo é realizado simplificando todos os fenômenos sintáticos presentes na sentença original e, para cada um desses fenômenos inicialmente identificados, uma sentença resultante é armazenada contendo apenas a simplificação deste fenômeno a partir da sentença original. Isto é feito para que, quando a ferramenta for utilizada em ambiente de produção textual, o autor possa escolher entre a sentença que mais atinge

suas expectativas: uma sentença contendo a simplificação de todos os fenômenos identificados ou uma sentença que teve apenas um dos fenômenos simplificado.

Considere a sentença contendo uma oração relativa e uma oração na voz passiva: *"A alta foi puxada principalmente pelas linhas de financiamento de veículos, que fecharam o mês passado com a maior inadimplência da série histórica do BC, iniciada em 1991."*, esta tem os dois fenômenos sintáticos simplificados, obtendo-se as sentenças: *"As linhas de financiamento de veículos puxaram a alta. As linhas de financiamento de veículos fecharam o mês passado com a maior inadimplência da série histórica do BC, iniciada em 1991."*.

3.2. Resultados parciais

Uma avaliação da ferramenta em detectar os fenômenos sintáticos foi realizada e é prevista uma avaliação da simplificação propriamente dita, em que será necessária a atuação humana para verificar as sentenças resultantes das operações de simplificação, verificando itens como gramaticalidade, coerência, delimitação correta dos sujeitos nas sentenças geradas, etc.

A avaliação da identificação dos casos de simplificação foi feita automaticamente utilizando o córpus de textos simplificados, que foi gerado por uma lingüista para uso do projeto. As sentenças desse córpus foram processadas pela ferramenta de simplificação, que marcou, para cada sentença as operações que executou, em um arquivo no mesmo formato do utilizado pelo córpus. Para obter os resultados da avaliação os dois arquivos foram confrontados. Os resultados encontram-se na Tabela 1. As medidas precisão, cobertura e medida-F foram empregadas dada sua grande utilização na avaliação de sistemas de Processamento de Língua Natural.

Tabela 1. Avaliação da identificação das operações de simplificação

Operação	Precisão	Cobertura	Medida F
Quebra de sentença	64.07	82.63	72.17
Inversão da ordem das orações	15.40	18.91	16.97
Transformação para voz ativa	44.29	44.00	44.14
Ordenar em Sujeito-Verbo-objeto	1.12	4.65	1.81
Todas	51.64	65.19	57.62

Vale salientar aqui que as operações de inversão da ordem das orações e ordenar em Sujeito-Verbo-Objeto realizadas pela lingüista não correspondem às mesmas operações realizadas pela ferramenta de simplificação. Por exemplo, a lingüista caracterizou como inversão de cláusulas a operação de mover adjuntos adnominais de lugar na sentença. Esses fatos levaram a uma baixa performance da ferramenta nas duas citadas operações, com valores 16.97 e 1.81, respectivamente, na medida-F.

O córpus será corrigido, e informações necessárias (como os fenômenos sintáticos presentes em cada sentença) serão adicionadas. Assim, esperamos melhores resultados em próximas avaliações.

4. Conclusões e trabalhos futuros

Como próximas etapas de trabalho sobre a ferramenta de simplificação está a geração de novas regras ou refinamento das existentes, para melhorar a performance

(precisão) da ferramenta. Regras, como o tratamento dos adjuntos adverbiais de uma sentença, quanto ao local onde devem ocorrer a fim de aumentar a inteligibilidade do texto também serão estudadas.

Referências

Liben-Nowell, D.: Syntactic Simplification. Thesis, University of Cambridge (2000). citeseer.ist.psu.edu/liben-nowell00syntactic.html (07/03/2008).

Inui, K., Fujita, A., Takahashi, T., Iida, R., Iwakura, T.: Text Simplification for Reading Assistance: A Project Note. In Proceedings of the Second International Workshop on Paraphrasing, 9 -16 (2003).

Siddharthan, A.: Syntactic Simplification and Text Cohesion. PhD Thesis. University of Cambridge (2003).

Max, A.: Writing for language-impaired readers. In Proc. of Computational Linguistics and Intelligent Text Processing (CICLing), volume LNCS 3878, 567–570 (2006).