Sumarização Automática para Simplificação de Textos: Experimentos e Lições Aprendidas

Paulo R. A. Margarido, Thiago A. S. Pardo e Sandra M. Aluísio
Núcleo Interinstitucional de Lingüística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação (ICMC), Universidade de São Paulo/São Carlos
R. Trabalhador São-carlense, 400 – São Carlos, SP, Brasil
pauloram@grad.icmc.usp.br, {taspardo,sandra}@icmc.usp.br

RESUMO
Relata-se, neste trabalho, um experimento piloto de simplificação textual com pessoas com limitações de leitura, mais especificamente, analfabetos funcionais. Para simplificar textos, nós utilizamos estratégias de sumarização automática para identificar e destacar partes importantes de um texto e gerar resumos. Demonstramos que cada técnica de simplificação tem efeitos diferentes sobre leitores de diversos níveis de letramento, mas todas auxiliam na compreensão de algum modo. Também relatamos algumas lições aprendidas.

Palavras-chave do autor
Sumarização automática, simplificação textual.

Palavras-chave da classificação da ACM
H.3.1 [Content Analysis and Indexing]: Linguistic processing, Abstracting methods. H.5.2 [User Interfaces]: Natural language, Evaluation/methodology.

INTRODUÇÃO

Este trabalho é focado na tarefa de sumarização automática, tema de pesquisa da área de Processamento de Língua Natural (PLN). A sumarização consiste em gerar automaticamente o resumo (também chamado sumário) de um texto.

Neste trabalho, a sumarização é realizada com fins de simplificação textual, tanto para pessoas com baixo nível de alfabetização [1] e pessoas com disfunções cognitivas (devido a doenças ou lesões cerebrais) quanto para outras operações automáticas (como recuperação e extração de informações).

Primeiramente, foram conduzidos alguns experimentos para determinar a capacidade de diversos métodos de sumarização de identificar a idéia principal de um texto e de gerar bons sumários. Com estes dados em mãos, foi feito um experimento piloto para avaliar o impacto da sumarização na tarefa de simplificação com leitores de diversos níveis de letramento. Várias estratégias de apresentação dos dados obtidos durante a sumarização foram testadas.

EXPERIMENTOS

Seleção de métodos de sumarização

Para encontrar um método ideal de sumarização automática, diversos métodos foram implementados e comparados. Alguns destes métodos são clássicos na área e alguns fornecem ótimos resultados para o português do Brasil. Todos os sistemas avaliados realizam sumarização extrativa, isto é, o resumo é composto por sentenças que já existem no texto original e são extraídas na íntegra para formar o resumo. Esta técnica é dominante em métodos para o português do Brasil. A seguir há uma breve descrição dos métodos investigados.

Métodos baseados em extração de palavras-chave

Os métodos de sumarização baseados em palavras-chave são os mais antigos e foram usados abundantemente desde seu desenvolvimento, devido a sua facilidade de implementação e uso.

Os métodos selecionados para representar esta classe são apresentados em [3] e são bastante simples: dado um texto e seu respectivo conjunto de palavras-chave, qualquer sentença pertencente ao texto que contenha pelo menos uma das palavras-chave fará parte do sumário gerado.

Inicialmente, dois algoritmos de extração de palavras-chave foram utilizados [4], denominados EPC-P (Extração de Palavras-Chave por Padrão) e EPC-R (Extração de Palavras-Chave por freqüência de Radicais). O EPC-P procura no texto por construções do tipo <Nome> ou <Nome+Preposição+Nome>, assim como suas versões adjetivadas, e elege essas ocorrências como palavras-chave. O EPC-R procura no texto pelos radicais mais freqüentes, atribuindo peso maior a grupos de radicais que ocorrem no

início ou no final do texto, e atribui peso maior a grupos maiores (até 3 radicais adjacentes), sendo que os radicais mais freqüentes são considerados palavras-chave.

Além destes, dois novos métodos foram implementados com base nestes dois: antes de gerar o resumo, o número de palavras-chave presente em cada sentença é computado e as sentenças são então ordenadas; o resumo é gerado pela seleção das sentenças mais bem ranqueadas. Estes métodos serão chamados EPC-P2 e EPC-R2.

Métodos baseados em identificação de Gist

Outro sistema que foi testado é o GistSumm [5], que cria um sumário baseado em uma sentença principal, chamada *Gist Sentence*. Este método foi um dos primeiros a serem implementados para o português, tendo a maior precisão (em termos das medidas tradicionais de precisão e cobertura) conhecida para esta língua [2].

O algoritmo calcula a freqüência de cada radical no texto, e então calcula o peso de cada sentença, que é a soma das freqüências dos *stems* que as compõem. A sentença de maior peso é a *Gist Sentence*. Em seguida, qualquer sentença que tenha um radical em comum com a *Gist Sentence* e que tenha um peso acima de um *threshold* empírico também fará parte do sumário.

Métodos baseados em aprendizado de máquina

O método SuPor-2 [7] é baseado em uma abordagem Naïve-Bayes É atualmente o melhor sumárizador para o português brasileiro [2][6].

Seu algoritmo se baseia em diversas características para ordenar as sentenças. Algumas características já são bem conhecidas na área, como tamanho e posição de sentenças, presença de frases importantes, entre outras, e algumas características mais complexas.

Métodos baseados em grafos

Métodos baseados em grafos têm recentemente recebido grande atenção [8], principalmente por sua elegância e pela influência do algoritmo PageRank da Google [9]. O método TextRank representa as sentenças como nós de um grafo, e computa os pesos das arestas através de uma medida de similaridade (sobreposição de palavras). Uma versão melhorada do TextRank também se baseia em relações de sinonímia e antonímia de palavras provenientes de um thesasurus.

Métodos baseline

Como foi feita uma comparação de diversos métodos, é importante que se tivesse uma base para comparações. Então foi implementado um sistema *baseline* forte (para o gênero jornalístico, principalmente) que seleciona as primeiras sentenças de um texto, e um *baseline* fraco, que seleciona sentenças aleatoriamente.

Avaliação dos métodos de sumarização

Uma vez que todos os métodos foram implementados/coletados, foram realizados três experimentos.

O primeiro experimento consistiu em determinar a capacidade de cada método em selecionar a sentença mais importante de um texto. Para isso, foi utilizado um corpus (conjunto de textos) de 187 textos, obtidos do jornal Folha de São Paulo, para verificar a porcentagem de sucesso de cada método. Presume-se que, uma vez que se trata de textos jornalísticos, a primeira sentença será a mais importante. Assim, foi gerado um sumário de uma sentença para cada texto e se verificou se o sumário consistia na primeira sentença. Os resultados estão dispostos na Tabela 1. Pode-se verificar que o EPC-P e o Primeira Sentenças têm ótimos resultados, mas é importante notar que o Primeira Sentenças é extremamente dependente de gênero.

Método	Precisão	Precisão manual	ROUGE
EPC-P	89%	85%	0.4722
EPC-P2	58%	60%	0.4640
EPC-R	60%	60%	0.4736
EPC-R2	39%	35%	0.4655
GistSumm	46%	50%	0.4185
SuPor-2	75%	85%	0.5839
TextRank	39%	50%	0.5426
TextRank+Thesaurus	27%	35%	0.5603
Primeiras Sentenças	96%	90%	0.4369
Aleatório	30%	25%	0.3121

Tabela 1. Performance dos métodos de sumarização

Para ter certeza que este resultado não é uma coincidência, foi feita uma verificação manual sobre uma amostra aleatória de 20 textos. Estes resultados também se encontram na Tabela 1. Neste caso, o EPC-P e o SuPor-2 se destacam, enquanto os outros métodos obtêm resultados mais limitados.

No entanto, também é importante que o método escolhido seja capaz de criar sumários informativos. Para verificar esta capacidade, foi utilizado um conjunto de métricas conhecidas como ROUGE [10], que compara um sumário automático com um sumário gerado por um humano, atribuindo-lhe uma nota entre 0 e 1. Foi utilizado um corpus diferente neste experimento, chamado TeMário [11], que consiste em 100 textos do jornal Folha de São Paulo de vários anos e seus respectivos sumários humanos. Foram utilizados sumários com 70% de compressão (ou seja, sumários com no máximo 30% do tamanho original). Neste quesito, os métodos SuPor-2 e as duas versões do TextRank

se saíram melhor, enquanto os outros métodos são semelhantes.

Após esta análise, é necessário determinar qual método será usado no sistema de simplificação. Para tanto, alguns critérios foram levados em conta: (i) os resultados para seleção de sentenças principais, (ii) a nota ROUGE do algoritmo, (iii) a disponibilidade do sistema, (iv) quão fácil o sistema é de se utilizar.

Pelos critérios, temos: (i) a melhor escolha é o EPC-P; (ii) o SuPor-2 é o melhor, mas o EPC-P tem um bom desempenho; (iii) e (iv) o SuPor-2 é o pior de todos, pois é um sistema grande e complexo. Com isso, o EPC-P foi escolhido para os experimentos seguintes.

Experimento piloto com usuários reais

Uma vez que o módulo de sumarização estava completo, foi realizado um experimento com usuários de diferentes níveis de escolaridade.

Três estratégias de apresentação dos sumários foram testadas: mostrar apenas o sumário, mostrar a sentença principal em destaque no texto e mostrar o texto com o sumário em destaque.

Foram selecionados dois textos: um sobre um apresentador americano que morreu e um sobre problemas de fronteira entre Israel e Jordânia. O primeiro texto é naturalmente simples e acessível, enquanto o segundo é mais complexo pela própria natureza do assunto. Foram gerados sumários com 70% de compressão dos dois. É importante dizer que foram escolhidos textos cujos sumários foram considerados bons.

O experimento foi realizado com 19 sujeitos: 3 tinham até 2 anos de estudo; 5 tinham até 5 anos de estudo e estão no nível de letramento focado pelo sistema de simplificação (ou seja, são considerados analfabetos funcionais); 6 tinham até 8 anos de estudo e 5 tinham 10 ou mais anos de estudo. Os sujeitos foram selecionados por uma ficha que preencheram com seus dados básicos, como nome, idade, se tinham ou não alguma deficiência visual, nível de escolaridade, etc. O fato de ter alguma deficiência visual pode atrapalhar a realização do experimento, já que o mesmo se baseia em leitura e escrita. Nenhum sujeito do experimento apresentou deficiência visual.

O experimento foi planejado de modo que ao menos um usuário de cada faixa recebesse uma das estratégias de apresentação. Cada pessoa recebeu dois pares de textos, cada par contendo um texto original e sua versão simplificada. A versão simplificada poderia ser o sumário de 70%, o texto com as sentenças do sumário de 70% em negrito, ou o texto com sua sentença principal (proveniente do sumário de uma sentença) em negrito. Após ler cada par de textos, a pessoa teve que preencher um formulário (em papel) respondendo questões de múltipla escolha (opções de 'sim' ou 'não'): se cada texto foi difícil de ler, se a leitura de cada texto foi cansativa, e se a pessoa compreendeu o texto. Se a resposta fosse 'sim' para a última questão para algum texto, a pessoa deveria escrever algumas linhas (até 5) dizendo qual era a idéia principal do texto. Finalmente, após ler os dois textos do grupo e responder às perguntas acima para cada uma, a pessoa deveria dizer (por questões de múltipla escolha) qual texto foi mais difícil de ler, qual foi mais cansativo de se ler e qual foi mais fácil de entender.

Para realizar a avaliação, foram distribuímos os pares aos sujeitos de modo que cada par de textos tivesse a mesma quantidade de julgamentos e fosse julgado por pessoas de todos os níveis de escolaridade. O experimento foi conduzido em um laboratório de computadores. Decidiu-se fazê-lo usando computadores (em vez de usar papel) para simular a situação real em que a simplificação textual será realizada. A interface para acessar os textos foi desenvolvida por pesquisadores de Interação Humano-Computador, sendo uma interface muito simples e intuitiva. O usuário tinha o texto exibido em um browser e um botão para ir ao próximo texto (simplificado). Tomou-se o cuidado de se apresentar os textos com fonte de tamanho razoavelmente grande (para evitar que pessoas com deficiências visuais tivessem problemas) e evitando-se o uso da barra de rolagem (pois isso já exige um grau adicional de conhecimento sobre o uso de computador).

Como exemplo das interfaces, as Figuras 1 e 2 exibem interfaces com um texto original e sua versão com as sentenças principais em destaque.

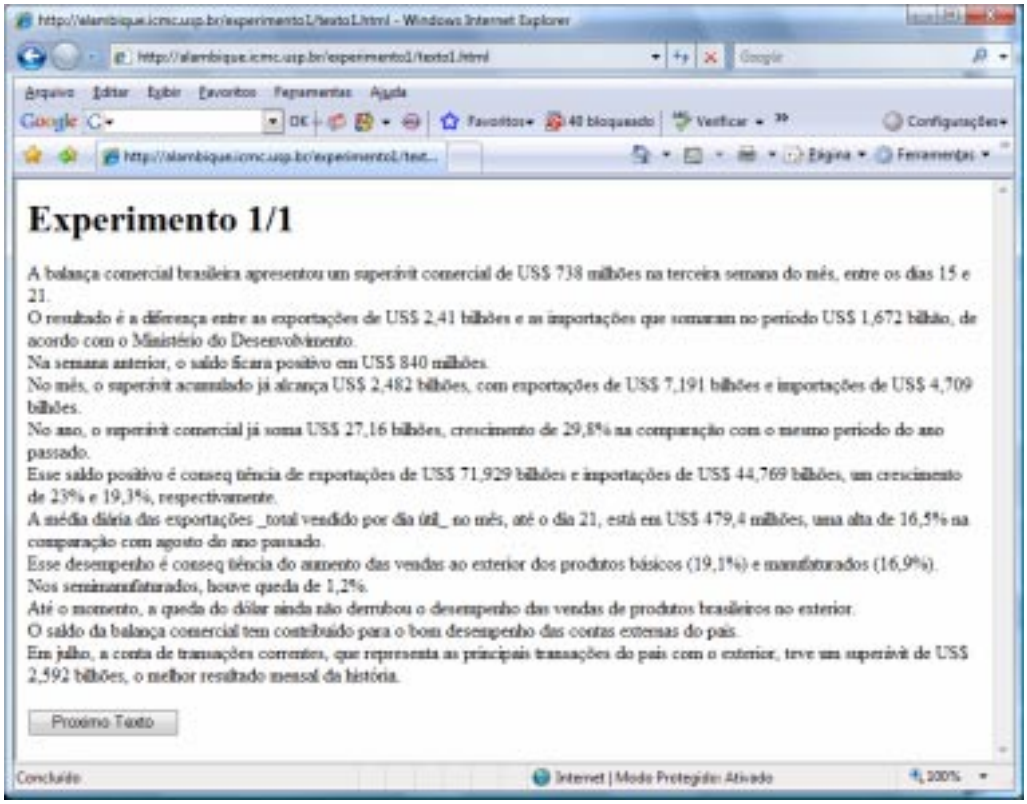

Figura 1. Interface com texto original

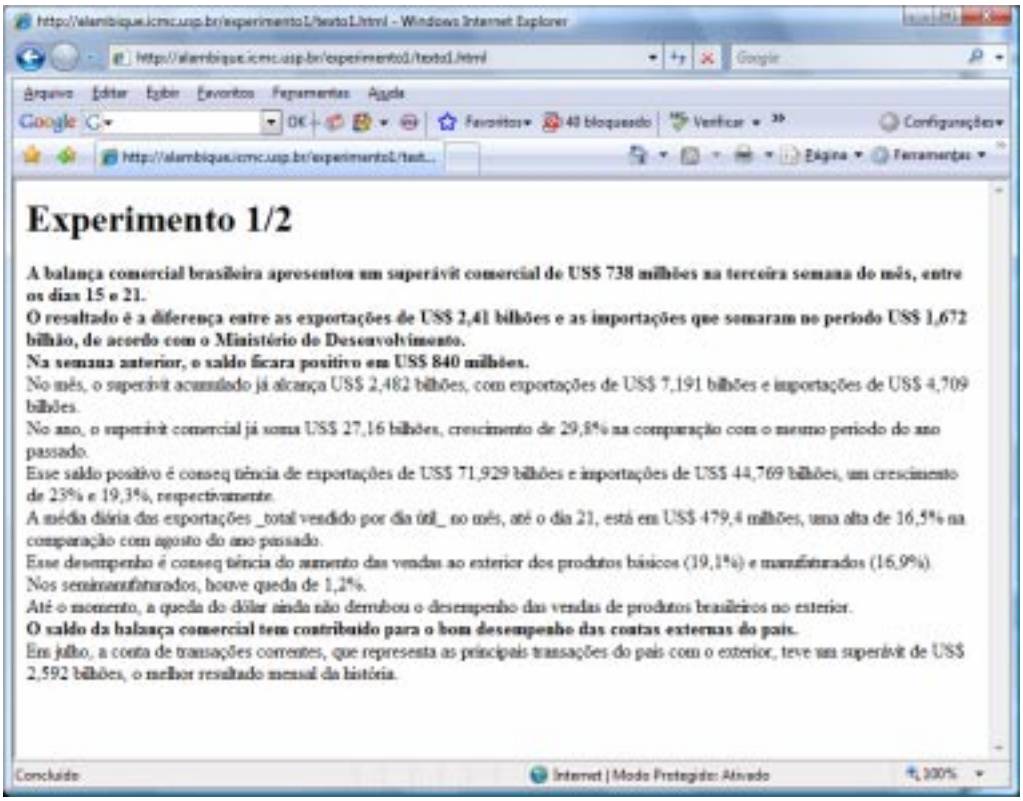

Figura 2. Interface com texto com sentenças principais em destaque

Antes de se executar o experimento, uma demonstração foi realizada com o intuito de se sanar possíveis dúvidas. O experimento tomou quase 40 minutos para ser conduzido (com todas as pessoas trabalhando em paralelo) e contou com diversos especialistas em computação para auxiliar às pessoas que não soubessem como usar um computador.

Das pessoas que tinham até 2 anos de estudo, 2 não conseguiram terminar o experimento: uma não podia ler e uma se cansou demasiadamente após ler o primeiro par de textos. A única pessoa que terminou o experimento considerou que as simplificações não ajudaram em nada.

Sobre as pessoas com até 5 anos de estudo: 66% consideraram que o sumário foi mais fácil de entender; 100% consideraram que o texto original e o texto com as sentenças importantes em negrito foram igualmente inteligíveis; e 60% consideraram que o texto com a sentença principal em negrito foi mais difícil de entender. Estes resultados indicam que o sumário ajuda, mas que a informação em negrito não ajuda. Nossa hipótese para este fato é que a informação destacada é mais uma informação a ser processada, o que, para essas pessoas, pode ser uma carga extra em sua memória de trabalho.

Sobre as pessoas com até 8 anos de estudo: 100% consideraram ler o sumário menos cansativo, mas ambos o texto original e o sumário igualmente inteligíveis; 75% consideraram que o texto com as sentenças importantes em negrito foram mais fáceis de entender; e 60% consideraram o texto com a sentença principal em destaque mais difícil de entender.

Sobre as pessoas com mais de 10 anos de estudo: 57% consideraram o texto mais fácil de entender do que o sumário, mas o sumário menos cansativo de ler; 100% consideraram o texto original mais fácil de entender do que o texto com todas as sentenças importantes em negrito; 80% consideraram o texto com a sentença principal em negrito mais fácil de entender.

Os resultados são interessantes, pois demonstram que a sumarização auxilia pessoas de diferentes níveis de escolaridade de diferentes modos: para pessoas com 2 anos de estudo, a simplificação foi irrelevante; textos mais curtos ajudam mais pessoas com 5 anos de estudo; para pessoas com 8 anos de estudo, o sumário em destaque ajuda; e apenas a sentença principal ajuda pessoas com 10 anos de estudo.

Várias lições puderam ser tiradas do experimento conduzido, lições estas que deverão guiar a realização de experimentos futuros na área.

Apesar dos resultados obtidos, acreditamos que um experimento maior, com mais sujeitos, deva ser realizado para confirmar as hipóteses levantadas. Em particular, o experimento deverá ser realizado com analfabetos funcionais, ou seja, pessoas com 4 ou 5 anos de estudo. Pensa-se em procurar sujeitos em programas de alfabetização de adultos, programas estes que em geral são mantidos pelos governos municipais.

A inclusão de pessoas com menor grau de escolaridade não foi proveitosa para o experimento. Tais sujeitos não foram capazes de terminar o experimento e, para preencher a ficha de seleção do experimento, pediram auxílio a colegas, pois se sentiram constrangidos em dizer que não podiam ler e escrever no nível mínimo pedido.

Uma questão importante que foi levantada pelo experimento piloto aqui relatado é a adequação do uso da tecnologia de simplificação para usuários com grandes dificuldades no manuseio de computadores. A questão central é se a tecnologia seria de fato utilizada no futuro, dada a falta de habilidade com computadores das pessoas que necessitariam de tal recurso. Interfaces ubíquas parecem ser a solução natural para isso, mas esta realidade é ainda muito distante dos usuários comuns visualizados para esta pesquisa.

Outra questão interessante na condução de futuros experimentos tem relação com o tipo de texto lido/julgado. Textos jornalísticos são fontes de conhecimento e de atualização para a maioria das pessoas, mas muitas vezes não são de fato necessários para analfabetos funcionais, o que pode causar certo desinteresse na realização dos experimentos. Por outro lado, textos instrucionais podem ser recursos importantes que despertem interesse nos usuários pretendidos. Por exemplo, texto com instruções para se obter um documento ou se chegar a um local. Nesta linha, a adequação da tecnologia proposta poderia ser medida diretamente pelo sucesso do usuário em realizar ou não a tarefa que pretende.

Futuramente, pretende-se investigar outras técnicas de sumarização aplicadas à simplificação. Por exemplo, outras estratégias consistem em adicionar um subtítulo significativo para cada parágrafo de um texto, remover informações redundantes e de senso comum de um texto (o que produz um texto mais curto, mas não necessariamente o que se chamaria de resumo), adicionar *highlights* ao texto[1], isto é, sentenças que expressam os fatos principais do texto, mas que não são coesas e coerentes de modo a formar o resumo, etc.

CONCLUSÃO

Estes experimentos constituem passos iniciais para um sistema de simplificação automática de textos para o português do Brasil. Até onde se sabe, esta é a primeira iniciativa nesse sentido. O projeto em que este trabalho se insere, chamado PorSimples (Simplificação Textual do Português para Inclusão e Acessibilidade Digital), visa

[1] Tal estratégia é utilizada nas notícias da CNN, por exemplo, como pode ser visto no site www.cnn.com

produzir sistemas tanto para o escritor que escreve seu texto quanto para o usuário final.

Além do módulo de sumarização, as ferramentas produzidas pelo projeto PorSimples também incluirão módulos de simplificação, propriamente dita, e de explicitação da estrutura textual. O módulo de simplificação realizará transformações léxicas e sintáticas no texto, de forma a torná-lo mais inteligível. O módulo de explicitação da estrutura textual visa adicionar marcadores discursivos no texto, tornando-o mais coesivo, portanto, e indicar ao usuário o relacionamento entre as partes do texto (por exemplo, relações de causa-efeito e contraste).

Mais detalhes sobre o projeto PorSimples podem ser obtidos tanto em [12], [13], [14] e [15] quanto na página oficial do projeto: caravelas.icmc.usp.br/wiki/index.php/Principal

AGRADECIMENTOS

Este trabalho foi possível graças ao apoio da FAPESP e da Microsoft Research (MSR). Além disso, agradecemos aos demais membros do projeto pelo auxílio ao desenvolvimento e aplicação dos experimentos relatados.

REFERÊNCIAS

1. Ribeiro, V. M.: Analfabetismo e alfabetismo funcional no Brasil. Boletim INAF. São Paulo: Instituto Paulo Montenegro (2006).
2. Rino, L.H.M., Pardo, T.A.S., Silla Jr., C.N., Kaestner, C.A., Pombo, M.: A Comparison of Automatic Summarization Systems for Brazilian Portuguese Texts. SBIA 2004, LNAI, vol. 3171, pp. 235-244. Springer, Heidelberg (2004).
3. Souza, C.F.R., Nunes, M.G.V.: Avaliação de Algoritmos de Sumarização Extrativa de Textos em Português. Relatório Técnico do ICMC-USP. NILC-TR-01-09 (2001).
4. Pereira, M.B., Souza, C.F.R., Nunes, M.G.V.: Implementação, Avaliação e Validação de Algoritmos de Extração de Palavras-Chave de Textos Científicos em Português. Revista Eletrônica de Iniciação Científica. Ano II, Volume II, Número I (2002).
5. Pardo, T.A.S., Rino, L.H.M., Nunes, M.G.V.: GistSumm: A Summarization Tool Based on a New Extractive Method. In N.J. Mamede, J. Baptista, I. Trancoso, M.G.V. Nunes (eds.), 6th Workshop on Computational Processing of the Portuguese Language - Written and Spoken - PROPOR (Lecture Notes in Artificial Intelligence 2721), pp. 210-218 (2003).
6. Leite, D.S., Rino, L.H.M., Pardo, T.A.S., Nunes, M.G.V.: Extractive Automatic Summarization: Does more linguistic knowledge make a difference? In C. Biemann, I. Matveeva, R. Mihalcea, and D. Radev (eds.), Proceedings of the HLT/NAACL Workshop on TextGraphs-2: Graph-Based Algorithms for Natural Language Processing, pp.17-24 (2007).
7. Leite, D.S., Rino, L.H.M.: Selecting a Feature Set to Summarize Texts in Brazilian Portuguese. In J. S. Sichman et al. (eds.), Proceedings of 18th Brazilian Symposium on Artificial Intelligence (SBIA'06) and 10th Ibero-American Artificial Intelligence Conference (IB- ERAMIA'06). Lecture Notes on Artificial Intelligence, No. 4140, Springer-Verlag, pp. 462-471 (2006).
8. Mihalcea, R., Tarau, P.: TextRank: Bringing Order into Texts. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (2004).
9. Brin, S., Page, L.: The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems 30, pp. 1-7 (1998) Webengine.
10. Lin, C., Hovy, E.H.: Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In Proceedings of Language Technology Conference (2003).
11. Pardo, T.A.S., Rino, L.H.M.: TeMário: Um Corpus para Sumarização Automática de Textos. NILC Tech. Report NILC-TR-03-09 (2003).
12. Aluísio, S.M., Specia, L., Pardo, T.A.S., Maziero, E.G., Caseli, H.M., Fortes, R.P.M.: A Corpus Analysis of Simple Account Texts and the Proposal of Simplification Strategies: First Steps towards Text Simplification Systems. In Proceedings of the 26th ACM International Conference on Design of Communication - SIGDOC. Lisboa, Portugal. September, 22-24. (2008).
13. Aluísio, S.M.; Specia, L.; Pardo, T.A.S.; Maziero, E.G.; Fortes, R.P.M.: Towards Brazilian Portuguese Automatic Text Simplification Systems. In Proceedings of the 8th ACM Symposium on Document Engineering - DocEng, pp. 240-248. São Paulo-SP, Brazil. September, 16-19. (2008).
14. Specia, L.; Aluísio, S.M.; Pardo, T.A.S.: Manual de Simplificação Sintática para o Português. Série de Relatórios do NILC. NILC-TR-08-06. São Carlos-SP, Junho, 27p. (2008).
15. Maziero, E.G.; Pardo, T.A.S.; Aluísio, S.M.: Ferramenta de Análise Automática de Inteligibilidade de Córpus (AIC). Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, no. 328. São Carlos-SP, Julho, 14p. (2008).