Modelando Textos como Redes Complexas*

Lucas Antiqueira[1], Maria das Graças V. Nunes[1], Osvaldo N. de Oliveira Jr.[2], Luciano da F. Costa[2]

[1]Núcleo Interinstitucional de Lingüística Computacional (NILC-ICMC)
Universidade de São Paulo (USP)
Caixa Postal 668 – 13.560-970 – São Carlos – SP – Brasil

[2]Instituto de Física de São Carlos (IFSC)
Universidade de São Paulo (USP)
Caixa Postal 369 – 13.560-970 – São Carlos – SP – Brasil

lantiq@nilc.icmc.usp.br, gracan@icmc.usp.br,
{chu,luciano}@if.sc.usp.br

Abstract. *This paper presents the use of complex networks for modelling texts. The results of some experiments have shown that there is a strong correlation between the networks parameters and the quality of texts. Therefore, this model is potentially useful to distinguish bad and good texts. The model using complex networks is language independent and makes simpler the automatic analysis of texts.*

Resumo. *Este artigo descreve e discute os resultados de experimentos de análise de textos quando estes são modelados como redes complexas. Os experimentos indicam que os parâmetros das redes complexas apresentam uma evidente correlação com a qualidade de textos e, portanto, são potencialmente úteis para distinguir textos bons e ruins. A modelagem em redes complexas é independente de língua e simplifica sobremaneira o trabalho de análise automática de textos, mostrando-se uma alternativa promissora aos métodos linguisticamente motivados.*

1. Introdução

Os sistemas de Processamento de Língua Natural (PLN), até a década de 1990, seguiam o formalismo conhecido como simbólico, no qual todo o conhecimento deve ser expresso em gramáticas e em léxicos, fazendo com que todas as possíveis situações devam ser previstas durante a fase de desenvolvimento para que possam ser resolvidas pelo sistema. Entretanto, o que se verifica na prática é uma queda no índice de sucesso desses sistemas na medida em que eles se tornam mais abrangentes e robustos. Recentemente, a área de PLN passou a estudar e a aplicar novos formalismos, devido principalmente às limitações do paradigma simbólico, mas também devido a um grande obstáculo que vem impulsionando os pesquisadores da área na busca por outras abordagens até os dias de hoje: a análise profunda de um texto visando à compreensão

* Este trabalho teve o apoio do CNPq, FAPESP e Human Frontier Science Program (RGP39/2002)

de seu significado. Entre os modelos que passaram a ser utilizados, destacam-se os estatísticos e os conexionistas, provenientes dos estudos em Aprendizagem Automática. Posteriormente, o uso dessas novas técnicas proporcionou a construção, com sucesso, de alguns sistemas de PLN a partir de corpora de textos reais, dispensando o conhecimento lingüístico explícito indispensável na abordagem simbólica. Os melhores etiquetadores (Part-of-Speech Taggers) disponíveis atualmente são resultados do uso de métodos estatísticos de aprendizado, fato que reforça a tendência em buscar novas alternativas na análise automática de línguas naturais. Nesse contexto, uma opção a ser considerada é a aplicação de redes complexas [Albert e Barabási 2002] [Newman 2003] no PLN, um conceito recente, proveniente da mecânica estatística e que faz uso intenso da teoria dos grafos, ainda pouco estudado no âmbito do processamento automático de línguas naturais.

Uma ampla gama de sistemas, naturais ou sociais, pode ser descrita por redes complexas. Como exemplos desses sistemas, temos as cadeias alimentares, a Internet, a World Wide Web, as redes neurais e as redes de relações sociais entre indivíduos. Um método primário para análise de uma rede é representá-la como uma figura com arcos e nós, para então responder determinadas questões através de um exame visual da figura. Como tem havido uma mudança significativa nessa área, em que o foco se move da análise de um único e pequeno grafo (e das propriedades de seus nós e arestas) para considerar propriedades estatísticas em larga escala, essa abordagem torna-se impraticável. Com essa mudança, tornou-se necessária uma correspondente transformação na tradicional abordagem analítica de redes.

Anteriormente, os grafos randômicos eram os mais utilizados para estudar sistemas modelados como redes (em um grafo randômico de N nós, cada par de nós é conectado com probabilidade p). Mas o interesse em sistemas complexos por parte dos cientistas estimulou uma reconsideração desse paradigma de modelagem. Como é cada vez mais evidente que a topologia e a evolução dessas redes são governadas por princípios robustos de organização (e não simplesmente randômicos, daí o nome "redes complexas"), sentiu-se a necessidade de desenvolver ferramentas para capturar quantitativamente esses princípios. Grande parte das recentes descobertas está relacionada à maneira como as redes do mundo real diferem das redes randômicas [Newman 2003]. Existe também uma crescente necessidade de entender o comportamento do sistema como um todo, movendo-se para além das abordagens reducionistas.

Pesquisadores têm desenvolvido uma diversidade de técnicas e modelos que auxiliam no entendimento e previsão do comportamento desses sistemas. Três conceitos merecem destaque no estado da arte em redes complexas [Albert e Barabási 2002]: as redes *small-world*, o coeficiente de aglomeração e as redes livres de escala (*scale-free*). O conceito *small-world* refere-se ao fato de que, mesmo enormes, a maioria das redes apresenta um caminho relativamente curto entre quaisquer dois nós. Já o coeficiente de aglomeração quantifica a tendência de agrupamento dos nós da rede. Por fim, uma rede é dita livre de escala se a probabilidade $P(k)$ de um nó possuir k arestas obedece uma distribuição por lei de potência (*power law*) $P(k) \sim k^{-\gamma}$. As redes livres de escala apresentam os chamados *hubs*, que são poucos nós altamente conectados que coexistem com um grande número de nós com poucas conexões (esse fato pode ser observado na

lei de potência) [Barabási 2003]. Esses conceitos deram, recentemente, início a uma remodelagem de redes, incentivando o estudo de novos paradigmas.

A linguagem humana também pode ser entendida como uma rede complexa. Cancho e Solé (2001) apresentam a análise de uma rede derivada do British National Corpus, sendo que os nós dessa rede representam as palavras, e suas arestas conectam palavras que aparecem no córpus pelo menos uma vez, em seqüência ou separadas por uma palavra. Essa rede contém 478.773 nós e $1{,}77 \times 10^7$ arestas. Outra rede foi construída, semelhante à anterior, com a diferença de que apenas são considerados os pares de palavras consecutivas (*i*,*j*) que ocorrem mais vezes do que seria esperado quando a independência entre as palavras é assumida, ou seja, quando $p_{ij} > p_i p_j$. Essa rede apresenta 460.902 nós e $1{,}61 \times 10^7$ arestas. Foi mostrado que as duas redes apresentam as características *small-world* e livre de escala, indicando que essa rede de palavras pertence à mesma classe, por exemplo, da Internet e da World Wide Web. Outro exemplo de estudo em linguagem natural envolvendo redes complexas foi desenvolvido por Sigman e Cecchi (2002). Nesse trabalho, o banco de dados Wordnet [Miller 1985] foi mapeado em uma rede, no qual os nós representam os diferentes significados das palavras (neste caso, apenas dos substantivos) e as arestas refletem as relações semânticas (como antonímia e hipernímia). É mostrado que a polissemia é responsável pela existência de *hubs*, os quais deixam os conceitos mais próximos entre si, tornando a rede *small-world*. Em outro estudo [Motter *et al.* 2002], um thesaurus da língua inglesa é modelado de modo que, na rede, duas palavras estão conectadas se representam conceitos similares. Essa rede apresentou a propriedade *small-world* e livre de escala (assintoticamente). Já Costa (2003) aplicou um modelo de rede em um experimento psicofísico, no qual uma pessoa fornece livremente uma palavra que julgue relacionada a outra palavra apresentada por um computador. Esse processo é repetido diversas vezes, e as arestas são criadas entre as palavras associadas pela pessoa. A distribuição dos graus de saída dessa rede indica que ela é uma rede livre de escala.

Nesse contexto de estudos recentes que evidenciam propriedades de redes complexas nas redes inspiradas na linguagem humana, este trabalho investiga sua potencial utilidade na análise de textos. Por meio do uso de medidas estatísticas independentes de língua, nosso objetivo é avaliar (julgar) automaticamente critérios como qualidade, legibilidade, autoria, etc. Este artigo apresenta algumas experiências nesse sentido, em especial na avaliação da qualidade de textos, e direciona para uma possível aplicação resultante do encontro entre o processamento automático de línguas, uma área multidisciplinar por excelência, e a efervescente área da Física chamada redes complexas.

Na seção 2 deste artigo, é apresentado o processo de modelagem de um texto até a obtenção de uma rede, da qual são retiradas diversas medidas estatísticas, explicadas na seção 3. Na seção 4 são apresentados três experimentos, juntamente com seus resultados, que foram conduzidos utilizando o modelo de rede, para então apresentar algumas conclusões e perspectivas na seção 5.

2. O Processo de Modelagem dos Textos

A modelagem utilizada neste trabalho requer um tratamento dos textos antes de representá-los como redes complexas. Nas seções 2.1 e 2.2 as fases de pré-processamento dos textos e de construção das redes são detalhadas.

2.1. O Pré-Processamento dos Textos

O objetivo dessa representação por redes complexas é codificar as relações entre os conceitos dos textos. Para tanto, cada texto tem suas *stopwords* removidas, eliminando assim as palavras com pouco significado. Além disso, as palavras restantes são lematizadas, a fim de agrupar conceitos de mesma forma canônica, mas com flexões diferentes. O texto é também etiquetado morfossintaticamente, neste caso pelo *tagger* MXPost ([Aires *et al.* 2000], baseado no modelo de Ratnaparkhi [Ratnaparkhi 1996]), o qual adiciona informações úteis na resolução de ambigüidades na fase de lematização. Esta, por sua vez, é feita acessando-se o léxico computacional do NILC [Nunes *et al.* 1996][1], no qual cada palavra tem uma regra associada para geração da forma canônica.

2.2. A Construção das Redes

A estrutura que representa a rede derivada de um texto é uma matriz de adjacências com pesos. Após o pré-processamento do texto, as N palavras distintas restantes passam a representar os nós da rede, e a seqüência de palavras resultante é utilizada na criação das arestas, de modo que, para cada par de palavras consecutivas, existe uma aresta direcionada correspondente na rede. As arestas apresentam pesos, os quais indicam o número de vezes que as respectivas associações de palavras aparecem no texto. Note que, devido à etapa anterior de lematização, as palavras, por exemplo, "aparecem" e "aparecerão" serão representadas por apenas um nó na rede, neste caso rotulado pelo infinitivo "aparecer". Todas as medidas estatísticas (seção 3) são obtidas da matriz de adjacências W (de dimensão NxN) que representa a rede. Ela apresenta inicialmente todos os elementos iguais a zero e, a cada par de palavras (i,j) lido do texto, faz-se $W(j,i)=W(j,i)+1$, incrementando desse modo o peso da associação $i \rightarrow j$. O grafo da Figura 2 é exemplo do resultado da modelagem do poema "No meio do caminho"[2], de Carlos Drummond de Andrade (Figura 1).

No meio do caminho tinha uma pedra
tinha uma pedra no meio do caminho
tinha uma pedra
no meio do caminho tinha uma pedra.

Nunca me esquecerei desse acontecimento
na vida de minhas retinas tão fatigadas.
Nunca me esquecerei que no meio do caminho
tinha uma pedra
tinha uma pedra no meio do caminho
no meio do caminho tinha uma pedra.

Figura 1. Poema "No meio do caminho", de Carlos Drummond de Andrade, cuja modelagem está representada na Figura 2

[1] O léxico do NILC (Núcleo Interinstitucional de Lingüística Computacional), conta com cerca de 1,5 milhão de palavras. É o maior léxico computacional para o português brasileiro.

[2] Poema retirado do livro: Andrade, C.D., "Alguma Poesia", Record, Rio de Janeiro, 2001. Ele é utilizado aqui apenas para fins ilustrativos, já que não foi utilizado nos experimentos relatados neste artigo.

3. Extração de Medidas Estatísticas

Foram computadas a média dos graus de entrada, dos graus de saída e do coeficiente de aglomeração para todos os nós, e também a média dos caminhos mínimos entre todos os pares de nós da rede (com exceção das auto-conexões). Como em um dígrafo a média dos graus de entrada é igual à dos graus de saída, no decorrer deste artigo só será mencionada a média dos graus de saída. Além disso, três tipos de caminhos mínimos foram calculados: (i) $CM1(i,j)$, o qual considera todos os pesos iguais a 1; (ii) $CM2(i,j)$, o qual faz uso do complemento dos pesos W_{max}-$W(j,i)$+1; e (iii) $CM3(i,j)$, que por sua vez emprega o inverso dos pesos, $1/W(j,i)$. *CM2* e *CM3* procuram dar ênfase às associações de palavras com maior peso.

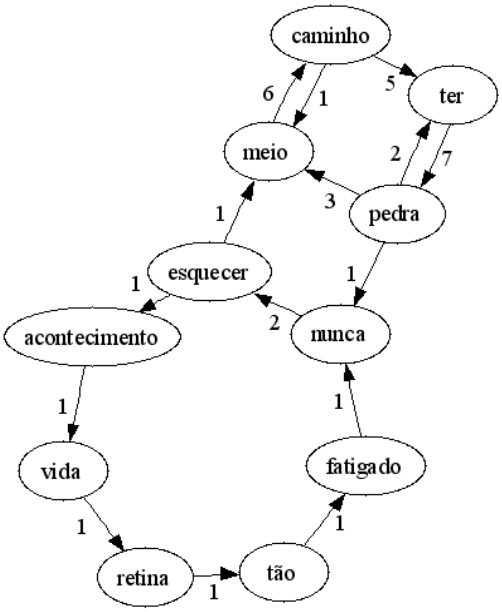

Figura 2. Rede que representa o texto da Figura 1

Grau de entrada, grau de saída e caminho mínimo são conceitos familiares da teoria dos grafos. Já o coeficiente de aglomeração tem uma definição mais recente. Considere que, para cada nó i, existem k_i arestas que o associam a k_i outros nós. Se esses k_i nós formassem um clique, ou seja, se cada nó estivesse diretamente conectado a qualquer outro nó do conjunto, haveria $k_i(k_i - 1)$ arestas entre eles. Seja E_i o número de arestas que realmente existem entre os k_i nós, então

$$CA_i = \frac{E_i}{k_i(k_i - 1)} \quad (1)$$

é o coeficiente de aglomeração do nó i, o qual reflete o quanto as palavras conectadas a esse nó também estão conectadas entre si. O coeficiente da rede inteira é a média de todos os CA_i.

4. Alguns Experimentos e suas Análises

Um possível uso desse tipo de modelagem em PLN é a avaliação automática de determinadas características textuais por meio das medidas estatísticas coletadas das redes complexas. Procurou-se inicialmente investigar se as medidas das redes

apresentam alguma indicação a respeito da complexidade, clareza, legibilidade ou qualidade de um texto, pois essas são características sem definições universalmente aceitas, e que podem porventura estar codificadas de alguma maneira na estrutura da rede. Em cada experimento realizado, juízes humanos avaliaram os textos escolhidos, dando notas de 0 a 10 para as características textuais selecionadas, usando seus próprios critérios, formando assim uma base para comparação com as medidas das redes. Esperava-se que os juízes pudessem refletir, por meio dessas notas, as diferenças de complexidade, clareza, legibilidade ou qualidade dos textos. Para se ter uma indicação dessa hipótese, em cada experimento foram fornecidos aos juízes dois grupos distintos de textos, pois, se realmente as pessoas têm uma definição comum não explícita para as quatro características selecionadas, as notas dos juízes poderiam refletir as diferenças entre os textos dos grupos escolhidos. Para que os critérios dos juízes emergissem naturalmente com as notas, não foi fornecida qualquer definição das características sob julgamento.

Em primeiro lugar, as notas dadas pelos juízes foram analisadas para averiguar se existe diferença significativa entre as notas dos dois grupos de texto de cada experimento. Quando isso acontece, é uma evidência de que a escolha dos grupos permitiu uma variabilidade nas notas. Esse é um ponto importante, pois é baseado nele que se investiga como as medidas retiradas da rede variam à medida que as notas variam. Nas seções seguintes são relatados três experimentos que foram realizados até a obtenção de alguns resultados importantes. Os experimentos foram conduzidos na mesma ordem em que são aqui apresentados.

4.1. Experimento 1

O Experimento 1 contou com seis juízes com bom conhecimento de português, que foram convidados a ler 10 textos do caderno Esporte e 10 textos do caderno Dinheiro, ambos do jornal Folha de São Paulo. Esses textos, de aproximadamente mesmo tamanho, receberam notas de 0 a 10 quanto à complexidade, clareza, legibilidade e qualidade. Além disso, as medidas mencionadas anteriormente foram calculadas a partir dos textos então modelados como redes complexas, conforme seção 2.2. O primeiro passo da análise do experimento foi estudar as notas dadas pelos juízes, separadas por caderno. Para cada característica textual, foi utilizado o teste t-student [Casella e Berger 2001], para averiguar se as médias dos cadernos Dinheiro e Esporte são iguais (hipótese nula). Em todos os casos a hipótese alternativa se confirmou (médias diferentes), mas os valores *p* dos testes não foram suficientemente pequenos para todos os critérios (

Tabela 1), fato necessário para atestar que a diferença entre as médias não tenha sido obra do acaso. Dessa maneira os valores *p* mostram que as notas não apresentaram uma boa diferença entre os dois tipos de textos, e, portanto, não proporcionaram a variabilidade necessária para verificar se existe alguma correlação entre as notas e as medidas retiradas das redes. Uma possível explicação para essa semelhança de notas é que os textos jornalísticos, mesmo tratando de assuntos diversos, seguem um padrão de escrita comum.

Tabela 1. Valores *p* do teste t-student entre as médias das notas dos cadernos Dinheiro e Esporte (Experimento 1)

Critério	Valor *p*
Legibilidade	0,0013
Clareza	0,0129
Complexidade	0,0193
Qualidade	0,2297

4.2. Experimento 2

O Experimento 1 serviu de base para que outro experimento fosse planejado e conduzido. Dessa vez, os textos foram selecionados dentre um conjunto de redações da Fuvest disponíveis no corpus NILC[3]. Foram utilizados 10 textos, todos sobre um mesmo assunto e com tamanhos semelhantes, separados em dois grupos, de 5 textos bons e de 5 textos ruins. Um grupo mais heterogêneo de 12 juízes foi convidado, sendo seis alunos do curso de Bacharelado em Ciências de Computação do ICMC, portanto falantes não especialistas na língua, mais seis pessoas do NILC, para dar notas aos textos de acordo com os quatro critérios selecionados para o estudo. Essa distinção entre os dois grupos de juízes poderia ajudar a verificar se as notas dos alunos da graduação apresentam diferenças marcantes em relação às notas dos pesquisadores do NILC. Como no primeiro experimento, o passo seguinte foi analisar as notas por meio do teste t-student. Os valores *p* do teste t-student para as médias das categorias "bons" e "ruins", foram calculados para as notas dadas pelos alunos da graduação (Tabela 2), para as dadas pelos membros do NILC (Tabela 3), e para todas as notas (Tabela 4). Embora alguns valores *p* sejam bem pequenos (como o do critério Legibilidade da Tabela 3), eles não o são para todos os critérios conforme esperado. Novamente, portanto, não foi possível verificar diferenças significativas entre as notas dadas aos textos bons e aos textos ruins para todos os critérios em estudo, tanto por um grupo de juízes quanto pelo outro. Logo, esse experimento também não permitiu analisar se as medidas retiradas das redes têm alguma correlação com a avaliação dos juízes, devido à baixa variabilidade das notas.

Tabela 2. Valores *p* do teste t-student entre as médias das notas das categorias "bons" e "ruins" (Experimento 2 - Notas dos alunos da graduação)

Critério	Valor *p*
Legibilidade	0,0468
Clareza	0,2498
Complexidade	0,0153
Qualidade	0,1246

Tabela 3. Valores *p* do teste t-student entre as médias das notas das categorias "bons" e "ruins" (Experimento 2 - Notas dos membros do NILC)

Critério	Valor *p*
Legibilidade	0,0004
Clareza	0,0014
Complexidade	0,3805
Qualidade	0,0012

[3] Veja http://www.nilc.icmc.usp.br/nilc/tools/corpora.htm

Tabela 4. Valores *p* do teste t-student entre as médias das notas das categorias "bons" e "ruins" (Experimento 2 - Todas as notas)

Critério	Valor *p*
Legibilidade	0,0004
Clareza	0,0051
Complexidade	0,0363
Qualidade	0,0009

4.3. Experimento 3

Os resultados dos experimentos anteriores motivaram a preparação de um experimento direcionado ao estudo apenas do critério 'qualidade' dos textos. Nele, 20 textos foram produzidos da seguinte forma: 10 textos do gênero informativo, por alunos experientes do curso de Letras da Universidade Mackenzie, SP, portanto, textos considerados bons; e 10 redações, por vestibulandos da Fuvest, retirados do corpus NILC, reconhecidamente ruins. Todos foram avaliados por seis alunos do curso de Letras da UFSCar, portanto, juízes com conhecimento da matéria. Os textos do primeiro grupo receberam notas acima da média, e foram classificados como textos da categoria "bons". Os textos da categoria "ruins" recaíram sobre as redações da Fuvest. Dessa vez, o teste t-student para as médias das notas das duas categorias (agora somente para o critério 'qualidade') indicou um valor *p* igual a zero. Tem-se assim uma confirmação de que os juízes puderam diferenciar os dois grupos de textos, o que permite uma investigação segura da correlação entre as notas e as medidas estatísticas provenientes das redes complexas.

A correlação entre as medidas estatísticas e as notas revelou vários fenômenos interessantes. A relação entre as médias dos graus de saída e as notas para o critério de qualidade tem um comportamento tal que, quando todos os textos são considerados, a qualidade tende a cair na medida em que os valores do grau de saída aumentam. Levando-se em consideração apenas os textos bons, percebeu-se que a qualidade quase independe dos graus de saída (na verdade, cresce levemente), enquanto que para os textos ruins, a qualidade aumenta à medida que o grau de saída aumenta. Em geral, os valores do grau de saída para os textos bons são menores do que os encontrados para os textos ruins, e a qualidade dos textos bons praticamente independe desses valores. Um número maior de associações aparece nos textos de menor qualidade e, dentro dessa classe, os textos com melhores notas parecem ser os que têm grau de saída maior. Os resultados para o grau de entrada são idênticos aos de saída e, portanto, foram aqui omitidos.

Já a relação entre o coeficiente de aglomeração e as notas revelou que a qualidade dos textos diminui com o coeficiente de aglomeração. Para os textos bons, o coeficiente de aglomeração é uniforme e independente da qualidade. Já nos textos ruins, o coeficiente de aglomeração tende a ser maior, com uma variação mais acentuada. Em particular, há um texto classificado como ruim com um coeficiente muito maior do que o dos outros. Esse texto, que trata da televisão e de sua influência na sociedade, repetidamente argumenta sobre as características maléficas e manipuladoras desse meio de comunicação. O coeficiente de aglomeração acentuado pode ter sido reflexo, como era de se esperar, desse argumento cíclico.

Quanto aos três tipos de caminhos mínimos, na medida em que aumentam, apresentam notas maiores. A qualidade é praticamente uniforme para os textos bons, sendo que, em média, seus caminhos mínimos são maiores do que os dos textos ruins. Esta classe por sua vez tem a qualidade diminuída conforme o tamanho do caminho mínimo aumenta. Isso indica que a qualidade do texto é prejudicada quando um escritor inexperiente tenta fazer caminhos longos, ou seja, que a manipulação de conexões longas exige uma maior habilidade do escritor [Koch 2002].

Uma verificação adicional foi realizada no Experimento 3, agora a respeito do comportamento temporal das redes. Esse comportamento diz respeito ao crescimento dinâmico de uma rede na medida em que as associações de palavras vão sendo adicionadas a ela, e é medido pelo seu número de componentes conexos [Antiqueira *et al.* 2005]. Essa abordagem mostrou que a variação temporal do número de componentes pode ser usada para distinguir os textos bons dos textos ruins utilizados no experimento.

5. Conclusões e Trabalhos Futuros

Este trabalho ressaltou a utilização dos novos conceitos de redes complexas na análise de dois grupos de textos de qualidade distinta, fato este sustentado por juízes humanos. As medidas estatísticas calculadas das redes foram confrontadas com as notas dadas pelos juízes, indicando que, para os textos bons, as medidas independem da qualidade. Os textos de baixa qualidade (categoria "ruins") são mais heterogêneos em qualidade, e uma correlação mais forte foi percebida para eles, de modo que as deficiências dos escritores menos experientes são mais aparentes nos parâmetros das redes complexas. Em especial, a qualidade diminui quando os caminhos mínimos aumentam, o que elucida a dificuldade dos escritores inexperientes em estabelecer conexões entre numerosos conceitos. Nos textos da categoria "bons", caminhos mínimos maiores foram observados. Entretanto, esse fato não correlacionou negativamente com as notas, ao contrário do que aconteceu com os textos da outra categoria. Trabalhos futuros devem replicar esses experimentos para outros tipos de textos, tendo em vista os resultados já obtidos. As redes complexas constituem assim um novo e promissor caminho para o processamento de textos, pois podem ser estudadas também na identificação de gêneros e autorias, na extração de terminologias, entre outras aplicações.

Referências Bibliográficas

Aires, R.V.X.; Aluísio, S.M.; Kuhn, D.C.S.; Andreeta, M.L.B.; Oliveira Jr., O.N. (2000) "Combining Multiple Classifiers to Improve Part of Speech Tagging: A Case Study for Brazilian Portuguese". In the Proceedings of the Brazilian AI Symposium (SBIA'2000), 20-22.

Albert, R.; Barabási, A.L. (2002) "Statistical Mechanics of Complex Networks", Rev. Modern Phys., 74, 47–97, cond-mat/0106096.

Antiqueira, L.; Nunes, M.G.V.; Oliveira Jr., O.N.; Costa, L.F. (2005) "Complex networks in the assessment of text quality", physics/0504033.

Barabási, A.L., "Linked: How Everything Is Connected to Everything Else and What It Means for Business, Science, and Everyday Life", Plume, New York, 2003.

Cancho, R.F.; Solé, R.V. (2001) "The Small World of Human Language", Proceedings of The Royal Society of London. Series B, Biological Sciences, 268, 2261-2265.

Casella, J.; R.L. Berger, “Statistical Inference”, Duxbury, Belmont, California, 2001.

Costa, L.F. (2003) “What’s in a Name?”, Int. J. Mod. Phys. C, Vol. 15, No. 1, 371-379, cond-mat/0309266.

Koch, I.G., “Desvendando os segredos do texto”, Cortez, São Paulo, 2002.

Miller, G.A. (1985) “Wordnet: a dictionary browser”, Proceedings of the First International Conference on Information in Data. University of Waterloo.

Motter, A.E.; Moura, A.P.S.; Lai, Y.C.; Dasgupta, P. (2002) “Topology of the Conceptual Network of Language”, Phys. Rev. E, 65, 065102.

Newman, M.E.J. (2003) “The Structure and Function of Complex Networks”, SIAM Review 45, 167-256, cond-mat/0303516.

Nunes, M.G.V.; Vieira, F.M.V.; Zavaglia, C.; Sossolote, C.R.C.; Hernandez, J. (1996) “O Processo de Construção de um Léxico para o Português do Brasil: Lições Aprendidas e Perspectivas”, II Encontro para o Processamento Computacional de Português Escrito e Falado, p.61-70. CEFET-PR, Curitiba.

Ratnaparkhi, A. (1996) “A Maximum Entropy Part-Of-Speech Tagger”. In the Proceedings of the Empirical Methods in Natural Language Processing Conference, University of Pennsylvania.

Sigman, M.; Cecchi, G.A. (2002) “Global Organization of the Wordnet Lexicon”, Proceedings of the National Academy of Sciences, 99, 1742-1747.