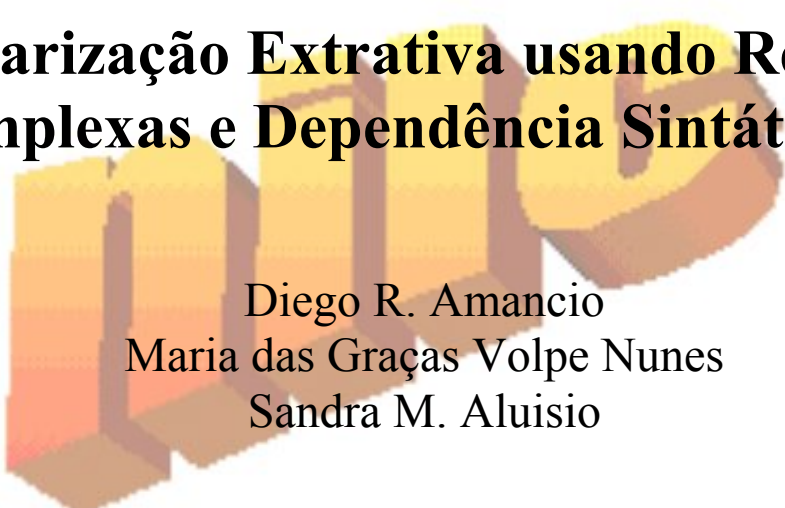


Universidade de São Paulo - USP
Universidade Federal de São Carlos - UFSCar
Universidade Estadual Paulista - UNESP

Sumarização Extrativa usando Redes Complexas e Dependência Sintática



Diego R. Amancio
Maria das Graças Volpe Nunes
Sandra M. Aluisio

NILC-TR-09-08

Dezembro, 2009

Série de Relatórios do Núcleo Interinstitucional de Linguística Computacional
NILC - ICMC-USP, Caixa Postal 668, 13560-970 São Carlos, SP, Brasil

Sumarização Extrativa Usando Redes Complexas e Dependência Sintática

Diego R. Amancio, Maria G. V. Nunes, Sandra M. Aluísio

1

***Resumo.** Este artigo estuda e avalia duas modelagens aplicadas em oito métodos baseados na análise de textos por redes complexas para a tarefa de sumarização extrativa. Especificamente, mostra-se que a adição de informação linguística mais profunda (dependência sintática entre núcleo de sujeitos e verbos) é capaz de promover um ganho de informatividade nos sumários. Os resultados mostram que a abordagem híbrida envolvendo o uso de redes e conhecimento linguístico proporciona maiores índices de qualidade em relação à abordagem baseada unicamente em redes.*

1. Introdução

Os conceitos e metodologias de redes complexas vêm sendo usados numa enorme variedade de áreas [1], incluindo a análise de textos escritos no que se denomina processamento de línguas naturais (PLN) [2]. Exemplos de ferramentas de PLN mais comuns são os tradutores automáticos [3], revisores gramaticais [4] e sumarizadores automáticos [5]. A adequação de redes complexas nesse tipo de análise foi demonstrada em várias instâncias, a partir da comprovação de que um texto pode ser representado por uma rede livre de escala [6, 7, 8, 9]. Neste contexto, o crescente número de metodologias para caracterização de redes tem permitido que tais aplicações da área de PLN sejam repensadas em termos de tal modelagem. Particularmente, este trabalho propõe sistemas de sumarização baseados em métricas desenvolvidas recentemente na teoria de redes [1]. Adicionalmente, uma nova modelagem baseada em dependência sintática é sugerida e sua influência na qualidade dos sumarizadores é avaliada.

Este artigo está organizado do seguinte modo: uma introdução à sumarização e à sua avaliação é descrita na Seção 2, uma introdução ao processo de *parsing* é ilustrada na Seção 3, as redes complexas e sua modelagem como textos são discutidas na Seção 4, seguida pela definição das métricas extraídas da rede, na Seção 5. Os sumarizadores propostos estão descritos na Seção 6 e o seu respectivo corpus de avaliação é descrito na Seção 7. Finalmente, os resultados dos experimentos e a conclusão são discutidos nas seções 8 e 9, respectivamente.

2. Sumarização extrativa e avaliação automática

Entende-se por sumarização automática extrativa a tarefa de reestruturação textual focada na seleção de fragmentos (usualmente sentenças) a fim de produzir uma versão mais limitada em tamanho [10]. Em outras palavras, a sumarização pretende compactar o texto fonte de forma que apenas a informação relevante esteja presente no sumário produzido. Ultimamente, tal aplicação tem atraído um grande interesse das pesquisas, uma vez que o crescente número de informação disponível exige algum tipo de filtragem da informação relevante. Neste contexto, pode-se dizer que uma das ambições da sumarização é o tratamento desse problema através da seleção da informação pertinente. Um outro exemplo

de aplicação é a geração de resumos, os quais possuem a função de indicar o assunto ou tema tratado por um dado texto, permitindo assim que uma leitura rápida indique se o dado texto é ou não de interesse do leitor.

A avaliação de sumários têm sido reconhecida como um caso à parte em relação à sua geração. Embora vários métodos de avaliação tenham sido desenvolvidos, ainda não há um consenso de qual é o melhor. Uma das dificuldades é a necessidade de avaliação de uma grande gama de critérios, tais como quantidade de informação presente, coerência, coesão, legibilidade, gramaticalidade e textualidade. Diferentemente dos métodos, as métricas de avaliação de métodos baseados em sumários de referências são mais padronizadas. Entre as métricas mais usadas estão a cobertura (ς) e precisão (Γ) [11], definidas como a porcentagem de sentenças do sumário referência que aparecem no sumário gerado e a porcentagem de sentenças do sumário gerado que estão no sumário referência. Adicionalmente, a métrica denominada medida-F (Φ) também costuma ser usada como um fator de ponderação entre precisão e cobertura, conforme equação 1.

$$\Phi = \frac{2 \cdot \Gamma \cdot \varsigma}{\Gamma + \varsigma} \quad (1)$$

Neste artigo, as métricas utilizadas referem-se à precisão, cobertura e medida-F presentes no pacote de avaliação automática denominado ROUGE¹ [12], baseadas na co-ocorrência de unidades (tais como n-gramas²) entre sumários criados automaticamente e sumários de referência. Optou-se por esta métrica uma vez que apresenta grande correlação com a avaliação humana [13, 14].

3. Análise sintática

Define-se a tarefa de análise sintática como a operação de determinação das estrutura de um texto. Tal estrutura consiste em uma hierarquia de fragmentos, que podem ir desde palavras até sentenças. Muitas vezes, associa-se tal tarefa com a estrutura de dados denominada árvore, sendo que os segmentos sintáticos mais simples são armanezados nos nós folhas, enquanto a sentença analisada está no vértice raiz.

A importância de tal tarefa se dá ao fato desta estabelecer as “regras” de formação da linguagem. Em outras palavras, dada uma gramática e uma linguagem, o analisador sintático procurar verificar se a sentença de entrada da linguagem segue as regras da gramática, além de identificar a função de cada segmento. Exemplos de analisadores sintáticos em linguagem de computadores são os compiladores e em linguagem natural humana são os *parsers*, como por exemplo o *parser* PALAVRAS [15].

Particularmente, nos experimentos deste artigo, utilizou-se o *parser* PALAVRAS, que fornece toda informação sintática³, isto é, trata-se de um parser completo. No entanto, apesar de todo aparato disponível, utilizou-se neste trabalho apenas o sintagma nominal (especificamente o seu núcleo) e os verbos, de forma a recuperar a dependência sintática entre núcleo de sintagma e verbo. Um exemplo de dependência sintática recuperada é exibida na Figura 1 à direita, conforme discussão na Seção 4.

¹Recall-Oriented Understudy for Gisting Evaluation.

²Particularmente neste trabalho foram utilizados apenas unigramas no processo de avaliação.

³O parser PALAVRAS também fornece algumas informações semânticas.

4. Redes Complexas e Modelagem de Textos

Redes complexas são grafos que apresentam alguma organização especial, isto é, grafos cuja estrutura segue princípios complexos de organização [16, 17]. Por exemplo, a *World Wide Web*, se vista como um grafo (em que os vértices são páginas e as arestas os *links* entre as páginas), segue padrões não aleatórios de organização, ou seja, sua estrutura não se dá ao acaso. A rede complexa correspondente à *World Wide Web* e outros sistemas naturais (tráfego em aeroportos, malha rodoviária, distribuição elétrica, etc.) apresentam propriedades importantes. As redes são livres de escala: seus nós seguem a lei de potência, ou seja, alguns poucos nós, chamados *hubs*, têm muitas conexões. As redes também podem apresentar a propriedade de pequeno mundo, caracterizada pela existência de um caminho relativamente curto entre quaisquer dois nós da rede. Tal propriedade contribui para que a rede apresente comunidades, isto é, grupos de nós altamente interconectados.

Recentemente, devido a suas propriedades interessantes, as redes começaram a ser aplicadas em estudos de PLN. Já foi provado que a rede de adjacência das palavras é livre de escala [6]. Apoiando-se neste estudo e em trabalhos como o de [7], que utilizaram redes para modelar recursos léxicos, [18] estudou a associação de palavras induzidas por humanos. Já em [19], investigou-se a evolução da linguagem por meio de redes complexas. Uma proposta de modelagem de textos via redes complexas foi apresentada, por exemplo, em [20]. Tal modelagem foi a base para trabalhos de atribuição de autoria [21] e de averiguação de qualidade de textos [22] e traduções [23].

Com relação às redes de dependência sintática, destaca-se o trabalho desenvolvido em [24]. Neste, mostra-se que a rede de dependência sintática de palavras para três línguas razoavelmente distantes (Tcheco, Alemão e Romeno) é complexa no sentido de apresentar as características de escala na distribuição dos graus e aglomeração de vértices, além da propriedade pequeno mundo⁴, de modo análogo à rede de adjacência de palavras [6]. Também, de especial interesse, mostra-se que estes padrões estatísticos na organização da sintaxe não são resultados de uma organização trivial das sentenças, mas uma característica intrínseca na estrutura global.

Particularmente, a modelagem proposta por [20] foi adotada neste trabalho com o nome de SUM-RC. Em tal modelagem, a rede correspondente a um texto é construída segundo relações de adjacência entre palavras. Inicialmente, em uma etapa de pré-processamento, o texto é lematizado e tem as *stopwords* removidas. A seguir, cada palavra é representada como um vértice na rede, sendo que palavras iguais são representadas por um mesmo vértice. As arestas são obtidas associando-se vértices cujas palavras correspondentes são imediatamente adjacentes (não se consideram limites de sentenças ou parágrafos). Caso uma associação seja repetida, incrementa-se em 1 o peso da aresta. Assim, obtém-se um grafo ponderado representado por uma matriz de adjacência quadrada W , em que cada $W(j,i)$ representa o número de associações $i \rightarrow j$ encontradas no texto, em que i e j são pares de palavras imediatamente adjacentes. Motivado pelo fato de que a adição de conhecimento lingüístico tende a melhorar a informatividade de sumarizadores extrativos [25], este trabalho propõe uma nova modelagem, denominada SUM-P. Particularmente, esta adiciona à modelagem SUM-RC arestas referentes à dependência sintática entre núcleos de sujeito e verbos, nesta ordem, com auxílio do *parser* PALA-

⁴A propriedade pequeno mundo refere-se à existência de caminhos curtos (em relação ao comprimento da rede) conectando quaisquer dois vértices.

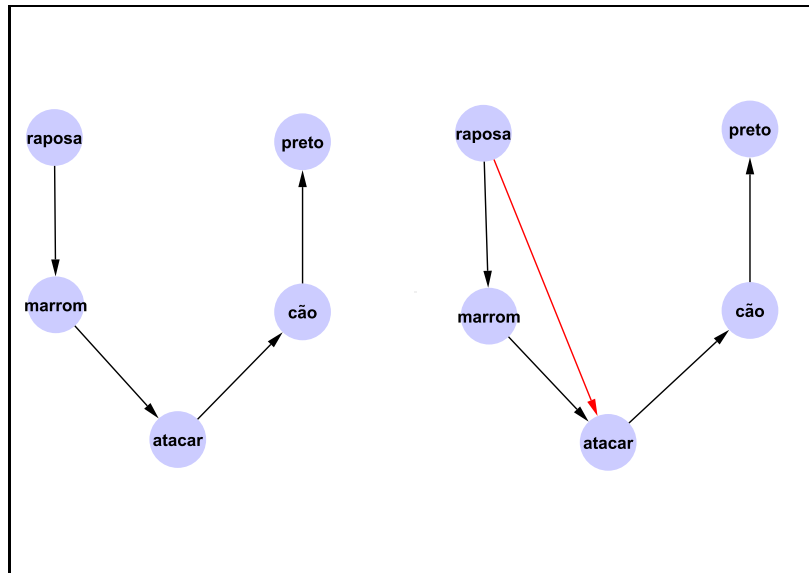


Figura 1. Exemplo de rede formada com a modelagem SUM-RC (esquerda) e SUM-P(direita). A aresta em vermelho existe devido à relação sintática extraída entre núcleo do sintagma nominal e verbo principal.

VRAS. Como exemplo de cada uma das modelagens, a Figura 1 ilustra a rede formada a partir da modelagem SUM-RC (esquerda) e SUM-P (direita) para a sentença “A raposa marrom ataca o cão preto”, destacando em vermelho a aresta referente à dependência sintática.

5. Métricas de Redes Complexas

5.1. Graus

As medidas de grau, também conhecidas como conectividade na literatura Física [1] podem ser divididas em dois tipos. O grau de saída de um nó (OD) corresponde à soma das ponderações das arestas que saem de um nó. Analogamente, define-se o grau de entrada de um nó (ID), como a soma das ponderações das arestas que incidem naquele nó. Ambas as medidas globais são calculadas como a média de todos os nós da medida em questão. É importante notar que ao se trabalhar com a medida global⁵ dos graus, OD e ID são sempre iguais, já que a soma de todas as arestas que saem é igual a soma de todas as arestas que incidem naquela rede. Em termos da matriz de adjacência utilizada para representar uma rede com ponderações [26] tem-se que para um dado nó i seu OD e seu ID são calculados pelas expressões abaixo, sendo que N representa o número total de nós⁶ :

$$ID(i) = \sum_{j=1}^N W_{ij} \quad (2)$$

$$OD(i) = \sum_{j=1}^N W_{ji} \quad (3)$$

⁵A medida global difere da medida local, pois esta última refere-se a um nó em específico, enquanto a primeira refere-se à média da medida para todos os nós da rede.

⁶Neste artigo, cada palavra distinta do texto pré-processado é modelada como um nó.

5.2. Diversidade

Com o crescimento do emprego das redes complexas como ferramenta de modelagem de sistemas reais, houve a necessidade de um aperfeiçoamento das técnicas relativas à identificação de alguns conceitos relevantes relativos a este tipo de modelagem. Um desses conceitos refere-se ao grau de importância de cada vértice. Dentre as abordagens padrões, pode-se dizer que as métricas de grau de entrada ou saída têm sido as principais no reconhecimento da relevância de vértices, já que em várias modelagens o grau está relacionado com a frequência do vértice, daí sua relação direta com a relevância. Adicionalmente, outras medidas tradicionais como coeficiente de aglomeração ou vulnerabilidade também podem ser utilizadas para esta mesma tarefa, dependendo do objetivo relacionado à aplicação.

Relativo à tarefa de reconhecimento da relevância de vértices, define-se um conceito recente na área: a identificação da interioridade dos vértices na rede. Mais especificamente, o estabelecimento da métrica denominada diversidade quantifica a proximidade em que um vértice se encontra da borda de forma a atribuir baixa relevância aos vértices marginais e conseqüentemente alta relevância aos vértices centrais.

Abordagens simples e intuitivas na identificação de bordas podem quantificar esta propriedade através de algoritmos que calculam a distância do vértice analisado a todos os vértices folhas. Com esta definição, aqueles vértices cuja distância média aos folhas é relativamente pequena têm grande probabilidade de estar próximo da borda. No entanto, este método possui suas limitações, uma vez que a média pode não ser uma boa abordagem para distribuições com grande desvio.

Métodos recentes abordam o problema de identificação de bordas com o uso de entropia de probabilidade das arestas da rede, baseando-se no fato de que se um vértice está próximo da borda a variedade de acesso aos outros vértices a partir deste é pequena. Um exemplo da metodologia é ilustrado na Figura 2. Neste exemplo, ilustra-se a variedade de acesso esperada para uma rede fictícia. Os vértices interiores (tons mais claros) apresentam uma maior homogeneidade de possibilidades para percorrer caminhos de um dado comprimento enquanto os vértices mais externos apresentam uma maior heterogeneidade de possibilidades para percorrer tais caminhos, o que caracteriza a baixa diversidade (representada pelos tons escuros). Então, a definição desta métrica deve fornecer altos valores para vértices centrais com distribuição de caminhos homogênea.

Formalmente, define-se neste artigo a diversidade de um dado vértice v como sendo a entropia de probabilidade de transição entre vértices vizinhos, através de caminhos aleatórios sobre a rede. Seja a diversidade do vértice em questão representada por δ_v , isto é, o valor que quantifica o quão diverso é o acesso deste nó a partir de caminhos aleatórios de tamanho h (o caminho envolve exatamente h vértices), iniciados a partir dos outros $(N - 1)$ vértices da rede. Calcula-se então δ_v como apresentado na equação 4:

$$\delta_v^h = -\frac{1}{\log(N - 1)} \sum_{j=1}^N P_h(j, v) \cdot \log(P_h(j, v)) \quad (4)$$

$$P_h(j, v) = \sum_{c=1}^{NC_h} \prod_{(x,y) \in c} \frac{W_{yx}}{\sum_k W_{kx}} \quad (5)$$

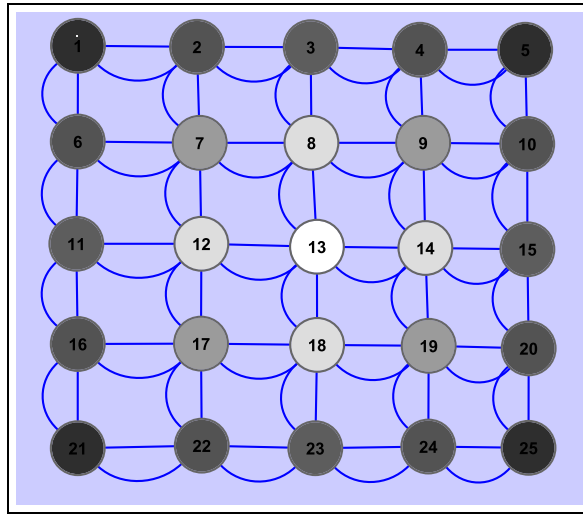


Figura 2. Identificação de bordas utilizando a diversidade de acesso ao vértice em caminhos de distância igual a 3. Quanto mais escuro, menor é o valor da diversidade e mais próximo da borda o vértice se apresenta.

com NC_h representando o número de caminhos de tamanho h desde o vértice v até o vértice j . A equação 5 representa a probabilidade do caminho aleatório seguir a aresta do vértice v ao vértice j . Caso $P_h(j,v)$ seja zero, toma-se o produto no interior da somatória na equação 4 como sendo zero.

A fim de validar a equação 4 ilustram-se a seguir os valores reais de diversidade aplicados a uma rede com 100 vértices, dispostos numa topologia semelhante à topologia da Figura 2, isto é, cada vértice na posição não periférica disposto na posição (x,y) possui ligações com seus vizinhos da posição $(x+1,y)$, $(x-1,y)$, $(x,y+1)$ e $(x,y-1)$. Os resultados são ilustrados na Figura 3 para $h=1$ até $h=6$. À esquerda, exemplifica-se o valor da diversidade da rede em questão de forma que o plano xy representa o vértice posicionado na posição (x,y) e o eixo z representa o valor da sua diversidade. De fato, à medida que se aproxima do centro $(5,5)$, maior é o valor da diversidade. A mesma conclusão pode ser deduzida observando o corresponde gráfico de curvas de nível na Figura 3 à direita.

5.3. Eficiência Global

A utilização de distância média entre vértices, apesar de bem intuitiva e bastante utilizada na teoria de grafos, apresenta a desvantagem de poder divergir caso exista na rede nós desconectados. Neste contexto, para superar esta situação, define-se a medida relativa à eficiência global (GE) da rede na equação 6, onde o termo d_{ij} representa a distância entre os vértices i e j da rede.

$$GE = \frac{1}{N(N-1)} \sum_{i \neq j} \frac{1}{d_{ij}} \quad (6)$$

A interpretação desta medida está relacionada com a capacidade da rede em trocar informações entre quaisquer dois nós, dado que se uma distância d_{ij} for pequena esta contribuirá de forma mais significativa em relação a uma distância d_{ij} grande. Nota-se que a fórmula acima é uma das formas que previne a divergência das medidas relacio-

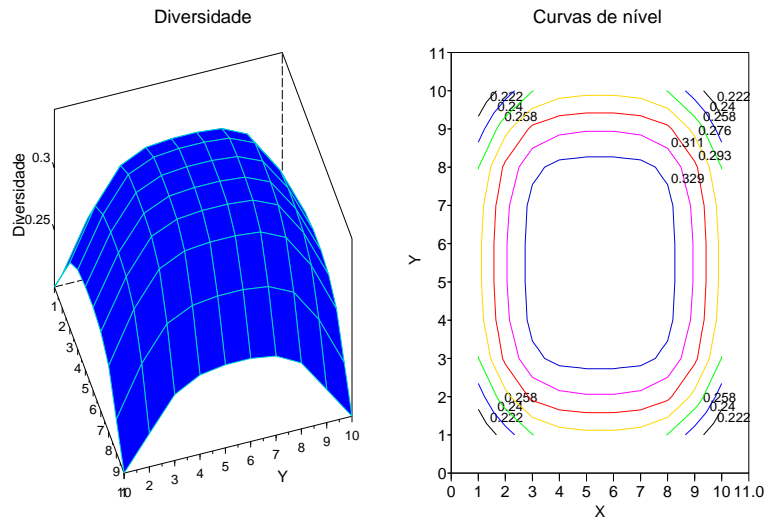


Figura 3. *Diversidade em função da posição do vértice em forma de função de duas variáveis (esquerda) com suas respectivas curvas de nível (direita) evidenciam que vértices internos possuem maiores valores da métrica de diversidade da entropia.*

nadas com distâncias, portanto, mostra-se útil no caso do grafo apresentar mais de uma componente⁷.

Pode-se ressaltar também que o inverso de GE também tem sido utilizado como medida em redes complexas, sendo conhecido também como *média harmônica das distâncias geodésicas*.

5.4. Vulnerabilidade

Sabe-se que em uma rede complexa, nem todos os nós apresentam a mesma importância na manutenção da estrutura e funcionamento da rede como um todo. Para perceber isto, basta observar intuitivamente os *hubs*⁸ e logo nota-se que estes são essenciais para a rede devido ao seu alto grau de conectividade. Mas os vértices importantes não se resumem apenas aos *hubs*: supondo um considerável subgrafo com característica muito próxima de uma árvore binária, a raiz desta árvore seria um ponto de vulnerabilidade, já que se trata de um único meio de ligação entre duas árvores e este vértice raiz não necessariamente seria um *hub*.

Para quantificar o grau de importância de um dado vértice na rede, deve-se verificar a variação da estrutura da rede quando o nó é removido da rede, isto é, todas as arestas referentes àquele vértice, além do próprio vértice são removidos. Uma maneira comum seria considerar como variação da rede uma medida de desempenho, tal como a eficiência global da rede. Assim, usualmente define-se a vulnerabilidade de um vértice i como sendo:

$$V_i = \frac{GE - GE_i}{GE} \quad (7)$$

⁷ caso da rede de adjacência de palavras obtida da modelagem do texto como rede complexa.

⁸ Ver seção 4.

na qual GE refere-se à eficiência global da rede com todos os nós e GE_i refere-se à eficiência global da rede após a retirada do vértice i . Nota-se que V_i representa a diferença relativa percentual (se multiplicado por 100) refletindo a remoção daquele nó. Diferentemente das outras medidas vistas até aqui, a medida global correspondente não é a média obtida da vulnerabilidade de cada nó. Neste caso o importante é conhecer o pior da caso da rede, então segue que a vulnerabilidade da rede é:

$$V = \max_i V_i \quad (8)$$

5.5. Closeness

Um caminho mínimo, uma das medidas relacionadas com distâncias dentro da rede, é também conhecido como caminho geodésico entre dois nós. Trata-se de um caminho conectando dois nós com distância mínima, levando-se em consideração a soma das ponderações das arestas percorridas por este caminho. Tal caminho, que não precisa ser único, é calculado entre um dado nó fixo e todos os outros nós. Para a rede como um todo, define-se o caminho mínimo como a média dos comprimentos de todos os caminhos mínimos existentes entre quaisquer dois vértices. Localmente, define-se o caminho mínimo de um vértice como :

$$CLS(i) = \frac{\sum_{j \neq i} d_{ij}}{N - 1} \quad (9)$$

5.6. Betweenness

Medidas de centralidade, como *betweenness centrality* (B_v), buscam classificar a importância dos vértices ou arestas da rede. Para isso, o *betweenness centrality* classifica um dado vértice utilizando o número de caminhos mínimos que passam por este vértice. Mais precisamente, a definição de *betweenness centrality* para um dado vértice v é:

$$B_v = \sum_i \sum_j \frac{\sigma(i, v, j)}{\sigma(i, j)} \quad (10)$$

na qual o numerador da fórmula representa o número de caminhos mínimos que passam pelos vértices i , v e j e o denominador significa o número de caminhos mínimos que passam pelos vértices i e j .

6. Sumarizadores propostos

6.1. Diversidade

A medida de diversidade (δ_v^x , equação 4) e seus vários níveis hierárquicos é utilizada na construção dos textos simplificados em forma de extrato conforme descrito a seguir. Inicialmente, calcula-se o número de sentenças que devem fazer parte do texto sumarizado, de acordo com o parâmetro referente à taxa de compressão τ do sumário. Em seguida, para cada vértice (palavra) presente na rede, calcula-se a métrica de diversidade até o nível desejado⁹. Em seguida, um ponderação (ω_s) é atribuída a cada sentença s presente

⁹O método denominado DIV x indica que utilizou-se a métrica de diversidade calculada no nível x , ou correspondentemente δ_v^x . Já o método denominado DM3 calcula para cada vértice os níveis hierárquicos $h = 1 \dots 3$ e armazena a média destes valores.

no texto, de acordo com o peso de cada palavra ρ_i^s presente nesta sentença s , como ilustra a equação 11.

$$\omega_s = \frac{1}{\eta_s} \sum_{i=1}^{\eta_s} \delta_{\rho_i^s}^x \quad (11)$$

onde o termo η_s representa o número de palavras desta sentença.

Em seguida, ordenam-se os valores de ω_s em ordem decrescente e determina-se o limiar de corte como o valor no vetor ordenado de ω_s na posição $(1 - \tau) \cdot \iota$, onde ι representa o tamanho do texto original, em número de sentenças. Por fim, as sentenças do texto original são percorridas uma a uma na mesma ordem em que aparecem no texto original, para que mantenham a ordem natural do texto-fonte, sendo eleita para fazer parte do sumário somente se seu valor ω_s for superior ao limiar definido anteriormente.

A intuição presente neste método pressupõe o fato de que sentenças que apresentem valores médios de diversidade relativamente alto possuem grande chance de reunir conceitos chaves e provavelmente são capazes de fornecer bom grau de informatividade ao texto simplificado. De fato, esta interpretação parece razoável, uma vez que valores altos de diversidade parecem estar correlacionados com as palavras chaves, neste tipo de modelagem.

6.2. Diversidade e Grau

Para este método, denominado DDG, também utiliza-se a modelagem da rede por palavras, de forma que as métricas (neste caso, diversidade e grau) sejam calculadas para cada palavra individualmente. Basicamente, tenta-se utilizar as vantagens de quantificação de centralidade de ambas as medidas. Para isso, inicialmente, segue-se o mesmo passo descrito na Seção 6.1 para a métrica de diversidade, de forma a obter dois vetores ordenados: o primeiro, de acordo com a métrica de diversidade e o segundo de acordo com a métrica de grau. Para cada sentença, verifica-se sua classificação para que se possa eleger como sentenças do texto simplificado aquelas que aparecem bem classificadas nos dois vetores. Então, uma nova ponderação (ω'_s) é atribuída para cada sentença: se a sentença s é classificada na posição ρ_1 no primeiro vetor (relativo à diversidade) e na posição ρ_2 no segundo vetor (relativo ao grau), então ω'_s é tomado simplesmente como descrito na equação 12.

$$\omega'_s = \rho_1 + \rho_2 \quad (12)$$

Finalmente são escolhidas as sentenças com menor valor de ω'_s , uma vez que quanto menor é tal ponderação, maior será a relevância (segundo as duas métricas utilizadas) atribuída a tal sentença. Por exemplo: se uma sentença é classificada na primeira e terceira posição, e outra é classificada na segunda e terceira posição, a primeira será considerada mais relevante, pois está mais bem classificada nas duas listas. A partir deste ponto, este método é análogo ao descrito na Seção 6.1, com a diferença que os menores valores do vetor de ponderações ω'_s são selecionados.

6.3. Closeness

Este método, denominado CLS, é o análogo ao método descrito na Seção 6.1, distinguindo-se pelo uso da métrica de centralidade denominada *closeness* (Seção 5.5).

Tal método pretende selecionar palavras chaves do seguinte modo: dado um vértice v , considera-se este como de alta relevância quando, partindo-se dele, é relativamente fácil chegar a qualquer outro vértice na rede através de caminhos mínimos, o que pode ser quantificado por altos valores da métrica *closeness*. Esta abordagem é justificável do ponto de vista de que vértices relativamente próximos a outros podem ser *hubs* ou, no pior caso, estão próximos aos *hubs*, aumentando portanto a possibilidade que tais vértices representem conceitos de alta informatividade para o texto simplificado. Da mesma forma, se um dado vértice é distante dos outros, então provavelmente está pouco relacionado com os demais e provavelmente não se trata de um vértice de conteúdo especial.

6.4. *Betweenness*

Assim como o método descrito na Seção 6.1, este método (BTW) utiliza uma métrica de centralidade para caracterizar as palavras chaves. Neste caso, a métrica denominada *betweenness* é utilizada para isto.

Analogamente a outras técnicas tradicionais de centralidade, tal métrica é baseada no cálculo de caminhos mínimos para classificar a proeminência de um vértice. Neste caso, um vértice é candidato a representar uma palavra chave caso tenha um alto valor de *betweenness*, ou seja, vários caminhos mínimos passa por ele. De fato, tal abordagem parece ser intuitiva, uma vez que vértices de alto *betweenness* tendem a ser mais acessados em caminhadas eficientes. Portanto, este método seleciona como palavras chaves aquelas com tendência em abreviar a distância entre conceitos, isto é, aquela que se apresenta como um ponto comum entre vários caminhos de agregação de conceitos.

6.5. Vulnerabilidade

Equivalente ao método descrito na Seção 6.1, este método utiliza a métrica de vulnerabilidade localmente, atribuindo uma relação diretamente proporcional entre relevância da palavra e seu valor de vulnerabilidade. A partir da definição de vulnerabilidade (Seção 5.4), este método considera que uma palavra é importante se a retirada do vértice correspondente na rede complexa diminui a eficiência global da rede¹⁰. Portanto, vértices que mantêm a estrutura da rede ou são pontos de articulação¹¹ [27] terão alto score quanto à classificação em palavras-chaves.

7. Corpus para sumarização

Esta seção descreve o corpus utilizado nos experimentos referentes à sumarização textual. Fundamentalmente, os textos dividem-se em dois conjuntos, segundo seu intento: textos fontes, a partir dos quais os sumários são construídos e textos de referência, úteis na identificação da qualidade do sumário gerado.

O corpus utilizado, denominado TeMário¹² e disponível via web¹³ foi construído justamente para auxiliar a tarefa de sumarização automática. Fazem parte da constituição

¹⁰A diminuição na eficiência global corresponde ao aumento da distância entre vértices da rede, conforme definição na Seção 5.3.

¹¹Pontos de articulação são vértices que quando retirados da rede aumentam o número de componentes conexos. Uma propriedade interessante com respeito a este tipo de vértice é que se uma rede não possui pontos de articulação então será biconexa, ou seja, existe pelo menos dois caminhos conectando dois vértices distintos na rede.

¹²TExtos com suMÁRIOS.

¹³<http://www.linguateca.pt/Repositorio/TeMario>

dos textos fontes um conjunto de 100 textos jornalísticos, sendo 60 pertencentes ao jornal Folha de São Paulo¹⁴ (versão on-line) e 40 pertencentes ao Jornal do Brasil¹⁵, totalizando mais de 60000 palavras. Quanto ao assunto dos textos, pode-se dizer que este é suficientemente variado, uma vez que os textos estão distribuídos igualmente nas Seções Mundo, Opinião, Especial (Folha de São Paulo), Política e Internacional (Jornal do Brasil).

A escolha de textos de tal gênero para composição do corpus se deve ao fato da sua heterogeneidade de complexidade linguística, uma vez que tais textos possuem um público leitor abrangente. Adicionalmente, justifica-se o uso do gênero jornalístico pelo fato deste ser amplamente utilizado na avaliação de sumarizadores automáticos em concursos internacionais, como por exemplo a TAC¹⁶ (*Text Analysis Conference*).

Os textos de referência, utilizados na comparação com os sumários automáticos, podem ser de natureza automática ou humana. Neste trabalho, utilizou-se sumários de referência de natureza humana, realizada por um professor e consultor de editoração de textos em português. Adicionalmente, restringiu-se o tamanho do sumário de forma a corresponder entre 25% a 30% do tamanho do texto original, ou seja, do texto fonte. Desta forma, sumários informativos manuais construídos de forma profissional complementam o TeMário a fim de constituir um repositório significativo (apesar de relativamente pequeno) de dados para as tarefas de sumarização automática e sua respectiva avaliação de qualidade.

8. Resultados e discussão

Os experimentos desenvolvidos concentraram-se em duas fases. No primeiro (Experimento 1), procurou-se verificar a eficiência da métrica de diversidade em selecionar palavras chaves e no segundo (Experimento 2), verificou-se a qualidade dos sumarizadores propostos e confrontou-se as modelagens baseadas em dependência sintática (SUM-P) e em redes complexas (SUM-RC).

8.1. Experimento 1

A fim de motivar a utilização de novas métricas recém introduzidas na análise de redes complexas pretende-se mostrar neste experimento que a métrica de diversidade definida na Seção 5.2 é capaz de identificar as palavras chaves em um texto, isto é, as palavras mais relevantes e que representam os conceitos centrais para o desenvolvimento textual adequado. Para isto, adotou-se a hipótese de que as palavras chaves devem se encontrar no núcleo da rede, ao mesmo tempo que as palavras de menor relevância provavelmente se encontram na periferia ou próximas à ela.

A metodologia consistiu em aplicar a definição desta métrica para vários textos relativos a um único tema a fim de verificar quais palavras apresentaram maiores e menores valores de diversidade e se tais palavras são adequadas ao tema proposto. Adicionalmente, verifica-se se esta medida é redundante, isto é, se há alguma correlação com as métricas padrões utilizadas. Tal verificação é importante uma vez que questiona a originalidade desta métrica para a tarefa de sumarização.

¹⁴<http://www.folha.com.br>

¹⁵<http://jbonline.terra.com.br>

¹⁶<http://www.nist.gov/tac/2009/Summarization/index.html>

Palavra	Frequência	Palavra	Frequência
ser	732	fazer	132
não	447	pessoa	104
votar	407	candidato	104
ter	384	estar	90
voto	379	ir	80
poder	255	político	84
país	246	melhor	63
direito	246	brasileiro	58
dever	162	povo	58
Brasil	162	saber	57

Tabela 1. Tabela representando o número de vezes que uma dada palavra foi considerada entre as 10% mais importantes pela estratégia de detecção de bordas.

Como corpus para este experimento utilizou-se um conjunto de 300 redações no formato dissertativo e argumentativo, cujo tema foi proposto no âmbito do Exame Nacional do Ensino Médio (ENEM) do ano de 2002. Especificamente, o tema escolhido deveria ser captado pelo candidato, que dentre outras fontes de informação (trechos extraídos de jornais ou revistas), possuía a seguinte pergunta a ser respondida dentro do seu texto: **“O direito de votar: como fazer dessa conquista um meio para promover as transformações sociais de que o Brasil necessita ?”**. Desta forma, espera-se que os conceitos chaves estejam ligados de algum modo correlacionados semânticamente com os termos votação, eleições e outros.

A estratégia desenvolvida aqui consistiu em averiguar para cada texto e para cada valor de h^{17} da métrica de diversidade quais são as palavras (ou simplesmente os vértices da rede) que se apresentam com maiores e menores valores de diversidade. Para isto, para cada texto e para cada valor de h , determinaram-se os vértices cujos valores da diversidade estavam entre os 10% maiores valores (ou menores valores, para avaliação dos menos relevantes). Desta forma, para cada texto, uma dada palavra pode aparecer até h vezes entre os 10% maiores valores, já que este cálculo é realizado para cada nível. Portanto, a frequência de cada palavra estará no intervalo de 0 a 900, uma vez que a contagem é realizada sobre as 300 redações do corpus.

As palavras de maior frequência na lista de alta diversidade estão sumarizados na Tabela 1. Nota-se que várias palavras de grande importância relacionada ao tema aparecem na tabela (destacado), além das palavras de uso comum na linguagem, como os verbos **ser**, **ter**, **fazer**, **estar** e **ir**. A presença de tais verbos, não relacionados aos tema (são verbos de uso geral na linguagem) estimula intuitivamente a adição destes e outros verbos na lista de *stopwords*, uma vez que são análogos aos artigos e às preposições quanto à questão semântica. Complementando a Tabela 1, a Tabela 2 exhibe as palavras mais frequentes na lista de menor diversidade. Pode-se notar que grande parte das palavras está fracamente correlacionada com o tema desenvolvido nas redações, o que confirma que os vértices próximos da borda referem-se aos vértices de pouca importância.

¹⁷Para maiores detalhes a respeito do parâmetro h (variando de 1 a 3) na métrica de diversidade consulte a Seção 5.2.

Palavra	Frequência	Palavra	Frequência
ano	125	agora	62
bem	99	antes	61
cidadão	82	analisar	61
ajudar	74	conquista	60
acabar	72	arma	52
bom	72	acontecer	52
ainda	66	através	50
acreditar	64	coisa	45

Tabela 2. Tabela representando o número de vezes que uma dada palavra foi considerada entre as 10% com menor valor de diversidade, representando portanto as bordas.

A fim de verificar a correlação entre a diversidade em vários níveis com outras métricas comuns na identificação de vértices centrais, algumas métricas de centralidade foram calculadas localmente para cada texto: OD, ID, CLS, BTW, além das métricas padrões de caminho mínimo (SP_1) [1] e coeficiente de aglomeração (CC) [1]. A Figura 4 ilustra a dinâmica da taxa média do módulo do coeficiente de correlação de Pearson (com indicação do desvio padrão) para tais medidas, para valores de h variando de 1 a 5. Como pode-se notar, em todos os casos, a correlação não é suficientemente grande, indicando que a utilização da métrica de diversidade não pode ser substituída com mesmo efeito. Adicionalmente, também é importante notar que tal correlação é inversamente proporcional à hierarquia analisada, sugerindo que métodos mais precisos de identificações de bordas possuem tal correlação com métricas de centralidade padrões ainda menores.

Enfim, os resultados mostram que as palavras localizadas mais internamente (com maiores valores de diversidade) na rede de adjacência de palavras possuem forte ligação com o tema principal, reforçando a hipótese de que as palavras (vértices) de alta diversidade da rede identificam de alguma maneira a centralidade dos conceitos no texto, caracterizando portanto uma abordagem original (devido à sua baixa correlação com outras métricas) na detecção de palavras chaves.

Enfim, pode-se perceber que este método tem grande tendência em eleger as palavras chaves do texto, favorecendo a estratégia de sumarização usando a técnica denominada de extração de palavras chaves.

8.2. Experimento 2

Neste experimento, os sumarizadores propostos foram gerados utilizando ambas as modelagens (SUM-RC e SUM-P) e comparados quanto à qualidade, primeiramente dentro de cada modelagem. Adicionalmente, verificou-se a correlação entre sumarizadores, no intuito de perceber semelhanças entre sumários gerados dentro de uma mesma modelagem. A Tabela 3 exibe os valores encontrados para o índice de qualidade ROUGE-1 na modelagem convencional. Tais resultados (principalmente os relativos à diversidade) podem ser considerados satisfatórios quando comparados com outros métodos baseados em medidas padrões, dado que o melhor sumarizador existente baseado em redes complexas apresenta um índice de cobertura de 0.5031 [28]. Como próximo passo, verificou-se a correlação entre os postos das sentenças geradas através do coeficiente de correlação de Spearman [29]. Tal coeficiente é máximo (igual a 1) quando a ordenação de sentenças elegíveis para com-

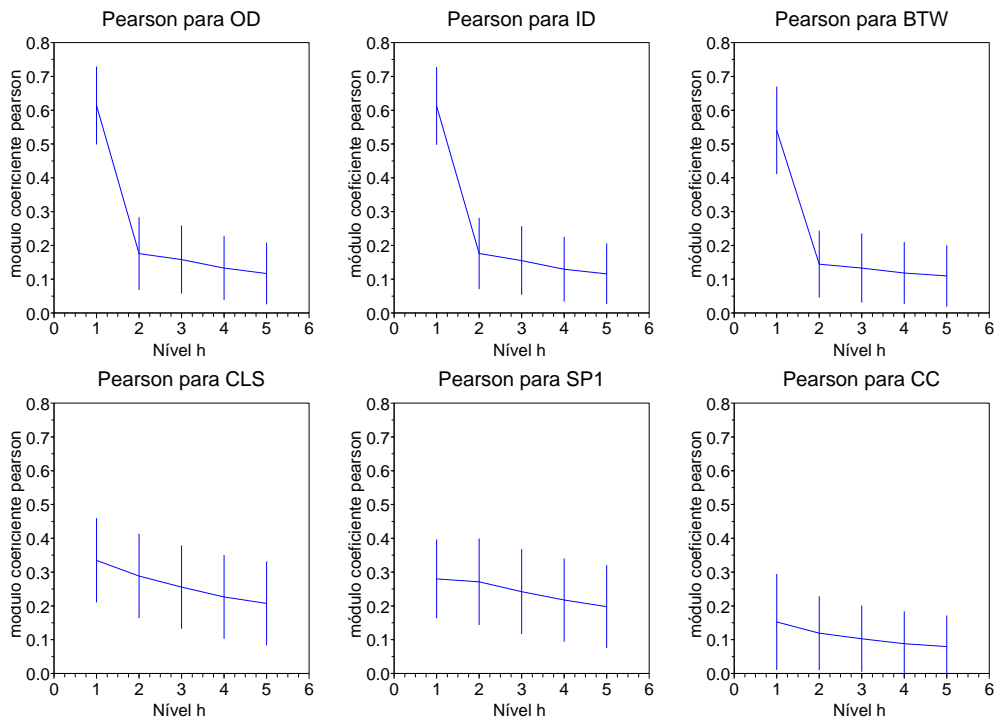


Figura 4. Dinâmica da correlação da diversidade com os graus de entrada e saída, de acordo com o nível h .

por o sumário for igual e é mínima (igual a -1) quando os postos resultantes da ordenação for invertido (a primeira sentença de um é a última do outro, a segunda sentença de um é a penúltima do outro e assim por diante). Os resultados do índice de similaridade de postos entre sumarizadores está ilustrada na Figura 5. Uma rápida observação de tal figura deixa claro que os métodos baseados em diversidade possuem alta correlação, assim como aqueles baseados em métricas geodésicas são correlatos entre si. No meio termo, o método DDG parece não apresentar uma correlação forte com qualquer outro método, provavelmente por ser uma abordagem híbrida que considera métricas de diversidade e métricas geodésicas.

Método	Cobertura	Precisão	Medida-F
DIV1	0.5085	0.3847	0.4305
DIV3	0.5058	0.3779	0.4245
DIV2	0.5046	0.3810	0.4265
DM3	0.5045	0.3796	0.4254
CLS	0.5032	0.3761	0.4228
DDG	0.4983	0.3942	0.4323
BTW	0.4954	0.3859	0.4268
VUL	0.4882	0.3766	0.4178

Tabela 3. Cobertura, precisão e medida-F para ROUGE-1 (unigramas) para métodos baseados na modelagem convencional.

Agora, usando a modelagem baseada em dependência sintática entre sujeito e verbo, percebe-se que em geral há um aumento na qualidade dos sumários, como ilus-

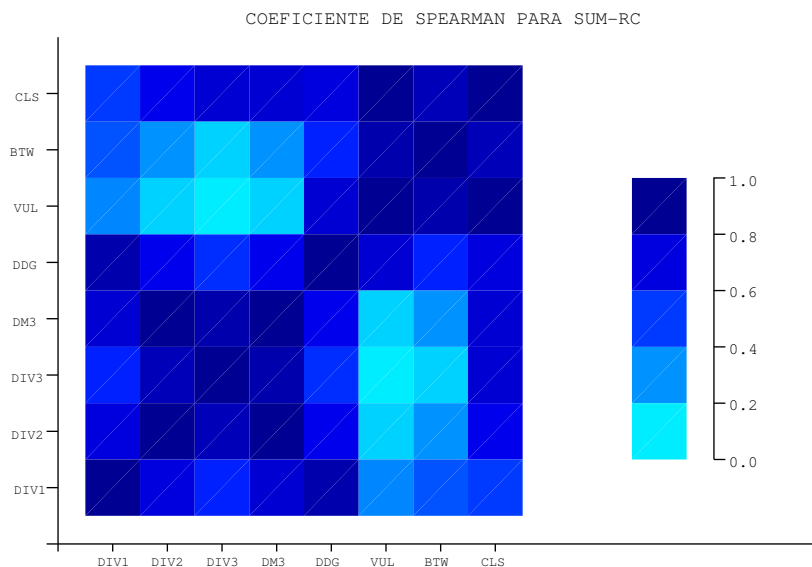


Figura 5. Coeficiente de correlação de Spearman para métodos referentes à modelagem por redes complexas (SUM-RC).

trado na Tabela 4. Pode-se perceber que em praticamente todos os sumariantes houve algum tipo de ganho: em precisão, cobertura ou em ambos. Apesar deste ganho, pode-se perceber que o melhor resultado pouco mudou (aumento de 0.004 em cobertura, 0.0018 em precisão e 0.0013 em medida-F). Também, podemos perceber que a correlação entre os sumariantes pouco mudou (Figura 6) em relação ao caso da modelagem convencional, o que mostra que a escolha de sentenças não muda muito. No entanto, esta mudança por pequena que seja, já consegue elevar os índices de qualidade.

Método	Cobertura	Precisão	Medida-F
DIV1	0.5089	0.3865	0.4318
DIV3	0.5073	0.3815	0.4272
DIV2	0.5051	0.3722	0.4210
CLS	0.5047	0.3800	0.4253
DM3	0.5041	0.3816	0.4265
DDG	0.4989	0.3899	0.4304
BTW	0.4901	0.3834	0.4223
VUL	0.4887	0.3807	0.4202

Tabela 4. Cobertura, precisão e medida-F para ROUGE-1 (unigramas) para métodos baseados na modelagem de dependência sintática.

Como análise subsequente, fixou-se os sumariantes e variou-se o tipo de modelagem, a fim de verificar as relações entre as duas modelagens de forma mais profunda. Primeiramente, verificou-se a correlação entre sumariantes, podendo-se concluir que as mudanças introduzidas pelas “arestas sintáticas” são pequenas, dado que a correlação foi maior que 0.90 em todos os casos, como ilustra a Figura 7.

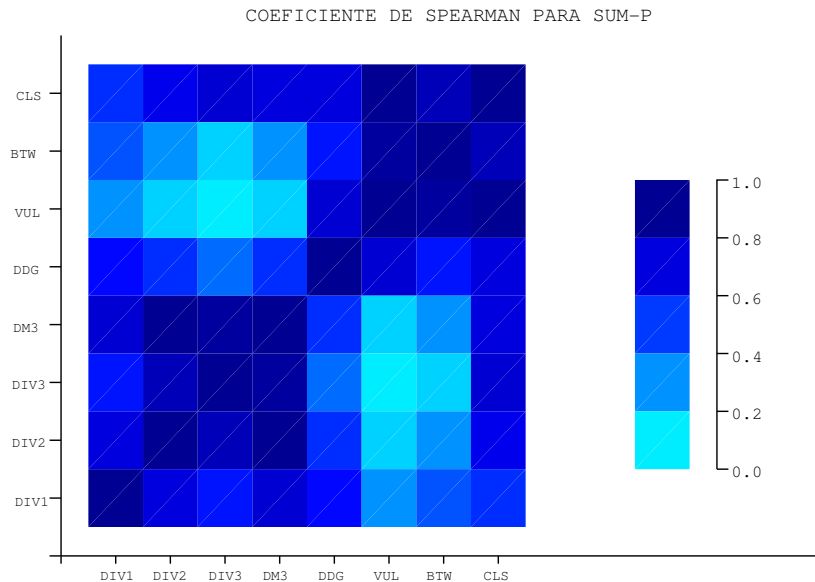


Figura 6. Coeficiente de correlação de Spearman para métodos referentes à modelagem de dependência sintática (SUM-P).

Por fim, verificou-se a significância estatística das diferenças encontradas entre os valores de ROUGE de um mesmo método usando modelagens distintas. As tabelas 5, 6 e 7 apresentam os valores da porcentagem dos métodos que apresentaram ganho de qualidade respectivamente para precisão, cobertura e medida-F, após a introdução da informação sintática. Tais valores de porcentagem são proporcionais à significância da diferença de qualidade. Em outras palavras, a porcentagem indica a probabilidade que uma amostra de sumários proveniente da modelagem SUM-P apresente um índice ROUGE maior que outra amostra de mesmo tamanho proveniente da modelagem SUM-RC. Analisando os resultados, percebe-se que a maioria das porcentagens não é maior que 70%, o que indica baixa significância. Mesmo assim, é interessante perceber uma significância razoável no método DDG (precisão e medida-F).

Método	Porcentagem
DDG	75.0 %
VUL	66.5 %
CLS	65.5 %
DIV3	64.0 %
DIV1	59.0 %

Tabela 5. Significância estatística do ganho encontrado para *precisão*.

9. Conclusão

Este trabalho pode ser dividido em dois resultados principais. No primeiro, constatou-se que as características topológicas através de identificação de bordas em redes complexas são capazes de identificar razoavelmente as palavras chaves de um texto, o que levou

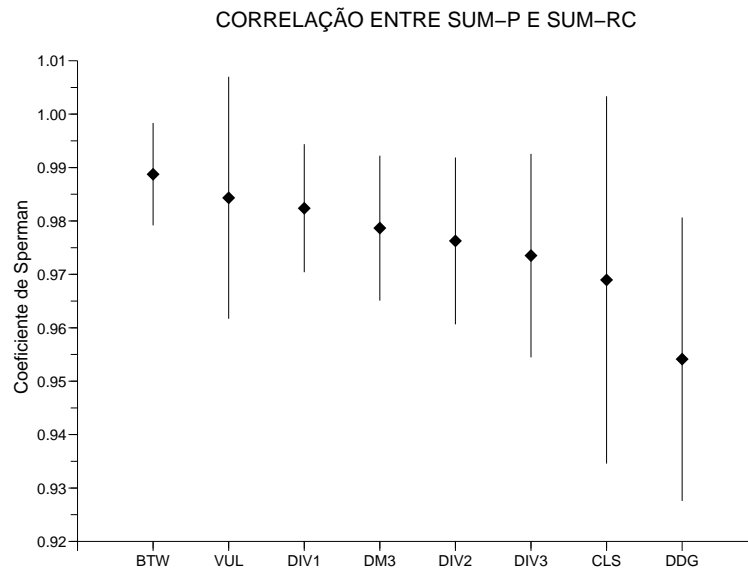


Figura 7. Média e desvio padrão para os valores de correlação de Spearman entre modelagens distintas. Os valores altos indicam que as sentenças escolhidas entre sumarizadores (de um mesmo método) provenientes de modelagens distintas são semelhantes.

Método	Porcentagem
DDG	58.5 %
DIV3	58.5 %
VUL	55.5 %
DIV2	54.5 %
DIV1	54.0 %

Tabela 6. Significância estatística do ganho encontrado para *cobertura*.

a bons índices de qualidade para sumarizadores baseados em redes complexas. Como segundo resultado principal, constatou-se que a adição de informação linguística mais profunda (análise sintática) fornece uma modelagem mais acurada que apenas o uso da modelagem tradicional baseada em redes complexas, assim como foi previsto em [25]. Apesar desta melhoria, a baixa significância estatística da diferença de qualidade sugere que novas avanços devem ser implementados para que esta seja notada naturalmente. Com isto, pretende-se implementar novas idéias para a modelagem dos textos, tal como a consideração de informação semântica na rede (através do uso de um *thesaurus*), reconhecimento de expressões multi-palavras e tratamento de anáforas.

Método	Porcentagem
DDG	75.0 %
VUL	65.0 %
CLS	65.0 %
DIV3	64.5 %
DIV1	60.0 %

Tabela 7. Significância estatística do ganho encontrado para *medida-F*.

Referências

- [1] *Costa, L. F.; Oliveira Jr., O. N.; Travieso, G.; Rodrigues, F. A.; Villas Boas, P. R.; Antiqueira, L.; Viana, M. P.; Rocha, L. E. C.* Analyzing and Modeling Real-World Phenomena with Complex Networks: A Survey of Applications. Physics and Society, 2008.
- [2] *Bates, M.* Models of natural language understanding. Proceedings of the National Academy of Sciences of the United States of America, Vol. 92, No. 22 (Oct. 24, 1995), pp. 9977-9982. 1995.
- [3] *Hutchins, W.J.; Somers, H.L.* An Introduction to Machine Translation. London: Academic Press. 1992.
- [4] *Martins, R.T.; Hasegawa, R.; Nunes, M.G.V.; Montilha, G.; Oliveira Jr., O.N.* Linguistic issues in the development of ReGra: a Grammar Checker for Brazilian Portuguese. Natural Language Engineering, Volume 4, p287-307; Cambridge University Press, 1998.
- [5] *Marcu, D.* The Theory and Practice of Discourse Parsing and Summarization, The MIT Press, A Bradford Book, 2000.
- [6] *Cancho, R.F.; Solé, R.V.* "The Small World of Human Language", Proceedings of The Royal Society of London. Series B, Biological Sciences, 268, 2261-2265, 2001.
- [7] *Sigman, M.; Cecchi, G.A.* "Global Organization of the Wordnet Lexicon", Proceedings of the National Academy of Sciences, 99, 1742-1747, 2002.
- [8] *Miller, G.A.* "Wordnet: a dictionary browser", Proceedings of the First International Conference on Information in Data. University of Waterloo, 1985.
- [9] *Motter, A.E.; Moura, A.P.S.; Lai, Y.C.; Dasgupta, P.* "Topology of the Conceptual Network of Language", Phys. Rev. E, 65, 065102, 2002.
- [10] *Spärck, K.J.* Automatic summarising: factors and directions. Advances in Automatic Text Summarization, MIT Press, pp. 1-12, 1999.
- [11] *Salton, G.; McGill, M. J.* Introduction to modern information retrieval. New York: McGraw-Hill, 1983.
- [12] *Lin, C.Y.; E.H., Hovy 2003.* Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In Proceedings of 2003 Language Technology Conference (HLT-NAACL 2003), Edmonton, Canada, May 27 - June 1, 2003.

- [13] *Lin, C. Y.* ROUGE: A package for automatic evaluation of summaries. In: Proceedings of the Workshop on Text Summarization Branches Out (WAS), Barcelona, Spain, 2004.
- [14] *Lin, C. Y.; Hovy, E.* Automatic evaluation of summaries using n-gram co-occurrence statistics. In: Proceedings of the 2003 Language Technology
- [15] *Bick, E.* The Parsing System Palavras - Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework , Arhus, 2000. Conference (HLT-NAACL-2003), Edmonton, Canada, 2003.
- [16] *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.
- [17] *Newman, M.E.J.* The structure and function of complex networks. SIAM Review, Vol. 45, pp. 167-256, 2003.
- [18] *Costa, L.F.* What's in a name? International Journal of Modern Physics C, Vol. 15, pp. 371-379, 2004.
- [19] *Dorogovtsev, S.N.; Mendes, J.F.F.* Evolution of networks. Advances in Physics, Vol. 51, N. 4, pp. 1079-1187,2002.
- [20] *Antiqueira, L.; Nunes, M.G.V.; Oliveira Jr., O.N.; Costa, L.F.* Modelando Textos como Redes Complexas. In Anais do III Workshop em Tecnologia da Informação e da Linguagem Humana, pp. 1-10, 2005.
- [21] *Antiqueira, L.; Pardo, T.A.S.; Nunes, M.G.V.; Oliveira Jr., O.N.* Some issues on complex networks for author characterization. Revista Iberoamericana de Inteligencia Artificial, V. 11, N. 36, pp. 51-58, 2007.
- [22] *Antiqueira, L.; Nunes, M.G.V.; Oliveira Jr.; O.N.; Costa, L.F.* Strong Correlations Between Text Quality and Complex Networks Features. Physica A - Statistical Mechanics and its Applications, Vol. 373, pp.811-820, 2007.
- [23] *Amancio, D.R.; Antiqueira, L.; Pardo, T.A.S.; Costa, L.F.; Oliveira Jr. O.N.; Nunes, M.G.V.* Complex networks analysis of manual and machine translations. International Journal of Modern Physics C - IJMPC, V. 19, N. 4, pp. 583-598, 2008.
- [24] *Ferrer i Cancho, R.; Solé, R.V.; Köhler, R.* Patterns in syntactic dependency networks. Physical Review E. 69 (5): pp. 1-8, 2004.
- [25] *Leite. D.; Rino ,L.; Pardo, T.A.S; Nunes, M.G.V.* Extractive Automatic Summarization: Does more linguistic knowledge make a difference? TextGraphs-2: Graph-Based Algorithms for Natural Language Processing, 2007.
- [26] *Barthélemy, M; Barrat, A.; Pastor-Satorras, R.; Vespignani, A.* Characterization and modeling of weighted networks. Physica A, 346:34-43, 2005.
- [27] *Sedgewick, R.* Algorithms in C, Part 5: Graph Algorithms, Terceira Edição, 1998.
- [28] *Antiqueira, L.; Oliveira Jr., O.N.; Costa, L.F.; Nunes, M.G.V* A Complex Network Approach to Text Summarization. Information Sciences, 179(5), 584-599, 2009.
- [29] *Spearman, C.* "The proof and measurement of association between two things". American Journal of Psychology, 15, pp 72-101, 1904.