
Desenvolvimento de técnicas baseadas
em redes complexas para sumarização
extrativa de textos

Lucas Antiqueira

Desenvolvimento de técnicas baseadas em redes complexas para sumarização extrativa de textos

Lucas Antiqueira

Orientadora: *Profa. Dra. Maria das Graças Volpe Nunes*

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP, como parte dos requisitos para obtenção do título de Mestre em Ciências - Ciências de Computação e Matemática Computacional.

“VERSÃO REVISADA APÓS A DEFESA”

Data da Defesa:	27 / 02 / 2007
Visto da Orientadora:	

USP – São Carlos
Março/2007

Agradecimentos

Aos meus pais, Gilberto e Valquíria, por moldarem minha personalidade, por sempre me aconselharem e por proporcionarem minha reabilitação em momentos difíceis. Procuro não decepcioná-los.

À minha namorada, Mariá, por ser tão companheira, compreensível e linda. Seu jeito de viver, com tanta garra e dedicação, me inspira. Valorizo sua companhia cada vez mais.

Ao meu irmão, Moisés, pelas incontáveis horas de conversa a respeito de música e futebol. Permanecemos amigos depois de adultos, o que me deixa muito feliz.

À minha orientadora, Graça Nunes, por me acompanhar desde os primeiros passos na iniciação científica e por confiar no meu modo de trabalhar.

A Luciano da F. Costa e a Osvaldo N. Oliveira Jr., pela boa vontade em me auxiliar em diversos pontos desta e de outras pesquisas.

A Thiago A. S. Pardo, pelas dicas e sugestões dadas a respeito de sumarização automática.

A Rada Mihalcea, a Daniel S. Leite, a Lucia H. M. Rino e a Carlos N. Silla Jr., pela prestatividade ao responder dúvidas quanto à avaliação automática de sumários.

A John Conroy, por fornecer o cópys com extratos manuais da DUC'2001.

Aos colegas do laboratório, pelas inúmeras vezes em que fui ajudado, não somente no mestrado, mas também durante esses cinco anos de NILC.

À USP, pela infraestrutura e pelo suporte técnico.

Ao CNPq e à FAPESP, pelo auxílio financeiro.

A Deus, pela vida.

Resumo

A Sumarização Automática de Textos tem considerável importância nas tarefas de localização e utilização de conteúdo relevante em meio à quantidade enorme de informação disponível atualmente em meio digital. Nessa área, procura-se desenvolver técnicas que possibilitem obter o conteúdo mais relevante de documentos, de maneira condensada, sem alterar seu significado original, e com mínima intervenção humana. O objetivo deste trabalho de mestrado foi investigar de que maneira conceitos desenvolvidos na área de Redes Complexas podem ser aplicados à Sumarização Automática de Textos, mais especificamente à sumarização extrativa. Embora grande parte das pesquisas em sumarização tenha se voltado para a utilização de técnicas extrativas, ainda é possível melhorar o nível de informatividade dos extratos gerados automaticamente. Neste trabalho, textos foram representados como redes, das quais foram extraídas medidas tradicionalmente utilizadas na caracterização de redes complexas (por exemplo, coeficiente de aglomeração, grau hierárquico e índice de localidade), com o intuito de fornecer subsídios à seleção das sentenças mais significativas de um texto. Essas redes são formadas pelas sentenças (representadas pelos vértices) de um determinado texto, juntamente com as repetições (representadas pelas arestas) de substantivos entre sentenças após lematização. Cada método de sumarização proposto foi aplicado no *cópus* TeMário, de textos jornalísticos em português, e em *cópus* das conferências DUC, de textos jornalísticos em inglês. A avaliação desse estudo foi feita por meio da realização de quatro experimentos, fazendo-se uso de métodos de avaliação automática (ROUGE-1 e Precisão/Cobertura de sentenças) e comparando-se os resultados com os de outros sistemas de sumarização extrativa. Os melhores sumarizadores propostos referem-se aos seguintes conceitos: *d*-anel, grau, *k*-núcleo e caminho mínimo. Foram obtidos resultados comparáveis aos dos melhores métodos de sumarização já propostos para o português, enquanto que, para o inglês, os resultados são menos expressivos.

Palavras-chave: Sumarização Automática, Redes Complexas, Processamento de Línguas Naturais, Inteligência Artificial.

Abstract

Automatic Text Summarization has considerably importance in tasks such as finding and using relevant content in the enormous amount of information available nowadays in digital media. The focus in this field is on the development of techniques that allow someone to obtain the most relevant content of documents, in a condensed way, preserving the original meaning and with little (or even none) human help. The purpose of this MSc project was to investigate a way of applying concepts borrowed from the studies of Complex Networks to the Automatic Text Summarization field, specifically to the task of extractive summarization. Although the majority of works in summarization have focused on extractive techniques, it is still possible to obtain better levels of informativity in extracts automatically generated. In this work, texts were represented as networks, from which the most significant sentences were selected through the use of ranking algorithms. Such networks are obtained from a text in the following manner: the sentences are represented as nodes, and an edge between two nodes is created if there is at least one repetition of a noun in both sentences, after the lemmatization step. Measurements typically employed in the characterization of complex networks, such as clustering coefficient, hierarchical degree and locality index, were used on the basis of the process of node (sentence) selection in order to build an extract. Each summarization technique proposed was applied to the TeMário corpus, which comprises newspaper articles in Portuguese, and to the DUC corpora, which comprises newspaper articles in English. Four evaluation experiments were carried out, by means of automatic evaluation measurements (ROUGE-1 and sentence Precision/Recall) and comparison with the results obtained by other extractive summarization systems. The best summarizers are the ones based on the following concepts: d -ring, degree, k -core and shortest path. Performances comparable to the best summarization systems for Portuguese were achieved, whilst the results are less significant for English.

Keywords: Automatic Summarization, Complex Networks, Natural Language Processing, Artificial Intelligence.

Índice

Lista de Figuras	ix
Lista de Tabelas	xi
1 Introdução	1
2 Sumarização Automática de Textos	5
2.1 Sumarização Extrativa	10
2.1.1 A Abordagem de Luhn	11
2.1.2 O Paradigma Edmundsoniano	11
2.1.3 O Uso de Frases Indicativas	13
2.1.4 A Flexibilidade do Aprendizado de Máquina	13
2.1.5 Identificando a Idéia Principal	16
2.1.6 Uma Extensão para a Métrica TF-IDF	17
2.1.7 Uma Abordagem Híbrida	18
2.2 Sumarização Extrativa com Redes	20
3 Redes Complexas	27
3.1 Redes Complexas e Língua Natural	30
4 Propostas de Geração de Extratos	33
4.1 Construção das Redes	33
4.2 Sumarizadores Propostos	36
4.2.1 Grau	38
4.2.2 Coeficiente de Aglomeração	39
4.2.3 Caminhos Mínimos	41
4.2.4 Índice de Localidade	42

4.2.5	Índice de Concordância	45
4.2.6	Grau Hierárquico	46
4.2.7	d -Anéis	48
4.2.8	k -Núcleos	50
4.2.9	w -Cortes	52
4.2.10	Comunidades	53
5	Avaliação	57
5.1	Técnicas de Avaliação Automática	58
5.2	Cópus Seleccionados	61
5.3	Definições dos Experimentos	64
5.4	Resultados Obtidos	66
5.4.1	TeMário com P , C e F	67
5.4.2	TeMário com ROUGE-1	72
5.4.3	DUC'2002 com ROUGE-1	75
5.4.4	DUC'2001 com P , C e F	79
5.5	Correlações entre Sumarizadores	83
5.6	Exemplos de Extratos Gerados	90
6	Conclusões	97
	Referências Bibliográficas	101

Lista de Figuras

2.1	Texto do caderno Opinião do jornal Folha de São Paulo, presente no corpus TeMário.	6
2.2	Resumo manual (presente no corpus TeMário) do texto da Figura 2.1. . . .	7
2.3	Extrato manual construído a partir da seleção de sentenças do texto da Figura 2.1.	8
2.4	Critérios para classificação de sistemas de sumarização automática. . . .	9
2.5	Diferentes tipos de avaliação de sistemas de sumarização automática. . . .	10
4.1	Sentenças extraídas do texto da Figura 2.1, que ilustram a construção de uma rede de sentenças.	36
4.2	Rede derivada do texto da Figura 4.1.	36
4.3	Redes obtidas a partir de dois textos do corpus TeMário.	37
4.4	Vértices 1 e 2 com graus $k_1 = 5$ e $k_2 = 2$	39
4.5	Vértices 1 e 2 com coeficientes de aglomeração $C_1 = 0,7$ e $C_2 = 0,2$	40
4.6	Vértices 1 e 2 com caminhos mínimos médios $sp_1 = 4,46$ e $sp_2 = 2,85$	41
4.7	Vértices 1 e 2 com índices de localidade $l_1 = 0,44$ e $l_2 = 0,73$	43
4.8	Arestas (1,2) e (3,4) com índices de concordância $\mu_{12} = 0$ e $\mu_{34} = 0,5$	45
4.9	Vértice 1 e suas duas primeiras hierarquias.	47
4.10	k -Núcleo com $k = 4$, identificado pelos vértices em cinza.	51
4.11	w -Corte com $w = 3$, identificado pelos vértices em cinza.	52
4.12	Exemplo de divisão de uma rede em três comunidades (áreas em cinza). . .	54
5.1	Distribuições do número de sentenças por texto-fonte nos corpus TeMário, DUC'2002 e DUC'2001.	63
5.2	Medida-F média (F) dos sumarizadores da Tabela 5.3.	69
5.3	Valores ROUGE-1 médios dos sumarizadores da Tabela 5.4.	75
5.4	Valores ROUGE-1 médios dos sumarizadores da Tabela 5.5.	77

5.5	Medida-F média (F) dos sumarizadores da Tabela 5.6.	81
5.6	Dois exemplos de correlações entre sumarizadores no corpus DUC'2002. . .	85
5.7	Exemplo de aplicação do algoritmo sp_i^{wc} em texto-fonte do corpus TeMário. . .	92
5.8	Resumo manual, retirado do corpus TeMário, construído para o texto-fonte da Figura 5.7.	93
5.9	Extrato para o texto-fonte da Figura 5.7, gerado por sp_i^{wc} , com tamanho similar (em número de palavras) ao do resumo manual da Figura 5.8. . . .	93
5.10	Exemplo de aplicação do algoritmo $r_i^{l,k}$ em texto-fonte do corpus DUC'2001. . .	94
5.11	Extrato manual, retirado do corpus DUC'2001, construído para o texto-fonte da Figura 5.10.	95
5.12	Extrato para o texto-fonte da Figura 5.10, gerado por $r_i^{l,k}$, com tamanho similar (em número de palavras) ao do extrato manual da Figura 5.11. . .	95

Lista de Tabelas

4.1	Lista de medidas utilizadas nos experimentos de sumarização.	56
5.1	Propriedades dos <i>corpus</i> utilizados nos experimentos de avaliação.	64
5.2	Métricas de avaliação aplicadas em cada <i>corpus</i>	65
5.3	Valores médios de Precisão (P), Cobertura (C) e Medida-F (F) para o <i>corpus</i> TeMário.	68
5.4	Valores médios da medida ROUGE-1 para o <i>corpus</i> TeMário.	74
5.5	Valores médios da medida ROUGE-1 para o <i>corpus</i> DUC'2002.	76
5.6	Valores médios de Precisão (P), Cobertura (C) e Medida-F (F) para o <i>corpus</i> DUC'2001.	80
5.7	Sistemas baseados em redes complexas que apresentaram os melhores desempenhos nos quatro experimentos realizados.	83
5.8	Coefficientes de correlação linear entre as medidas do Grupo-1 (<i>corpus</i> TeMário).	86
5.9	Coefficientes de correlação linear entre as medidas do Grupo-1 (<i>corpus</i> DUC'2002).	88
5.10	Coefficientes de correlação linear entre as medidas do Grupo-1 (<i>corpus</i> DUC'2001).	89

Introdução

Vivemos tempos em que a quantidade de informação disponível, já enorme, cresce vertiginosamente. Um estudo realizado em Berkeley indica que, em 2002, foram criados cinco milhões de terabytes de informação, ou seja, duas vezes mais dados do que foi gerado em 1999, o que resulta em uma taxa de crescimento de aproximadamente 30% ao ano (Lyman e Varian, 2003). É de se esperar, portanto, que áreas como Extração de Informação e Sumarização Automática tenham considerável importância nas tarefas de localização e utilização de conteúdo relevante em meio a essa avalanche de dados. Particularmente, a Sumarização Automática de Textos pode ser útil de várias maneiras. Os sumários podem ser empregados, por exemplo, para indexar documentos: ao invés de se utilizar o documento original, pode-se utilizar seu sumário, diminuindo a carga de trabalho tanto para o humano quanto para um indexador automático. Outro exemplo é o uso de sumários na exibição dos resultados de uma ferramenta de busca de documentos. Os resultados do Google certamente seriam muito mais úteis se, ao invés de trechos de texto incoerentes, fosse disponibilizado um pequeno sumário de cada documento selecionado, permitindo que a escolha do documento mais relevante possa ser realizada em menor tempo. Outra utilidade é a sumarização de artigos de jornais em versão digital, separados por tópico. Nesse caso, as informações mais importantes de vários artigos, todos sobre um mesmo assunto e possivelmente de diferentes jornais, são condensadas em um único sumário, evitando assim que todos os textos sejam consultados.

Na Sumarização Automática de Textos, procura-se desenvolver técnicas que possibilitem obter o conteúdo mais relevante de documentos, de maneira condensada, sem alterar

seu significado original, e com mínima intervenção humana. As técnicas empregadas em Sumarização Automática são usualmente divididas em dois grandes grupos: as que adotam uma abordagem superficial e as que utilizam uma abordagem profunda. Os sistemas superficiais tipicamente limitam-se a considerar apenas uma representação textual nos níveis morfológico e sintático e, geralmente, produzem sumários por meio da seleção e justaposição de sentenças do texto original (sumarização extrativa). Já na abordagem profunda, costuma-se construir ao menos uma representação semântica do documento, e geralmente envolve geração de língua natural, por meio de paráfrases, especializações, generalizações ou rearranjos das informações selecionadas. Embora a sumarização superficial geralmente produza sumários problemáticos (um exemplo de problema, nesse caso, de coesão, é a ausência de referentes anafóricos), ela é mais robusta e simples que a abordagem profunda. Conseqüentemente, a maior parte dos sistemas construídos até então adotam a abordagem superficial (Mani, 2001).

Existem diversas técnicas superficiais que costumam ser adotadas na sumarização extrativa (ou seja, na construção de extratos), os quais são formados pela seleção, cópia e reorganização dos segmentos (sentenças, por exemplo) mais importantes de um texto. Entre elas, encontram-se o método baseado na representatividade das palavras-chave (Luhn, 1958; Edmundson, 1969), o método baseado na localização das sentenças (Baxendale, 1958) e o método baseado na presença de frases indicativas (Paice, 1981). Técnicas de aprendizado de máquina são comumente utilizadas na sumarização extrativa (Kupiec et al., 1995). Outras abordagens também são empregadas na construção de extratos, como a determinação da idéia central de um texto (Pardo et al., 2003a) e a utilização de representações para textos baseadas em grafos (Skorochod'ko, 1971; Mihalcea, 2005).

Embora grande atenção seja dada pela comunidade de Processamento de Línguas Naturais (PLN) à sumarização extrativa, a construção de extratos ainda precisa ser aperfeiçoada quando o objetivo for gerar sumários coerentes e coesos. Tais limitações são aceitáveis em algumas aplicações nas quais os sumários não são utilizados diretamente por humanos, como por exemplo, na recuperação de informação. Esses problemas, portanto, não invalidam a utilização da abordagem superficial. Outro desafio da sumarização extrativa, que pode ser enfrentado com uma abordagem superficial, é o desequilíbrio no nível de informatividade dos sumários, fruto da redundância ou da falta de informações importantes.

Objetivou-se nesta pesquisa de mestrado investigar de que maneira conceitos da área de Redes Complexas (Albert e Barabási, 2002; Dorogovtsev e Mendes, 2002; Newman, 2003; Boccaletti et al., 2006) podem ser aplicados à Sumarização Automática de Textos, mais especificamente à sumarização extrativa. A hipótese aqui levantada é a de que, uma

vez que um texto seja modelado como uma rede¹, é possível reconhecer suas sentenças mais informativas, ou relevantes, para compor um sumário. O reconhecimento dessas sentenças seria possível por meio do uso de conceitos desenvolvidos e/ou utilizados na área de Redes Complexas. Foram propostas 26 versões de sumarizadores, baseadas em 10 desses conceitos: (i) grau, (ii) coeficiente de aglomeração, (iii) caminhos mínimos, (iv) índice de localidade, (v) índice de concordância, (vi) grau hierárquico, (vii) *d*-anéis, (viii) *k*-núcleos, (ix) *k*-cortes e (x) comunidades. A pesquisa sobre redes complexas aumentou consideravelmente nos últimos anos, depois que os conceitos de redes pequeno-mundo (*small-world*) (Watts e Strogatz, 1998) e redes livre de escala (*scale-free*) (Faloutsos et al., 1999; Barabási e Albert, 1999) foram introduzidos, dando novo impulso à área.

As redes utilizadas neste projeto foram construídas da seguinte maneira: para um dado texto, cada sentença representa um vértice (também chamado nó), e as arestas indicam repetição de substantivos entre sentenças, após aplicação do processo de lematização. A frequência de repetição de palavras entre duas sentenças dá origem ao peso da respectiva aresta. Dessa maneira, é codificado na rede um tipo de similaridade entre sentenças, dado pela co-ocorrência de substantivos. Cada texto é representado por uma rede, onde os conceitos de Redes Complexas são aplicados com o objetivo de construir um extrato composto por um subconjunto de sentenças do texto original. Embora tenham sido utilizadas ferramentas de PLN dependentes de língua para o pré-processamento dos textos antes de modelá-los como redes (como etiquetadores morfossintáticos e lematizadores), as técnicas estudadas para a construção de extratos não levam em consideração a língua, pois baseiam-se unicamente na estrutura da rede que representa um dado texto. A avaliação dos sumarizadores aqui propostos foi feita por meio da comparação com outros sistemas de sumarização extrativa e do uso de técnicas de avaliação automática de sumários. Os sistemas cujos resultados foram comparados com os resultados obtidos neste projeto são:

1. Os que participaram de uma avaliação comparativa de sumarizadores para a língua portuguesa (do Brasil) (Rino et al., 2004). O corpus utilizado nessa avaliação foi o TeMário (Pardo e Rino, 2003), formado por textos jornalísticos.
2. Os que participaram da DUC (Document Understanding Conference)² de 2002 (Over e Liggett, 2002), uma conferência de grande escala que tem a finalidade de avaliar sistemas de sumarização automática. O corpus utilizado nessa edição da conferência também é formado por textos jornalísticos, desta vez em língua inglesa.

¹Consideramos aqui rede e grafo sinônimos, embora o termo rede seja utilizado com maior frequência, seguindo a tendência das pesquisas em Redes Complexas. Cabe ressaltar que nem todo grafo é uma rede complexa, como veremos no Capítulo 3.

²<http://duc.nist.gov>

3. Outros sistemas cujos resultados foram comparados aos divulgados em 1 e 2, como os propostos por Mihalcea (2005) e Leite e Rino (2006a).

Um outro *córpus* de textos jornalísticos em inglês, criado para treinamento dos sistemas participantes da DUC'2001 (Over, 2001), também foi utilizado, mas sem haver comparação de resultados com os de outros sistemas, justamente por não existir divulgação de tais números. A utilização do referido *córpus* é interessante, pois ele apresenta extratos de referência *golden standard*, ou seja, criados manualmente. Os outros *córpus* utilizados neste projeto apresentam sumários criados manualmente, mas que não são do tipo extrativo.

A avaliação da informatividade dos sumários foi feita automaticamente, por meio do sistema ROUGE (Lin, 2004; Lin e Hovy, 2003) e das medidas de Precisão e Cobertura de sentenças (Salton e McGill, 1983), técnicas de avaliação comumente empregadas na área de Sumarização Automática. Os melhores métodos de sumarização propostos baseiam-se nos d -anéis, no grau, nos k -núcleos e nos caminhos mínimos. Particularmente, os resultados obtidos para os textos em português são próximos dos apresentados pelos sistemas SuPor-v2 (Leite e Rino, 2006a), SuPor (Rino e Módolo, 2004), ClassSumm (Larocca Neto et al., 2002), PageRank e HITS (Mihalcea, 2005). Entretanto, para os textos em inglês, raramente superou-se o Top-Baseline (o qual apenas seleciona as primeiras sentenças do texto-fonte), e diversos sistemas participantes da DUC'2002 apresentaram resultados sensivelmente superiores. As causas dessas diferenças de desempenho entre português e inglês ainda devem ser investigadas. Adicionalmente, realizou-se neste trabalho uma análise de correlação entre os melhores sumarizadores propostos, possibilitando a identificação de métodos muito semelhantes, que geram extratos parecidos (como os baseados no grau dos vértices, que consideram ou não os pesos das arestas), ou métodos complementares, que geram extratos bem diferentes (como os baseados nos caminhos mínimos, quando comparados aos demais métodos).

Esta dissertação está organizada da seguinte forma. No Capítulo 2, é dada uma introdução à área de Sumarização Automática de Textos, juntamente com uma revisão de alguns dos sistemas de sumarização extrativa já propostos. O Capítulo 3 contém uma breve introdução à área de Redes Complexas, de maneira a acompanhar a explicação dos conceitos utilizados neste projeto, os quais são intimamente relacionados a essa linha de pesquisa. No Capítulo 4, explica-se em detalhes como as redes para textos são construídas e como funcionam os sumarizadores propostos. Já no Capítulo 5, todos os resultados obtidos por avaliação automática são relatados e discutidos, tendo em vista a análise dos métodos propostos e sua comparação com outros sistemas de sumarização. Por fim, no Capítulo 6, são apresentadas as conclusões e algumas perspectivas relacionadas a uma possível continuação deste trabalho.

Sumarização Automática de Textos

Spärck Jones (1999) define um sumário como sendo fruto da redução de um texto-fonte por meio da seleção e/ou generalização de suas informações mais importantes. Na Sumarização Automática de Textos objetiva-se, como o próprio nome da área indica, construir sumários de maneira automatizada. Spärck Jones (1999) argumenta que o processo de sumarização envolve três estágios:

1. *Interpretação*: Criação de uma representação do texto-fonte por meio de sua interpretação.
2. *Transformação*: Passagem da representação do texto-fonte para uma representação do sumário.
3. *Geração*: Construção do sumário a partir de sua representação.

Segundo Mani (2001), um sumário pode ser chamado de *extract* (extrato) ou *abstract* (resumo). Um extrato é um sumário cujo material foi completamente copiado do texto-fonte, e pode ser formado, por exemplo, por um subconjunto de sentenças. Já um resumo envolve reescrita, e não se limita à simples cópia de trechos do texto original. Esse tipo de sumário pode conter paráfrases, rearranjos, generalizações ou especializações das informações contidas no texto-fonte, o que teoricamente possibilita um grau mais alto de compressão. Normalmente, os sumarizadores humanos produzem resumos, e não extratos.

A principal restrição da sumarização é a não transgressão do significado do texto

O LAMENTÁVEL COMPASSO DE ESPERA
ANTONIO ERMÍRIO DE MORAES

No passado, o Brasil parava antes do Natal e só recomeçava depois do Ano Novo. Mais tarde, a retomada passou para o Carnaval. Agora é após a Páscoa. Logo logo, vai ficar para depois da Copa do Mundo. Com um agravante: se vencermos, serão mais uns dez ou 15 dias para as celebrações e comentários; se perdermos, outros tantos para amargar a derrota e fazer as críticas.

E assim vai. Depois da Copa, virão as supereleições -em dois turnos-, o que, na prática, "mata" setembro, outubro e novembro. E aí, chega outra vez a hora de nos prepararmos para as festas de Natal e Ano Bom, pois ninguém é de ferro...

Para quem não gosta de trabalhar, este ano de 1994 é um prato cheio. Ele reúne, num só tempo, as melhores justificativas para adiar tudo para 1995 -e olhe lá...

A revisão constitucional está nesse ritmo. Raramente há quórum e, quando isso acontece, falta a vontade de votar. Bem diferente foi a conduta do deputado William Natcher, falecido na semana passada. Durante 40 anos de mandatos consecutivos, ele não faltou uma única vez às sessões do Congresso dos Estados Unidos.

O mais interessante é que o deputado Natcher conseguiu se reeleger, repetidas vezes, desde 1953, visitando muito pouco as suas "bases" -no Estado de Kentucky- e gastando a irrisória quantia de US\$ 10 mil por campanha. Com isso, ele provou que as tais bases gostam de ver os seus representantes trabalhando em benefício da coletividade lá no Congresso, não havendo a menor justificativa para faltarem ao seu trabalho. Uma vez presentes, ativos e atuantes, o reconhecimento é imediato. A reeleição é garantida. E com pouco dinheiro. É o triunfo dos que fazem sobre aqueles que falam.

A maioria dos nossos parlamentares está demonstrando não querer a revisão constitucional. Para eles, os problemas da pátria não merecem regime de urgência. Só os pessoais. Se quisessem tudo seria votado rapidamente -como o fizeram na aprovação do aumento de seus vencimentos. Os interesses pessoais falam mais alto do que a estabilização da moeda, a retomada do desenvolvimento, a criação de empregos etc.

Por essa razão, "enrolar" a revisão tem sido a palavra de ordem que os gigolôs de partidos vêm passando aos seus vassalos. É dessa forma que eles pretendem sabotar a resolução dos nossos problemas para fazer crescer a sua candidatura no meio do caos.

Tudo isso pode até ser lógico. Mas, os que assim agem, ignoram que o eventual fracasso do plano econômico jogará este país na mais pavorosa hiperinflação. Sem revisão, não haverá plano econômico -é verdade. Mas correremos o risco de não haver tampouco eleições e regime democrático. Tudo irá para o espaço. E quem responderá por mais essa irresponsabilidade?

Figura 2.1: Texto do caderno Opinião do jornal Folha de São Paulo, presente no corpus TeMário (Pardo e Rino, 2003).

original (Rino e Pardo, 2003). O campo lingüístico da análise do discurso explora vários aspectos do que faz um texto ser não apenas a simples justaposição de suas sentenças. Da extração seguida da junção das sentenças de um texto (a sentença é a unidade básica mais comum na sumarização extrativa), pode surgir o problema da perda de contexto, resultando freqüentemente em sumários incoerentes. Na Figura 2.1 está presente um texto cujos resumo e extrato, ambos construídos manualmente, são apresentados nas Figuras 2.2 e 2.3, respectivamente. Tomando como exemplo o extrato da Figura 2.3, a passagem do trecho 2 para o trecho 3 é feita de maneira problemática, pois ocorre uma mudança brusca na

Com o passar dos tempos, o Brasil vem inaugurando progressivamente formas de dar um “break” para descansar. Quando não são os feriados tradicionais, é a Copa, são as eleições.

Este ano de 1994 é propício para essa estagnação. Infelizmente, pois a revisão constitucional em pauta vem sendo protelada pelos deputados, acostumados a uma semana curta para visitas às bases. William Natcher, deputado norte-americano recém-falecido, depois de 40 anos de mandatos consecutivos sem nenhuma falta, poderia servir de lição.

No entanto, os deputados patricios preferem empurrar com a barriga a revisão, mais preocupados com os próprios interesses do que com os da pátria. A estabilidade da moeda, a retomada do desenvolvimento, a criação de empregos podem esperar. Que se dane um plano econômico para melhorar o país.

Figura 2.2: Resumo manual (presente no corpus TeMário) do texto da Figura 2.1.

progressão temática, passando de comentários a respeito de feriados e festividades para uma crítica sobre a falta de quórum em uma votação. Outro problema comum da sumarização dita extrativa é o aparecimento de anáforas sem o respectivo referente (um problema de coesão, exemplificado pela falta do referente “deputado William Natcher” da anáfora “ele”, em negrito, no trecho 4 da Figura 2.3). Note que o resumo da Figura 2.2 não apresenta problemas desse tipo.

No entanto, a construção automática de extratos é menos custosa que a construção de resumos, pois, para este último caso, são necessários recursos sofisticados, tais como complexos interpretadores para inferir o significado das sentenças ou ontologias para prover generalizações. Portanto, um sumarizador extrativo é mais facilmente portátil para diversas línguas, além de ser mais propício ao uso de algoritmos de aprendizado de máquina. Além disso, imitar o modo como a sumarização é feita por humanos para construir resumos de qualidade é uma tarefa por demais complexa. Estratégias de sumarização humana, quando existentes, raramente são racionalizadas e formalizadas. É justificável, portanto, a maior atenção que tem sido dada à produção automática de extratos (Luhn, 1958; Edmundson, 1969; Paice, 1981; Kupiec et al., 1995; Barzilay e Elhadad, 1999; Erkan e Radev, 2004).

Um sistema de sumarização pode empregar uma abordagem superficial (empírica) ou profunda (fundamental), de acordo com os níveis de conhecimento lingüístico (morfológico, sintático, semântico ou pragmático) contemplados em seu projeto (Mani, 2001). Os sistemas superficiais geralmente não ultrapassam o nível de representação sintática e, tipicamente, produzem extratos. Esse tipo de sistema pode até realizar uma análise das palavras no nível semântico, contudo, a análise sentencial geralmente não ultrapassa o nível sintático. A robustez é a principal vantagem da abordagem superficial. Já a abordagem profunda assume ao menos uma representação semântica no nível sentencial. Ela envolve

-
- 1) No passado, o Brasil parava antes do Natal e só recomeçava depois do Ano Novo. Mais tarde, a retomada passou para o Carnaval. Agora é após a Páscoa. Logo logo, vai ficar para depois da Copa do Mundo.
 - 2) E aí, chega outra vez a hora de nos prepararmos para as festas de Natal e Ano Bom, pois ninguém é de ferro...
 - 3) Raramente há quórum e, quando isso acontece, falta a vontade de votar.
 - 4) Com isso, **ele** provou que as tais bases gostam de ver os seus representantes trabalhando em benefício da coletividade lá no Congresso, não havendo a menor justificativa para faltarem ao seu trabalho.
 - 5) A maioria dos nossos parlamentares está demonstrando não querer a revisão constitucional. Para eles, os problemas da pátria não merecem regime de urgência. Só os pessoais.
 - 6) Mas correremos o risco de não haver tampouco eleições e regime democrático. Tudo irá para o espaço. E quem responderá por mais essa irresponsabilidade?
-

Figura 2.3: Extrato manual construído a partir da seleção de sentenças do texto da Figura 2.1. Os trechos estão numerados a fim de apoiar os comentários presentes no texto.

geração de língua natural, utilizando um nível de representação do discurso, o que permite a criação de resumos. Contudo, os sistemas que utilizam a abordagem profunda são restritos a domínios particulares, pois a construção de bases de conhecimento de propósito geral para análise e síntese semântica é altamente complexa (Martins e Rino, 2002; Pardo e Rino, 2002).

Outros dois critérios para classificação de sumários merecem destaque. De acordo com sua função, um sumário pode ser classificado como indicativo, se fornecer apenas uma referência para uma leitura mais profunda do documento original, ou como informativo, se contemplar toda a informação saliente do texto-fonte, guardadas as restrições de compressão e de nível de detalhamento do sumário. Já um sumário é dito genérico se não for direcionado às necessidades de um tipo específico de usuário. Esses critérios para classificação de sistemas de sumarização automática estão esquematizados na Figura 2.4.

A avaliação é uma fase de grande importância em uma disciplina prática tal como a Sumarização Automática. Tão importante quanto construir experimentos é avaliar os seus resultados e, além de ser parte integrante do método científico, a avaliação é ainda decisiva na confirmação ou refutação de uma teoria ou método. Não são poucos, tampouco triviais, os desafios enfrentados na avaliação de sumarizadores automáticos, pois eles provêm da subjetividade humana normalmente envolvida na sumarização. Entre eles, destacam-se (Mani, 2001):

- A complexidade em se definir a noção do que seja um sumário correto, pelo fato de envolver comunicação em língua natural. Sempre existe a possibilidade de um

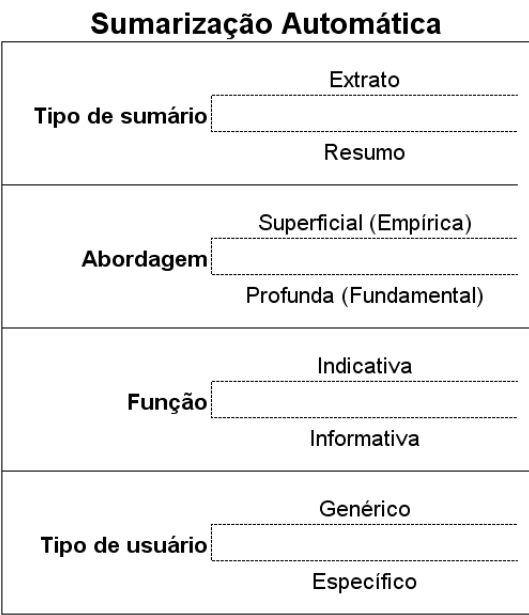


Figura 2.4: Critérios para classificação de sistemas de sumarização automática.

sistema gerar um bom sumário que é diferente de qualquer outro sumário tomado como referência produzido por um humano. Além disso, os humanos costumam não concordar muito bem entre si quanto ao que seja um bom sumário.

- Frequentemente, é necessário utilizar trabalho manual para julgar o resultado dos sumarizadores, o que encarece a avaliação.
- Como a sumarização envolve compressão, é importante avaliar sumários em diferentes taxas de compressão. Isso implica que os sumários de referência criados por humanos também tenham que se adequar a essas taxas, aumentando assim a complexidade da avaliação.
- Devem ser levadas em consideração as necessidades do usuário e da aplicação do sistema de sumarização, o que implica mais restrições na avaliação.

A avaliação pode ser classificada como intrínseca ou extrínseca. Na intrínseca, o sistema é avaliado de acordo com a qualidade dos sumários automáticos. Na extrínseca, é mensurado o quanto o sumarizador automático é útil para alguma outra tarefa que o utiliza. Se o sistema é avaliado observando-se apenas sua entrada e sua saída, a avaliação é dita *black-box*. A avaliação será do tipo *glass-box* se concentrar-se também nos módulos internos do sumarizador, e não apenas no seu funcionamento global. Algumas avaliações podem ainda ser conduzidas utilizando-se um examinador automático, caracterizando-as assim como avaliações *off-line*. As avaliações *on-line*, por sua vez, requerem o auxílio

de pessoas para testar o sistema. Se os resultados são comparados com os resultados de um outro sistema, a avaliação é comparativa (caso contrário, é dita autônoma). Na Figura 2.5 encontram-se esquematizados esses quatro critérios para avaliação de sistemas de sumarização automática.

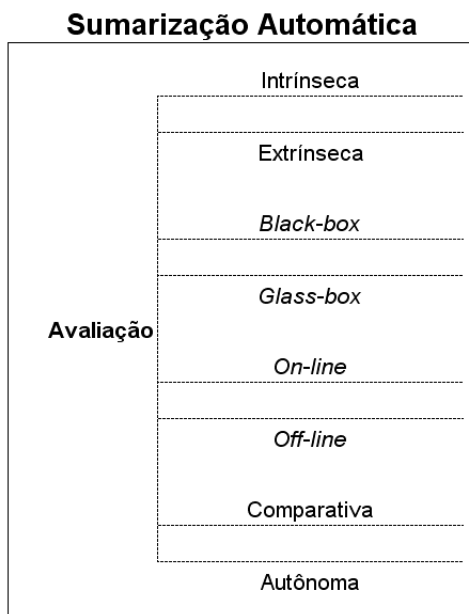


Figura 2.5: Diferentes tipos de avaliação de sistemas de sumarização automática.

Recentemente, algumas métricas de avaliação automáticas, isto é, que não fazem uso de juízes humanos, foram desenvolvidas. Exemplos são as métricas contidas no pacote ROUGE (Lin, 2004; Lin e Hovy, 2003) de avaliação automática de sumários, utilizado neste trabalho de mestrado (detalhes na Seção 5.1). Já a avaliação comparativa de sistemas de sumarização costuma ser foco de grandes conferências internacionais, como a DUC (Over, 2001; Over e Liggett, 2002).

Nas seções seguintes são apresentadas algumas das pesquisas que se relacionam a este projeto, começando pelos trabalhos de sumarização extrativa (Seção 2.1) e finalizando em uma revisão dos sistemas que utilizam o conceito de rede (ou grafo) na geração de sumários (Seção 2.2).

2.1 Sumarização Extrativa

Os métodos empregados na sumarização extrativa adotam o seguinte procedimento: (i) identificar os segmentos relevantes do texto, (ii) extrair do texto-fonte as unidades mínimas (orações, sentenças ou parágrafos) que contêm esses segmentos e (iii) justapor essas

unidades para produzir o sumário final (Rino e Nunes, 2005). Os trabalhos em sumarização extrativa costumam diferenciar-se uns dos outros ao proporem diferentes técnicas para resolver a etapa (i) do procedimento de construção de um extrato. A seguir, encontra-se um resumo de algumas das pesquisas já realizadas nessa área. A fim de ilustrar diferentes técnicas propostas ao longo de décadas de estudos em sumarização extrativa, procurou-se aqui relatar alguns dos trabalhos pioneiros e alguns dos trabalhos mais recentes, tanto para a sumarização de textos em inglês quanto em português.

2.1.1 *A Abordagem de Luhn*

Luhn (1958) deriva o que chama de fator de significância de uma sentença por meio da análise da frequência das palavras que a constituem. Como um escritor usualmente repete certas palavras conforme desenvolve seu raciocínio ao elaborar um texto, Luhn determina que a frequência de uso de cada palavra pode ser útil no cálculo do fator de significância das sentenças. Para tanto, uma lista de palavras em ordem decrescente de frequência deve ser compilada para cada texto. São então estabelecidos limites superior e inferior de frequência para essa lista, de modo que palavras muito ou pouco frequentes não sejam consideradas como pertencentes à lista de palavras-chave (também chamadas de palavras significantes), pois são palavras que adicionam ruído ao sistema. Quanto mais próximas essas palavras-chave estiverem umas das outras, com mais ênfase um determinado tópico do texto é tratado. A pontuação dada a cada sentença leva em consideração o número de palavras-chave em uma sentença e a distância entre elas devido à presença de outras palavras. Para cada sentença, grupos de no máximo quatro palavras não significantes delimitados por duas palavras-chave são selecionados. É importante lembrar que é possível haver sobreposição de palavras entre os grupos. Para cada grupo é calculado um fator de significância, dado pelo quadrado do número de palavras-chave dividido pelo total de palavras presentes no grupo, sempre incluindo as palavras delimitadoras do grupo. O fator de significância de uma sentença, utilizado como critério para formar o extrato, é igual ao maior fator de significância obtido entre seus grupos. Luhn relata que obteve resultados encorajadores em um experimento com 50 artigos, mas não fornece maiores detalhes a respeito. Entretanto, o método proposto por Luhn é precursor, e influenciou outras pesquisas subseqüentes (Edmundson, 1969; Pardo et al., 2003a).

2.1.2 *O Paradigma Edmundsoniano*

O modelo definido por Edmundson (1969)¹ é uma extensão da abordagem de Luhn. Na pontuação das sentenças nesse modelo, para posterior geração de extratos indicativos em

¹Apud (Mani, 2001).

inglês, foram considerados os seguintes atributos (*features*): (i) palavras indicativas (*cue words*), (ii) palavras-chave, (iii) palavras de título e (iv) localização da sentença. As palavras indicativas foram obtidas de um subconjunto do corpus utilizado (artigos de química), e consistiam de substantivos superlativos, advérbios de conclusão e termos de causalidade, entre outros (tais como “significant” e “impossible”). O dicionário de palavras indicativas foi dividido em três subdicionários: o de Bonus Words (que aumenta a pontuação da sentença), o de Stigma Words (que diminui a pontuação da sentença) e o de Null Words (com palavras irrelevantes). O atributo de palavras-chave baseia-se no princípio proposto por Luhn (1958), de que palavras de alta frequência (aqui apenas as palavras de conteúdo - *content words* - são consideradas) são importantes na indicação do conteúdo principal de um texto. Foi selecionado um determinado número de palavras mais frequentes, excluindo-se as palavras indicativas. As palavras de título foram obtidas do título, dos subtítulos e dos cabeçalhos, assumindo-se que os títulos são informativos e excluindo-se desse conjunto as Null Words. Cada palavra-chave ou palavra de título presente em uma sentença aumenta a chance de que ela seja selecionada para formar o extrato. O atributo de localização aumenta a pontuação de uma sentença que esteja no primeiro ou último parágrafo do texto, ou ainda seja a primeira ou a última sentença de qualquer outro parágrafo². Além disso, para esse último atributo, sentenças que contenham palavras que costumam aparecer em cabeçalhos (tais como “introduction” e “conclusions”) recebem um aumento em sua pontuação.

O método de pontuação das sentenças para extração foi baseado em uma combinação linear, denotada por $W(s)$, dos quatro atributos citados ($C(s)$ = palavras indicativas na sentença s , $K(s)$ = palavras-chave em s , $T(s)$ = palavras de título em s e $L(s)$ = localização de s),

$$W(s) = \alpha C(s) + \beta K(s) + \gamma T(s) + \delta L(s). \quad (2.1)$$

Os atributos α , β , γ e δ foram ajustados manualmente por meio de comparações com extratos gerados por humanos. Em suas avaliações, Edmundson percebeu que as palavras-chave não eram tão boas quanto os outros três atributos na seleção de sentenças, enquanto que a localização da sentença era o melhor atributo entre os quatro. A melhor combinação dos parâmetros era formada por palavras indicativas, palavras de título e localização da sentença.

Embora o trabalho de Edmundson seja importante, influenciando os estudos em su-

²Baxendale (1958) também propôs a seleção de sentenças para a formação de extratos de acordo com a sua posição no texto. Em um experimento com 200 parágrafos de textos científicos, Baxendale notou que em 85% dos casos a sentença mais importante do parágrafo era a primeira, enquanto que em 7% a sentença mais relevante era a última. Ao selecionar a primeira e a última sentença de cada parágrafo, tem-se que, para 92% dos parágrafos, a principal sentença é escolhida.

marização extrativa por anos a fio, a Equação 2.1 apresenta alguns problemas, como a não contemplação da taxa de compressão, e o uso exclusivo de características superficiais das sentenças. Além disso, os resultados obtidos por Edmundson são válidos somente para o *córpus* de textos científicos utilizado em seu estudo, sendo que a importância dos atributos escolhidos para calcular o peso de cada sentença pode variar com o *córpus* utilizado. Conseqüentemente, algoritmos de aprendizado de máquina passaram a ser utilizados na sumarização extrativa (Kupiec et al., 1995). Para tanto, um *córpus* de textos com seus respectivos extratos é necessário para treinar o sumarizador.

2.1.3 O Uso de Frases Indicativas

Passagens importantes de um texto podem ser identificadas por certas estruturas comuns das quais um escritor lança mão ao redigir seu texto. Paice (1981) chamou essas estruturas de *indicators* (frases indicativas), e propôs que elas fossem utilizadas em um gerador de extratos indicativos para a língua inglesa. Exemplos dessas frases, para textos científicos, são “The principal aim of this paper is to investigate. . .” e “In the present paper, a method is described for. . .”. Paice argumenta que relatórios e artigos técnicos podem não conter frase indicativa alguma, o que forçaria o uso de outros atributos como critério para seleção de sentenças. Foi proposta uma divisão do conjunto de frases indicativas em grupos, sendo que cada qual teria um peso associado diferente a ser utilizado na pontuação das sentenças. Note que o método proposto por Paice tem estreita relação com as palavras indicativas de Edmundson (1969). Um teste desse método na geração manual de extratos de um grupo de artigos científicos mostrou que um trabalho de refinamento do algoritmo ainda deveria ser realizado, principalmente devido às complexas regras utilizadas para selecionar outras sentenças além daquelas que contivessem frases indicativas. Além disso, a construção de uma tabela de frases indicativas torna o sistema altamente dependente da língua e do domínio dos textos a serem sumarizados, o que dificulta a portabilidade do sistema.

2.1.4 A Flexibilidade do Aprendizado de Máquina

Kupiec et al. (1995) basearam-se grandemente na pesquisa de Edmundson (1969), pois utilizaram um conjunto de atributos para pontuar as sentenças, e não somente um atributo, como o fez Luhn (1958). Entretanto, Kupiec et al. não empregaram o modelo de combinação linear de atributos da Equação 2.1, e sim transformaram o problema de ajuste manual dos pesos dos atributos em um problema de aprendizado de máquina (ou de classificação). O classificador utilizado é o Naive Bayes (Mitchell, 1997), que assume independência de probabilidade entre os atributos, e constrói uma função de classificação que estima a pro-

bilidade de uma dada sentença do texto-fonte pertencer ao extrato. Dado um conjunto de atributos e um *cópus* de documentos de treinamento, com os respectivos extratos, o algoritmo escolhe uma combinação de atributos de modo que um bom esquema de pontuação seja produzido. Foi proposto um conjunto de cinco atributos discretos:

- *Comprimento da sentença*: dado um limite de palavras, o atributo é verdadeiro para sentenças com um número de palavras acima desse limite, e é falso, caso contrário.
- *Frases fixas*: similar às frases indicativas de Paice (1981), utiliza frases pré-definidas em sua maioria de no máximo duas palavras de comprimento (por exemplo, “This letter...” e “In conclusion...”). Esse atributo é verdadeiro quando uma sentença contém uma frase fixa ou ainda quando é a primeira sentença de determinadas seções (“Results” e “Conclusions”, por exemplo).
- *Localização da sentença*: para sentenças presentes nos dez primeiros ou nos cinco últimos parágrafos, indica se ela é a primeira sentença do parágrafo, a última, ou se está entre essas duas.
- *Palavras temáticas*: palavras de conteúdo de mais alta frequência são utilizadas para pontuar as sentenças. Esse atributo é similar ao de palavras-chave, e indica se a sentença está entre as sentenças mais bem pontuadas de acordo com a frequência de suas palavras de conteúdo.
- *Nomes próprios*: é similar ao atributo anterior, mas somente considera palavras cuja primeira letra seja maiúscula e não esteja no início de uma sentença. A intenção é capturar nomes próprios e definições para acrônimos.

Em seus experimentos, Kupiec et al. utilizaram um *cópus* de 188 artigos técnicos/científicos, juntamente com os respectivos resumos (não extratos, em sua maioria indicativos) feitos à mão. Como os resumos manuais não necessariamente faziam uso literal das sentenças dos documentos originais, foi preciso realizar um emparelhamento entre suas sentenças. Em geral, a performance foi de 42% de Cobertura (*Recall*) com relação às sentenças presentes nos resumos manuais. A Precisão não foi calculada e foi utilizada uma estratégia *cross-validation*³. Assim como no trabalho de Edmundson, Kupiec et al. obtiveram uma melhor performance para o atributo de localização das sentenças. A melhor combinação de atributos foi: localização, palavras temáticas e comprimento.

Algoritmos de aprendizado de máquina continuaram a ser aplicados em sumarização extrativa. Larocca Neto et al. (2002) desenvolveram uma abordagem muito parecida com

³Mais detalhes a respeito das métricas Precisão e Cobertura na Seção 5.1.

a de Kupiec et al. Seu sistema, chamado ClassSumm, além de empregar o classificador Naive Bayes, também pode utilizar o algoritmo de árvores de decisão C4.5 (Quinlan, 1993) para determinar os segmentos mais relevantes de um texto. Ele associa 13 atributos a cada sentença, entre eles o comprimento da sentença, sua posição no documento, a ocorrência de nomes próprios ou de anáforas e a semelhança com o título, dada pela similaridade entre vetores de palavras. Foram realizados dois tipos de experimentos com 200 textos em língua inglesa de revistas técnicas: no primeiro, foram considerados extratos produzidos automaticamente para as fases de treino e de teste dos dois algoritmos e, no segundo, foram considerados extratos produzidos automaticamente para a fase de treino e extratos produzidos manualmente para a fase de teste. Os extratos automáticos foram obtidos a partir de resumos fornecidos pelos próprios autores dos textos, de maneira não explicitada por Larocca Neto et al. (2002). Os extratos manuais foram feitos por pessoa especialista contratada especialmente para tanto. Embora Larocca Neto et al. não tenham realizado um experimento somente com extratos manuais nas fases de treino e de teste (pois são supostamente melhores que os automáticos), nos dois experimentos reportados o algoritmo Naive Bayes foi superior ao algoritmo C4.5. Além disso, ambos tiveram melhor desempenho que o método considerado *baseline*, o qual seleciona as primeiras sentenças do documento a ser sumarizado.

Outro sistema, o NeuralSumm (NEURAL network for SUMMarization) (Pardo et al., 2003b) utiliza uma rede neural do tipo SOM (Self-Organizing Map) (Kohonen, 1990) para classificar as sentenças do texto a ser sumarizado, com base em um conjunto de atributos pré-selecionado. A rede neural do tipo SOM organiza as informações aprendidas na fase de treino em grupos de similaridade, e as sentenças do texto-fonte são classificadas de acordo com esses grupos da rede. Uma sentença pode receber uma das seguintes classificações no NeuralSumm: essencial, complementar ou supérflua. As sentenças essenciais devem estar no extrato, as supérfluas não. Já as complementares podem ou não fazer parte do sumário. Foi utilizado um conjunto de oito atributos, entre eles a posição da sentença, a presença de palavras-chave e a presença de palavras indicativas (tais como “avaliação”, “objetivo” e “solução”). A rede foi treinada com um corpus de dez textos científicos em português, anotado por juízes humanos de acordo com as três classificações possíveis para cada sentença. Em sua avaliação, baseada em comparações com o corpus anotado manualmente, o NeuralSumm apresentou desempenho superior aos algoritmos Naive Bayes e C4.5, também treinados com o mesmo corpus de textos em língua portuguesa. Outra avaliação foi realizada, comparando-se os extratos gerados automaticamente com extratos de referência. Nesse caso, as medidas de Precisão e de Cobertura do NeuralSumm mostraram-se relativamente próximas dos resultados obtidos em outras pesquisas. Como em qualquer aplicação de técnicas de aprendizado de máquina, o desempenho do algoritmo

utilizado é altamente dependente dos corpúscos de treino e de teste e do conjunto de atributos escolhidos para representar as instâncias do problema. Além disso, o NeuralSumm é ainda bastante influenciado pela arquitetura de sua rede neural (número de neurônios e precisão de treinamento).

São inúmeras as técnicas de aprendizado de máquina aplicadas em sumarização. Na DUC realizada em 2002, três dos quatro melhores sistemas que participaram da tarefa de sumarização de textos jornalísticos empregaram algoritmos de aprendizado de máquina. A classificação referente ao desempenho desses sistemas baseia-se na avaliação feita por Mihalcea (2005), utilizando a métrica ROUGE-1 (detalhes dessa métrica são fornecidos na Seção 5.1). Um desses sistemas, identificado por ntt.duc02, utiliza o algoritmo Support Vector Machines (SVM) (Vapnik, 2000) treinado com atributos superficiais, tais como localização e comprimento das sentenças (Hirao et al., 2002). Esse sistema obteve a melhor classificação entre os sistemas participantes da conferência. O terceiro melhor sistema, chamado ccsnsa.v2, une as técnicas de aprendizado Hidden Markov Model (HMM) (Rabiner, 1989) e Logistic Regression Model (LRM) (Hosmer e Lemeshow, 2000), e também utiliza atributos superficiais das sentenças (Schlesinger et al., 2002). Já o algoritmo de aprendizado Weighted Probability Distribution Voting (WPDV) (van Halteren, 2000) foi utilizado no sistema wpdv-xtr.v1, quarto colocado na avaliação segundo a métrica ROUGE-1 (van Halteren, 2002). Os atributos utilizados nesse sistema, assim como nos dois outros sistemas citados neste parágrafo, são superficiais.

2.1.5 Identificando a Idéia Principal

O GistSumm (GIST SUMMarizer) (Pardo et al., 2003a) é um sumarizador que determina a idéia central (*gist*) do texto-fonte utilizando técnicas estatísticas. Faz uso do método das palavras-chave ou da métrica TF-ISF (vide detalhes dessa métrica a seguir), a critério do usuário, para escolher a sentença mais bem pontuada (*gist sentence*). O GistSumm então seleciona as sentenças com as maiores pontuações para compor o extrato, com a restrição de que possuam ao menos uma palavra em comum com a *gist sentence*. Essa proposta foi avaliada com relação à escolha da *gist sentence* e à produção do extrato. No primeiro caso, foi utilizado um corpúscos de dez textos científicos em língua portuguesa, para o qual a identificação da *gist sentence* baseada em palavras-chave apresentou desempenho superior à baseada na métrica TF-ISF. Na segunda avaliação, 20 textos jornalísticos em inglês foram selecionados e, novamente, a utilização das palavras-chave teve melhor desempenho quando comparada à utilização da medida TF-ISF. Uma limitação do GistSumm é a correspondência da idéia principal a somente uma sentença. Contudo, a proposta é inovadora, pois procura garantir maior coerência aos extratos por meio da identificação da

gist sentence e posterior seleção de sentenças relacionadas a ela.

2.1.6 Uma Extensão para a Métrica TF-IDF

O TF-ISF-Summ (TF-ISF-based SUMMarizer) é um sumarizador automático que utiliza a métrica TF-ISF para selecionar as sentenças de um texto (Larocca Neto et al., 2000b). A métrica TF-ISF não é comprovadamente eficaz para a sumarização, embora a utilidade da medida na qual ela se baseia, a TF-IDF, seja bem fundamentada na área de Recuperação de Informação (Salton e McGill, 1983). TF-IDF significa Term Frequency-Inverse Document Frequency e, em uma coleção de documentos e em sua forma mais simples, é calculada para cada palavra tomando-se sua frequência de ocorrência em um dado documento e dividindo-a pelo número de documentos em que ocorre. É uma medida de frequência normalizada, que procura dar menos ênfase a termos muito freqüentes que não ajudam a discriminar os documentos entre si. Se a noção de documento for substituída pela de sentença, a métrica TF-IDF passa a se chamar TF-ISF, e seu valor dá a importância de uma palavra com relação a um único documento, e não a uma coleção de documentos. Sendo assim, ao introduzir essa idéia, Larocca Neto et al. determinam que cada sentença tem uma pontuação associada dada pela média aritmética do valor TF-ISF de todas as suas palavras. Esse valor é, portanto, considerado como critério para selecionar as sentenças que devem formar um extrato. A avaliação dessa técnica foi realizada comparando-se manualmente extratos produzidos para textos em inglês (quantidade não fornecida) pelo TF-ISF-Summ e pelo CGI/CMU (sistema que obteve os melhores resultados na tarefa *ad hoc* de sumarização⁴ na conferência SUMMAC⁵). A conclusão foi a de que ambos os sistemas produzem extratos de qualidade similar.

Já em (Larocca Neto et al., 2000a), a métrica TF-ISF é utilizada em conjunto com uma versão modificada do algoritmo TextTiling (Hearst, 1997), o qual procura segmentar textos em trechos coerentes formados por grupos de sentenças (tópicos). Os autores computam a importância relativa de cada tópico, que é dada pela soma das médias dos valores TF-ISF de cada sentença presente no tópico. O número de sentenças selecionadas de cada tópico para formar o extrato é diretamente proporcional à sua importância relativa. As sentenças escolhidas em cada tópico são as que apresentam maior similaridade com o respectivo centróide, definido como sendo o vetor das palavras presentes no tópico com valores TF-ISF médios, considerando-se todas as sentenças do tópico. A similaridade é dada pelo cosseno do ângulo entre o vetor de uma sentença e o vetor centróide. Essa

⁴Objetiva-se nessa tarefa determinar se a relevância de um texto, com relação a um determinado tópico, pode ser avaliada apenas pela leitura de seu extrato do tipo indicativo.

⁵SUMMAC (TIPSTER Text Summarization Evaluation Conference, http://www.itl.nist.gov/iaui/894.02/related_projects/tipster_summac/).

abordagem foi comparada com os extratos produzidos para sete textos em inglês pelos sistemas CGI/CMU e Cornell da conferência SUMMAC (os dois apresentaram os melhores resultados na tarefa *ad hoc*). Um juiz humano classificou os três grupos de extratos produzidos automaticamente, em ordem decrescente de qualidade, considerando se o extrato captura as idéias principais do texto-fonte e se é inteligível para alguém que não tem acesso ao texto-fonte. O método proposto obteve desempenho similar ao CGI/CMU, enquanto o sistema Cornell apresentou resultados inferiores aos dos dois outros sistemas.

2.1.7 Uma Abordagem Híbrida

O SuPor (SUMmarization of texts in PORtuguese) é um ambiente que permite que seja escolhida a combinação de atributos que se deseja utilizar no sumarizador (Módolo, 2003; Rino e Módolo, 2004). Ele utiliza quatro técnicas anteriormente propostas, nesse caso orientadas para a língua portuguesa:

1. *Classificador*: segue a proposta de Kupiec et al. (1995) já apresentada nesta seção, ou seja, utiliza um classificador Naive Bayes para treinar o sistema de acordo com um conjunto de atributos pré-selecionado (frequência das palavras, localização e comprimento da sentença e ocorrência de nomes próprios).
2. *Cadeias lexicais*: calcula a coesão lexical por meio de redes de relações entre substantivos (Barzilay e Elhadad, 1999). Das cadeias mais proeminentes, usa três heurísticas para selecionar sentenças (mais detalhes na Seção 2.2).
3. *Mapa de relações*: utiliza uma rede de conexões entre parágrafos para selecionar os nós mais relevantes, de acordo com três diferentes maneiras de se percorrer os nós da rede (Salton et al., 1997). Mais detalhes são dados na Seção 2.2.
4. *Importância de tópicos*: aplica a técnica desenvolvida por Larocca Neto et al. (2000a), resumida nesta seção. A idéia é dividir o texto em tópicos, e então selecionar as sentenças mais relevantes de cada tópico (por meio da métrica TF-ISF) em número proporcional à sua respectiva importância.

Em uma avaliação do tipo *cross-validation*, utilizando 51 textos jornalísticos e respectivos extratos/resumos de referência, Rino e Módolo procuraram avaliar, por meio de medidas de Precisão e Cobertura, quais configurações do sistema SuPor conduzem a melhores resultados. Os grupos de atributos que se destacaram foram: (i) cadeias lexicais, comprimento da sentença e ocorrência de nomes próprios; (ii) cadeias lexicais, comprimento da sentença e frequência das palavras; e (iii) cadeias lexicais e mapa de relações.

Em outro experimento, relatado também por Rino et al. (2004), apenas uma configuração do SuPor foi utilizada, com cinco atributos (localização da sentença, frequência das palavras, comprimento da sentença, ocorrência de nomes próprios e cadeias lexicais), a fim de compará-lo com outros sistemas de sumarização. O atributo de localização foi adicionado ao grupo pois é um atributo bastante utilizado, embora não figure entre as configurações com melhor desempenho. Os outros sistemas avaliados foram: TF-ISF-Summ (Larocca Neto et al., 2000b), NeuralSumm (Pardo et al., 2003b), GistSumm (Pardo et al., 2003a) e ClassSumm (Larocca Neto et al., 2002), já apresentados nesta seção. Todos os sistemas foram avaliados para a língua portuguesa, entretanto, o GistSumm e o TF-ISF-Summ empregam métricas totalmente independentes de língua. Já o NeuralSumm, o SuPor e o ClassSumm requerem que um corpus de treinamento para a língua em questão seja fornecido. Adicionalmente, dois métodos *baseline* foram também aplicados ao corpus: o que seleciona as primeiras sentenças do texto-fonte (Top-Baseline), e o que as seleciona aleatoriamente (Random-Baseline). Foram utilizados 100 textos jornalísticos do corpus TeMário (Pardo e Rino, 2003), juntamente com extratos de referência. Aplicando-se a estratégia *cross-validation*, e calculando-se as métricas de Precisão, Cobertura e Medida-F (que associa Precisão e Cobertura), os sistemas SuPor e ClassSumm apresentam os melhores resultados. A classificação dos sistemas de acordo com a Medida-F foi a seguinte, em ordem decrescente: SuPor, ClassSumm, Top-Baseline, TF-ISF-Summ, GistSumm, NeuralSumm e Random-Baseline. A performance do SuPor pode estar relacionada ao uso de cadeias lexicais, técnica não utilizada pelos outros sistemas. Contudo, o ClassSumm, segundo sistema na classificação, emprega um tipo de coesão (dada por similaridade entre sentenças) que tem certa relação com as cadeias lexicais do SuPor. Adicionalmente, os dois primeiros sistemas foram treinados por meio de um classificador Bayesiano. Por fim, Rino e Mólolo sugerem que uma análise mais extensiva das diferentes configurações do SuPor ainda é necessária, já que o sistema oferece inúmeras possibilidades de personalização.

Baseando-se na arquitetura do SuPor, Leite e Rino (2006a) realizaram um conjunto de experimentos que culminou na elaboração do SuPor-v2. Os autores procuraram aumentar o detalhamento dos atributos do SuPor, de forma que, muitos deles, antes binários, passaram a aceitar diversos valores. Um exemplo é o atributo de cadeias lexicais, que, no SuPor, indica apenas se uma dada sentença foi selecionada por uma das três heurísticas propostas (vide Seção 2.2). No SuPor-v2, o atributo de cadeias lexicais indica quais dessas heurísticas selecionaram determinada sentença, fornecendo, portanto, mais informações ao algoritmo de treinamento. Outra proposta interessante desse trabalho foi a utilização do ambiente WEKA (Witten e Frank, 2005), que implementa diversas facilidades para o uso de algoritmos de aprendizado de máquina. Adicionalmente, os autores utilizaram o algoritmo CFS (Correlation Feature Selection) para diminuir o espaço de atributos do SuPor

(Hall, 2000), e também empregaram o algoritmo C4.5, além do Naive Bayes, para treinar e gerar o modelo do sumariizador. Em testes com o corpus TeMário, nos mesmos moldes do experimento descrito no parágrafo anterior, Leite e Rino verificaram que o sumariizador com melhor desempenho (chamado de SuPor-v2) foi o que utilizou o algoritmo Naive Bayes, sem seleção de atributos. Além disso, o desempenho do SuPor-v2, de acordo com a Medida-F, foi 6,5% superior ao do SuPor.

2.2 *Sumarização Extrativa com Redes*

As pesquisas em Redes Complexas apóiam-se firmemente nas definições e algoritmos da Teoria dos Grafos. No contexto deste projeto de sumarização, em que um tratamento de textos inspirado nessas novas pesquisas em redes é proposto, os estudos prévios de sumarização que utilizam o conceito de grafo (ou rede) ganham singular importância. Nesta seção, serão comentadas as pesquisas em sumarização extrativa que procuram modelar o texto-fonte como um grafo e, a partir dessa estrutura, selecionam os segmentos relevantes a fim de formar o extrato.

No trabalho de Skorochochod'ko (1971), nós representam sentenças, e arestas indicam relações entre sentenças, as quais baseiam-se nas relações semânticas entre as palavras das sentenças. As relações semânticas entre palavras não são definidas em detalhes; Skorochochod'ko indica que qualquer tipo de relação semântica pode ser utilizada. Além disso, se duas palavras são importantes para um dado texto, de acordo com algum critério, elas também podem ser utilizadas para ligar duas sentenças. Skorochochod'ko também sugere que as relações semânticas podem ter diversas intensidades, dependendo, por exemplo, do número de relações entre duas sentenças e do número de palavras relacionadas semanticamente a uma dada palavra. Skorochochod'ko define dois critérios para identificar a saliência de uma sentença, nomeados por Mani (2001) como (i) critério de conectividade, o qual define que a saliência de uma sentença é proporcional ao número de sentenças relacionadas a ela, e (ii) critério de indispensabilidade, o qual define a saliência como sendo proporcional ao grau de mudança que ocorre na rede ao se excluir uma sentença. Esses dois critérios foram combinados em uma fórmula que determina a saliência de uma sentença, dada por,

$$F_i = k_i(N - N_i), \quad (2.2)$$

onde F_i é a saliência da sentença i , k_i é o grau da sentença i (vide Seção 4.2.1), N é o número de sentenças da rede e N_i é o número máximo de nós em qualquer componente conexo que resta na rede após a exclusão da sentença i . A Equação 2.2 pode então ser utilizada para selecionar as sentenças mais salientes na construção de um extrato. Uma

medida de ligação semântica de um texto (chamada por Mani (2001) de medida de coesão) também foi fornecida por Skorochood'ko, conforme a equação

$$C = \frac{2E}{N(N-1)}, \quad (2.3)$$

onde E é o número de arestas na rede e N é o número de nós da rede. Skorochood'ko ainda defende que o tipo de sumarização a ser empregada em um texto depende da estrutura que sua rede apresenta. Por exemplo, para textos cujos valores de saliência de suas sentenças pouco diferem entre si, Skorochood'ko evidencia que métodos estatísticos são empiricamente comprovados pouco eficientes.

Em outro trabalho, Benbrahim e Ahmad (1994) modelam nas arestas de uma rede as ligações de coesão (tais como repetição, sinonímia, antonímia e hiponímia) entre as palavras das sentenças, as quais por sua vez representam os nós. Benbrahim e Ahmad sugerem que sentenças que iniciam um tópico são as que possuem um número de arestas com sentenças que aparecem posteriormente no texto maior do que com sentenças que aparecem anteriormente. As sentenças que finalizam um tópico têm uma definição oposta. Sentenças que têm um número de arestas acima de um determinado limiar são consideradas centrais ao texto. Sentenças marginais são as que possuem um número de arestas abaixo de um dado limite. Benbrahim e Ahmad definem então três maneiras de se construir um extrato a partir de sua rede de sentenças: (i) selecionando apenas as sentenças que iniciam um tópico, (ii) selecionando as sentenças centrais, as que iniciam e as que finalizam um tópico e (iii) selecionando apenas as sentenças não-marginais. O procedimento (i) procura construir extratos do tipo indicativo, enquanto os procedimentos (ii) e (iii) se concentram em extratos informativos. Os autores apresentam um exemplo de execução de sua proposta, mas não reportam uma avaliação em maior escala.

Salton et al. (1997) interligam parágrafos de um documento em termos de uma medida de similaridade. Cada parágrafo é representado por um vetor de termos, e a medida de similaridade entre eles é dada pelo produto escalar entre seus vetores (normalizado entre 0 e 1). Em sua representação na forma de uma rede de conexões entre parágrafos, as arestas foram rotuladas de acordo com a pontuação de similaridade entre os parágrafos. Após calcular a similaridade entre todos os pares de parágrafos da coleção, os $1,5N$ maiores valores de similaridade são selecionados para representar as arestas (N é o número de nós/parágrafos). Além disso, Salton et al. trabalham com a noção de segmentos de texto, os quais são definidos como sendo trechos contíguos de texto cujos parágrafos são fortemente conectados entre si, mas são fracamente conectados aos outros parágrafos. Essa definição tem certa relação com o conceito de comunidade, apresentado na Seção 4.2.10. Para detectar mudanças de segmentos, as arestas que interligam parágrafos muito distantes

(mais de cinco parágrafos separando-os) são eliminadas. Na rede obtida, os autores aplicam três algoritmos de percurso para extrair os parágrafos mais salientes:

- *Global Bushy (Central) Path*: um *bushy node* é um nó com alto grau (muitas arestas a ele relacionadas). Nesse algoritmo, os nós com os maiores graus são selecionados para compor o extrato. Além disso, esses nós são percorridos na ordem em que aparecem no texto. Esse algoritmo é idêntico a um dos métodos propostos neste projeto (Seção 4.2.1), embora as redes utilizadas sejam diferentes.
- *Depth First Path*: de acordo com esse algoritmo, primeiramente um nó importante é selecionado (o primeiro parágrafo ou um *bushy node*). A seguir, o nó mais similar ao nó atual (maior peso da aresta que os une) é visitado, contanto que esteja em uma posição posterior no texto. Como este algoritmo seleciona sequencialmente os nós mais similares entre si, ele tende a formar extratos mais coerentes.
- *Segmented Bushy Path*: alguns segmentos podem tratar de um tópico muito específico, e seus parágrafos podem ter poucas conexões com os outros segmentos do texto. Como, nesse caso, os outros dois algoritmos tenderiam a selecionar as sentenças de um único tópico, esse algoritmo constrói *bushy paths* para cada segmento, e os concatena mantendo a ordem original. Dessa maneira, todos os segmentos são contemplados.

A avaliação da proposta de Salton et al. foi feita utilizando-se um corpus de 50 textos da enciclopédia Funk and Wagnalls. Para cada texto, foram construídos manualmente dois extratos (por pessoas diferentes), e foram gerados automaticamente extratos utilizando cada um dos três algoritmos de percurso propostos. O sistema *baseline* utilizado foi um extrator aleatório de parágrafos. *Global Bushy Path* apresentou os melhores resultados: 45,60% dos parágrafos selecionados foram também escolhidos em um dos extratos manuais. Os outros dois algoritmos apresentaram desempenho um pouco melhor do que o *baseline*. Considerou-se que essa proposta tem desempenho aceitável, pois é próxima do nível de concordância entre os juízes (de 45,81%). A abordagem de Salton et al. é simples, mas tende a ser limitada fortemente pela taxa de compressão dos extratos, pois utiliza grandes trechos de texto (parágrafos) como unidade mínima de extração.

Abraços e Lopes (1997) utilizaram a medida de poder de resolução e de informação mútua para definir as arestas em uma rede de parágrafos. A medida de poder de resolução para um par de palavras (separadas por até dez palavras) é dada pela seguinte equação:

$$\rho = -P_d \log P_c, \quad (2.4)$$

onde P_d é a probabilidade de ocorrência do par no documento d , P_c é a probabilidade de

ocorrência do par no corpus e $-\log P_c$ é a quantidade de informação associada ao par. ρ é diretamente proporcional à frequência do par no documento e inversamente proporcional à sua frequência no corpus. O conceito de informação mútua é definido pela equação,

$$I(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)}, \quad (2.5)$$

onde x e y são as duas palavras do par, $P(z)$ denota a probabilidade da palavra z ocorrer no corpus e $P(x,y)$ indica a probabilidade do par de palavras x,y ocorrer no corpus (também dentro de uma janela de até dez palavras). Somente são selecionados os pares de palavras com poder de resolução e informação mútua acima de um determinado limiar. Esses pares de termos são comparados ao longo do texto, de modo que a co-ocorrência de pares entre parágrafos define as arestas na rede. Os parágrafos são então selecionados para formar o extrato baseando-se no número de arestas que possuem. A avaliação dessa proposta foi feita utilizando-se um corpus de 537.000 palavras de notícias em português (de Portugal) e, em sete dos dez textos selecionados foi possível identificar o parágrafo mais significativo. Contudo, o método é propenso a encontrar poucos pares por documento e, conseqüentemente, propenso a definir um grafo com nenhuma aresta.

Barzilay e Elhadad (1999) implementaram um algoritmo para computar as chamadas cadeias lexicais, que são seqüências de palavras inter-relacionadas que caracterizam um tópico de um texto. As relações de repetição, sinonímia, hipernímia, antonímia e holonímia compõem as cadeias lexicais. O número de relações em uma cadeia lexical, e seus respectivos pesos, são utilizados para que a melhor cadeia seja selecionada para cada segmento. Esses segmentos são delimitados pelo TextTiling, o qual segmenta um texto em grupos coerentes de sentenças (Hearst, 1997). As cadeias dos diferentes segmentos são unidas quando têm um termo em comum (de mesmo sentido), o que dá origem a uma rede de relações semânticas entre os termos do texto-fonte. Três heurísticas foram definidas para gerar um extrato: (i) seleciona, para cada cadeia, a sentença que contém a primeira ocorrência de um membro da cadeia; (ii) para cada cadeia, escolhe a sentença que contém a primeira ocorrência de um membro representativo da cadeia (membro que tem sua frequência de ocorrência na cadeia maior ou igual à média das frequências das palavras da cadeia); e (iii) procura extrair sentenças de um tópico que é discutido em diversos segmentos do texto. Barzilay e Elhadad avaliaram seu método em um experimento utilizando sumários construídos por humanos, e obtiveram melhores resultados de Precisão e Cobertura do que o sumariizador AutoSummarize da Microsoft, embutido no processador de textos MS-Word. Os autores apontam que, em alguns casos, uma cadeia lexical formada por palavras de baixa frequência pode ser mais indicativa para um tópico do que palavras de alta frequência, devido à relação semântica entre as palavras que formam a cadeia. O

uso de cadeias lexicais na sumarização costuma ser vantajoso, como mostram os sistemas SuPor e SuPor-v2, já apresentados neste capítulo, e o sistema ULeth131m (Brunn et al., 2002), participante da DUC'2002 e segundo colocado na conferência, segundo avaliação feita por meio da métrica ROUGE-1 (Mihalcea, 2005). O sistema ULeth131m ainda aplica algumas heurísticas de reparo nos extratos gerados por cadeias lexicais, como a inclusão da sentença imediatamente anterior a uma sentença do extrato que contenha alguma anáfora sem referente.

Mani e Bloedorn (1999) criaram uma representação de documentos em forma de uma rede que interliga termos, e não sentenças ou parágrafos. Suas arestas indicam relações de coesão entre os termos (proximidade, repetição, sinonímia, hipernímia e co-referência). Sendo assim, cada nó, que representa uma instância de uma palavra, pode estar ligado a outro nó por meio de diversos tipos de arestas que representam as relações de coesão entre os termos. O algoritmo de sumarização de Mani e Bloedorn recebe como entrada um tópico fornecido pelo usuário e produz um extrato que satisfaz esse tópico. Primeiramente, os termos presentes no tópico são selecionados na rede e, a seguir, um algoritmo de ativação por espalhamento (*spreading activation*) percorre outros nós relacionados aos nós do tópico. Conforme o sinal de ativação percorre a rede, ele associa pesos aos termos (o que define um contorno de saliência do texto) e perde sua intensidade de acordo com os níveis de importância associados aos diferentes tipos de arestas. Os picos desse contorno de saliência são utilizados para detectar segmentos no texto-fonte e, somente então, as sentenças são extraídas baseando-se nos pesos dos termos presentes nos segmentos. Um experimento conduzido por Mani et al. (1998) para detectar a saliência de orações em cinco textos mostrou que o algoritmo de Mani e Bloedorn obteve desempenho superior (i) ao uso da métrica TF-IDF e (ii) ao uso do grau dos nós (soma dos pesos das arestas) para associar pesos aos termos. Além disso, esse algoritmo correlacionou-se bem com o julgamento de humanos, de acordo com o nível de saliência, em três dos cinco textos utilizados.

Mihalcea (2005) propõe um sistema de sumarização extrativa no qual aplica algoritmos de pontuação de nós desenvolvidos para classificar páginas da Web. Nesse trabalho, foram utilizados os algoritmos PageRank⁶ (Page et al., 1998) e HITS (Kleinberg, 1999) para selecionar os nós mais bem pontuados em uma rede cujos nós representam sentenças e arestas indicam termos em comum entre elas. O número de interseções entre duas sentenças dá o peso de uma aresta, normalizado pelo tamanho das sentenças. PageRank é calculado para um vértice i da seguinte maneira:

$$PR(i) = (1 - d) + d \sum_{j \in In(i)} \frac{PR(j)}{\|Out(j)\|}, \quad (2.6)$$

⁶PageRank é utilizado para classificar páginas Web no mecanismo de busca Google.

onde d é um parâmetro definido entre 0 e 1 (tem a função de integrar saltos aleatórios no modelo de caminhada aleatória), $In(i)$ é o conjunto de vértices com arestas que apontam para i , $Out(i)$ é o conjunto de vértices que recebem arestas de i , considerando uma rede direcionada. Já o algoritmo HITS usa duas expressões para distinguir os vértices que recebem um grande número de arestas dos que apontam para um grande número de outros vértices. O primeiro é chamado de *authority*, e o último, de *hub*. Existem dois tipos de pontuação HITS:

$$HITS_A(i) = \sum_{j \in In(i)} HITS_H(j) \quad (2.7)$$

$$HITS_H(i) = \sum_{j \in Out(i)} HITS_A(j), \quad (2.8)$$

onde a primeira refere-se a *authorities*, e a segunda, a *hubs*. Mihalcea, além de adaptar essas três equações para redes com pesos, define três tipos de redes para textos: a (i) não-direcionada, a (ii) direcionada tipo-1, cujas arestas seguem o fluxo de leitura do texto (arestas *forward*) e a (iii) direcionada tipo-2, cujas arestas seguem o fluxo contrário de leitura do texto (arestas *backward*). Os textos utilizados nos experimentos foram as reportagens em inglês da DUC'2002 e as reportagens em português do corpus TeMário, e o desempenho dos algoritmos utilizados por Mihalcea foi avaliado pelo sistema ROUGE. Nas redes com arestas *forward* e *backward*, o algoritmo HITS obteve melhor desempenho que o melhor sistema classificado na DUC'2002 (o PageRank ficou um pouco abaixo no modelo *backward*). Na rede com arestas *backward*, o algoritmo PageRank foi o que obteve melhor desempenho com o TeMário, próximo ao desempenho do PageRank e do HITS para os textos em inglês, o que indica uma certa independência de língua na proposta de Mihalcea.

Erkan e Radev (2004) também introduzem um método baseado em redes para calcular a importância de sentenças em textos. Utilizam o modelo *bag-of-words* para representar cada sentença (vetor n -dimensional com o TF-IDF de cada uma das n palavras). Empregam também um modelo de rede que representa a conectividade entre sentenças, dada pelo cosseno do ângulo entre os vetores de cada par de sentenças, de maneira que valores acima de um dado limite definem as arestas, sem pesos. Erkan e Radev procuram mensurar a centralidade, ou importância, de cada sentença em uma coleção de documentos, a fim de realizar sumarização multi-documento⁷ sobre um mesmo tópico, não especificado. São definidos três tipos de pontuação para cada sentença, (i) *degree centrality* (ou grau de um nó), (ii) LexRank e (iii) LexRank contínuo, sendo que LexRank é o algoritmo PageRank aplicado à rede de sentenças e LexRank contínuo é o LexRank aplicado na rede com pesos dados pela similaridade de cosseno. Em seus experimentos, Erkan e Radev utilizaram corpus em inglês das DUC's de 2003 e 2004, cujas tarefas envolviam sumarização genérica de

⁷Coleções de documentos servem como entrada em sistemas de sumarização multi-documento.

coleções de notícias (30 coleções na DUC'2003 e 50 coleções na DUC'2004). O sistema de avaliação automática ROUGE (vide Seção 5.1) também foi empregado nos experimentos. As três novas métricas foram agrupadas em uma combinação linear, nos moldes da abordagem de Edmundson (1969), com outros dois atributos (posição e comprimento da sentença), de modo que o peso das novas métricas foi variado. Dois sistemas *baseline* também foram utilizados para comparação: um extrator de sentenças aleatório (Random-Baseline) e um seletor das primeiras sentenças (Top-Baseline). Para todos os conjuntos de dados selecionados, os novos métodos foram os que obtiveram os melhores resultados, bem acima dos obtidos para os *baselines*. Entretanto, não foi possível distinguir o desempenho do grau e do LexRank, o que indica que o grau já é uma boa medida para mensurar a importância de uma sentença. Comparando as novas abordagens com os sistemas participantes da DUC, LexRank se apresentou como o segundo melhor método na maioria dos testes realizados com os dados da DUC'2003. Com relação à DUC'2004, pelo menos uma das três novas abordagens obteve o primeiro lugar nos experimentos realizados.

Tendo sido apresentada neste capítulo uma visão da área de Sumarização Automática de Textos, encontra-se, no próximo capítulo, uma introdução aos estudos em Redes Complexas, uma área de estreita relação com os métodos de sumarização propostos e avaliados neste projeto.

Redes Complexas

É fornecida, a seguir, uma breve introdução à área de Redes Complexas. Esta introdução foi aqui incluída pois os conceitos apresentados na Seção 4.2 são provenientes dos (ou freqüentemente utilizados nos) estudos em Redes Complexas. Pretende-se, portanto, mais motivar o uso desses conceitos do que proporcionar uma introdução abrangente à área de Redes Complexas. A própria Seção 4.2 serve como uma introdução à área, pois tem exemplos de ferramentas utilizadas na caracterização de redes complexas. Note que este capítulo e o Capítulo 2 introduzem as duas áreas de pesquisa relacionadas a este projeto: Redes Complexas e Sumarização Automática de Textos. Entretanto, o capítulo de sumarização é muito mais extenso que este capítulo, por se tratar do foco principal deste trabalho.

As redes, ou grafos, são estruturas formadas por um conjunto de nós e um conjunto de arestas que conectam esses nós, e podem ser utilizadas para modelar praticamente qualquer estrutura discreta. É possível representar os mais diversos fenômenos presentes em nosso mundo, incluindo relações sociais entre indivíduos, rotas de vôo entre aeroportos e sinonímia entre palavras de um texto, empregando-se, para tanto, técnicas desenvolvidas na Teoria dos Grafos (Harary, 1969). Sendo uma subdisciplina madura¹ da Matemática, a Teoria dos Grafos apresenta estudos extensivos a respeito de diversos problemas teóricos e práticos em grafos estáticos, como a coloração de vértices e o percurso mínimo de um caixeiro viajante. Já os estudos em grafos dinâmicos recaem sob a Teoria dos Grafos Aleatórios, desenvolvida principalmente por Erdős e Rényi (1959). Esta teoria concentra-

¹A solução de Euler para o problema das pontes de Königsberg, em 1736, é considerada o marco inicial da Teoria dos Grafos (Barabási, 2003).

se em propriedades de modelos de formação de grafos regidos por probabilidades, onde a chance de existir uma conexão entre qualquer par de nós é a mesma. A Teoria dos Grafos Aleatórios foi considerada, por muito tempo, a principal explicação para a formação de redes reais.

Outra linha de pesquisa, desta vez em ciências sociais, proporcionou uma verificação prática a respeito da estrutura das redes presentes em nosso mundo. Na década de 60, Stanley Milgram, um psicólogo experimental, estudou como os cidadãos dos Estados Unidos estavam conectados entre si (Milgram, 1967). Ele realizou um experimento baseado no envio coordenado de cartas a uma pessoa pré-determinada, de modo que as cartas deveriam passar de mãos em mãos a partir de remetentes, escolhidos aleatoriamente, que não conheciam pessoalmente o destinatário. Cada indivíduo que recebia uma dessas cartas deveria repassá-la a outra pessoa de seu círculo de amizades, supostamente mais apta a encaminhar a carta ao destinatário escolhido. Ao analisar os resultados desse experimento, Milgram notou que cada carta passou por aproximadamente 6 pessoas, em média, antes de chegar ao destinatário final. Desse experimento surgiu a denominação Seis Graus de Separação (Six Degrees of Separation). Essa constatação experimental é coerente com o que se verifica na Teoria dos Grafos Aleatórios.

Os estudos sobre redes receberam novo impulso recentemente, quando foram descobertas diversas características que fazem as redes do mundo real serem diferentes das redes aleatórias, aceitas até então por décadas como o principal modelo de redes (Barabási, 2003). Watts e Strogatz mostraram que várias redes têm distâncias curtas entre seus nós (Seis Graus de Separação) em conjunto com alto coeficiente de aglomeração (Watts e Strogatz, 1998). O coeficiente de aglomeração (definido na Seção 4.2.2) mede o quão conectados estão os vizinhos de um nó, ou seja, o quão os amigos de um determinado indivíduo também são amigos entre si (tomando como exemplo a rede utilizada no experimento de Milgram). É natural que em redes sociais o coeficiente de aglomeração seja alto, devido aos grupos de amigos, mas os grafos aleatórios não refletiam essa característica. Watts e Strogatz criaram o modelo pequeno-mundo (*small-world*), unindo duas propriedades importantes: distância curta entre nós e alto agrupamento local. A dinâmica dos processos que ocorrem em uma rede é diretamente influenciada pelo efeito pequeno-mundo. Por exemplo, um boato se espalha muito mais rápido se, ao invés de mil passos, levar em média apenas seis para chegar de qualquer pessoa a outra.

Várias redes do mundo real costumam apresentar a propriedade livre de escala (*scale-free*), descoberta por Faloutsos et al. (1999) ao analisar a distribuição dos graus na Internet. Barabási e Albert (1999) mostraram que a distribuição dos graus em outras redes reais também é livre de escala, como na WWW (World Wide Web), e criaram um modelo para

a formação de tais redes. Nessas redes, a distribuição do número de arestas por nó (grau), segue uma lei de potência, ao contrário das redes aleatórias, que seguem uma distribuição de Poisson. A distribuição dos graus em redes livre de escala é igual a

$$P(k) \sim k^{-\gamma}, \quad (3.1)$$

sendo que $P(k)$ é a probabilidade de um dado nó ter grau igual a k , e γ é uma constante. As redes livre de escala apresentam os chamados *hubs*, nós que têm um número elevado de conexões. Os *hubs* aparecem em pequeno número, enquanto que a maior parte dos nós têm grau bem menor. Em contrapartida, nas redes aleatórias não existem nós com grau muito acima ou muito abaixo da média, pois, na distribuição de Poisson, existe um valor médio característico para o grau, que é o número de arestas k para o ponto de máximo global da curva, e é em torno desse valor médio que se concentra a maior parte dos graus da rede.

Os estudo de sistemas complexos modelados como grafos são chamados atualmente de estudos em Redes Complexas (Albert e Barabási, 2002; Dorogovtsev e Mendes, 2002; Newman, 2003; Boccaletti et al., 2006), e têm grande influência da Mecânica Estatística (Pathria, 1996), além, é claro, da Teoria dos Grafos. É cada vez mais evidente que a estrutura, a função e a evolução dessas redes não são uniformes, e sim, são governadas por princípios robustos, o que conduz a uma crescente necessidade de se desenvolver ferramentas para que esses princípios possam ser entendidos. Uma abordagem dos estudos em redes é a criação de modelos de formação de redes, como o pequeno-mundo (Watts e Strogatz, 1998) e o livre de escala (Barabási e Albert, 1999)², os quais permitem que as propriedades de redes sejam estudadas analiticamente ou por meio de simulações em computador. Modelos são úteis também na análise de atributos que não podem ser observados na prática. Por exemplo, é impossível obter o histórico de construção da WWW, enquanto que, utilizando um modelo de construção, é possível realizar uma simulação. Entretanto, existe uma certa limitação dos modelos, pois eles são aproximações e não capturam por completo as características dos objetos de estudo reais.

Outra abordagem em Redes Complexas é o uso de medidas que ajudam a caracterizar as propriedades de um determinado sistema. Costuma-se empregar, para tanto, uma ou mais métricas disponíveis para a análise de redes complexas (Costa et al., 2006b). Três das principais métricas que têm sido tradicionalmente aplicadas em redes são (i) grau, (ii) coeficiente de aglomeração e (iii) caminho mínimo, todas definidas, entre outras, na Seção 4.2. Medidas desse tipo geralmente são utilizadas para se associar um valor numérico a vértices, a pares de vértices, a arestas ou a toda a rede. Cada medida tem uma interpretação parti-

²Esses modelos são utilizados na construção, respectivamente, (i) de redes com distribuição dos graus livre de escala e (ii) de redes com distância curta entre vértices e alto agrupamento local.

cular e, ao ser aplicada, permite que uma determinada propriedade da rede seja analisada. Portanto, as medidas contidas na Seção 4.2 são apresentadas de forma a motivar seu uso na Sumarização Automática de Textos. Cada nó da rede descrita na Seção 4.1 é um objeto onde essas medidas são aplicadas, de maneira que exista uma pontuação que sirva como subsídio para a inclusão, ou não, de determinada sentença no sumário.

3.1 *Redes Complexas e Língua Natural*

Redes derivadas de manifestações lingüísticas também costumam ser estudadas na área de Redes Complexas. Um exemplo dessas redes é a de co-ocorrência de palavras, onde palavras que aparecem em seqüência em um dado texto são interligadas por arestas (Ferrer i Cancho e Solé, 2001). Outro exemplo, é a rede de palavras que são conectadas se expressam os mesmos conceitos ou se pertencem ao mesmo campo semântico (Motter et al., 2002). Outra rede, também de relações semânticas, é a que representa a estrutura da Wordnet (Sigman e Cecchi, 2002). Relações sintáticas, baseadas em uma gramática de dependência, foram utilizadas por Ferrer i Cancho et al. (2004) na construção de uma rede de palavras. Por fim, o fluxo de associações mentais entre palavras também é utilizado na construção de redes, de maneira que, quando uma palavra é apresentada a uma pessoa, ela fornece outra palavra, livremente, que esteja associada à anterior (cada aresta representa o número de vezes que um par de palavras foi associado pelo indivíduo) (Costa, 2004). Todas essas redes apresentam características não triviais, como o efeito pequeno-mundo e a distribuição dos graus livre de escala.

As redes inspiradas em língua natural também podem ser utilizadas em pesquisas relacionadas ao processamento de língua natural. Ferrer i Cancho et al. (2005) analisaram uma rede complexa em que os nós representam palavras e as arestas indicam relações sintáticas entre elas. Mostrou-se que é possível agrupar as palavras de acordo com a classe morfológica por meio de métodos espectrais utilizados para ordenar os nós. Dorow et al. (2005) introduziram métodos baseados na curvatura de grafos e no agrupamento de arestas para determinar o significado de substantivos e para detectar a ocorrência de ambigüidades. Foi utilizado um modelo de rede no qual cada nó é um substantivo do British National Corpus (BNC) e duas palavras estão interligadas se ocorrem no cópús separadas por ‘ou’, ‘e’ ou vírgula. Métricas extraídas de redes de co-ocorrência de palavras são utilizadas no processamento e análise de textos. Alguns resultados referentes à qualidade de textos foram publicados por Antigueira et al. (2005, 2007). Os autores observaram que, conforme o grau e o coeficiente de aglomeração aumentam, a qualidade dos textos tende a diminuir. Essa constatação indica que, quando o número de conexões entre as

palavras de um texto é excessivo, sua qualidade tende a cair. No caso da dinâmica do número de componentes³, pôde-se perceber que, quanto mais cedo novos conceitos são apresentados no texto, pior o texto fica. Dessa maneira, o escritor repete, no decorrer do texto, conceitos já apresentados anteriormente. Pardo et al. (2006a,b) também aplicaram métricas obtidas de redes complexas na avaliação automática de sumários. Os resultados dessa pesquisa indicam que é possível separar sumários de acordo com sua qualidade, nos moldes dos resultados obtidos por Antiqueira et al. (2005, 2007). A tarefa de extração de terminologia foi estudada por Antiqueira (2005a,b), utilizando uma rede de co-ocorrência de palavras derivada de um corpúsculo de textos científicos da área de Nanotecnologia. Nesse trabalho, o grau dos nós mostrou-se um bom parâmetro para extrair termos do referido corpúsculo. Por fim, Antiqueira et al. (2006) estudaram o problema de identificação de autoria, também transformando cada texto em uma rede de co-ocorrência de palavras. Os autores mostraram que medidas obtidas dessas redes, como grau e coeficiente de aglomeração, têm potencial para serem aplicadas no agrupamento de textos de acordo com a autoria.

No próximo capítulo, são detalhados o modelo de representação de textos na forma de redes e a metodologia proposta para geração de sumários.

³Essa medida quantifica a velocidade com que novas palavras são utilizadas em um texto.

Propostas de Geração de Extratos

O foco deste projeto foi a produção automática de extratos informativos e genéricos, empregando-se uma abordagem superficial (empírica), por meio de um modelo de rede para textos e da aplicação de conceitos da área de Redes Complexas na seleção das sentenças mais relevantes do texto original. Para tanto, foram propostas diversas técnicas que possibilitam a escolha de um subconjunto de sentenças em uma rede (derivada do texto-fonte), e posterior construção de um extrato pela justaposição dessas unidades de texto selecionadas. Nas próximas seções, a metodologia de construção de redes é explicada (Seção 4.1), seguida pela apresentação dos métodos de sumarização propostos (Seção 4.2).

4.1 *Construção das Redes*

As possibilidades para representação de um texto na forma de nós interligados por arestas são inúmeras. Quanto à estrutura do grafo utilizado, pode-se ter arestas direcionadas ou não-direcionadas, arestas com pesos associados ou ainda diferentes tipos de nós e arestas. Além disso, é preciso definir o que nós e arestas representam na rede. Nós podem indicar, por exemplo, palavras, orações, sentenças ou parágrafos. Arestas podem representar relações de coesão, relações sintáticas ou semânticas. O nível de representação lingüística considerado no modelo pode variar bastante, desde um que considera apenas características superficiais (como repetição de palavras) até um que utiliza estruturas resultantes de uma análise sintática ou discursiva do texto. Conforme já salientado anteriormente, foi

adotada neste projeto uma representação superficial de textos. Portanto, o modelo de rede utilizado não emprega, por exemplo, *parsers* ou teorias de representação retórica.

A rede aqui empregada segue a tendência iniciada por Skorochod'ko (1971), ou seja, representa um texto na forma de sentenças interligadas. As relações entre as sentenças são definidas por meio da co-ocorrência de palavras em sentenças diferentes (após os processos de lematização e de exclusão de *stopwords*). Todas as fases do pré-processamento de um texto contemplam as línguas inglesa e portuguesa (do Brasil), a fim de possibilitar uma avaliação bilíngüe dos sistemas de sumarização aqui propostos. Essas fases são detalhadas a seguir:

1. *Segmentação*: Fase em que o início e o fim de cada sentença do texto-fonte são identificados. Para o inglês, foi utilizado o *software* MXTerminator (Reynar e Ratnaparkhi, 1997) e, para o português, o *software* Sentencer¹.
2. *Etiquetagem morfossintática*: Associa a cada *token* do texto uma identificação morfossintática (preposição, verbo ou substantivo, por exemplo). Para textos em inglês e em português é utilizado o *tagger* MXPost (Ratnaparkhi, 1996; Aluisio e Aires, 2000).
3. *Eliminação de stopwords*: Qualquer palavra que não seja um substantivo é considerada uma *stopword*, ou seja, é eliminada das análises posteriores. Para tanto, a saída da etiquetagem morfossintática é utilizada.
4. *Lematização*: A transformação das palavras remanescentes em suas respectivas formas canônicas (lemas) é útil, pois serve para agrupar nas estatísticas as diferentes desinências das palavras. As informações provenientes do etiquetador são utilizadas como fonte de desambiguação no processo de lematização (a palavra “casa” é um substantivo singular-feminino ou uma forma flexionada do verbo “casar”?). A lematização de textos em inglês é feita pela chamada a uma função da biblioteca C do projeto WordNet (Miller, 1995). Já para textos em português, é utilizado um script Perl² que faz acesso ao léxico KLS (Kowaltowski et al., 1998; Nunes et al., 1996).

As palavras restantes, modificadas ou não pelo processo de lematização, servem para definir as arestas na rede de sentenças. É possível que haja uma aresta entre qualquer par de sentenças distintas, basta que exista uma palavra em comum entre elas. Além disso,

¹Desenvolvido no NILC (Núcleo Interinstitucional de Linguística Computacional) por Jorge Marques Pelizzoni.

²Desenvolvido no NILC por Jorge Marques Pelizzoni e Valéria Delisandra Feltrim, e posteriormente adaptado por Lucas Antikeira.

o número de repetições de palavras entre duas sentenças indica o peso da aresta que as une na rede. Mais formalmente, seja $P_i = \{p_1, p_2, \dots, p_{\|P_i\|}\}$ o conjunto de $\|P_i\|$ palavras contidas na i -ésima sentença de um texto-fonte de N sentenças, após o pré-processamento do texto. As sentenças são numeradas sequencialmente da primeira até a última e, como P_i é um conjunto, não há repetição de palavras entre seus elementos (ou seja, repetições são descartadas). O peso da aresta que liga a sentença i à sentença j é dado pelo número de elementos contidos na interseção entre P_i e P_j , ou seja, $w_{ij} = w_{ji} = \|P_i \cap P_j\|$, caso $i \neq j$, e $w_{ij} = 0$, caso $i = j$. Se w_{ij} for igual a zero, não existe aresta entre os nós i e j . Os pesos w_{ij} (com $i = 1, \dots, N$ e $j = 1, \dots, N$), são elementos da matriz simétrica W de ordem $N \times N$, utilizada em todos os métodos de sumarização aqui propostos (Seção 4.2). Essa matriz representa completamente a rede, não-direcionada, obtida de um dado texto-fonte.

Procurou-se, com essa metodologia de processamento de textos, codificar um tipo básico de coesão lexical entre sentenças (repetição), de modo que sentenças com conteúdo similar tenham grande chance de estar conectadas na rede, possivelmente com alto peso. A lematização é interessante, pois evita que diferenças de gênero e número, entre substantivos de mesma forma canônica, acrescentem um ruído indesejável na definição das arestas. Somente são considerados substantivos para evitar que exista um número demasiado de arestas na rede, como pode-se imaginar ao se considerar todas as palavras das sentenças na definição das arestas, o que provavelmente dificultaria o propósito de discriminar as sentenças. Os substantivos, neste caso, são tomados como bons indicadores do conteúdo de uma dada sentença. Acreditamos que este processo de modelagem de textos seja satisfatório do ponto de vista da complexidade de sua implementação (utiliza recursos facilmente encontrados para diversas línguas) e da sua utilidade para a sumarização de textos (como mostram os resultados contidos no Capítulo 5).

Na Figura 4.1 está incluído um trecho de texto, com o propósito de ilustrar a montagem das redes. Considerando as 4 sentenças da Figura 4.1, obtém-se uma rede com 4 nós e 2 arestas (Figura 4.2). Os substantivos “Natal” e “Ano” fazem com que uma aresta (de peso 2) seja criada entre as sentenças 1 e 2, e a forma canônica “justificativa” resulta em uma aresta (de peso 1) entre as sentenças 3 e 4. Na Figura 4.3, são mostradas duas redes obtidas a partir de textos maiores, utilizados nos experimentos do Capítulo 5. Esses exemplos servem para ilustrar o potencial que o tipo de rede proposto tem na discriminação dos vértices para posterior geração de extratos. Em ambos os exemplos da Figura 4.3 é possível identificar vértices com muitas arestas, vértices com poucas arestas e até mesmo vértices isolados. Nota-se também uma diversidade de pesos associados às arestas, representados nos exemplos pelas espessuras das linhas que unem dois vértices. Além disso, percebe-se, na Figura 4.3a, que existe um grupo de vértices bem conectados entre si, o que pode indicar uma forte coesão entre algumas sentenças. Se as redes apresentassem vértices muito

parecidos uns com os outros, levando-se em conta sua conectividade (quantidade de arestas, pesos das arestas), a tarefa de escolha das sentenças mais relevantes para formar um extrato seria prejudicada, pois, ao analisar a rede, todas as sentenças pareceriam similares.

-
- 1) No passado, o Brasil parava antes do **Natal** e só recomeçava depois do **Ano** Novo.
 - 2) E aí, chega outra vez a hora de nos prepararmos para as festas de **Natal** e **Ano** Bom, pois ninguém é de ferro...
 - 3) Ele reúne, num só tempo, as melhores **justificativas** para adiar tudo para 1995 -e olhe lá...
 - 4) Com isso, ele provou que as tais bases gostam de ver os seus representantes trabalhando em benefício da coletividade lá no Congresso, não havendo a menor **justificativa** para faltarem ao seu trabalho.
-

Figura 4.1: Sentenças extraídas do texto da Figura 2.1, que ilustram a construção de uma rede de sentenças.

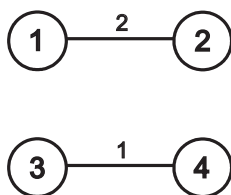


Figura 4.2: Rede derivada do texto da Figura 4.1.

Métricas da área de Redes Complexas, comumente empregadas na caracterização dos mais diversos tipos de redes (Costa et al., 2006b), foram aplicadas neste projeto sob a ótica da construção de extratos. Isso não implica que o modelo aqui utilizado seja necessariamente uma rede complexa. Inclusive, não há concordância na literatura a respeito do que seja uma rede complexa, embora as propriedades livre de escala, pequeno-mundo e tendência de aglomeração sejam amplamente aceitas como indicadores de complexidade em redes (vide Capítulo 3), e alguns tipos de redes, como os grafos regulares, não sejam considerados redes complexas. Vale ressaltar que algumas redes derivadas de manifestações em língua natural, como as apresentadas na Seção 3.1, são redes complexas.

4.2 Sumarizadores Propostos

As técnicas apresentadas nesta seção foram desenvolvidas visando a construção de sumários genéricos e informativos do tipo extrativo, por meio de uma abordagem superficial. Todas as medidas definidas a seguir fazem uso da matriz simétrica W de pesos, de ordem $N \times N$,

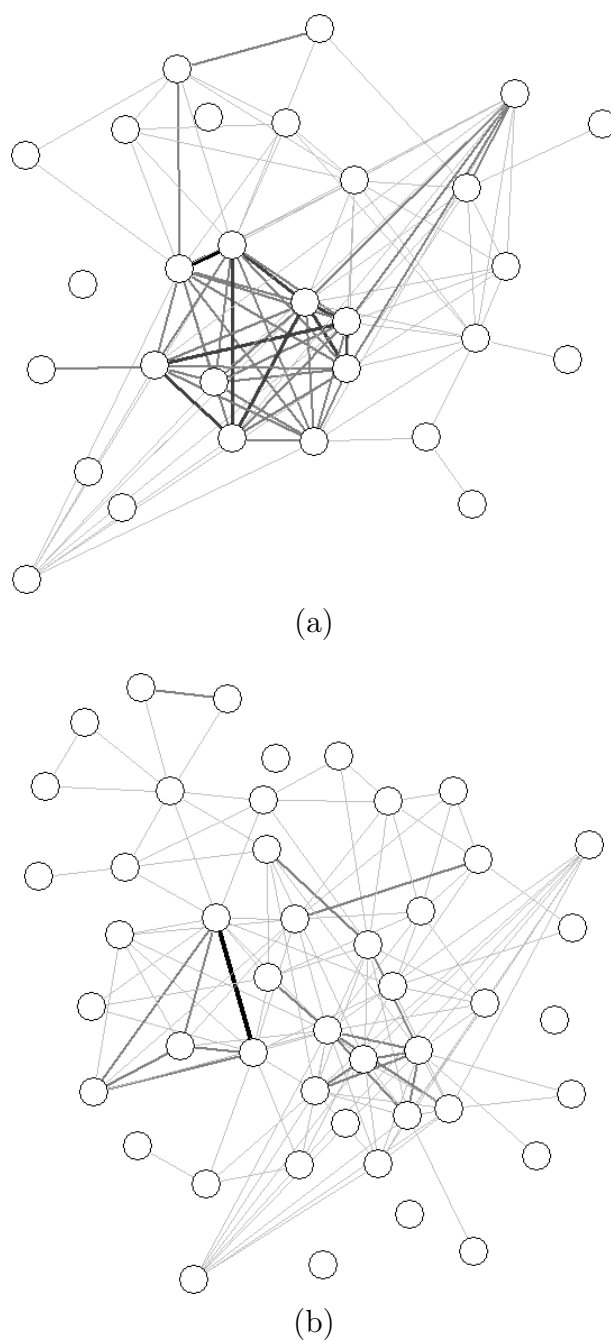


Figura 4.3: Redes de 29 (a) e de 46 (b) nós obtidas a partir de dois textos do corpus TeMário (Pardo e Rino, 2003). A espessura das arestas nos diagramas é proporcional ao respectivo peso.

obtida segundo o método de montagem de redes descrito na Seção 4.1, onde N é o número de sentenças contidas no texto-fonte. Se um elemento w_{ij} da matriz W for igual a zero, não existe aresta ligando os vértices i e j . Se $w_{ij} > 0$, então o peso da aresta que associa as sentenças i e j é igual a w_{ij} . A matriz de adjacências A , útil também nos métodos de sumarização aqui implementados, é derivada da matriz W de maneira que, se $w_{ij} = 0$ então

$a_{ij} = 0$, e, se $w_{ij} > 0$ então $a_{ij} = 1$, onde a_{ij} é elemento de A . Essa matriz indica apenas se existe uma aresta entre os nós i e j , desconsiderando o peso. No decorrer desta seção são apresentadas diversas métricas (tomadas dos estudos na área de Redes Complexas) que associam um valor a cada nó de uma rede, dando, assim, embasamento à escolha das sentenças que devem compor um extrato. Ao explicar uma determinada medida (10 no total, ou 26, se consideradas variações), explica-se como ela foi utilizada nos experimentos de sumarização extrativa relatados no Capítulo 5. Na Tabela 4.1, no final desta seção (página 56), pode ser consultada uma lista com os símbolos e nomes de todas as técnicas de sumarização utilizadas neste projeto.

4.2.1 Grau

O grau (*degree*) de um nó i é o número de arestas a ele associadas, ou, em outras palavras, é o número de outros nós associados a i . Mais especificamente, o grau de um nó i é dado por

$$k_i = \sum_{j=1}^N a_{ij} = \sum_{j=1}^N a_{ji}. \quad (4.1)$$

A Figura 4.4 mostra dois nós, 1 e 2, cujos respectivos graus são $k_1 = 5$ e $k_2 = 2$. Note que, nessa figura, os pesos das arestas não são considerados. Caso contrário, se considerarmos a matriz W no cálculo do grau, ao invés da matriz A , teremos uma variante do grau que considera a somatória dos pesos das arestas associadas ao nó i (Costa et al., 2006b). Essa variação do grau, conhecida como *strength*, é dada por

$$s_i = \sum_{j=1}^N w_{ij} = \sum_{j=1}^N w_{ji}. \quad (4.2)$$

As medidas k_i e s_i foram utilizadas nos experimentos de sumarização da seguinte maneira. Dado um número x de sentenças que devam compor o extrato, as x sentenças com os maiores valores de k_i , ou s_i , são selecionadas. Essas duas medidas são aplicadas separadamente, cada uma delas funciona como um sumarizador independente. A seguir, as x sentenças são ordenadas de acordo com a seqüência em que aparecem no texto-fonte, para, enfim, serem reagrupadas na forma textual³. É obtido, portanto, um extrato composto por um subconjunto das sentenças do texto-fonte. Considerou-se que sentenças com alto valor de k_i , ou s_i , possam contribuir positivamente para a informatividade de um extrato, pois são concentradoras de conexões, e, possivelmente, compartilham informações (por repetição lexical) com diversas outras sentenças. Percebe-se, no tipo de rede aqui utilizado, que

³Esse procedimento de rearranjo das x sentenças selecionadas é realizado em todos os algoritmos de sumarização aqui descritos.

o grau de um nó tem estreita relação com a frequência de palavras utilizada por Luhn (1958) na sumarização (vide Seção 2.1), já que as arestas são definidas pela co-ocorrência de palavras (mais precisamente substantivos) entre duas sentenças. Em outras palavras, o grau é diretamente influenciado pela frequência dos substantivos, embora a maneira como as arestas sejam construídas faça com que k_i e s_i não sejam idênticos à frequência dos substantivos presentes em uma sentença i .

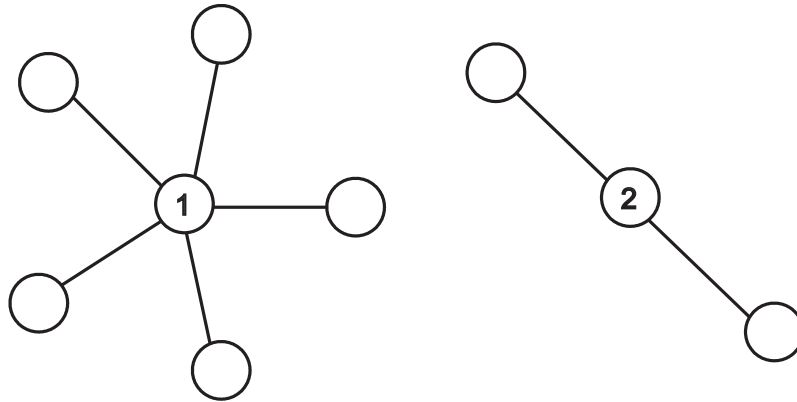


Figura 4.4: Vértices 1 e 2 com graus $k_1 = 5$ e $k_2 = 2$.

4.2.2 Coeficiente de Aglomeração

A tendência intrínseca de algumas redes formarem agrupamentos (*clustering* ou *transitivity*) é quantificada pelo coeficiente de aglomeração (*clustering coefficient*) (Watts e Strogatz, 1998). Quando um vértice i está conectado a um vértice j , e o vértice j a um vértice k , essa medida verifica se o vértice i também está conectado ao vértice k . Em vértices com alto coeficiente de aglomeração, significa que seus vizinhos estão bem conectados entre si. Para obter uma definição do coeficiente de aglomeração, considere que, para cada nó i da rede, 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 $k_i(k_i - 1)/2$ arestas entre eles. Seja E_i o número de arestas que realmente existem entre os k_i nós, então,

$$C_i = \frac{2E_i}{k_i(k_i - 1)}, \quad (4.3)$$

é o coeficiente de aglomeração do nó i ($0 \leq C_i \leq 1$) em uma rede não direcionada (Albert e Barabási, 2002), o qual reflete o quanto os nós conectados a esse nó também estão conectados entre si⁴. Se $k_i \leq 1$, então $C_i = 0$. Na Figura 4.5 encontram-se em destaque

⁴A Equação 4.3 tem estreita relação com a Equação 2.3, definida na Seção 2.2, a qual fornece uma medida de ligação semântica para um texto. Se adicionarmos um nó i fictício (que não representa sentença

dois vértices (1 e 2), cujos coeficientes de aglomeração são $C_1 = 0,7$ e $C_2 = 0,2$. Nesse caso, os graus k_i são iguais para os vértices 1 e 2 ($k_1 = k_2 = 5$), mas os vizinhos de 1 são mais conectados entre si do que os vizinhos de 2, como pode ser verificado na diferença entre C_1 e C_2 .

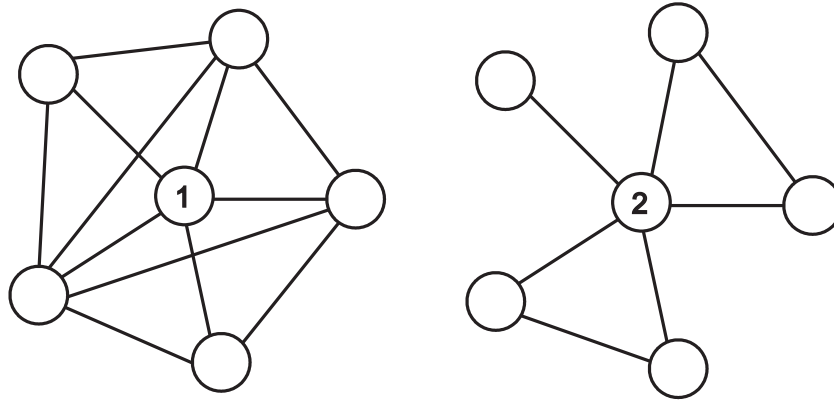


Figura 4.5: Vértices 1 e 2 com coeficientes de aglomeração $C_1 = 0,7$ e $C_2 = 0,2$.

Note que a Equação 4.3 e os exemplos da Figura 4.5 não levam em consideração os pesos das arestas, e sim, apenas a existência de determinadas conexões. O coeficiente de aglomeração com pesos (*weighted clustering coefficient*) (Barthélemy et al., 2005)⁵ é dado pela equação

$$C_i^w = \frac{1}{s_i(k_i - 1)} \sum_{(j,k)} \frac{w_{ij} + w_{ik}}{2} a_{ij} a_{ik} a_{jk}, \quad (4.4)$$

sendo que $0 \leq C_i^w \leq 1$. Essas duas medidas, C_i e C_i^w , refletem o nível de concentração de arestas entre os k_i vizinhos de um nó. Se um determinado nó i tem alto coeficiente de aglomeração (considerando ou não os pesos), ele e seus vizinhos formam um agrupamento coeso, com um compartilhamento elevado de informações, o que poderia ser um bom indicador da utilidade do nó i na sumarização. Esse nó central pode ser tomado como um representante do agrupamento todo, de modo que sumarie o conteúdo de seus vizinhos. Além disso, se esse nó central for um bom representante do agrupamento do qual faz parte, como espera-se que seja, é possível que a sentença representada por ele tenha um bom nível de informatividade. De acordo com esse raciocínio, nos testes relatados no Capítulo 5, dá-se prioridade às sentenças com alto C_i , ou C_i^w , na confecção de um extrato.

alguma) à rede de Skorochod'ko, e o ligamos a todos os outros nós já existentes na rede, o cálculo do coeficiente de aglomeração do nó i é idêntico ao cálculo da ligação semântica de um texto.

⁵Apud (Costa et al., 2006b).

4.2.3 Caminhos Mínimos

Caminhos mínimos, ou mais geralmente, medidas relacionadas à distância entre vértices, são importantes pois consideram a estrutura global de uma rede (Costa et al., 2006b). Um caminho entre dois vértices é uma seqüência de arestas que leva um vértice a outro, e o comprimento do caminho é o número de arestas contidas na seqüência. Um caminho mínimo que parte do nó i ao j , denotado por d_{ij} , é aquele com comprimento mínimo, e pode ser calculado por meio da matriz A . Se tomarmos todos os caminhos mínimos associados a um determinado nó i , temos a medida de distância média

$$sp_i = \frac{1}{N-1} \sum_{i \neq j} d_{ij} = \frac{1}{N-1} \sum_{i \neq j} d_{ji}, \quad (4.5)$$

de maneira que, se N é o número total de vértices e se o caminho entre i e j não existir, então $d_{ij} = N$. Quanto menor o valor de sp_i , mais próximo ele está, em média, dos outros nós da rede. Para exemplificar essa medida, a rede da Figura 4.6 apresenta dois vértices em destaque. O vértice 1 está mais distante da maior parte dos outros vértices, e apresenta $sp_1 = 4,46$, ou seja, partindo-se de qualquer vértice são necessários, em média, 4,46 passos para se chegar ao vértice 1. Já o vértice 2 não está tão distante do restante da rede, e apresenta $sp_2 = 2,85$.

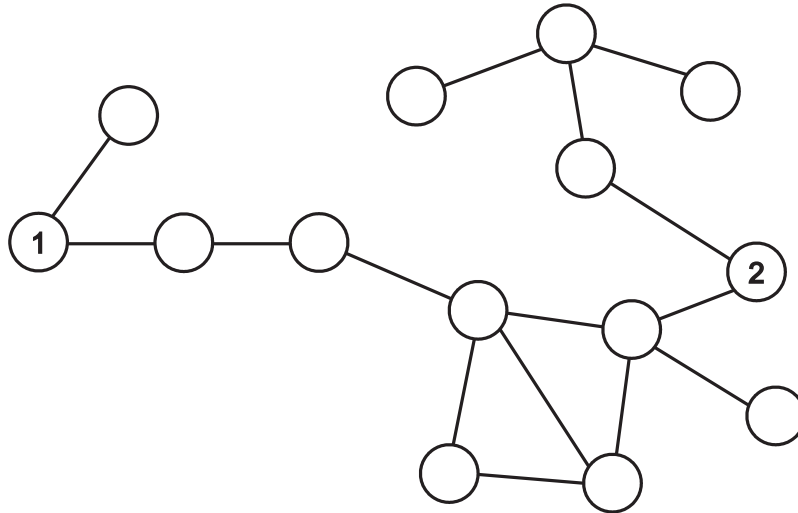


Figura 4.6: Vértices 1 e 2 com caminhos mínimos médios $sp_1 = 4,46$ e $sp_2 = 2,85$.

Em redes com pesos, a distância pode ser considerada igual ao somatório dos pesos das arestas que formam o caminho. Neste caso, arestas com alto peso tornam o caminho custoso, o que entra em contradição com a definição do peso de uma aresta dada na Seção 4.1. Consideramos que uma aresta com alto peso indica forte relação entre duas

sentenças e, portanto, deveria ser considerada mais vantajosa do que arestas com baixo peso. Para solucionar esse problema, foram utilizadas duas variações de sp_i que utilizam o peso das arestas no cálculo dos caminhos mínimos. A primeira delas utiliza uma matriz W^{wc} com elementos $w_{ij}^{wc} = 0$ se $w_{ij} = 0$, e $w_{ij}^{wc} = w_{max} - w_{ij} + 1$ se $w_{ij} > 0$, ou seja, utiliza o maior peso de W (denotado por w_{max}) para complementar os valores w_{ij} . As distâncias mínimas d_{ij}^{wc} , baseadas na rede representada por W^{wc} , são utilizadas no cálculo da medida de distância

$$sp_i^{wc} = \frac{1}{N-1} \sum_{i \neq j} d_{ij}^{wc} = \frac{1}{N-1} \sum_{i \neq j} d_{ji}^{wc}, \quad (4.6)$$

onde $d_{ij}^{wc} = N \bar{w}^{wc}$ quando o caminho entre i e j não existe (a média dos pesos de W^{wc} é denotada por \bar{w}^{wc}). A outra variação de sp_i aqui proposta considera o inverso dos pesos de W , ou seja, utiliza uma matriz W^{wi} com elementos $w_{ij}^{wi} = 0$ se $w_{ij} = 0$, e $w_{ij}^{wi} = 1/w_{ij}$ se $w_{ij} > 0$. Portanto, as distâncias mínimas d_{ij}^{wi} , baseadas na rede representada por W^{wi} , dão origem a outra medida:

$$sp_i^{wi} = \frac{1}{N-1} \sum_{i \neq j} d_{ij}^{wi} = \frac{1}{N-1} \sum_{i \neq j} d_{ji}^{wi}, \quad (4.7)$$

onde $d_{ij}^{wi} = N \bar{w}^{wi}$ quando o caminho entre i e j não existe.

As três medidas de caminhos mínimos, sp_i , sp_i^{wc} e sp_i^{wi} , servem para mensurar o quão distante um determinado nó i está dos demais nós da rede. Altos valores para essas medidas indicam que, partindo do nó i , é custoso chegar até outro nó percorrendo as arestas da rede. Esse tipo de vértice está, de certa maneira, afastado do restante da rede (e das informações veiculadas nas demais sentenças), o que foi considerado neste projeto como algo ruim para a informatividade de sumários. Ao contrário, um vértice que está mais próximo dos demais pode tratar de idéias relacionadas a boa parte do texto e, portanto, representaria uma sentença mais informativa e útil do ponto de vista da sumarização. Dessa maneira, considera-se que as sentenças com os mais baixos valores de sp_i , sp_i^{wc} ou sp_i^{wi} devam compor o extrato. Por fim, é importante observar que as medidas baseadas em distância mínima são bastante sensíveis, no sentido de que uma pequena alteração na conectividade da rede pode acarretar grandes mudanças nos comprimentos dos caminhos mínimos.

4.2.4 Índice de Localidade

Assim como o coeficiente de aglomeração, o índice de localidade (*locality index*) é utilizado na análise das conexões existentes entre os vizinhos de um determinado nó (Costa et al., 2006a). No entanto, o índice de localidade leva em consideração todas as conexões

desses nós vizinhos, e não somente as conexões existentes entre eles (identificadas por E_i na definição do coeficiente de aglomeração). O número de conexões contidas na sub-rede formada pelo nó i e seus k_i vizinhos é denotada por N_i^{int} (são as chamadas conexões internas). O nó i é incluído no cômputo de N_i^{int} , evitando assim uma singularidade quando $k_i = 1$. O número de conexões externas, simbolizado por N_i^{ext} , é igual ao número de conexões que os k_i vizinhos do nó i têm com os demais nós da rede. O índice de localidade é igual a

$$l_i = \frac{N_i^{int}}{N_i^{int} + N_i^{ext}}, \quad (4.8)$$

onde $0 < l_i \leq 1$. Se o número de conexões externas for nulo, o índice de localidade é máximo. Por outro lado, se as conexões externas existirem em número bem maior que as conexões internas, l_i tende a zero. A Figura 4.7 mostra dois vértices com índices de localidade variados ($l_1 = 0,44$ e $l_2 = 0,73$). Note que os vertices 1 e 2 dessa figura têm o mesmo grau k_i e o mesmo coeficiente de aglomeração C_i . O que os diferencia é o número de conexões externas (indicadas na Figura 4.7 por linhas tracejadas), que é menor no caso do vértice 2, fazendo com que seu índice de localidade seja maior que o do vértice 1.

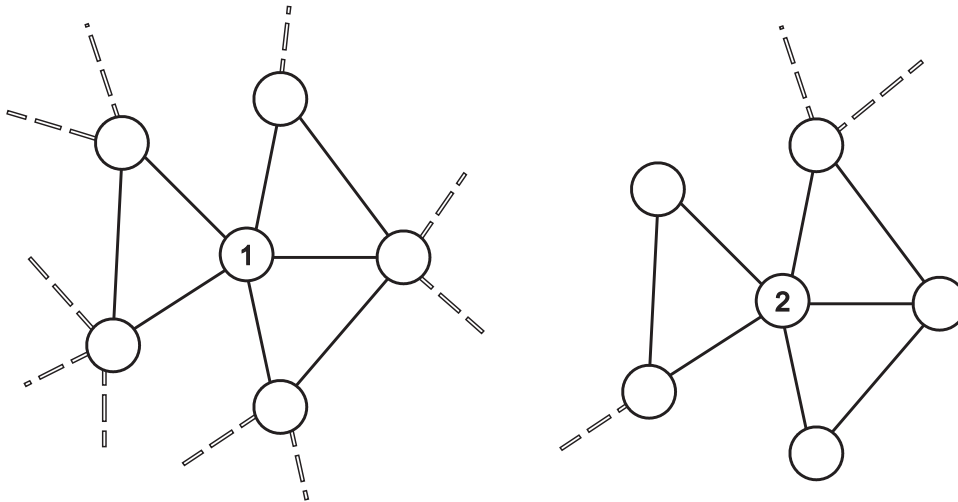


Figura 4.7: Vértices 1 e 2 com índices de localidade $l_1 = 0,44$ e $l_2 = 0,73$. As arestas tracejadas conectam os vizinhos de 1 e de 2 aos demais nós da rede (conexões externas).

Quanto à tarefa de sumarização, considerou-se que sentenças com alto índice de localidade são mais aptas a formar um extrato do que as com baixo índice de localidade. As sentenças com alto l_i formam pequenos agrupamentos que compartilham poucas arestas (ou nenhuma) com o restante da rede. Esses agrupamentos locais podem ser bem representados por seu nó central, pois ele apresenta conexões com todos os nós do grupo. Se cada grupo desses for considerado um conjunto coeso de sentenças, a sentença central pode ser tomada como informativa de todo o agrupamento. Dessa maneira, considera-se que uma sentença

com alto l_i deva compor um extrato. Em alguns casos, entretanto, mais de um vértice do agrupamento centrado no vértice i pode ter um alto índice de localidade. Tomemos dois nós, um nó j vizinho de um nó i , ambos com altos l_i e l_j . Levando-se em consideração apenas o índice de localidade, a probabilidade de i e j serem incluídos em um extrato é alta. O fato de i e j fazerem parte de um mesmo extrato gera uma certa redundância, pois os dois nós estão contidos em um mesmo agrupamento. Com o intuito de evitar esse comportamento, foi proposta uma variação da técnica de sumarização que utiliza a medida l_i , que funciona da seguinte maneira:

- Um nó i com alto l_i somente é adicionado ao extrato se nenhum de seus k_i vizinhos já estiver no extrato.
- Se um nó com alto índice de localidade for descartado, ele é armazenado em uma fila L .
- Quando todos os nós já tiverem sido analisados, e ainda for necessário incluir alguma sentença no extrato, as sentenças da fila L são utilizadas.

Com essa variação, ao selecionar-se um nó i para compor o extrato, associa-se a ele um número inteiro positivo z_i , indicando que esse nó representa a z_i -ésima sentença a ser adicionada a um extrato. Desconsiderando-se a taxa de compressão, todas as sentenças são numeradas sequencialmente de 1 até N . Portanto, a modificação da primeira técnica definida com o índice de localidade usa a numeração sequencial z_i :

$$l_i^{mod} = z_i, \quad (4.9)$$

sendo que os nós com os menores valores de l_i^{mod} são considerados prioritários na construção de um extrato.

À primeira vista, a utilização da numeração z_i pode parecer desnecessária, pois o algoritmo de sumarização não depende de z_i . Contudo, associar um valor a cada vértice permite que os diversos sumarizadores propostos nesta seção sejam comparados entre si por meio da análise de correlação entre medidas, como é mostrado na Seção 5.5. A numeração z_i pode, portanto, ser considerada uma medida derivada de um algoritmo de seleção de vértices. Além disso, a numeração z_i é utilizada diversas vezes nesta dissertação, sem o propósito, entretanto, de ser redundante. Seu uso é justificado pelo fato de z_i destacar que uma numeração sequencial é aplicada na pontuação dos vértices, ao invés de fórmulas como nos casos do grau e do coeficiente de aglomeração.

4.2.5 Índice de Concordância

O índice de concordância (*matching index*) é usado para comparar a conectividade de dois nós ligados por uma aresta (Kaiser e Hilgetag, 2004)⁶. Portanto, esta é uma medida aplicada a cada aresta (i,j) da rede. Ao comparar-se a conectividade de i e j , calcula-se a quantidade de nós que estão conectados simultaneamente a i e a j , e divide-se a mesma pelo número total de conexões de ambos os vértices, excluindo-se a conexão entre i e j ,

$$\mu_{ij} = \frac{\sum_{k \neq i,j} a_{ik} a_{jk}}{\sum_{k \neq j} a_{ik} + \sum_{k \neq i} a_{jk}}, \quad (4.10)$$

onde $0 \leq \mu_{ij} \leq 0,5$. Um valor baixo de μ_{ij} indica que a aresta (i,j) une duas regiões distintas da rede, pois os nós i e j compartilham um número relativamente pequeno de vizinhos. Ao contrário, quando μ_{ij} é alto, significa que i e j têm um padrão semelhante de conexões. A Figura 4.8 ilustra dois casos extremos do índice de concordância. O primeiro deles refere-se à aresta $(1,2)$, com índice de concordância $\mu_{12} = 0$, pois nenhum vizinho do vértice 1 é também vizinho do vértice 2, e vice-versa. O segundo caso, referente à aresta $(3,4)$, apresenta índice de concordância máximo ($\mu_{34} = 0,5$), pois todos os vizinhos do vértice 3 são também vizinhos do vértice 4.

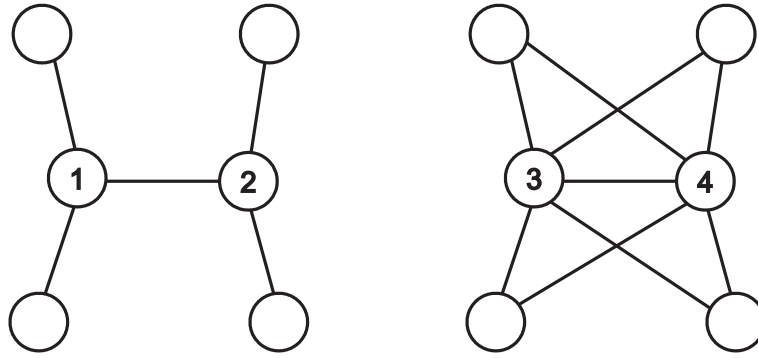


Figura 4.8: Arestas $(1,2)$ e $(3,4)$ com índices de concordância $\mu_{12} = 0$ e $\mu_{34} = 0,5$.

Ao aplicar esta medida na sumarização, preferiu-se incluir em um extrato pares de nós com baixo índice de concordância. Dessa maneira, preferência é dada a vértices que, apesar de conectados entre si, ligam-se a diferentes grupos de vértices, proporcionando um sumário teoricamente mais abrangente e informativo. Assume-se que, quando μ_{ij} é alto, i e j representam sentenças redundantes, pois são representadas por vértices que apresentam praticamente os mesmos vizinhos. Entretanto, μ_{ij} é uma medida aplicada a arestas. Para utilizá-la na sumarização, e também transformá-la em uma medida aplicada a vértices, define-se o seguinte procedimento:

⁶Apud (Costa et al., 2006b).

- Percorre-se a lista de arestas (i,j) , ordenada crescentemente pelos valores μ_{ij} .
- Para cada aresta visitada, adicionam-se os nós i e j ao extrato. Se um desses nós já estiver incluso no extrato, ele não é novamente inserido.
- Para cada vértice i que acaba de ser inserido no extrato, associa-se um número inteiro seqüencial z_i , iniciado em 1.

Se desconsiderarmos a taxa de compressão, é possível fazer com que a numeração z_i siga de 1 até N . Conseqüentemente, o sumarizador baseado no índice de concordância é fundamentado na medida

$$m_i = z_i, \quad (4.11)$$

onde os extratos são construídos selecionando-se sentenças com os menores valores de m_i .

4.2.6 Grau Hierárquico

A noção de grau hierárquico está relacionada à operação chamada dilatação (*dilation*) (Costa e da Rocha, 2006). A dilatação $\delta(g)$ de um subgrafo g é o subgrafo que contém os vértices de g mais os vértices conectados aos vértices de g . A d -dilatação de um subgrafo g é a aplicação de $\delta(g)$ por d vezes:

$$\delta_d(g) = \underbrace{\delta(\delta(\dots(g)\dots))}_d, \quad (4.12)$$

sendo que $\delta_0 = g$. O d -anel (*d-ring*) de um subgrafo g é um subgrafo $R_d(g)$ de vértices

$$\mathcal{N}(\delta_d(g)) \setminus \mathcal{N}(\delta_{d-1}(g)), \quad (4.13)$$

onde \setminus é a operação diferença de conjuntos, $\mathcal{N}(G)$ é o conjunto de vértices de um grafo G e $R_0 = g$. O d -anel de g é a hierarquia nível d , obtida a partir de d dilatações do subgrafo g . Quando g é formado por um único vértice i , então usa-se $R_d(i)$ ao invés de $R_d(g)$. A Figura 4.9 ilustra os dois primeiros níveis hierárquicos ($R_1(1)$ e $R_2(1)$) do vértice em destaque (o de índice 1).

O grau hierárquico de um nó i ao nível d , denotado por k_i^d , é definido como o número de arestas da rede original que conectam os anéis $R_{d-1}(i)$ e $R_d(i)$, onde $d \geq 1$. Note que $k_i^1 = k_i$, ou seja, o grau hierárquico nível 1 é igual ao grau tradicional. Na Figura 4.9, tem-se que $k_1^1 = 5$ e $k_1^2 = 9$ (as arestas que ligam vértices de um mesmo nível não são utilizadas no cálculo do grau hierárquico). Ao somarmos os pesos das arestas da rede original que

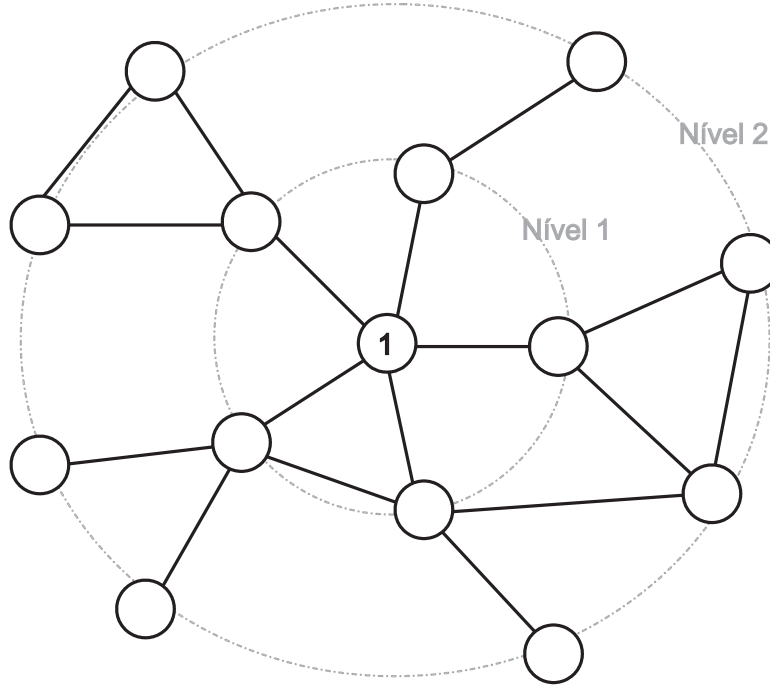


Figura 4.9: Vértice 1 e suas duas primeiras hierarquias.

conectam os anéis $R_{d-1}(i)$ e $R_d(i)$, temos o *strength* hierárquico s_i^d . Analogamente, $s_i^1 = s_i$. Por fim, são definidos os graus hierárquicos cumulativos nível d :

$$k_i^{d,c} = \sum_{n=1}^d k_i^n \quad (4.14)$$

e

$$s_i^{d,c} = \sum_{n=1}^d s_i^n, \quad (4.15)$$

onde $k_i^{1,c} = k_i^1 = k_i$ e $s_i^{1,c} = s_i^1 = s_i$. Essas medidas utilizam a soma dos graus de todos os níveis hierárquicos anteriores. Como exemplo de grau hierárquico cumulativo, tem-se que, na rede da Figura 4.9, o vértice 1 apresenta $k_1^{2,c} = 14$.

As medidas de grau hierárquico aplicadas neste projeto foram $k_i^2, k_i^{2,c}, k_i^3, k_i^{3,c}, s_i^2, s_i^{2,c}, s_i^3$ e $s_i^{3,c}$, ou seja, foram computadas métricas hierárquicas até o nível 3. Vértices com grau hierárquico elevado não necessariamente têm um grande número de vizinhos, pois boa parte das conexões podem estar presentes em níveis hierárquicos mais altos. Com a aplicação dessas oito medidas de grau hierárquico, objetiva-se complementar os graus tradicionais k_i e s_i , e oferecer maneiras de se capturar a conectividade dos nós em vizinhanças mais distantes. Como os níveis hierárquicos considerados não são muito distantes (níveis 2 e 3), considera-se que as vizinhanças capturadas pelos graus hierárquicos tenham algum tipo

de relação, mesmo que indireta, com o nó central da hierarquia. Da mesma maneira que para o grau tradicional, considera-se que sentenças com alto grau hierárquico em níveis 2 e 3 possam ser mais informativas, por estarem relacionadas a uma grande quantidade de sentenças em vizinhanças próximas. Portanto, ao construir extratos utilizando essas medidas, nós com alto grau hierárquico têm preferência.

4.2.7 d -Anéis

O conceito de d -anel, utilizado na definição dos graus hierárquicos, foi empregado na elaboração de um outro algoritmo de sumarização. Nele, são computados todos os anéis $R_d(i)$ para o nó mais conectado (com maior k_i , chamado *hub*). Como $i = \text{hub}$, denota-se esses anéis particulares por $R_d(\text{hub})$, lembrando que $\mathcal{N}(R_0(\text{hub})) = \{\text{hub}\}$. No cálculo de todos os d -anéis do *hub*, obtém-se uma tupla

$$\mathcal{T} = (\mathcal{N}(R_0(\text{hub})), \mathcal{N}(R_1(\text{hub})), \dots, \mathcal{N}(R_{d_{\max}}(\text{hub}))), \quad (4.16)$$

onde $\mathcal{N}(G)$ é o conjunto de vértices de uma rede G . Possivelmente, o subconjunto de vértices

$$\tau = \{1, \dots, N\} \setminus \bigcup_{i=0}^{d_{\max}} \mathcal{N}(R_i(\text{hub})) \quad (4.17)$$

não é vazio, indicando que um ou mais vértices não fazem parte de algum anel $R_d(\text{hub})$ (isso acontece quando a rede é desconexa). Utilizando \mathcal{T} e τ , o algoritmo de sumarização funciona da seguinte maneira:

- Cada conjunto $\mathcal{N}(R_d(\text{hub}))$ tem seus elementos ordenados crescentemente, onde cada elemento é o índice i do vértice. Dessa maneira, as sentenças que aparecem primeiro no texto-fonte ocupam as primeiras posições na ordenação (vide Seção 4.1), e receberão tratamento prioritário no algoritmo de sumarização⁷.
- A seguir, percorre-se os elementos de \mathcal{T} , partindo de $\mathcal{N}(R_0(\text{hub}))$ a $\mathcal{N}(R_{d_{\max}}(\text{hub}))$, selecionando uma sentença por vez, de cada um dos anéis, na ordem definida no passo anterior. Cada sentença selecionada recebe uma numeração inteira seqüencial z_i , iniciada em 1.
- Se $\tau \neq \{\}$, seus elementos são selecionados, e recebem uma numeração z_i , dando seqüência aos vértices já numerados de \mathcal{T} .

⁷Edmundson (1969) e Kupiec et al. (1995) (Seção 2.1) mostram que o atributo de localização é bastante útil na sumarização extrativa.

Essa numeração dá origem à medida inspirada nos d -anéis e na localização das sentenças:

$$r_i^l = z_i, \quad (4.18)$$

de maneira que as sentenças com os menores valores de r_i^l são escolhidas na formação do extrato. Com essa técnica, primeiramente são colocados no extrato o *hub* e seus vizinhos mais próximos. Se considerarmos que o *hub* seja a sentença mais importante do texto, é natural que adicionemos ao extrato seus vizinhos, com o intuito de complementar as informações contidas no *hub* e assim deixar o extrato mais informativo. Como outros níveis hierárquicos são utilizados, sentenças relacionadas aos vizinhos dos vizinhos (e assim por diante) do *hub* possivelmente são inseridas no extrato, fazendo com que todas as sentenças tenham alguma tipo de relação com o *hub*, o que pode contribuir para a coesão do sumário. Se, devido à taxa de compressão, for necessário escolher um subconjunto das sentenças contidas em algum $R_d(\text{hub})$, o passo inicial de ordenação garante que as sentenças desse conjunto que aparecem primeiro no texto-fonte sejam selecionadas. A fase de ordenação é importante, pois, em alguns casos, somente parte de um anel pode ser incluída no extrato. Propôs-se, alternativamente, que os vértices de todos os anéis fossem ordenados pelo grau k_i , de forma decrescente, fazendo com que seja dada preferência aos vértices mais conectados dos anéis quando houver impossibilidade de incluir todo o anel no extrato. A numeração z_i adquire então outro sentido, e passa a ser utilizada na definição da medida inspirada nos d -anéis e no grau dos nós:

$$r_i^k = z_i, \quad (4.19)$$

sendo que os vértices com os valores mais baixos de r_i^k devem compor o sumário.

Uma última medida inspirada nos d -anéis foi proposta, desta vez utilizando tanto a localização das sentenças quanto os graus. Calcula-se inicialmente o grau médio \bar{k} e divide-se cada $R_d(\text{hub})$ em duas partes, tal que

$$\mathcal{N}(R_d(\text{hub})) = \mathcal{N}(R_d^K(\text{hub})) \cup \mathcal{N}(R_d^k(\text{hub})), \quad (4.20)$$

onde $R_d^K(\text{hub})$ contém os nós $i \in R_d(\text{hub})$ tal que $k_i \geq \bar{k}$, e $R_d^k(\text{hub})$ contém os nós $i \in R_d(\text{hub})$ tal que $k_i < \bar{k}$. A tupla \mathcal{T} é então redefinida da seguinte forma:

$$\mathcal{T} = (\mathcal{N}(R_0^K(\text{hub})), \dots, \mathcal{N}(R_{d_{\max}}^K(\text{hub})), \mathcal{N}(R_0^k(\text{hub})), \dots, \mathcal{N}(R_{d_{\max}}^k(\text{hub}))). \quad (4.21)$$

Os vértices contidos nos conjuntos que formam a tupla \mathcal{T} são então selecionados e numerados sequencialmente (z_i passa a ter outro significado), da mesma maneira que o primeiro algoritmo definido nesta seção, ou seja, utilizando a ordenação por localização das senten-

ças no texto-fonte. Com a alteração de \mathcal{T} , nós com grau abaixo da média dão lugar aos nós com grau acima da média, mesmo que estes últimos estejam em hierarquias mais distantes do *hub* que os primeiros. O conjunto τ é redefinido e utilizado de maneira análoga. Portanto, a medida inspirada nos d -anéis, na localização das sentenças e no grau dos vértices é

$$r_i^{l,k} = z_i, \quad (4.22)$$

utilizada na construção de extratos da mesma maneira que r_i^l e r_i^k .

É importante ressaltar que os sumarizadores baseados nos d -anéis são intimamente relacionados à técnica utilizada no sistema GistSumm (Pardo et al., 2003a) (vide Seção 2.1). Ambas as propostas primeiramente selecionam a sentença (teoricamente) mais importante do texto-fonte (*gist sentence* ou *hub*) e, a seguir, selecionam as sentenças que estejam a ela relacionadas. GistSumm e d -anéis apresentam, portanto, uma grande similaridade conceitual, guardadas as devidas diferenças de implementação.

4.2.8 k -Núcleos

O k -núcleo (k -core) de um grafo G é o subgrafo $core_k(G)$ tal que para todo vértice i de $core_k(G)$, $k_i \geq k$, ou seja, todos seus vértices têm grau no mínimo k . Além disso, $core_k(G)$ é o maior subgrafo de G com essa propriedade (Batagelj e Zaversnik, 1999). Para obter o k -núcleo, elimina-se da rede, recursivamente, todos os vértices com grau abaixo de k . A Figura 4.10 mostra um k -núcleo com $k = 4$. Não necessariamente $core_k(G)$ forma um componente conexo, por isso, denota-se o maior componente conexo do k -núcleo pelo subgrafo $core'_k(G)$ de $\mathcal{N}(core_k(G))$ vértices. Considera-se aqui, para fins de sumarização, que o $core'_k(G)$ não vazio de maior k contém vértices importantes do texto representado por G . Tal núcleo é interessante, pois representa um subgrafo conectado com vértices de, possivelmente, alto grau (ou seja, representa um grupo de sentenças fortemente coesas). A informatividade dos sumários seria garantida pela presença de vértices com alto grau, isto é, de sentenças que possuem diversas conexões com outras sentenças. Um ponto negativo do $core'_k(G)$ não vazio de maior k , para a sumarização, seria a possibilidade de haver redundância de informações entre suas sentenças, pelo fato de haver muitas conexões entre elas. Entretanto, como dificilmente duas sentenças de um mesmo texto veiculam as mesmas informações, tal $core'_k(G)$ deve ainda conter sentenças complementares.

Variando-se o índice k de k_{max} até 1 (onde k_{max} é o maior grau presente na rede), diminui-se a importância de $core'_k(G)$, de acordo com a suposição aqui feita. Em outras palavras, vértices que aparecem apenas em subgrafos $core'_k(G)$ de k baixo têm menor prioridade no processo de construção de um sumário. Seguindo essa sequência de diminuição

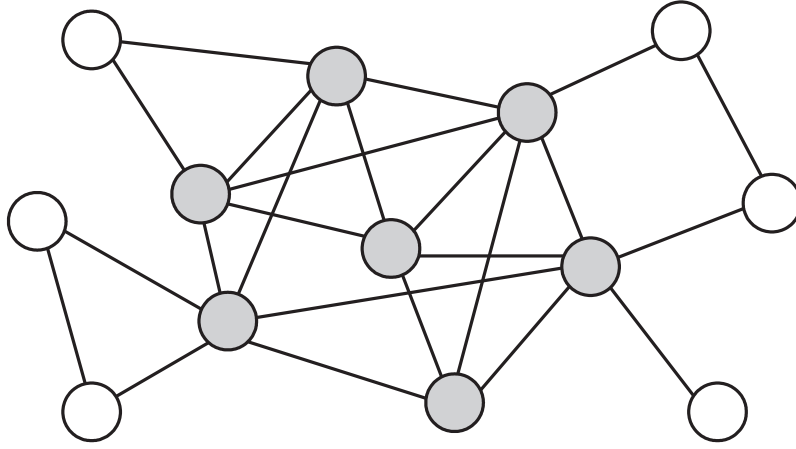


Figura 4.10: k -Núcleo com $k = 4$, identificado pelos vértices em cinza.

do índice k , o conjunto

$$\lambda_k(G) = \mathcal{N}(\text{core}'_k(G)) \setminus \bigcup_{i=k+1}^{k_{\max}} \mathcal{N}(\text{core}'_i(G)), \quad (4.23)$$

é formado pelos vértices de $\text{core}'_k(G)$ menos os vértices dos k -núcleos anteriores. Define-se então a tupla

$$\mathcal{T} = (\lambda_{k_{\max}}(G), \lambda_{k_{\max}-1}(G), \dots, \lambda_1(G)), \quad (4.24)$$

formada pelos conjuntos de vértices provenientes dos $\text{core}'_k(G)$. Os vértices que não estão contidos em algum conjunto de \mathcal{T} são definidos como

$$\tau = \{1, \dots, N\} \setminus \bigcup_{i=1}^{k_{\max}} \lambda_i(G). \quad (4.25)$$

Utiliza-se \mathcal{T} e τ do mesmo modo que para as medidas inspiradas nos d -anéis, r_i^l e r_i^k . Seleciona-se cada nó, de cada conjunto da tupla \mathcal{T} , em ordem de prioridade por localização da sentença no texto-fonte ou por grau k_i , associando a cada sentença um número z_i . Se τ não for vazio, seus nós também são selecionados. Portanto, se cada elemento de \mathcal{T} e de τ for ordenado pela localização, de forma crescente, então obtém-se a medida

$$n_i^l = z_i, \quad (4.26)$$

enquanto que, se cada elemento de \mathcal{T} for ordenado pelo grau, de forma decrescente, tem-se

$$n_i^k = z_i, \quad (4.27)$$

sendo que z_i adquire outro sentido quando a ordenação dos elementos de cada $\lambda_k(G)$ é alterada. Por fim, supondo que os k -núcleos de maior k são mais interessantes para a sumarização, dá-se prioridade aos nós de menor n_i^l , ou n_i^k , na construção de um extrato.

4.2.9 w -Cortes

O w -corte de um grafo G , denotado por $cut_w(G)$, foi aqui definido como sendo o maior componente conexo de G após a eliminação das arestas (i,j) com $w_{ij} < w$. A Figura 4.11 mostra um w -corte com $w = 3$. Note que, nessa figura, dois vértices que estão unidos por uma aresta de peso 4 não fazem parte do w -corte, pelo fato de não estarem incluídos no maior componente conexo após a eliminação das arestas com peso menor que 3.

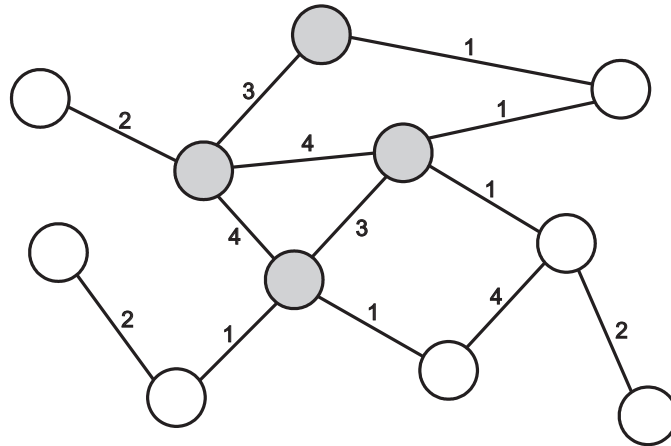


Figura 4.11: w -Corte com $w = 3$, identificado pelos vértices em cinza.

Quanto à sumarização, $cut_w(G)$ tem papel semelhante a $core'_k(G)$, pois ambos apresentam grupos de vértices coesos quando k ou w é alto. Portanto, analogamente, ao variarmos w de w_{max} até 1 (w_{max} é o maior peso da matriz W), obtém-se w -cortes cada vez maiores, com vértices que não figuram nos w -cortes de alto w . Temos então a definição do conjunto $\lambda_w(G)$:

$$\lambda_w(G) = \mathcal{N}(cut_w(G)) \setminus \bigcup_{i=w+1}^{w_{max}} \mathcal{N}(cut_i(G)), \quad (4.28)$$

restando, portanto, apenas vértices que não aparecem em w -cortes mais restritos. A tupla \mathcal{T} , cuja seqüência de conjuntos é aplicada diretamente à sumarização, passa a ser igual a

$$\mathcal{T} = (\lambda_{w_{max}}(G), \lambda_{w_{max}-1}(G), \dots, \lambda_1(G)). \quad (4.29)$$

Os nós isolados, que não entram em w -corte algum, formam o conjunto

$$\tau = \{1, \dots, N\} \setminus \bigcup_{i=1}^{w_{max}} \lambda_i(G). \quad (4.30)$$

Analogamente às medidas anteriores, ordena-se os conjuntos que formam \mathcal{T} e o conjunto τ , e aplica-se uma numeração z_i seqüencial. As novas medidas p_i^l e p_i^k são obtidas a partir de z_i , respectivamente, por ordenação guiada por localização ou por grau, e são utilizadas na sumarização da mesma maneira que n_i^l e n_i^k .

4.2.10 Comunidades

Outro conceito bastante utilizado nos estudos em Redes Complexas é o de comunidades, grupos de vértices arranjados de maneira que exista uma maior densidade de conexões dentro dos grupos do que entre grupos (Clauset et al., 2004). A Figura 4.12 mostra uma rede dividida em três comunidades. Não existe definição precisa do que seja uma comunidade. Uma divisão em comunidades adquire sentido ao se analisar o significado dos vértices presentes em cada grupo. Para uma rede considerada neste trabalho, considera-se que uma boa divisão em comunidades possa refletir a divisão de tópicos do texto, sendo que uma boa partição do conjunto de nós seja a que apresente modularidade alta (detalhes a seguir) e um tópico do texto seja formado por sentenças que tratam do mesmo assunto. O algoritmo de sumarização baseado em comunidades procura selecionar sentenças de todos os tópicos, em número proporcional ao tamanho de cada tópico. Dessa maneira, pretende-se obter um sumário bem informativo, que cubra os tópicos de todas as comunidades. Entretanto, a associação comunidade-tópico é uma suposição, e não foi realizada uma avaliação intrínseca da divisão dos textos em tópicos. A avaliação é sim extrínseca, ou seja, é realizada dentro da tarefa de sumarização automática. O algoritmo TextTiling (Hearst, 1997) também serve o propósito de dividir um texto em tópicos, e foi utilizado por Larocca Neto et al. (2000a) na construção de extratos (vide Seção 2.1).

A modularidade é uma medida a respeito da divisão de uma rede em comunidades, e serve para analisar o número de arestas dentro das comunidades com relação ao número de arestas presentes entre comunidades. A seguinte fração é utilizada como ponto de partida na definição da modularidade:

$$\frac{\sum_{ij} a_{ij} \delta(c_i, c_j)}{\sum_{ij} a_{ij}} = \frac{1}{2M} \sum_{ij} a_{ij} \delta(c_i, c_j), \quad (4.31)$$

onde c_i é o número da comunidade a que o nó i pertence, $\delta(a, b)$ é igual a 1 se $a = b$ ou igual a 0 se $a \neq b$, e M é o número de arestas presentes na rede ($M = \frac{1}{2} \sum_{ij} a_{ij}$). Esta

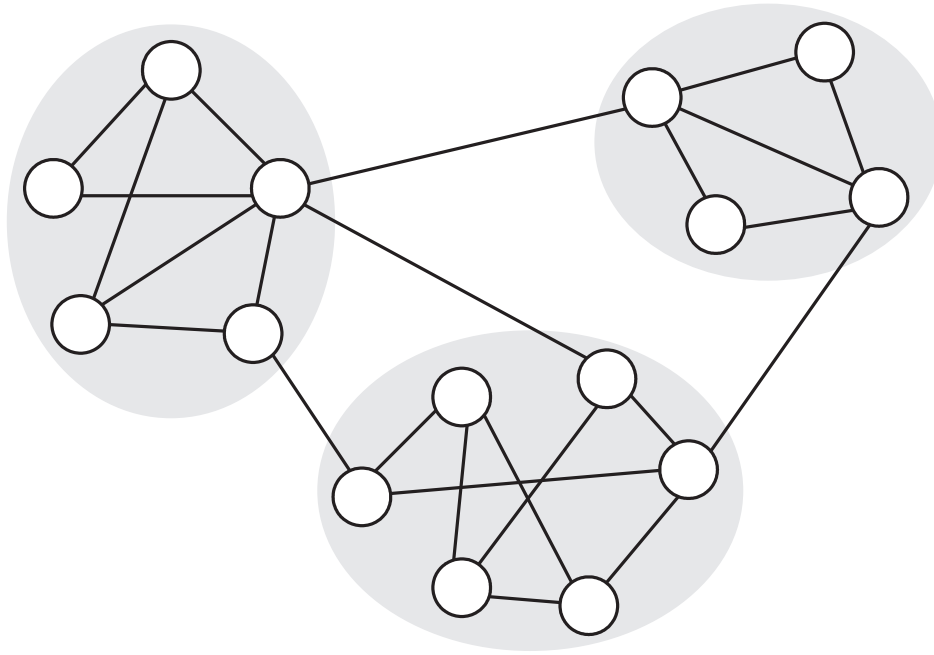


Figura 4.12: Exemplo de divisão de uma rede em três comunidades (áreas em cinza).

fração é grande se a divisão em comunidades é boa, mas ela apresenta um problema se for tomada como uma medida de modularidade pois, se considerarmos uma única comunidade que contém todos os nós da rede, o valor desta fração é máximo (igual a 1). Para contornar esse problema, a medida Q de modularidade em uma rede G é definida subtraindo-se da fração 4.31 o valor esperado dessa mesma quantidade em uma rede aleatória:

$$Q = \frac{1}{2M} \sum_{ij} \left[a_{ij} - \frac{k_i k_j}{2M} \right] \delta(c_i, c_j), \quad (4.32)$$

onde k_i é o grau do nó i e $k_i k_j / 2M$ é a probabilidade de existir uma aresta (i, j) em uma rede aleatória que preserva os graus dos vértices da rede G . Quando $Q > 0$ a modularidade é maior do que a esperada em uma versão aleatória de G , e um valor de Q acima de 0,3 indica que a rede em questão apresenta uma estrutura modular significativa (Clauset et al., 2004). A modularidade Q pode ser utilizada da seguinte maneira na identificação de comunidades: inicialmente, considera-se que cada nó esteja em uma comunidade diferente, para, a seguir, unir duas dessas comunidades em uma única comunidade, de tal maneira que o novo valor de Q seja o maior possível. As uniões de comunidades são realizadas até que exista uma única comunidade que contenha todos os vértices da rede. Clauset et al. (2004) definiram um algoritmo baseado nessa idéia, o qual é eficiente para redes grandes e esparsas (quando $M \approx N$), e disponibilizam uma implementação⁸ do mesmo.

⁸<http://cs.unm.edu/~aaron/research/fastmodularity.htm>

A divisão de uma rede G em comunidades é aquela com maior Q , onde cada comunidade c com mais de um nó é denotada pelo conjunto de nós ρ_c . A tupla

$$\mathcal{K} = (\rho_1, \rho_2, \dots, \rho_{n_c}), \quad (4.33)$$

onde n_c é o número de comunidades não unitárias, está ordenada da comunidade ρ_1 com o maior número de vértices para a comunidade ρ_{n_c} com o menor número de vértices. O tamanho da comunidade i , com relação à comunidade de menor tamanho ρ_{n_c} , é dado por

$$h_i = \text{round} \left(\frac{\|\rho_i\|}{\|\rho_{n_c}\|} \right), \quad (4.34)$$

onde $\|\rho_i\|$ denota o número de elementos do conjunto ρ_i e a função $\text{round}(a)$ faz o arredondamento de um número real a . Esse número inteiro indica que a comunidade ρ_i é, aproximadamente, h_i vezes maior que a menor comunidade não unitária. O algoritmo de sumarização funciona da seguinte maneira, onde τ é o conjunto de vértices que formam comunidades unitárias:

- Ordena-se os elementos dos conjuntos ρ_i de forma decrescente, de acordo com o grau k_i dos nós.
- Percorre-se as comunidades da tupla \mathcal{K} , selecionando os primeiros h_i elementos de cada comunidade ρ_i (ordenada pelo grau), com i variando de 1 até n_c , e numerando cada nó sequencialmente (numeração denotada por z_i). A seguir, cada um dos h_i nós é excluído de cada conjunto ρ_i .
- A tupla \mathcal{K} é percorrida enquanto houver algum conjunto ρ_i não vazio.
- Por fim, os elementos de τ são selecionados e numerados, dando preferência aos vértices de maior grau.

A medida inspirada na divisão de comunidades é

$$g_i = z_i, \quad (4.35)$$

e os vértices com os menores valores de g_i são escolhidos para compor um extrato. Dessa maneira, cada comunidade fornece ao extrato, aproximadamente, um número de sentenças proporcional ao seu tamanho. Além disso, cada comunidade contribui com seus vértices mais conectados. Por fim, a localização das sentenças não foi utilizada na ordenação dos elementos das comunidades ρ_i , pois implicaria em um algoritmo próximo ao que seleciona as primeiras sentenças do texto-fonte para compor um sumário, utilizado como sistema

baseline nas avaliações do Capítulo 5. Se, ao selecionar sentenças de cada comunidade, cada uma delas contribuir com seus vértices de índice i mais baixo, então é provável que o extrato seja formado pelas primeiras sentenças do texto-fonte.

Na Tabela 4.1 encontram-se listadas todas as medidas utilizadas nos experimentos de sumarização automática relatados no próximo capítulo.

Tabela 4.1: Lista de medidas utilizadas nos experimentos de sumarização, com símbolo e nome. Cada medida associa um valor ξ_i a cada nó i de uma rede, e pode ter sentido de aplicação crescente (\Uparrow), com prioridade para os nós de baixo valor ξ_i , ou decrescente (\Downarrow), com prioridade para os nós de alto valor ξ_i .

	Símbolo	Nome	Sentido
1	k_i	Grau	\Downarrow
2	s_i	Grau (<i>com Pesos</i>)	\Downarrow
3	C_i	Coefficiente de Aglomeração	\Downarrow
4	C_i^w	Coefficiente de Aglomeração (<i>com Pesos</i>)	\Downarrow
5	sp_i	Caminhos Mínimos	\Uparrow
6	sp_i^{wc}	Caminhos Mínimos (<i>Complemento dos Pesos</i>)	\Uparrow
7	sp_i^{wi}	Caminhos Mínimos (<i>Inverso dos Pesos</i>)	\Uparrow
8	l_i	Índice de Localidade	\Downarrow
9	l_i^{mod}	Índice de Localidade (<i>Modificado</i>)	\Uparrow
10	m_i	Índice de Concordância	\Uparrow
11	k_i^2	Grau Hierárquico (<i>Nível 2</i>)	\Downarrow
12	$k_i^{2,c}$	Grau Hierárquico (<i>Nível 2, Cumulativo</i>)	\Downarrow
13	k_i^3	Grau Hierárquico (<i>Nível 3</i>)	\Downarrow
14	$k_i^{3,c}$	Grau Hierárquico (<i>Nível 3, Cumulativo</i>)	\Downarrow
15	s_i^2	Grau Hierárquico (<i>Nível 2, com Pesos</i>)	\Downarrow
16	$s_i^{2,c}$	Grau Hierárquico (<i>Nível 2, com Pesos, Cumulativo</i>)	\Downarrow
17	s_i^3	Grau Hierárquico (<i>Nível 3, com Pesos</i>)	\Downarrow
18	$s_i^{3,c}$	Grau Hierárquico (<i>Nível 3, com Pesos, Cumulativo</i>)	\Downarrow
19	r_i^l	d -Anéis (<i>Ordenados por Localização</i>)	\Uparrow
20	r_i^k	d -Anéis (<i>Ordenados por Grau</i>)	\Uparrow
21	$r_i^{l,k}$	d -Anéis (<i>Ordenados por Localização, com Corte de Grau</i>)	\Uparrow
22	n_i^l	k -Núcleos (<i>Ordenados por Localização</i>)	\Uparrow
23	n_i^k	k -Núcleos (<i>Ordenados por Grau</i>)	\Uparrow
24	p_i^l	w -Cortes (<i>Ordenados por Localização</i>)	\Uparrow
25	p_i^k	w -Cortes (<i>Ordenados por Grau</i>)	\Uparrow
26	g_i	Comunidades	\Uparrow

Avaliação

As técnicas de sumarização apresentadas na Seção 4.2 foram aplicadas a três corpos de textos jornalísticos, após transformação dos textos-fonte em redes de sentenças, conforme metodologia apresentada na Seção 4.1. A qualidade dos extratos gerados, em termos de informatividade, foi avaliada pelas métricas ROUGE-1, Precisão, Cobertura e Medida-F, obtidas automaticamente e introduzidas neste capítulo, na Seção 5.1. Já os corpos utilizados são apresentados na Seção 5.2. A avaliação conduzida pode ser classificada como: intrínseca, pois os sumários são avaliados isoladamente, independentemente de alguma aplicação específica; *black-box*, pois apenas a entrada e a saída dos sumarizadores é avaliada, ignorando seus módulos internos; *off-line*, pois a avaliação é realizada de forma automática; e comparativa, pois os resultados de outros sistemas de sumarização são considerados¹ (veja Figura 2.5). Os resultados dos experimentos de avaliação, definidos na Seção 5.3, são relatados e discutidos na Seção 5.4. Na Seção 5.5, encontra-se uma análise das correlações entre os sumarizadores propostos, ou seja, verifica-se o caso de sumarizadores diferentes selecionarem as mesmas sentenças na construção de um extrato. Por fim, na Seção 5.6, são fornecidos e analisados alguns exemplos de extratos gerados por algumas das técnicas de sumarização propostas neste projeto.

¹Se apenas os sistemas aqui propostos forem analisados, considera-se que a avaliação não é comparativa.

5.1 Técnicas de Avaliação Automática

A avaliação de sumários é uma tarefa demasiadamente complexa e não padronizada, devido ao alto grau de subjetividade nela envolvida. Ela geralmente utiliza trabalho manual, o que demanda tempo e disponibilidade de mão-de-obra. A fim de se minimizar trabalho e tempo despendidos na avaliação de sumários, e também com o intuito de padronizar as métricas de avaliação de modo que diversos sistemas de sumarização automática sejam comparados de maneira mais justa, grande atenção tem sido voltada à criação e utilização de métodos de avaliação automática de sumários. Pode-se perceber que existe uma grande disparidade entre os métodos de avaliação aplicados aos sumarizadores já propostos (uma revisão desses sistemas, com suas respectivas avaliações, pode ser consultada nas Seções 2.1 e 2.2). Tendo em vista esse problema, procurou-se aqui utilizar técnicas de avaliação que permitissem que os resultados obtidos fossem comparados aos de outros sistemas. Duas abordagens de avaliação automatizada foram empregadas: (i) métricas de Precisão, Cobertura e Medida-F, e (ii) métrica ROUGE-1.

As métricas de Precisão (*Precision*) e Cobertura (*Recall*) são freqüentemente utilizadas na avaliação de sistemas de recuperação de informação (Salton e McGill, 1983). A unidade básica considerada nessas métricas, no caso da sumarização extrativa aqui realizada, é a sentença. Ao se avaliar um extrato automático por meio de Precisão e Cobertura, é preciso obter um outro extrato, considerado de boa qualidade, sobre o qual serão aplicadas as medidas. Para definir Precisão e Cobertura, o extrato de boa qualidade, chamado de extrato ideal ou de referência, é denotado por $E_r = \{s_1^r, s_2^r, \dots, s_{n_r}^r\}$, onde s_i^r é a i -ésima sentença do total $n_r = \|E_r\|$. O extrato automático é denotado por $E_a = \{s_1^a, s_2^a, \dots, s_{n_a}^a\}$, formado por $n_a = \|E_a\|$ sentenças. A Precisão do extrato automático é igual a

$$P(E_a) = \frac{\|E_r \cap E_a\|}{\|E_a\|}, \quad (5.1)$$

e expressa a proporção de sentenças coincidentes entre os dois extratos em relação ao número de sentenças do extrato automático. Já a Cobertura do extrato automático é dada por

$$C(E_a) = \frac{\|E_r \cap E_a\|}{\|E_r\|}, \quad (5.2)$$

e expressa a proporção de sentenças coincidentes entre os dois extratos em relação ao número de sentenças do extrato de referência. $P(E_a)$ e $C(E_a)$ variam de 0 a 100%, sendo que $P(E_a) = 100\%$ indica que todas as sentenças do extrato automático estão presentes no extrato de referência, e $C(E_a) = 100\%$ mostra que todas as sentenças do extrato de referência estão presentes no extrato automático. Precisão e Cobertura são inversamente

relacionadas, de maneira que uma tende a diminuir quando a outra sofre um aumento. Como as duas medidas são complementares, costuma-se utilizar uma outra medida que as agrupa em um único valor (entre 0 e 100%), chamada Medida-F (*F-Measure*), a qual é dada por

$$F_{\alpha}(E_a) = \frac{(1 + \alpha)P(E_a)C(E_a)}{\alpha P(E_a) + R(E_a)}, \quad (5.3)$$

onde α é uma constante não-negativa de balanceamento entre Precisão e Cobertura, de modo que, quanto maior α , maior o peso dado à Cobertura. Se tomarmos $\alpha = 1$, o peso dado à Precisão é igual ao dado à Cobertura, e

$$F_1(E_a) = F(E_a) = \frac{2P(E_a)C(E_a)}{P(E_a) + R(E_a)}. \quad (5.4)$$

$F(E_a)$ foi o caso particular da Medida-F adotado nas avaliações realizadas neste projeto. Quando se tratar de resultados referentes a um conjunto de sumários automáticos, Precisão, Cobertura e Medida-F serão tomados como valores médios e serão denotados por, respectivamente, P , C e F .

As métricas presentes no pacote de avaliação automática ROUGE² apresentam grande correlação com a avaliação humana (Lin e Hovy, 2003; Lin, 2004). ROUGE inclui quatro tipos de métricas (ROUGE-N, ROUGE-L, ROUGE-W e ROUGE-S) baseadas na co-ocorrência de unidades (tais como n -gramas) entre sumários criados automaticamente e sumários de referência. ROUGE foi utilizado nas DUC's de 2004, 2005 e 2006 para comparar o desempenho dos sistemas participantes da conferência, e apresentou correlação significativa com as avaliações manuais realizadas nas DUC's de 2001, 2002 e 2003. A seguir, será apresentada a métrica ROUGE-N, única utilizada neste projeto por ser uma medida amplamente aplicada. Isso possibilita uma comparação de desempenho com diferentes técnicas já propostas em Sumarização Automática, cujos resultados, com a medida ROUGE-N ($N = 1$), já foram divulgados.

ROUGE-N é uma medida de cobertura³ de n -gramas, e não de sentenças, entre um sumário candidato criado automaticamente e um conjunto de sumários de referência criados manualmente, sendo que os sumários de referência não costumam ser do tipo extrativo. Lin (2004) define ROUGE-N da seguinte maneira,

$$\text{ROUGE-N} = \frac{\sum_{S \in R} \sum_{n\text{-grama} \in S} \text{Total}_{inter}(n\text{-grama})}{\sum_{S \in R} \sum_{n\text{-grama} \in S} \text{Total}(n\text{-grama})}, \quad (5.5)$$

²ROUGE (Recall-Oriented Understudy for Gisting Evaluation, <http://haydn.isi.edu/ROUGE>).

³A medida BLEU, utilizada na avaliação de traduções, é baseada na *precisão* de n -gramas (Lin e Hovy, 2003). ROUGE baseia-se no método BLEU, sendo que este último não apresenta resultados tão bons para a avaliação de sumarização quanto o primeiro.

onde S é um sumário, R é o conjunto de sumários de referência, $Total(n\text{-grama})$ é a quantidade de um determinado n -grama presente no sumário $S \in R$, e $Total_{inter}(n\text{-grama})$ é o número de co-ocorrências de um determinado n -grama no sumário candidato e no sumário $S \in R$. Ou seja, é a divisão do número de n -gramas que co-ocorrem no sumário candidato e nos sumários de referência, pelo número total de n -gramas presentes no conjunto de sumários de referência. Conforme mais sumários de referência são adicionados à avaliação, o número de n -gramas presentes no denominador da Equação 5.5 aumenta, expandindo assim o número de sumários alternativos. Como seu numerador também considera todos os sumários de referência, um sumário candidato que contenha n -gramas presentes em muitas referências é favorecido pela ROUGE-N. Neste projeto, somente foram considerados unigramas no cálculo de ROUGE-N, ou seja, foi utilizada a métrica ROUGE-1.

O pacote ROUGE permite que suas medidas sejam aplicadas de diversas maneiras, por meio da alteração de um conjunto de parâmetros pré-definidos. O que guiou a configuração desses parâmetros, neste trabalho, foi a possibilidade de comparação com outros sistemas de sumarização conhecidos (Mihalcea, 2005), de maneira que todos os resultados relatados sejam provenientes de experimentos compatíveis entre si. A seguir, estão relacionados os parâmetros escolhidos para as avaliações realizadas neste projeto (com ROUGE versão 1.5.5):

- *Sem eliminação de stopwords*: todos os unigramas são considerados no cômputo da métrica ROUGE-1.
- *Stemming*: um processo semelhante à lematização (Seção 4.1) é aplicado às palavras dos sumários automáticos e de referência. Grosso modo, as palavras são reduzidas a seu radical (De Lucca e Nunes, 2002). Foi aplicado *stemming* somente aos textos em inglês, pois o pacote ROUGE não disponibiliza um *stemmer* para a língua portuguesa.
- *Média entre referências*: um sumário automático com n sumários de referência têm n valores ROUGE-1 calculados, um para cada referência. A média desses n valores é tomada como a medida ROUGE-1 do sumário automático em questão.

Como a medida ROUGE-1 é calculada para cada sumário automático isoladamente, o pacote ROUGE permite que seja obtido um único valor para um corpus, por meio do método estatístico chamado *bootstrapping* (Duda et al., 2000). *Bootstrapping* serve, neste caso, para estimar a média da amostra juntamente com um intervalo de confiança, por meio da seleção de valores da amostra original seguida de reposição (*resampling*). Os valores médios ROUGE-1 exibidos neste documento, quando calculados para mais de um sumário automático, foram obtidos com o uso de *bootstrapping*.

Com as medidas P , C , F e ROUGE-1, procura-se avaliar o grau de informatividade dos sumários automáticos, com relação a um conjunto de sumários de referência. Em outras palavras, as medidas devem indicar se o conteúdo que se espera em um bom sumário está contido no sumário automático. É preciso ter em mente que a coesão, a coerência, a gramaticalidade, ou qualquer outra característica dos sumários automáticos diferente da informatividade, não são consideradas pelas medidas apresentadas nesta seção.

5.2 *Cópus Seleccionados*

Os critérios de escolha dos cópus seleccionados para este projeto foram: (i) disponibilidade de sumários de referência, o que permite uma avaliação *off-line* dos sumarizadores propostos, (ii) ter sido utilizado na avaliação de outros sistemas de sumarização, possibilitando, portanto, uma avaliação comparativa e (iii) ser formado por textos em inglês ou em português, contemplando assim a língua local e a língua com mais estudos em Sumarização Automática. A seguir, são descritos em detalhes os três cópus seleccionados, sendo que o critério (ii) não pôde ser satisfeito para um dos cópus em questão. Embora todos sejam compostos por textos jornalísticos, o gênero informativo não foi um pré-requisito para a escolha dos cópus, pois os métodos propostos na Seção 4.2 não foram criados visando a sumarização de textos de gênero ou domínio específicos.

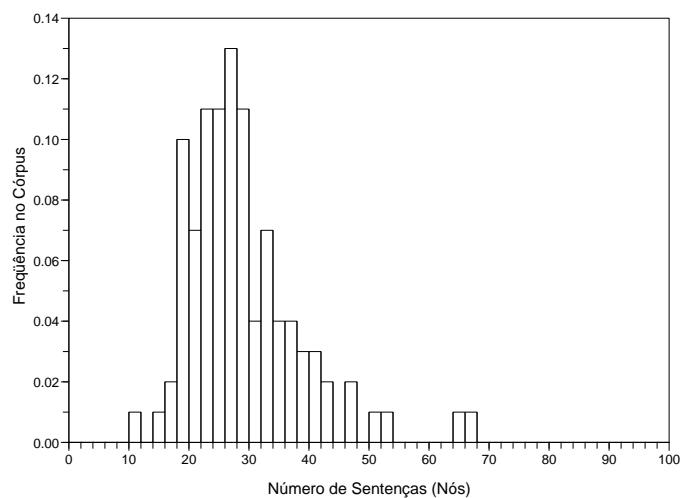
- *TeMário*: acrônimo para “TExtos com suMÁRIOS” (Pardo e Rino, 2003), reúne um conjunto de 100 textos jornalísticos com seus respectivos resumos⁴, em português (um resumo para cada texto-fonte), construídos por um sumarizador humano profissional. Os textos-fonte provêm dos jornais Folha de São Paulo e Jornal do Brasil, enquanto que os resumos foram construídos observando-se a restrição de que deveriam ter de 25 a 30% do tamanho de seus respectivos textos-fonte. O TeMário é ainda composto por extratos de referência gerados automaticamente pelo sistema GEI⁵ (Gerador de Extratos Ideais) (Pardo e Rino, 2004). Esses extratos de referência são vantajosos do ponto de vista do custo/benefício de sua construção, e, apesar de não serem criados manualmente, baseiam-se nos resumos criados por um sumarizador humano. Um grupo de sistemas de sumarização para o português já foi testado com o cópus TeMário (Rino et al., 2004; Leite e Rino, 2006a).

⁴A diferença entre extrato e resumo é explicada no Capítulo 2.

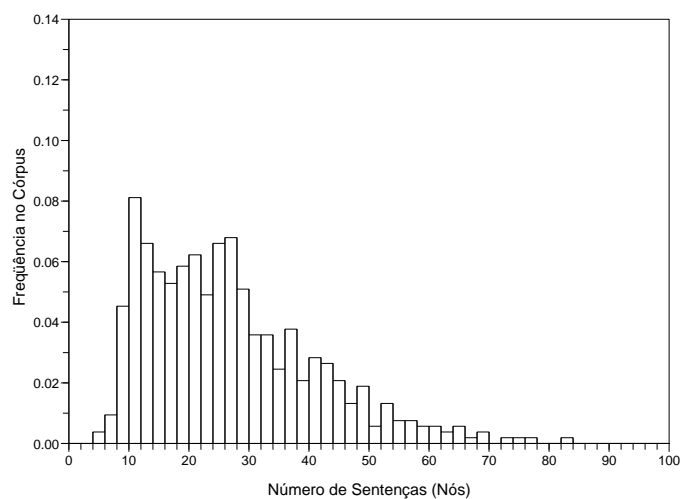
⁵O GEI constrói um vetor de palavras para cada sentença, e cada posição do vetor contém a frequência de uma dada palavra na sentença. O cosseno do ângulo entre dois vetores dá a similaridade entre duas sentenças, e as sentenças mais similares do texto-fonte, com relação às do resumo de referência, são escolhidas para compor o extrato de referência. Em outras palavras, cada sentença do resumo de referência é usada para seleccionar uma sentença do texto-fonte (a que apresenta menor ângulo).

- *DUC'2002*: na conferência DUC de 2002 foram avaliados sumarizadores automáticos de textos jornalísticos (Over e Liggett, 2002). Essa foi a Tarefa 1 da conferência, já que outros tipos de sistemas foram avaliados na DUC'2002, como os para sumarização multi-documento. Foram disponibilizados 567 textos, em inglês, retirados das seguintes fontes: Wall Street Journal, AP Newswire, San Jose Mercury News, Financial Times, LA Times e FBIS. Resumos manuais (dois, em média) de aproximadamente 100 palavras (não mais do que 100) acompanham cada documento, sendo que cada palavra é definida como uma sequência de caracteres separados por espaços.
- *DUC'2001*: Córpus de treinamento disponibilizado aos participantes da DUC de 2001 (Over, 2001), formado por 104 textos jornalísticos em inglês, e retirado das seguintes fontes: AP Newswire, San Jose Mercury News, Financial Times, LA Times e FBIS. Cada texto é acompanhado de um extrato criado manualmente, de tal modo que cubra o mesmo conteúdo dos respectivos resumos manuais (o GEI, utilizado para criar os extratos de referência do TeMário, simula esse comportamento). Os resumos manuais têm aproximadamente 100 palavras, e os extratos manuais são, em média, 60% maiores. Embora não existam resultados conhecidos de avaliações que utilizaram esse córpus, a presença de extratos construídos manualmente torna seu uso interessante.

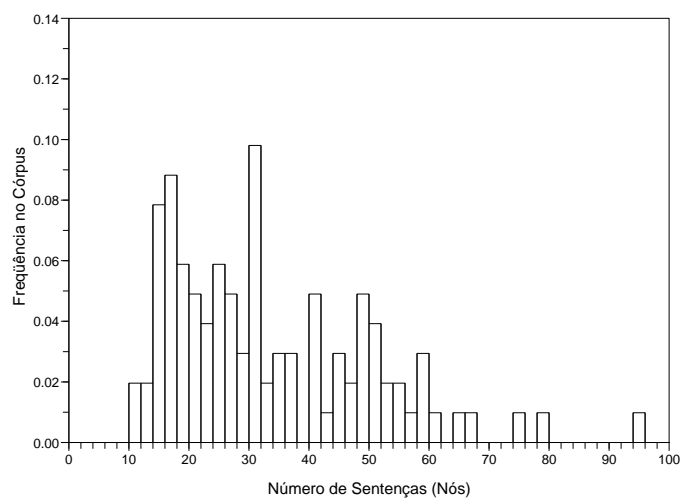
Na Figura 5.1 encontram-se os histogramas do número de sentenças por texto-fonte nos três córpus supracitados. O TeMário apresenta, em média, 29,37 sentenças, enquanto que os córpus DUC'2002 e DUC'2001 têm, respectivamente, 27,92 e 35,69 sentenças em média. Isso significa que as redes derivadas dos textos-fonte têm, em média, tamanhos parecidos (em número de nós). Entretanto, ao analisar as distribuições da Figura 5.1, percebe-se que, para os córpus em inglês, existe uma maior dispersão no número de sentenças do que em relação ao TeMário. Nesse último, a maior parte dos textos-fonte têm por volta de 30 sentenças, enquanto que, para os demais córpus, o número de sentenças se afasta mais da média. Essa constatação é importante, pois as redes geradas apresentam um número de nós igual ao número de sentenças do texto-fonte, e é sabido que algumas medidas aplicadas em redes, como o coeficiente de aglomeração, costumam ser sensíveis a pequenas mudanças na mesma (Margarido, 2007). Portanto, em redes muito díspares (com relação ao número de vértices) os algoritmos propostos podem sofrer uma variação sensível de desempenho.



(a)



(b)



(c)

Figura 5.1: Distribuições do número de sentenças por texto-fonte nos corpú (a) TeMário, (b) DUC'2002 e (c) DUC'2001. As médias são (a) 29,37, (b) 27,92 e (c) 35,69 sentenças.

5.3 Definições dos Experimentos

Na Tabela 5.1, as características de cada um dos corpúse selecionados encontram-se resumidas. Tais propriedades desempenharam papel importante na definição dos experimentos de avaliação dos sumarizadores propostos. A primeira delas é o tipo dos sumários de referência que acompanham um determinado corpúse: resumos de referência são mais propícios à aplicação da métrica ROUGE-1, enquanto que extratos de referência são apropriados à utilização das medidas P , C e F ⁶. O fato de um corpúse ter sido utilizado em alguma avaliação prévia também influenciou a definição dos experimentos realizados neste projeto, como na escolha das métricas de avaliação, na definição da taxa de compressão dos extratos e até mesmo na utilização dos parâmetros do pacote ROUGE (vide Seção 5.1). Dessa maneira, definindo experimentos que sejam compatíveis com outros já publicados (Over e Liggett, 2002; Rino et al., 2004; Mihalcea, 2005; Leite e Rino, 2006a), permite-se que uma comparação confiável entre diversos métodos de sumarização seja realizada. Já o fato de um corpúse ser composto por textos em língua portuguesa ou inglesa não influenciou a construção dos experimentos, por serem utilizadas técnicas de avaliação independentes de língua.

Tabela 5.1: Propriedades dos corpúse utilizados nos experimentos de avaliação.

	TeMário	DUC'2002	DUC'2001
Resumos de Referência (Manuais)	•	•	•
Extratos de Referência (Manuais)			•
Extratos de Referência (Automáticos)	•		
Textos em Português	•		
Textos em Inglês		•	•
Permite Avaliação Comparativa	•	•	

A lista de métricas de avaliação aplicadas em cada corpúse pode ser consultada na Tabela 5.2. Os dois corpúse em inglês apresentam sumários de referência, mas decidiu-se aplicar a métrica ROUGE-1 apenas no corpúse da DUC'2002, pelo fato de ter sido utilizado em avaliação comparativa e por ter um número maior de documentos (Over e Liggett, 2002). Vale ressaltar, novamente, que o corpúse da DUC'2001 foi incluído por apresentar extratos de referência construídos manualmente, e sua utilização refere-se apenas à aplicação das métricas P , C e F . Já no TeMário, os dois tipos de métricas de avaliação foram aplicados, por ser um corpúse bastante utilizado em análises de sistemas de sumarização para o português (Rino et al., 2004; Mihalcea, 2005; Leite e Rino, 2006a). A Tabela 5.2

⁶Nada impede que uma métrica ROUGE seja aplicada tendo como referência extratos, entretanto, as medidas de Precisão e Cobertura de sentenças só podem ser aplicadas utilizando extratos de referência.

indica, portanto, que foram realizados quatro experimentos de avaliação dos sumarizadores propostos:

- *TeMário com P , C e F* : nesse experimento, foram gerados extratos com 30% do tamanho (em número de sentenças) dos textos-fonte do corpus TeMário, nos moldes de avaliações como as de Rino et al. (2004) e Leite e Rino (2006a). As métricas P , C e F foram aplicadas na avaliação da informatividade dos extratos gerados automaticamente, tendo como referência os extratos ideais do corpus TeMário.
- *TeMário com ROUGE-1*: o tamanho do extrato automático foi definido como sendo próximo ao tamanho do resumo manual, em número de palavras (Mihalcea, 2005). Embora a taxa de compressão seja dada em número de palavras, somente sentenças completas foram selecionadas. A métrica ROUGE-1 foi aplicada utilizando-se os resumos de referência do corpus TeMário, e empregando-se os mesmos parâmetros utilizados por Mihalcea (2005), apresentados na Seção 5.1.
- *DUC'2002 com ROUGE-1*: nesse caso os extratos automáticos têm tamanho absoluto (não relativo ao tamanho do texto-fonte) de aproximadamente 100 palavras, conforme definição utilizada na DUC'2002 (Over e Liggett, 2002). Novamente, somente sentenças completas foram selecionadas. A métrica ROUGE-1 foi aplicada na avaliação, tendo como referência os resumos manuais da DUC'2002, e empregando os mesmos parâmetros utilizados por Mihalcea (2005).
- *DUC'2001 com P , C e F* : a taxa de compressão foi definida em 30% do número de sentenças do texto-fonte, da mesma maneira que no experimento com os extratos de referência do corpus TeMário. Os extratos manuais da DUC'2001 foram utilizados na aplicação das métricas de avaliação P , C e F .

Tabela 5.2: Métricas de avaliação aplicadas em cada corpus. Quando a avaliação for comparativa, indica-se com parênteses.

	TeMário	DUC'2002	DUC'2001
Precisão, Cobertura e Medida-F	(•)		•
ROUGE-1	(•)	(•)	

Quando a taxa de compressão é definida em número de sentenças, sumarizadores distintos podem dar origem a extratos com tamanhos muito variados entre si (considerando-se o mesmo texto-fonte), devido aos diferentes tamanhos das sentenças. Já a compressão em número de palavras permite uma definição mais exata do tamanho dos sumários. Contudo, optou-se por utilizar a compressão em número de sentenças em alguns experimentos, devido

à divulgação de experimentos anteriores que utilizam os mesmos parâmetros, o que possibilitaria uma avaliação comparativa dos métodos aqui propostos. Apesar da existência desse problema com relação ao tamanho dos extratos, a taxa de compressão dada em número de sentenças não beneficia extratos que selecionam muitas sentenças grandes, considerando, nesse caso, a aplicação das métricas de avaliação P , C e F . Como esse tipo de avaliação também é realizada com sentenças, o que importa é se determinada sentença está, ou não, contida no extrato de referência, e o tamanho da sentença não é levado em conta. Se todos os sumarizadores selecionarem o mesmo número de sentenças, as métricas P , C e F são, portanto, imparciais. Por outro lado, a métrica ROUGE-1 beneficiaria sentenças grandes, por ser justamente baseada em unigramas. Os experimentos realizados empregam os dois tipos de taxa de compressão (com as métricas de avaliação mais propícias para cada caso), tanto para o português quanto para o inglês, o que permite uma análise menos tendenciosa.

Vale ressaltar também que, quando a taxa de compressão é definida em número de palavras, os algoritmos de seleção de sentenças devem ser adaptados. Como cada medida ξ_i da Tabela 4.1 fornece uma ordem de importância para as sentenças, quando se considera uma taxa de compressão dada pelo número de sentenças, as x sentenças mais importantes são utilizadas na formação de um extrato. Entretanto, ao selecionar uma determinada sentença, ela pode ultrapassar o limite de compressão dado pelo número de palavras. Nesse caso, desconsidera-se essa sentença e procura-se selecionar a próxima de acordo com a pontuação ξ_i , nunca ultrapassando o limite de palavras. Por fim, em cada um dos experimentos realizados, foram utilizados dois sistemas do tipo *baseline*: o Top-Baseline e o Random-Baseline. O Top-Baseline seleciona as primeiras sentenças de um texto-fonte (atributo de localização) na formação de um extrato. Já o Random-Baseline seleciona as sentenças de forma aleatória. Esses dois sistemas são extremamente simples, e servem de base na avaliação dos sistemas aqui propostos. Um sistema com desempenho próximo dos obtidos para os *baselines* é considerado crítico, pois pouco ou nada acrescenta a sistemas simples tomados como referência. Frequentemente, o Top-Baseline, apesar de pouco complexo, apresenta bom desempenho em textos jornalísticos (Over e Liggett, 2002; Rino et al., 2004).

5.4 Resultados Obtidos

A seguir, são apresentados e discutidos os resultados obtidos nos quatro experimentos de avaliação realizados. Os recursos utilizados nesses experimentos, tais como corpus e métricas de avaliação, já foram detalhados em seções anteriores.

5.4.1 TeMário com P , C e F

Na Tabela 5.3 encontram-se listados os resultados obtidos para todos os 26 sumarizadores propostos neste projeto, avaliados segundo a aplicação das medidas Precisão, Cobertura e Medida-F no corpus TeMário. Além dos resultados dos métodos *baseline*, foram adicionados os resultados referentes a outros sumarizadores anteriormente propostos: SuPor-v2 (Leite e Rino, 2006a), SuPor, ClassSumm e TF-ISF-Summ (Rino et al., 2004). Todos esses sistemas foram apresentados na Seção 2.1. Os sumarizadores estão ordenados de forma decrescente na Tabela 5.3, de maneira que os primeiros sistemas sejam os que apresentam maiores valores para F (pois é a medida que une P e C). Na Figura 5.2, estão dispostos os resultados referentes à Medida-F na mesma ordem definida na Tabela 5.3, possibilitando uma visão complementar para as diferenças de desempenho entre os sistemas.

Considerando-se somente os métodos baseados em redes complexas, tem-se que o melhor sumarizador é o baseado nos caminhos mínimos, com complemento dos pesos (sp_i^{wc}). Os outros tipos de caminhos mínimos têm desempenho inferior, mas ainda acima do Top-Baseline, bem como os sumarizadores baseados nos d -anéis, no grau, nos w -cortes, nos k -núcleos, nas comunidades e no índice de localidade (excluindo-se a versão modificada l_i^{mod}). Do 17º para o 18º sistema ocorre uma queda mais brusca no desempenho dos sumarizadores baseados em redes complexas (vide Figura 5.2), e os métodos baseados no grau hierárquico, no índice de concordância, no índice de localidade (apenas l_i^{mod}) e no coeficiente de aglomeração figuram em um grupo de sistemas com Medida-F mais próxima das obtidas para o Top-Baseline e para o Random-Baseline. Essa queda no desempenho dos sumarizadores divide os sistemas aqui propostos em dois grupos: o primeiro, chamado de Grupo-1, contém os sistemas com melhor desempenho (até o 17º), e o segundo, chamado de Grupo-2, contém os sistemas com resultados inferiores (a partir do 18º). Essa divisão terá ainda maior sentido quando for mostrado, nas próximas seções, que seus membros praticamente não mudam de experimento para experimento.

O método baseado em caminhos mínimos sp_i^{wc} apresenta $F = 42,4\%$, a maior Medida-F obtida neste experimento para os métodos propostos. A sugestão de que vértices próximos dos outros vértices da rede seria importante para a informatividade dos extratos mostrou-se válida. Além disso, as outras variações dos caminhos mínimos têm desempenho um pouco inferior, com $F = 41,4\%$. As variações dos d -anéis, também contidas no Grupo-1, apresentam Medidas-F iguais a 42,2%, 40,8% e 39,3%, sendo que o melhor desempenho recai sobre a variação com ordenação por grau (r_i^k). Lembrando que os d -anéis são calculados a partir do nó mais conectado da rede, chamado de *hub*, e o extrato é formado por esse nó mais os nós contidos em suas hierarquias mais próximas. O conceito de *hub* parece influenciar positivamente a informatividade dos extratos gerados a partir do corpus TeMário, pois

Tabela 5.3: Valores médios de Precisão (P), Cobertura (C) e Medida-F (F), obtidos comparando-se os extratos gerados automaticamente com os extratos de referência do corpus TeMário. Os sistemas estão ordenados decrescentemente por F . Os métodos *baseline* estão identificados por (\Rightarrow), enquanto que os sumarizadores propostos em outros trabalhos estão identificados por (\rightarrow).

	Sistemas	P (%)	C (%)	F (%)
\rightarrow	1 SuPor-v2	47,4	43,9	45,6
\rightarrow	2 SuPor	44,9	40,8	42,8
	3 Caminhos Mínimos sp_i^{wc}	47,4	39,9	42,4
\rightarrow	4 ClassSumm	45,6	39,7	42,4
	5 d -Anéis r_i^k	47,2	39,8	42,2
	6 Grau k_i	47,0	39,7	42,1
	7 Grau s_i	47,0	39,3	41,8
	8 w -Cortes p_i^k	46,5	39,2	41,6
	9 Caminhos Mínimos sp_i^{wi}	46,6	38,8	41,4
	10 Caminhos Mínimos sp_i	46,4	39,0	41,4
	11 k -Núcleos n_i^k	46,2	38,9	41,3
	12 w -Cortes p_i^l	46,0	38,7	41,1
	13 d -Anéis $r_i^{l,k}$	45,7	38,6	40,8
	14 k -Núcleos n_i^l	44,6	37,1	39,6
	15 Índice de Localidade l_i	44,6	37,0	39,6
	16 Comunidades g_i	44,1	37,0	39,4
	17 d -Anéis r_i^l	44,3	37,0	39,3
	18 Grau Hierárquico $k_i^{2,c}$	41,6	35,3	37,3
\Rightarrow	19 Top-Baseline	42,9	32,6	37,0
\rightarrow	20 TF-ISF-Summ	39,6	34,3	36,8
	21 Grau Hierárquico $s_i^{2,c}$	40,2	34,1	36,1
	22 Grau Hierárquico $k_i^{3,c}$	40,0	34,0	36,0
	23 Grau Hierárquico k_i^2	39,2	33,5	35,3
	24 Grau Hierárquico $s_i^{3,c}$	38,5	32,6	34,5
	25 Grau Hierárquico s_i^2	37,1	31,5	33,3
\Rightarrow	26 Random-Baseline	34,0	28,5	31,0
	27 Índice de Concordância m_i	33,0	28,0	29,6
	28 Índice de Localidade l_i^{mod}	32,2	26,2	28,2
	29 Grau Hierárquico s_i^3	30,4	25,0	26,8
	30 Grau Hierárquico k_i^3	29,9	24,6	26,3
	31 Coeficiente de Aglomeração C_i^w	28,1	23,4	24,9
	32 Coeficiente de Aglomeração C_i	27,9	23,2	24,7

os sumarizadores baseados no grau apresentam resultados um pouco abaixo do método sp_i^{wc} , com Medidas-F iguais a 42,1% e 41,8%. Grupos de sentenças coesas, representados pelos w -cortes e pelos k -núcleos, também mostraram-se bons candidatos a compor um extrato, pois os sumarizadores baseados nesses conceitos têm Medida-F por volta de 40%,

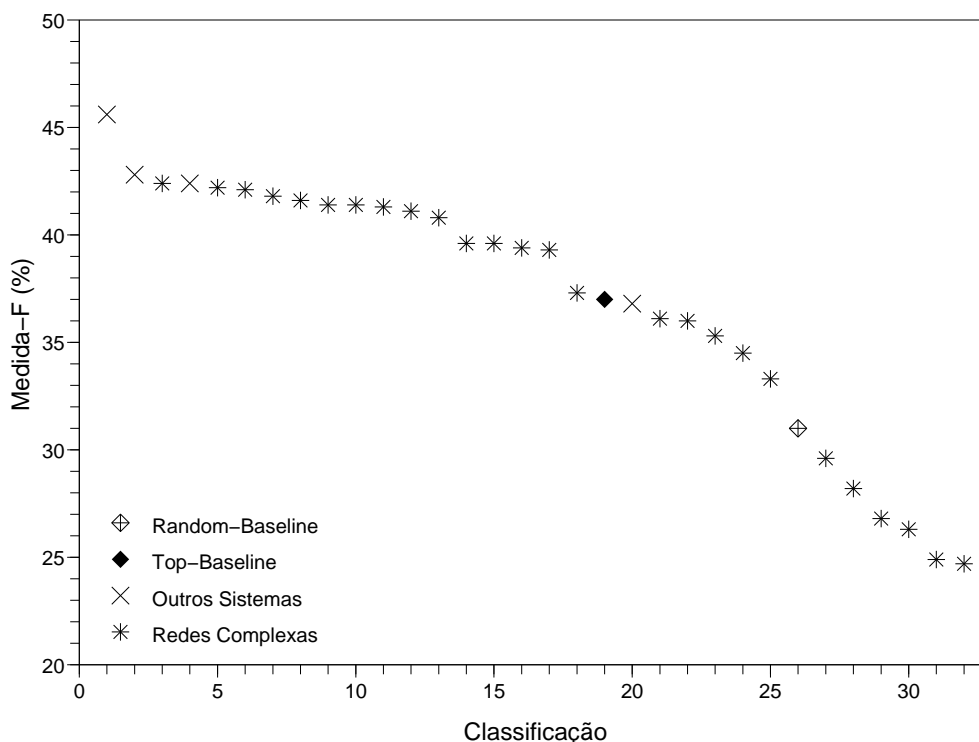


Figura 5.2: Medida-F média (F) dos sumarizadores da Tabela 5.3 (cópus TeMário). Os sistemas estão ordenados de forma decrescente de acordo com F .

com destaque para as variantes que utilizam a ordenação por grau. Note que, novamente, o grau tem papel positivo no desempenho de um sumarizador, mesmo que indiretamente (como nos d -anéis). A idéia contida no índice de localidade, que dá importância aos vértices cujos vizinhos não compartilham muitas arestas com o restante da rede, tem desempenho razoável ($F = 39,6\%$), já um tanto distante dos melhores sistemas. Entretanto, nesse ponto, a queda de desempenho nos sumarizadores baseados em redes complexas ainda não é brusca, e, além disso, o Top-Baseline ainda está abaixo do índice de localidade. Por isso considera-se que essa técnica, excluindo-se a versão não modificada, ainda está contida no grupo de sumarizadores mais promissores (Grupo-1). Da mesma maneira, o algoritmo baseado na divisão de comunidades ainda está contido no Grupo-1, com $F = 39,4\%$. Não se sabe ao certo se a identificação de comunidades separa as sentenças do texto-fonte de acordo com o tópico; a modularidade máxima obtida nas redes do cópus TeMário é, em média, igual a 0,26, o que indica que as redes não apresentam vértices claramente agrupados, prejudicando a divisão em comunidades⁷ (vide Seção 4.2). Se a rede para textos for alterada de maneira que seja possível aumentar sua modularidade⁸, talvez as

⁷Para o cópus DUC'2002 a modularidade máxima é, em média, igual a 0,25, enquanto que para o cópus DUC'2001 é igual a 0,26.

⁸Cabe aqui ressaltar que os resultados referentes ao sumarizador baseado na divisão de comunidades foram obtidos em uma rede ligeiramente diferente da definida na Seção 4.1. A fim de aumentar a modularidade das redes de sentenças, as arestas com peso igual a 1 foram eliminadas.

comunidades representem melhor a estrutura de tópicos do texto-fonte, e os extratos sejam mais informativos devido à seleção de sentenças de cada tópico.

Alguns dos métodos do Grupo-1 estão muito próximos dos sistemas SuPor e ClassSumm, inclusive com melhores resultados de Precisão⁹. Esses dois sistemas são baseados em algoritmos de aprendizado de máquina, e calculam diversos atributos para cada sentença, sendo necessária ainda a realização de uma fase de treinamento do sumariizador, possivelmente com seleção de atributos. No caso do SuPor, são agrupadas diferentes técnicas de sumarização, como as cadeias lexicais, cujo cômputo emprega semântica ao nível das palavras. O ClassSumm, inclusive, faz uso de uma aproximação da estrutura argumentativa do texto. Os sistemas aqui propostos obtêm apenas uma medida (atributo) para cada sentença e não necessitam de treinamento. Além disso, o pré-processamento dos textos-fonte, antes da montagem das redes, emprega apenas um etiquetador morfossintático e um lematizador, sem envolver semântica. Desse ponto de vista, o algoritmo sp_i^{wc} , por exemplo, é mais simples que o SuPor e o ClassSumm, com desempenho muito próximo aos obtidos para esses dois sistemas. Quanto ao SuPor-v2, primeiro colocado na avaliação comparativa, o desempenho é substancialmente melhor (vide Figura 5.2), por meio do aperfeiçoamento dos diversos recursos utilizados em sua primeira versão, o SuPor.

Já no Grupo-2, percebe-se que os sumariizadores baseados no grau hierárquico tendem a ficar, em sua maioria, abaixo do Top-Baseline, e, em alguns casos, abaixo até mesmo do Random-Baseline. O grau hierárquico usa o número de arestas contidas em d -anéis mais distantes do nó tomado como referência. Nas redes obtidas para os textos-fonte do corpus TeMário, a partir da 4ª hierarquia o número de vértices passa a ser bastante escasso¹⁰. Isso provoca uma equalização nos valores dos graus hierárquicos, pois já na 3ª hierarquia praticamente todos os vértices da rede são considerados, o que possivelmente dificultaria uma discriminação das sentenças na geração dos extratos. Outro possível problema dos graus hierárquicos é a representação das conexões presentes nos diversos d -anéis por apenas um valor, o que acabaria prejudicando vértices com alto grau tradicional (k_i ou s_i) quando comparados a vértices de baixo grau tradicional, mas com um alto grau em níveis hierárquicos mais distantes. Isso parece atrapalhar o desempenho do sumariizador, pois os vértices com um alto grau tradicional tendem a influenciar positivamente a informatividade dos extratos, como nos métodos do Grupo-1. O índice de concordância, por sua vez, está abaixo do Random-Baseline. Essa medida associa um índice a cada aresta, e tem valor alto quando os vértices unidos pela aresta em questão compartilham um grande número

⁹Geralmente, quanto maior a Medida-F, maiores os valores para Precisão e Cobertura (considerando os experimentos aqui relatados). Os casos extremos que fogem à essa regra são comentados no texto.

¹⁰Em média, 78,31% dos vértices das redes obtidas a partir dos textos-fonte do corpus TeMário apresentam a 4ª hierarquia nula. Esse valor é 86,22% para o corpus DUC'2002, e 88,42% para o corpus DUC'2001.

de vizinhos. Na construção dos extratos, entretanto, dá-se prioridade à seleção de pares de vértices que apresentem baixo índice de concordância, selecionando-se assim vértices complementares, associados a regiões distintas da rede. Essa idéia mostrou-se problemática para a informatividade dos extratos gerados com o TeMário. Da mesma maneira, com desempenho muito baixo, encontra-se o índice de localidade modificado (l_i^{mod}). Com essa medida, procura-se descartar vértices redundantes na formação do extrato, de modo que os vizinhos de um vértice com alto índice de localidade sejam desconsiderados, mesmo que também possuam um alto índice de localidade. Essa medida parece ser muito restritiva, descartando sentenças importantes para a sumarização, o que acarreta em baixa Cobertura. A última medida considerada, o coeficiente de aglomeração, teve o pior desempenho de todas as propostas, tanto em sua versão tradicional quanto na que considera os pesos das arestas. Um vértice “aglomerado” é o que possui vizinhos bem conectados entre si, contudo, isto não implica que o vértice em questão seja bem conectado e compartilhe informações com diversos outros vértices. Esse pode ser um problema dos sumarizadores baseados no coeficiente de aglomeração.

O outro sistema que participou da avaliação comparativa, o TF-ISF-Summ, figura entre os sumarizadores do Grupo-2, um pouco abaixo do Top-Baseline. Nesse sistema, não são utilizados recursos lingüísticos sofisticados, apenas a métrica TF-ISF (definida na Seção 2.1), o que, por si só, já é uma abordagem inovadora. Outros sistemas foram avaliados em um experimento muito parecido com o comentado nesta seção. O GistSumm e o NeuralSumm (vide Seção 2.1) foram aplicados ao corpus TeMário, e também foram avaliados por meio de P , C e F . Contudo, aplicou-se uma taxa de compressão de 30% proporcional ao número de palavras do texto-fonte, e não ao número de sentenças. Isso acarreta em diferentes quantidades de sentenças selecionadas pelo GistSumm e pelo NeuralSumm, o que prejudica a comparação com os sistemas da Tabela 5.3. As Medidas-F obtidas foram 33,8% para o GistSumm e 32,4% para o NeuralSumm, entretanto, não é confiável compará-las com as aqui obtidas. Esses dois sistemas resultam de propostas interessantes: o GistSumm é parecido com os algoritmos baseados nos d -anéis, pois ambos escolhem primeiro a sentença considerada a mais importante do texto-fonte (chamada de *gist sentence* ou *hub*, respectivamente), para logo após selecionar sentenças a ela relacionadas; já o NeuralSumm faz uso de uma rede neural do tipo SOM, e une as sentenças em grupos de similaridade (o que lembra a divisão das redes em comunidades). Uma avaliação uniforme que englobe todos os sistemas aqui citados deve fornecer uma indicação mais forte a respeito da informatividade dos extratos gerados pelo GistSumm e pelo NeuralSumm.

5.4.2 TeMário com ROUGE-1

Os resultados referentes à aplicação da métrica ROUGE-1 no corpus TeMário podem ser consultados na Tabela 5.4 e na Figura 5.3. Note que os sistemas estão ordenados de forma decrescente pelo valor ROUGE-1. Nesse experimento foram utilizados, para fins de comparação, os resultados publicados por Mihalcea (2005), referentes aos métodos de sumarização baseados nos algoritmos PageRank (ou PR), $HITS_A$ e $HITS_H$, todos definidos na Seção 2.2. Como a autora utilizou três variações para cada um desses algoritmos, as quais referem-se aos tipos das arestas (não direcionadas, *forward* e *backward*), somente foram reproduzidos aqui os resultados referentes à melhor variação, para o português, de cada uma dessas técnicas. São elas: PageRank Backward, $HITS_A$ Backward e $HITS_H$ Forward.

Neste experimento com a métrica ROUGE-1, a divisão dos métodos propostos em Grupo-1 e Grupo-2 é praticamente a mesma do experimento com a Medida-F da seção anterior. Na Figura 5.3, pode-se perceber que do 17º para o 18º sistema ocorre uma queda acentuada no desempenho dos sumarizadores de redes complexas, de modo que até o 17º sistema os valores ROUGE-1 são mais próximos do Top-Baseline, e do 18º sistema em diante os resultados são mais próximos do Random-Baseline. O método p_i^k , baseado nos k -cortes, agora faz parte do Grupo-2, e o Top-Baseline, diferentemente do experimento anterior, faz parte do Grupo-1. O melhor sistema dessa vez, entre os baseados em redes complexas, é o grau com pesos s_i , com $ROUGE-1 = 0,5020$. O grau k_i tem desempenho próximo, com $ROUGE-1 = 0,5003$. Novamente, as medidas baseadas no grau apresentam bons resultados, assim como as medidas inspiradas nos d -anéis, nos caminhos mínimos e nos k -núcleos. Em especial, a medida $r_i^{l,k}$, que usa os d -anéis com ordenação por localização das sentenças e corte de grau, apresenta um valor ROUGE-1 muito próximo do obtido para o grau s_i (igual a 0,5019). Note que, dessa vez, a melhor variação dos k -núcleos é a que usa a ordenação por localização das sentenças, diferentemente do experimento anterior. Já os w -cortes apresentam resultados relativamente inferiores quando comparados com os resultados do experimento com a Medida-F (tanto que, dessa vez, considera-se que uma das variações dos w -cortes faz parte do Grupo-2). O algoritmo baseado em comunidades g_i e o índice de localidade l_i continuam entre os piores sistemas do Grupo-1. De maneira geral, os melhores sistemas baseados em redes complexas são os mesmos da outra avaliação com o corpus TeMário.

Uma importante característica deste experimento é o aumento relativo no desempenho do Top-Baseline, com relação aos resultados da Tabela 5.3. Isso indica que as primeiras sentenças de um texto-fonte ganham maior importância ao se utilizar os resumos de referência no lugar dos extratos de referência do corpus TeMário, e ao se aplicar uma métrica

ROUGE ao invés de métricas baseadas na co-seleção de sentenças (P , C e F). Isso exemplifica a dificuldade envolvida mesmo em uma avaliação automática. Dependendo do tipo de métrica de avaliação e dos sumários de referência (geralmente criados por humanos), determinados tipos de extratos ganham ou perdem importância, apesar de gerados pelos mesmos algoritmos aplicados aos mesmos textos-fonte.

Os sumarizadores baseados nos algoritmos PageRank e HITS figuram entre os 6 melhores sistemas do experimento. O PageRank Backward, particularmente, está acima de qualquer um dos métodos aqui propostos, com $\text{ROUGE-1} = 0,5121$. Como neste projeto, esses três sistemas fazem uso de uma rede de sentenças, cujas arestas são criadas de acordo com o número de termos em comum entre as sentenças. Entretanto, Mihalcea (2005) não filtra os termos do texto-fonte (eliminação de *stopwords*) e não os lematiza, mas realiza uma normalização dos pesos das arestas de acordo com os tamanhos das sentenças. Além disso, a autora trabalha com três tipos de arestas, o que parece ter grande influência na sumarização. Se tomarmos como exemplo o valor $\text{ROUGE-1} = 0,4574$ da variação PageRank Forward (não listada na Tabela 5.4), é possível perceber que seu desempenho está consideravelmente abaixo da variação PageRank Backward, e abaixo de qualquer um dos sumarizadores aqui propostos. Nos algoritmos definidos neste projeto, somente redes com arestas não direcionadas são utilizadas. Ainda não está claro se os algoritmos da autora atingem bons resultados pelas diferenças nas redes utilizadas ou pela natureza dos algoritmos PageRank e HITS.

Os sumarizadores pertencentes ao Grupo-2 são, novamente, os baseados no grau hierárquico, no índice de localidade modificado (l_i^{mod}), no coeficiente de aglomeração e no índice de concordância. Basicamente, os problemas desses sistemas parecem ser os mesmos discutidos na seção anterior. Percebe-se que, agora, os graus hierárquicos cumulativos apresentam melhores resultados que os não cumulativos, justamente por considerarem também as conexões mais próximas dos vértices no cômputo dos graus hierárquicos em níveis 2 e 3. O Grupo-2 contém também o sumador p_i^k (w -cortes ordenados grau), antes classificado em 8º lugar na avaliação com a Medida-F.

Por fim, é importante mencionar que Leite e Rino (2006b) avaliaram outros sumarizadores por meio da métrica ROUGE-1, aplicada também no corpus TeMário, com resultados bem interessantes. Entretanto, preferiu-se não incluí-los na Tabela 5.4, por ter sido utilizada pelos autores uma taxa de compressão diferente da aqui utilizada, o que prejudica uma comparação mais exata entre as propostas (30% do número de sentenças do texto-fonte vs. tamanho dos extratos próximo do tamanho do resumo manual, em número de palavras). Os autores avaliaram o Supor-v2, e duas variações do PageRank em redes com arestas não direcionadas: uma delas utilizando um *thesaurus* para considerar sinonímia e antonímia

Tabela 5.4: Valores médios da medida ROUGE-1, obtidos comparando-se os extratos gerados automaticamente com os resumos de referência do corpus TeMário. Os sistemas estão ordenados decrescentemente por ROUGE-1. Os métodos *baseline* estão identificados por (\Rightarrow), enquanto que os sumarizadores propostos em outros trabalhos estão identificados por (\rightarrow).

	Sistemas	ROUGE-1
\rightarrow	1 PageRank Backward	0,5121
	2 Grau s_i	0,5020
	3 d -Anéis $r_i^{l,k}$	0,5019
	4 Grau k_i	0,5003
\rightarrow	5 HITS _A Backward	0,5002
\rightarrow	6 HITS _H Forward	0,5002
	7 Caminhos Mínimos sp_i^{wi}	0,4995
	8 d -Anéis r_i^k	0,4994
	9 k -Núcleos n_i^l	0,4992
\Rightarrow	10 Top-Baseline	0,4984
	11 Caminhos Mínimos sp_i^{wc}	0,4982
	12 k -Núcleos n_i^k	0,4978
	13 Caminhos Mínimos sp_i	0,4975
	14 d -Anéis r_i^l	0,4968
	15 Comunidades g_i	0,4959
	16 w -Cortes p_i^l	0,4940
	17 Índice de Localidade l_i	0,4935
	18 w -Cortes p_i^k	0,4889
	19 Grau Hierárquico $k_i^{2,c}$	0,4861
	20 Grau Hierárquico $s_i^{2,c}$	0,4844
	21 Índice de Localidade l_i^{mod}	0,4830
	22 Grau Hierárquico $k_i^{3,c}$	0,4785
	23 Grau Hierárquico $s_i^{3,c}$	0,4770
	24 Grau Hierárquico k_i^2	0,4770
\Rightarrow	25 Random-Baseline	0,4765
	26 Grau Hierárquico s_i^2	0,4758
	27 Grau Hierárquico s_i^3	0,4676
	28 Grau Hierárquico k_i^3	0,4671
	29 Coeficiente de Aglomeração C_i^w	0,4663
	30 Coeficiente de Aglomeração C_i	0,4647
	31 Índice de Concordância m_i	0,4604

na definição das arestas na rede de sentenças (PageRank+Thesaurus), e a outra com *stemming* e eliminação de *stopwords* em uma fase de pré-processamento dos textos-fonte (PageRank+Stem+StopRem). A avaliação dessas propostas resultou em: ROUGE-1 = 0,5839 para SuPor-v2, ROUGE-1 = 0,5603 para PageRank+Thesaurus e ROUGE-1 = 0,5426 para PageRank+Stem+StopRem. Embora esses resultados sejam fruto de um experimento um

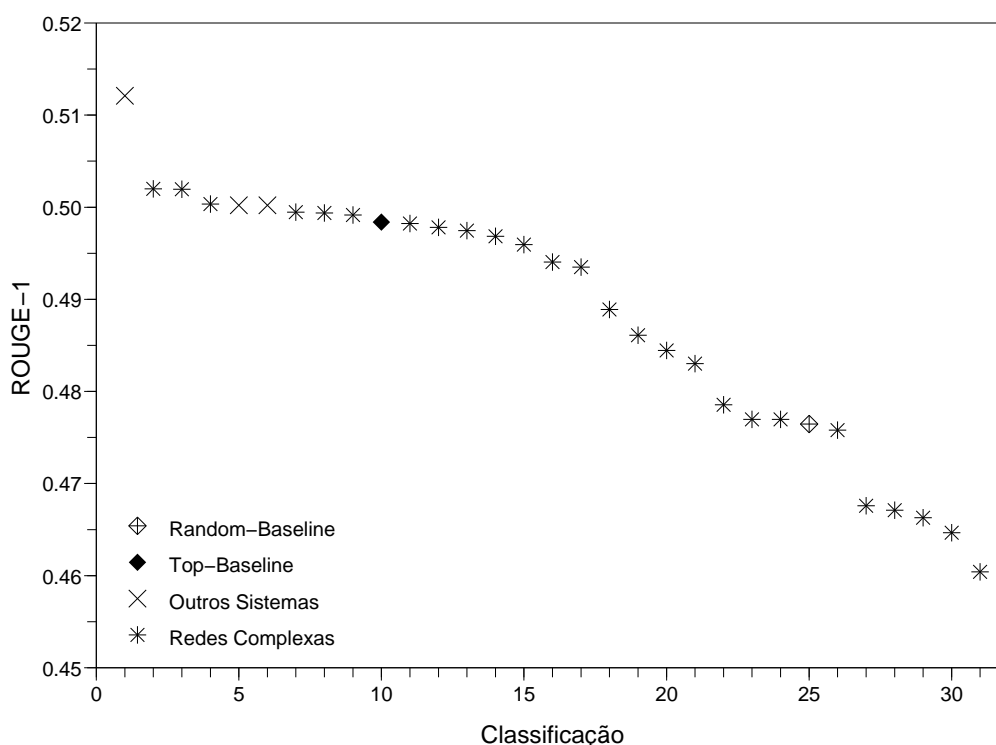


Figura 5.3: Valores ROUGE-1 médios dos sumarizadores da Tabela 5.4 (cópus TeMário). Os sistemas estão ordenados de forma decrescente de acordo com ROUGE-1.

pouco diferente do aqui realizado, a avaliação de Leite e Rino já mostra que as variações propostas no sumarizador PageRank original, utilizando agora recursos específicos para o processamento do português, são úteis para a geração de extratos. Além disso, era de se esperar que o Supor-v2 apresentasse bons resultados com a métrica ROUGE-1, pois já foi mostrado que seu desempenho tem destaque quando utilizada a Medida-F (Leite e Rino, 2006a).

5.4.3 DUC'2002 com ROUGE-1

Na Tabela 5.5 e na Figura 5.4 podem ser consultados os resultados referentes à avaliação baseada no cópus DUC'2002 e na métrica ROUGE-1. Os sistemas aqui propostos são comparados com os dois *baselines*, com as propostas de Mihalcea (2005) e com todos os 13 sistemas participantes da DUC'2002 (Over e Liggett, 2002). Lembrando que os sistemas propostos por Mihalcea são o PageRank, o HITS_A e o HITS_H, cujas melhores variações são, no caso do inglês: PageRank Backward, HITS_A Backward e HITS_H Forward. Como o pacote ROUGE não havia ainda sido criado na época da realização da DUC'2002, a classificação original dos sistemas participantes da conferência refere-se a uma avaliação manual. Entretanto, as métricas ROUGE têm alta correlação com a avaliação manual

realizada na conferência (Lin, 2004; Lin e Hovy, 2003), e fornecem uma classificação dos sistemas participantes muito próxima da obtida na época (Over e Liggett, 2002). Como o corpus da DUC'2002 é acompanhado pelos sumários automáticos gerados pelos sistemas participantes da conferência, foi possível calcular a métrica ROUGE-1 para cada um deles. Quatro desses sistemas foram apresentados nas Seções 2.1 e 2.2: ntt.duc02 (Hirao et al., 2002), ULeth131m (Brunn et al., 2002), ccsnsa.v2 (Schlesinger et al., 2002) e wpdv-xtr.v1 (van Halteren, 2002), todos com ROUGE-1 acima do Top-Baseline.

Tabela 5.5: Valores médios da medida ROUGE-1, obtidos comparando-se os extratos gerados automaticamente com os resumos de referência do corpus DUC'2002. Os sistemas estão ordenados decrescentemente por ROUGE-1. Os métodos *baseline* estão identificados por (\Rightarrow), enquanto que os sumarizadores propostos em outros trabalhos estão identificados por (\rightarrow). Os participantes da DUC'2002 estão acompanhados do nome da instituição onde o sistema foi desenvolvido.

	Sistemas	ROUGE-1
\rightarrow 1	HITS _A Backward	0,5023
\rightarrow 2	HITS _H Forward	0,5023
\rightarrow 3	ntt.duc02 - NTT	0,5013
\rightarrow 4	PageRank Backward	0,5008
\rightarrow 5	ULeth131m - Univ. of Lethbridge	0,4911
\rightarrow 6	ccsnsa.v2 - CCS-NSA	0,4889
\rightarrow 7	wpdv-xtr.v1 - Catholic Univ. Nijmegen	0,4865
\Rightarrow 8	Top-Baseline	0,4774
\rightarrow 9	kul.2002 - Catholic Univ. Leuven	0,4679
10	d -Anéis $r_i^{l,k}$	0,4625
11	d -Anéis r_i^l	0,4616
12	k -Núcleos n_i^l	0,4612
\rightarrow 13	uottawa - Univ. of Ottawa	0,4589
\rightarrow 14	lcc.duc02 - LCC	0,4561
\rightarrow 15	imp_col - Imperial College	0,4517
16	Caminhos Mínimos sp_i	0,4512
17	d -Anéis r_i^k	0,4511
18	Grau k_i	0,4509
19	Grau s_i	0,4497
20	k -Núcleos n_i^k	0,4490
21	Caminhos Mínimos sp_i^{wi}	0,4474
22	Caminhos Mínimos sp_i^{wc}	0,4471
23	Comunidades g_i	0,4421
24	Índice de Localidade l_i	0,4417
25	w -Cortes p_i^l	0,4384

Continua na próxima página...

	Sistemas	ROUGE-1
26	w -Cortes p_i^k	0,4339
→ 27	MICHIGAN - Univ. of Michigan	0,4336
→ 28	MSRC - Microsoft	0,4270
29	Índice de Localidade l_i^{mod}	0,4100
→ 30	gleans.v1 - ISI/Gleans	0,4099
31	Grau Hierárquico $k_i^{2,c}$	0,4052
32	Grau Hierárquico $s_i^{2,c}$	0,4052
33	Grau Hierárquico k_i^2	0,3985
⇒ 34	Random-Baseline	0,3945
35	Grau Hierárquico $s_i^{3,c}$	0,3945
36	Grau Hierárquico s_i^2	0,3945
37	Grau Hierárquico $k_i^{3,c}$	0,3908
38	Coeficiente de Aglomeração C_i^w	0,3776
39	Coeficiente de Aglomeração C_i	0,3768
40	Grau Hierárquico k_i^3	0,3676
41	Grau Hierárquico s_i^3	0,3665
42	Índice de Concordância m_i	0,3553
→ 43	SumUMFAR - Univ. of Montreal	0,1258
→ 44	bbn.headln - BBN	0,0651

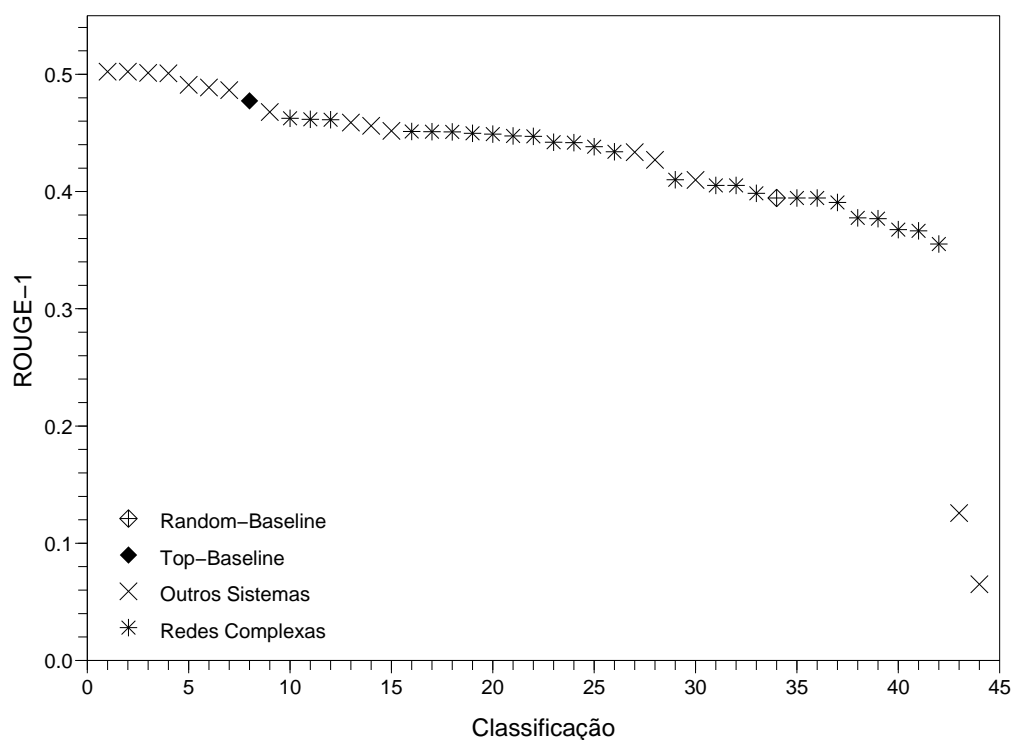


Figura 5.4: Valores ROUGE-1 médios dos sumariantes da Tabela 5.5 (córpus DUC'2002). Os sistemas estão ordenados de forma decrescente de acordo com ROUGE-1.

Novamente, ao dividirmos os sumarizadores propostos neste projeto em dois grupos, verifica-se que os melhores sistemas para o português também o são para o inglês. A Figura 5.4 mostra que, a partir do 29º sistema, a queda de desempenho para os sumarizadores baseados em redes complexas é mais acentuada. Considera-se, portanto, que o Grupo-1 de sumarizadores compreende os 14 primeiros sistemas (somente os aqui propostos, até a 26ª posição), e o Grupo-2 é formado pelos 12 sistemas a partir da 29ª posição. O Grupo-1 contém, portanto, os sumarizadores baseados nos d -anéis, nos k -núcleos, os caminhos mínimos, no grau, nas comunidades, no índice de localidade (não a versão modificada) e nos w -cortes. O índice de localidade modificado faz parte do Grupo-2, juntamente com o grau hierárquico, o coeficiente de aglomeração e o índice de concordância. A quase que constante divisão entre Grupos 1 e 2, desde o primeiro experimento com textos em português até este experimento com o corpus DUC'2002, é, por si só, interessante.

Verifica-se que agora o Top-Baseline tem desempenho superior com relação aos desempenhos obtidos nos experimentos com o TeMário (inclusive superior a todos os métodos propostos neste projeto). Parece haver uma mudança significativa na importância das primeiras sentenças, antes não tão relevantes de acordo com os resultados que o Top-Baseline vinha apresentando. Inclusive, o melhor sistema de redes complexas é agora o $r_i^{l,k}$, que usa os d -anéis e dá importância às primeiras sentenças do texto-fonte. O próximo experimento, com um corpus diferente de textos jornalísticos em inglês, reforça essa tendência. Contudo, não se sabe o porquê dessa maior relevância dada às primeiras sentenças. Ela pode ser creditada a uma ligeira diferença no estilo de escrita adotado nos jornais de língua inglesa, como reforça o experimento relatado na próxima seção. Por outro lado, verifica-se que os extratos gerados com o corpus DUC'2002 têm um número de sentenças pequeno: 5,47, em média¹¹. O primeiro experimento com o TeMário produz extratos 88% maiores, em número de sentenças, e o segundo, 52% maiores. Talvez os algoritmos propostos apresentem uma maior dificuldade em selecionar um pequeno número de vértices das redes, e os resultados abaixo do Top-Baseline no presente experimento podem ser reflexo disso, ao invés de uma diferença entre línguas. Além disso, o corpus DUC'2002 apresenta uma variação maior no tamanho dos textos-fonte do que o corpus TeMário (veja Figura 5.1), o que pode influenciar os algoritmos baseados em rede complexas.

Os sumarizadores de Mihalcea (2005) continuam figurando entre os melhores, com a diferença de que agora os algoritmos HITS_A e HITS_H ocupam as primeiras posições. Conforme ressaltado na seção anterior, ainda não é claro se as melhores performances obtidas para esses sistemas resultam dos algoritmos de classificação de páginas Web, ou das diferenças nas redes utilizadas pela autora. Sabe-se, por outro lado, que o tipo das

¹¹Esse número é calculado para os extratos gerados pelo Random-Baseline, quando a taxa de compressão é dada em número de palavras. Caso contrário, ele é fixo para todos os sumarizadores.

arestas influencia fortemente o algoritmo PageRank, pois sua variação em redes com arestas *forward* apresenta $\text{ROUGE-1} = 0,4202$. Quanto aos sistemas participantes da conferência DUC de 2002, considera-se principalmente os que figuram acima do Top-Baseline. O sistema ULeth131m faz uso de cadeias lexicais, um diferencial com relação aos sistemas baseados em redes complexas. Já os sistemas ntt.duc02, ccsnsa.v2 e wpdv-xtr.v1 empregam técnicas de aprendizado de máquina em atributos superficiais das sentenças, o que pode ser uma vantagem já que diversos atributos são considerados para cada sentença. Nesse caso, a maior complexidade desses sistemas é justificada pelos melhores resultados obtidos. Na outra ponta da Tabela 5.5, destacam-se dois sistemas da DUC'2002, justamente pelos valores ROUGE-1 extremamente baixos. Isso é explicado pelo fato dos sumários gerados por esses sistemas serem menores que os gerados pelos demais, o que influencia a métrica ROUGE-1. Na conferência de 2002 foi utilizado também um tipo de avaliação que fornece um bônus a sumários mais concisos (*length adjustment*) (Over e Liggett, 2002), fazendo com que esses sistemas apresentassem resultados substancialmente melhores.

Por fim, os problemas discutidos na Seção 5.4.1 a respeito do índice de localidade modificado, dos graus hierárquicos, do coeficiente de aglomeração e do índice de concordância, parecem também influenciar negativamente os extratos em língua inglesa. Os sistemas baseados nessas medidas (Grupo-2) continuam ocupando posições próximas à do Random-Baseline.

5.4.4 DUC'2001 com P , C e F

Os resultados obtidos segundo a aplicação das métricas Precisão, Cobertura e Medida-F no corpus DUC'2001 estão contidos na Tabela 5.6 e na Figura 5.5. Nesse caso, apenas os *baselines* e os métodos aqui propostos são considerados. Pode-se perceber que agora dois sumarizadores, o $r_i^{l,k}$ e o r_i^l , têm desempenho superior ao obtido para o método Top-Baseline. No experimento com o corpus DUC'2002, esses sumarizadores baseados nos d -anéis também ocupam as duas melhores posições (quando considera-se somente os sistemas baseados em redes complexas), entretanto, nenhum deles superou o Top-Baseline. Note que as primeiras sentenças de um texto-fonte continuam tendo grande importância na sumarização de textos em inglês e, conforme comentado na seção anterior, não se sabe o porquê dessa diferença de comportamento com relação às duas avaliações realizadas com o corpus de textos em português. Talvez isso se deva à grande variabilidade de tamanhos de textos-fonte no corpus DUC'2001, ao contrário do corpus TeMário, conforme já notado no outro experimento com textos em inglês. Além disso, os extratos gerados com o corpus DUC'2001 são 87% maiores, em média, e em número de sentenças, que os gerados com o corpus DUC'2002. São, portanto, mais próximos dos gerados para o português, o que

pode explicar o fato de agora dois métodos superarem o Top-Baseline. Isso mostra que tomar como definitivos os resultados de um único experimento é perigoso, pois as variáveis envolvidas em tais avaliações são inúmeras.

Tabela 5.6: Valores médios de Precisão (P), Cobertura (C) e Medida-F (F), obtidos comparando-se os extratos gerados automaticamente com os extratos de referência do corpus DUC'2001. Os sistemas estão ordenados decrescentemente por F . Os métodos *baseline* estão identificados por (\Rightarrow).

	Sistemas	P (%)	C (%)	F (%)
1	d -Anéis $r_i^{l,k}$	41,7	50,9	42,8
2	d -Anéis r_i^l	41,0	51,2	42,6
\Rightarrow 3	Top-Baseline	39,2	49,0	40,8
4	k -Núcleos n_i^l	39,6	48,1	40,6
5	w -Cortes p_i^l	39,1	48,0	40,3
6	Caminhos Mínimos sp_i	38,8	46,1	39,4
7	Grau k_i	38,5	46,1	39,2
8	d -Anéis r_i^k	38,3	46,0	39,1
9	w -Cortes p_i^k	38,0	45,7	38,8
10	Grau s_i	37,3	44,7	38,1
11	k -Núcleos n_i^k	37,4	44,6	38,0
12	Caminhos Mínimos sp_i^{wi}	37,1	44,0	37,7
13	Comunidades g_i	37,2	43,5	37,6
14	Caminhos Mínimos sp_i^{wc}	36,9	44,0	37,5
15	Índice de Localidade l_i	35,9	43,5	36,7
16	Índice de Localidade l_i^{mod}	27,7	31,7	27,8
17	Índice de Concordância m_i	25,2	28,5	25,2
18	Grau Hierárquico $s_i^{2,c}$	23,4	26,7	23,7
19	Grau Hierárquico $k_i^{2,c}$	23,5	26,5	23,5
\Rightarrow 20	Random-Baseline	23,5	26,8	23,4
21	Grau Hierárquico $s_i^{3,c}$	22,4	25,5	22,7
22	Grau Hierárquico k_i^2	22,2	25,7	22,6
23	Grau Hierárquico s_i^2	22,0	25,7	22,5
24	Grau Hierárquico $k_i^{3,c}$	22,3	25,3	22,4
25	Grau Hierárquico k_i^3	21,3	22,5	20,9
26	Grau Hierárquico s_i^3	21,0	22,3	20,6
27	Coeficiente de Aglomeração C_i	18,1	18,9	17,5
28	Coeficiente de Aglomeração C_i^w	17,9	18,7	17,3

No mais, percebe-se, pela quarta vez, uma clara divisão entre dois grupos de sumarizadores. Do 15º para o 16º sistema, o decaimento da Medida-F é visível (Figura 5.5), o que acarreta a divisão dos sistemas propostos entre Grupo-1 e Grupo-2, nos moldes das divisões realizadas nos três experimentos anteriores. O desempenho dos sumarizadores baseados em redes é, de maneira geral, constante, quando são consideradas as variações

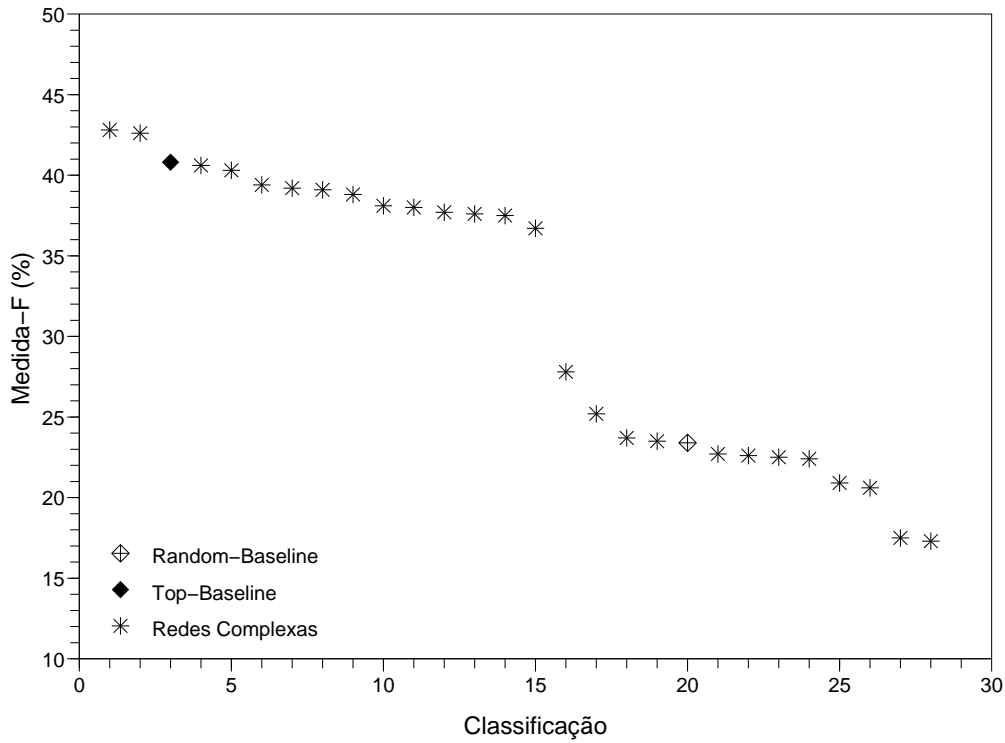


Figura 5.5: Medida-F média (F) dos sumarizadores da Tabela 5.6 (cópus DUC'2001). Os sistemas estão ordenados de forma decrescente de acordo com F .

de língua, de cópus e de métricas de avaliação registradas nos experimentos realizados. À primeira vista parece haver uma grande permutação quanto às posições dos sumarizadores, quando são comparados os resultados das Tabelas 5.3–5.6; contudo, quando são levadas em consideração as separações entre Grupo-1 e Grupo-2, o padrão¹² de desempenho dos sistemas fica mais claro. Como esse comportamento praticamente constante é fruto da análise de quatro experimentos diferentes, pode-se afirmar com uma certa segurança que, entre os métodos propostos, existem os (i) promissores, com alguns resultados próximos do estado da arte (para o TeMário), e existem os (ii) problemáticos, com resultados próximos aos do Random-Baseline. Relembrando, o grupo dos promissores (Grupo-1) é formado pelos métodos baseados nos seguintes conceitos: Grau (k_i e s_i), Caminhos Mínimos (sp_i , sp_i^{wc} e sp_i^{wi}), Índice de Localidade (l_i), d -Anéis (r_i^l , r_i^k e $r_i^{l,k}$), k -Núcleos (n_i^l e n_i^k), w -Cortes (p_i^l e p_i^k) e Comunidades (g_i). Já o grupo dos sumarizadores problemáticos (Grupo-2) é composto pelos métodos baseados nos conceitos: Coeficiente de Aglomeração (C_i e C_i^w), Índice de Localidade Modificado (l_i^{mod}), Índice de Concordância (m_i) e Grau Hierárquico (k_i^2 , $k_i^{2,c}$, k_i^3 , $k_i^{3,c}$, s_i^2 , $s_i^{2,c}$, s_i^3 e $s_i^{3,c}$).

Alguns dos métodos do Grupo-1 dão prioridade a vértices, ou grupos de vértices,

¹²Somente no experimento da Seção 5.4.2 ocorre uma pequena diferença na divisão entre Grupo-1 e Grupo-2, referente a um dos métodos baseados nos w -cortes.

concentradores de conexões (graus, k -núcleos e w -cortes), ou seja, são selecionadas as sentenças que compartilham vários termos com diversas outras sentenças. O uso dos d -anéis também dá importância à concentração de conexões, pois utiliza o *hub* como ponto de partida na identificação e seleção dos vértices contidos em suas hierarquias. O índice de localidade já analisa as conexões de um determinado nó com relação ao restante da rede, ou seja, procura por vértices cujos vizinhos compartilhem poucas arestas com os demais vértices da rede. Com essa medida, procura-se escolher sentenças que centralizem outras sentenças do texto-fonte e sejam representativas desse grupo de sentenças. Possivelmente, os extratos são mais informativos por conter várias dessas sentenças representativas. O algoritmo baseado em comunidades também segue essa linha de construção de extratos formados por sentenças representativas de grupos de sentenças. Já os caminhos mínimos derivam de uma análise global da rede, pois servem para calcular as mínimas distâncias entre todos os pares de vértices. Sentenças próximas das demais são escolhidas na formação de um extrato pelos sumarizadores baseados em distância mínima. Os conceitos embutidos nesses sumarizadores parecem ser úteis para a informatividade de extratos (considerando textos jornalísticos como fonte de entrada), com destaque para os sistemas listados na Tabela 5.7. Note que os sumarizadores baseados nos d -anéis figuram entre os três melhores sistemas nos quatro experimentos realizados, com maior relevância nos testes feitos com os corpúscos em inglês. Já os métodos baseados no grau e um dos que usam caminhos mínimos têm destaque nas avaliações para o português, enquanto que um dos métodos inspirados nos k -núcleos figura entre os três melhores apenas nas avaliações para o inglês. A maior parte desses métodos utiliza o conceito de grau de maneira clara (d -anéis, k -núcleos e, é claro, os graus k_i e s_i). Métodos em que se procura construir sumários que contenham sentenças representativas de grupos de sentenças (índice de localidade e comunidades) não figuram entre os melhores sistemas, apesar de estarem no Grupo-1. Esses métodos devem ser mais bem elaborados, principalmente com o uso de redes que apresentem índices de modularidade mais altos. Cabe aqui ressaltar que, com a criação do índice de localidade modificado, procurou-se construir sumários ainda mais informativos que os construídos pelo seu antecessor, o índice de localidade não modificado. Contudo, seus resultados são ainda piores (ele pertence ao Grupo-2), o que indica que a construção de extratos abrangentes, ou seja, que cobrem todos os tópicos do texto-fonte, deva ser redefinida. Outro método, de baixo desempenho, e com o mesmo objetivo do índice de localidade, é o baseado no índice de concordância. Ainda a respeito do Grupo-2, parece haver uma maior consideração, por parte de alguns métodos, pelas conexões dos vizinhos de um dado vértice do que pelas conexões que o próprio vértice em questão possui. Os sumarizadores baseados no grau hierárquico e no coeficiente de aglomeração comportam-se claramente dessa maneira.

Quando os resultados são analisados levando-se em consideração a língua dos textos-

Tabela 5.7: Sistemas baseados em redes complexas que apresentaram os melhores desempenhos nos quatro experimentos realizados. Cada coluna refere-se a um experimento, de maneira que os três melhores sistemas em cada um deles esteja marcado por •. O primeiro colocado em cada experimento está marcado também por colchetes.

Sumarizadores	TeMário		DUC'2002	DUC'2001
	P, C e F	ROUGE-1	ROUGE-1	P, C e F
d -Anéis r_i^l			•	•
d -Anéis r_i^k	•			
d -Anéis $r_i^{l,k}$		•	[•]	[•]
Grau k_i	•	•		
Grau s_i		[•]		
k -Núcleos n_i^l			•	•
Caminhos Mínimos sp_i^{wc}	[•]			

fonte, percebe-se que, para textos em português, os métodos propostos neste projeto ficam próximos dos melhores métodos propostos em outros trabalhos (SuPor-v2, SuPor, ClassSumm, PageRank, HITS_A e HITS_H). Provavelmente os sistemas SuPor-v2, SuPor e ClassSumm atinjam bons resultados pelo uso de aprendizado de máquina e de diversos atributos para as sentenças, envolvendo inclusive semântica. PageRank e HITS são algoritmos criados também para grafos, e são aplicados em redes de sentenças semelhantes às aqui utilizadas; contudo, para textos em inglês, apresentam resultados sensivelmente melhores do que os obtidos para os métodos baseados em redes complexas. Além disso, diversos outros sistemas que participaram da DUC'2002 conseguem resultados melhores para o inglês do que os aqui obtidos. Deve ser melhor investigado o porquê dos métodos baseados em redes complexas terem desempenho apenas razoável com os corpúscos em inglês, já que, superar o Top-Baseline, nesses casos, mostrou-se um fato incomum.

5.5 Correlações entre Sumarizadores

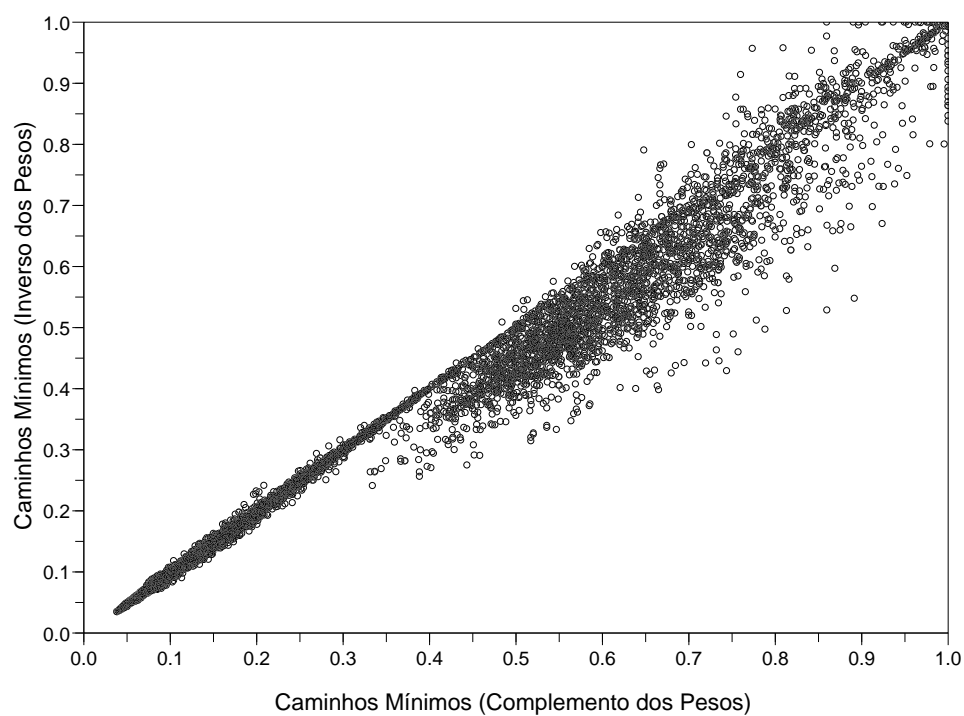
Visando complementar a avaliação descrita nas seções anteriores, foi realizada também uma análise da similaridade entre os sumarizadores propostos. Em outras palavras, verificou-se, por meio da correlação entre medidas, quais sumarizadores tendem a escolher as mesmas sentenças na construção de um extrato, e quais tendem a formar extratos complementares. Como todos os sumarizadores associam um valor a cada vértice da rede, é possível analisar as semelhanças e diferenças entre os extratos sem que eles sejam literalmente construídos. Uma ferramenta que possibilita essa análise é o coeficiente de correlação de Pearson (Casella e Berger, 1990). Esse coeficiente quantifica o nível de correlação linear entre duas amostras

(X_1, X_2, \dots, X_n) e (Y_1, Y_2, \dots, Y_n) , das variáveis aleatórias X e Y , da seguinte maneira:

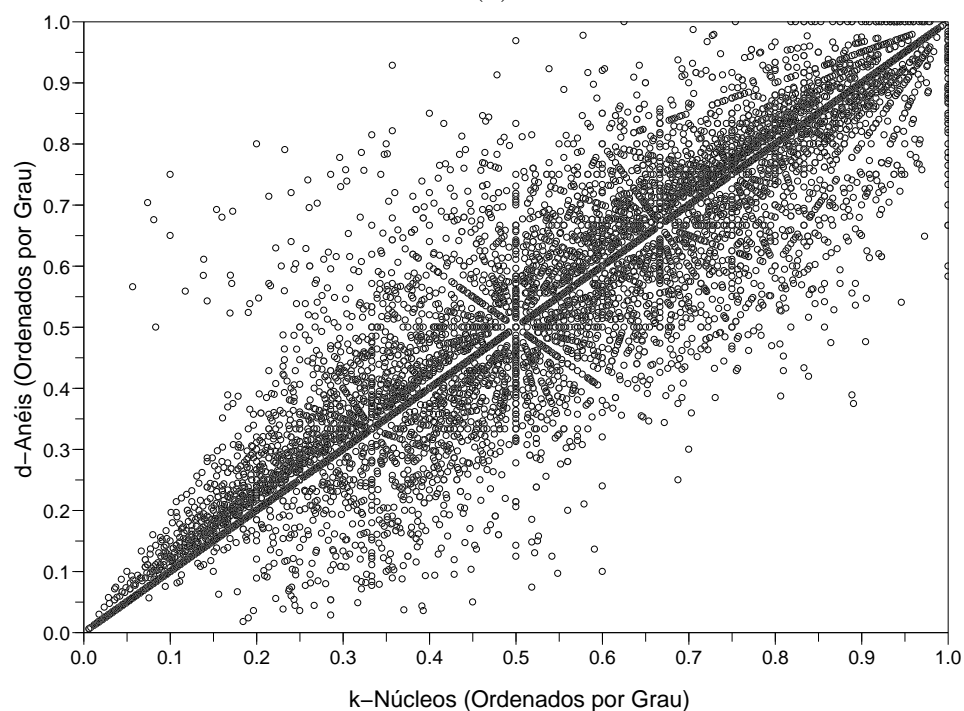
$$r_{X,Y} = \frac{\sum_{t=1}^n (X_t - \bar{X})(Y_t - \bar{Y})}{\sqrt{\sum_{t=1}^n (X_t - \bar{X})^2 \sum_{t=1}^n (Y_t - \bar{Y})^2}}, \quad (5.6)$$

onde \bar{X} é a média amostral de X , e \bar{Y} é a média amostral de Y . O coeficiente $r_{X,Y}$ varia de -1 a 1 e, quanto maior for, em módulo, mais forte a correlação linear entre X e Y . Quando $|r_{X,Y}|$ é 1, a variação entre X e Y pode ser perfeitamente representada por uma equação do primeiro grau $Y_t = mX_t + c$, onde m é o coeficiente angular da reta, e c é seu coeficiente linear. Por outro lado, coeficientes de correlação próximos de zero indicam que a correlação entre X e Y está longe de ser linear. Outra propriedade de $r_{X,Y}$ refere-se ao sinal, de maneira que, se $r_{X,Y} < 0$, então a inclinação da correlação é negativa (com $m < 0$ na reta que aproxima a relação entre X e Y), e, se $r_{X,Y} > 0$, então a inclinação da correlação é positiva (com $m > 0$). Ao selecionar as medidas da Tabela 4.1, duas a duas, e, a seguir, calcular os respectivos coeficientes de correlação de Pearson, obtêm-se valores que podem ser utilizados na análise das semelhanças e diferenças entre os sumarizadores propostos. Além disso, não é necessário analisar uma rede por vez, pois é possível obter essas medidas de redes complexas para todo um córpus, com um único coeficiente de correlação que considere todos os vértices de todas as redes. Para tanto, é necessário normalizar as medidas utilizadas pelos sumarizadores, de modo que o maior valor de uma medida em uma determinada rede seja igual a 1.

O coeficiente de Pearson pode substituir a análise de gráficos como os da Figura 5.6, em que cada vértice (sentença) do córpus DUC'2002 representa um ponto (em cada um desses gráficos, todas as sentenças do córpus estão representadas). Na Figura 5.6a, relacionam-se as medidas sp_i^{wc} (caminhos mínimos com complemento dos pesos) e sp_i^{wi} (caminhos mínimos com inverso dos pesos). Aproximadamente, essas duas medidas tendem a gerar valores próximos para os mesmo vértices, fato visível na Figura 5.6a, principalmente para valores próximos de 0. O coeficiente de Pearson, nesse caso, é 0,99, e realmente os dois tipos de caminhos mínimos considerados relacionam-se de maneira aproximadamente linear. Como os sumarizadores baseados em caminhos mínimos selecionam os vértices que apresentam os menores valores para tais medidas, percebe-se que sp_i^{wc} e sp_i^{wi} dão origem a extratos muito parecidos. Já na Figura 5.6b, as medidas utilizadas foram n_i^k (k -núcleos ordenados por grau) e r_i^k (d -anéis ordenados por grau). Com coeficiente de Pearson igual a 0,96, percebe-se que existe uma maior quantidade de pontos que se afastam de uma reta hipotética que corta a diagonal do plano. Entretanto, existe uma maior concentração de pontos próximos a essa reta, e o coeficiente de Pearson indica essa forte relação linear. Perceba, novamente, que sentenças com baixos valores de n_i^k e r_i^k são utilizadas na formação de extratos e, por conseguinte, tendem a selecionar extratos semelhantes (observe a menor



(a)



(b)

Figura 5.6: Dois exemplos de correlações entre sumarizadores no corpus DUC'2002: (a) entre sp_i^{wc} (caminhos mínimos com complemento dos pesos) e sp_i^{wi} (caminhos mínimos com inverso dos pesos), com coeficiente de Pearson igual a 0,99, e (b) entre n_i^k (k -núcleos ordenados por grau) e r_i^d (d -anéis ordenados por grau), com coeficiente de Pearson igual a 0,96. As medidas foram normalizadas separadamente em cada texto, e cada ponto refere-se a uma das 14.881 sentenças do corpus DUC'2002.

Tabela 5.8: Coeficientes de correlação linear entre as medidas do Grupo-1 (cópus TeMário).

	k_i	s_i	sp_i	sp_i^{wc}	sp_i^{wi}	l_i	r_i^l	r_i^k	$r_i^{l,k}$	n_i^l	n_i^k	p_i^l	p_i^k	g_i
k_i	-													
s_i	0,94	-												
sp_i	-0,41	-0,38	-											
sp_i^{wc}	-0,42	-0,39	1,00	-										
sp_i^{wi}	-0,44	-0,41	0,99	1,00	-									
l_i	0,86	0,81	-0,46	-0,47	-0,49	-								
r_i^l	-0,62	-0,61	0,39	0,40	0,41	-0,54	-							
r_i^k	-0,84	-0,83	0,42	0,42	0,43	-0,74	0,80	-						
$r_i^{l,k}$	-0,78	-0,77	0,41	0,42	0,43	-0,69	0,88	0,91	-					
n_i^l	-0,77	-0,74	0,41	0,41	0,42	-0,74	0,76	0,84	0,86	-				
n_i^k	-0,88	-0,86	0,42	0,43	0,43	-0,82	0,69	0,93	0,86	0,90	-			
p_i^l	-0,61	-0,69	0,39	0,41	0,43	-0,56	0,68	0,69	0,71	0,67	0,67	-		
p_i^k	-0,80	-0,84	0,41	0,43	0,44	-0,73	0,66	0,86	0,80	0,77	0,87	0,87	-	
g_i	-0,61	-0,71	0,26	0,27	0,29	-0,52	0,55	0,68	0,65	0,60	0,66	0,80	0,78	-

dispersão na origem do plano na Figura 5.6b).

Ao analisar visualmente todas as correlações entre os 26 sumarizadores propostos, são necessários 325 gráficos nos moldes dos da Figura 5.6. Optou-se, portanto, por utilizar somente os coeficientes de Pearson na análise que se segue. Adicionalmente, somente os 14 sumarizadores do Grupo-1 foram analisados¹³, o que acarretou em 91 valores de correlação para cada córpis utilizado neste projeto, todos listados nas Tabelas 5.8–5.10 (uma para cada córpis). Nessas tabelas, as correlações maiores ou iguais a 0,85, em módulo, são destacadas em negrito. Se um valor menor que 0,85 estiver em destaque, quer dizer que a correlação ultrapassa esse limite em pelo menos um dos outros córpis. Isso facilita a análise das correlações e, como pode ser observado nas Tabelas 5.8–5.10, nenhuma correlação abaixo de 0,80 está em negrito, ou seja, todas as correlações em destaque têm módulo alto. Note também que esses valores em destaque são negativos quando o sentido de aplicação entre as duas medidas em questão é diferente, ou seja, os sentidos \uparrow e \downarrow listados para os sumarizadores na Tabela 4.1 são opostos.

Primeiramente, observa-se uma alta correlação entre os dois tipos de grau, k_i e s_i , com coeficientes de Pearson acima de 0,93. Isso indica que, quando extratos são gerados por esses dois métodos, a tendência é que sejam muito semelhantes, pois vértices com altos valores de k_i tendem a apresentar também altos valores de s_i (próximos de 1, quando normalizados). Outras correlações altas são as que envolvem os três tipos de caminhos mínimos, sp_i , sp_i^{wc} e sp_i^{wi} , com valores¹⁴ acima de 0,98. Novamente, as variações que levam em conta os pesos no cálculo de uma determinada medida não acarretam em grandes mudanças. Considerando-se correlações um pouco mais baixas, agora no grupo dos métodos derivados dos d -anéis, percebe-se que r_i^l e r_i^k correlacionam-se bem com $r_i^{l,k}$, o que pode ser explicado pelo fato de $r_i^{l,k}$ ser uma fusão de r_i^l e r_i^k . Entre os sumarizadores baseados nos k -núcleos, n_i^l e n_i^k , existe também uma forte correlação, bem como entre os sumarizadores baseados nos w -cortes, p_i^l e p_i^k . Vale sempre lembrar que, mesmo em correlações altas, pequenas diferenças nos grupos de sentenças selecionadas por um ou outro método de sumarização podem ser importantes, o que justifica a definição de todas essas variações de sumarizadores.

Considerando-se agora correlações entre sumarizadores que não sejam variantes de um mesmo conceito principal, percebe-se que os dois tipos de grau correlacionam-se bem com o índice de localidade (l_i), com um dos d -anéis (r_i^k), com um dos k -núcleos (n_i^k) e com um dos w -cortes (p_i^k). Isso significa que os *hubs* tendem a apresentar altos índices de

¹³Optou-se por não analisar os sumarizadores do Grupo-2, por já se saber que tais métodos tendem a gerar extratos diferentes dos contruídos pelos métodos do Grupo-1, como mostram os resultados relatados em seções anteriores.

¹⁴Em alguns casos, esses valores são tão próximos da correlação total, que, quando escritos em duas casas decimais, são aproximados para 1,00 (vide Tabelas 5.8 e 5.9).

Tabela 5.9: Coeficientes de correlação linear entre as medidas do Grupo-1 (corpus DUC' 2002).

	k_i	s_i	sp_i	sp_i^{wc}	sp_i^{wi}	l_i	r_i^l	r_i^k	$r_i^{l,k}$	n_i^l	n_i^k	p_i^l	p_i^k	g_i
k_i	-													
s_i	0,94	-												
sp_i	-0,30	-0,28	-											
sp_i^{wc}	-0,32	-0,31	1,00	-										
sp_i^{wi}	-0,35	-0,34	0,98	0,99	-									
l_i	0,91	0,86	-0,36	-0,38	-0,41	-								
r_i^l	-0,60	-0,58	0,39	0,40	0,42	-0,56	-							
r_i^k	-0,85	-0,84	0,42	0,44	0,45	-0,79	0,75	-						
$r_i^{l,k}$	-0,77	-0,75	0,41	0,43	0,44	-0,73	0,87	0,90	-					
n_i^l	-0,78	-0,75	0,41	0,42	0,44	-0,77	0,77	0,86	0,89	-				
n_i^k	-0,86	-0,84	0,42	0,44	0,45	-0,82	0,68	0,96	0,87	0,91	-			
p_i^l	-0,63	-0,72	0,39	0,42	0,45	-0,60	0,66	0,71	0,72	0,70	0,70	-		
p_i^k	-0,78	-0,84	0,41	0,44	0,46	-0,74	0,63	0,86	0,78	0,78	0,86	0,90	-	
g_i	-0,66	-0,76	0,30	0,32	0,35	-0,61	0,55	0,73	0,68	0,66	0,73	0,81	0,82	-

Tabela 5.10: Coeficientes de correlação linear entre as medidas do Grupo-1 (córpus DUC' 2001).

	k_i	s_i	sp_i	sp_i^{wc}	sp_i^{wi}	l_i	r_i^l	r_i^k	$r_i^{l,k}$	n_i^l	n_i^k	p_i^l	p_i^k	g_i
k_i	-													
s_i	0,93	-												
sp_i	-0,25	-0,24	-											
sp_i^{wc}	-0,28	-0,27	0,99	-										
sp_i^{wi}	-0,31	-0,30	0,98	0,99	-									
l_i	0,95	0,87	-0,27	-0,30	-0,33	-								
r_i^l	-0,58	-0,57	0,30	0,32	0,35	-0,57	-							
r_i^k	-0,87	-0,87	0,35	0,37	0,39	-0,84	0,68	-						
$r_i^{l,k}$	-0,78	-0,76	0,33	0,35	0,38	-0,77	0,85	0,87	-					
n_i^l	-0,80	-0,77	0,33	0,35	0,38	-0,82	0,75	0,86	0,89	-				
n_i^k	-0,88	-0,87	0,35	0,37	0,39	-0,86	0,63	0,97	0,85	0,90	-			
p_i^l	-0,65	-0,75	0,32	0,35	0,39	-0,62	0,65	0,72	0,71	0,69	0,71	-		
p_i^k	-0,80	-0,86	0,34	0,37	0,40	-0,77	0,59	0,87	0,77	0,77	0,87	0,91	-	
g_i	-0,70	-0,80	0,26	0,29	0,33	-0,67	0,53	0,77	0,67	0,67	0,76	0,80	0,84	-

localidade, ou seja, os vizinhos de um *hub* não apresentam muitas conexões com os demais nós da rede. Já os referidos métodos baseados nos d -anéis, nos k -núcleos e nos w -cortes são variações que utilizam a ordenação por grau, o que acarreta em forte correlação com os próprios graus. O índice de localidade também apresenta boa correlação com os k -núcleos ordenados por grau (n_i^k), o que pode ser explicado pelo fato desses dois métodos serem também correlacionados com os graus. Os dois tipos de k -núcleos, por sua vez, têm alta correlação com dois dos três tipos de d -anéis (r_i^k e $r_i^{l,k}$), métodos aparentemente bem diferentes. Esse fato indica que os vértices selecionados em d -anéis calculados a partir do *hub* tendem a ser bem conectados entre si, já que os métodos baseados nos k -núcleos dão prioridade a grupos de vértices coesos. O grau parece ter influência também nas fortes correlações que p_i^k (um dos w -cortes) tem com r_i^k (d -anéis) e n_i^k (k -núcleos), já que todos utilizam a ordenação por grau.

Enquanto que sumarizadores bastante correlacionados tendem a gerar extratos parecidos, sumarizadores pouco correlacionados podem dar origem a extratos que apresentam visões bastante distintas de um mesmo texto-fonte. Quando as correlações mais baixas são consideradas, nota-se que os três tipos de caminhos mínimos apresentam sistematicamente correlações abaixo de 0,50, em módulo, quando comparados aos demais métodos. Tem-se indícios de que os caminhos mínimos geram extratos bem diferentes dos gerados pelos outros sumarizadores, e portanto, podem ser considerados métodos complementares aos demais. Um novo método poderia ser investigado procurando-se unir dois sumarizadores complementares, de maneira a obter extratos mais informativos. O sumarizador baseado nas comunidades, por sua vez, também apresenta baixas correlações com vários outros métodos. Entretanto, a complementaridade desse método resulta na seleção de sentenças menos importantes para os extratos, como mostra seu desempenho razoável quando comparado aos outros sumarizadores do Grupo-1, tendo como referência as medidas ROUGE-1, P , C e F .

5.6 Exemplos de Extratos Gerados

Para fechar este capítulo de avaliação dos sumarizadores propostos, são fornecidos e comentados, a seguir, dois exemplos de extratos construídos automaticamente neste projeto. Cada exemplo contempla uma das línguas avaliadas (português ou inglês), e é acompanhado do respectivo resumo ou extrato manual para referência. O primeiro exemplo refere-se à aplicação do sumarizador sp_i^{wc} (baseado nos caminhos mínimos) em um texto-fonte do corpus TeMário, enquanto que, no outro exemplo, aplicou-se o sumarizador $r_i^{l,k}$ (baseado nos d -anéis) em um texto-fonte do corpus DUC'2001. Esses dois métodos figuram entre os

melhores propostos neste trabalho.

A Figura 5.7 mostra um texto-fonte do corpus TeMário já segmentado, em que as sentenças aparecem isoladamente e acompanhadas de um número. Esse número indica a prioridade dada a cada sentença pelo sumariador sp_i^{wc} na construção de um extrato. Por exemplo, a primeira sentença da Figura 5.7 aparece em 3º lugar na lista de prioridades, ou seja, existem duas outras sentenças que devem ser inseridas no extrato antes que a referida sentença seja utilizada. Tem-se, portanto, uma visão global da aplicação do algoritmo sp_i^{wc} , independentemente de taxas de compressão. Esse algoritmo é de difícil interpretação quando somente o texto-fonte é analisado, pois são utilizadas informações globais da rede no cálculo dos caminhos mínimos. Sabe-se, ao menos, que a sentença com prioridade 1 está mais próxima das outras sentenças, na média. Aliás, essa sentença fornece uma boa noção a respeito do principal assunto veiculado no texto (“reescrever os anos de chumbo da história brasileira”), o que indica que o sumariador sp_i^{wc} já consegue, nesse caso, dar máxima prioridade a uma importante sentença.

A fim de fornecer uma indicação mais ampla de quais sentenças do texto-fonte são realmente importantes para a sumarização, encontra-se na Figura 5.8 o respectivo resumo manual, retirado do corpus TeMário. As sentenças do resumo são acompanhadas por um ou mais números, os quais referem-se às prioridades dadas por sp_i^{wc} às sentenças do texto-fonte. Cada sentença do resumo manual cobre informações veiculadas por um conjunto de sentenças do texto-fonte, as quais são identificadas justamente por esses números. A primeira sentença do resumo manual, por exemplo, apresenta informações contidas nas sentenças com prioridades 3, 10 e 13. Nota-se, na Figura 5.8, que as sentenças com alta prioridade pouco influenciaram a construção do resumo manual (somente as sentenças de prioridade 2, 3 e 6). Como o sumariador sp_i^{wc} apresenta bons desempenhos nas avaliações com o TeMário, pode parecer que o exemplo dado nesta seção seja um caso em que sp_i^{wc} gera extratos de baixa qualidade.

Contudo, ao ler o extrato gerado por sp_i^{wc} (de tamanho aproximadamente igual ao do resumo manual, em número de palavras - Figura 5.9), percebe-se que ele não deixa muito a desejar. Embora o extrato automático não contemple diversas informações veiculadas pelo resumo manual (as opiniões de Suzana Lisboa, pedidos de indenização), o próprio resumo manual deixa de inserir informações importantes (como os possíveis processos envolvendo médicos legistas). A análise é delicada, e, considerando a construção de sumários genéricos, é difícil dizer em alguns casos quais partes do texto-fonte são mais importantes que outras. Mas, no caso do texto-fonte da Figura 5.7, sabe-se qual informação deve constar em um sumário (a existência de uma comissão especial que procura reescrever a história da ditadura), e, tanto o resumo manual como o extrato automático o contemplam de forma clara.

3	Porto Alegre - A Comissão Especial dos Desaparecidos Políticos, que terá sua quarta reunião na quinta-feira quando irá estudar 20 novos casos, está reescrevendo para a História do Brasil as páginas mais obscuras da ditadura militar de 64: as das torturas, mortes e dos desaparecidos.
13	Documentos e versões divulgados na época pelas autoridades estão sendo desmentidos, formal e legalmente, um a um.
10	"Caem por terra todas as versões oficiais da época", afirma o representante das Forças Armadas na comissão, general Osvaldo Pereira Gomes, 65 anos.
1	O general Gomes se diz tranqüilo e isento para ajudar a reescrever, com seus votos, os anos de chumbo da história brasileira.
5	Por isso, também com tranqüilidade, o general Gomes apóia totalmente a segunda e futura etapa dos trabalhos da comissão na tentativa de localização dos corpos por "estar na lei de indenização dos desaparecidos.
4	Além disso, é uma questão humanitária, o direito de os familiares enterrarem os restos mortais dos seus entes queridos".
18	Contraditório - "Nas reuniões da comissão, faço o contraditório, pois há muita paixão política envolvida.
19	Alguns militares podem não gostar como atuo, mas ajo com isenção e independência.
16	Nunca recebi pressões de quem quer que seja, nem aceitaria", garante o general.
2	O presidente da comissão, Miguel Reale Jr, 52 anos, destaca que "o regime militar autorizou e deu guarida a todas as violências, mesmo quando a versão oficial era mentirosa, até bisonha.
7	Ao reescrevermos essa parte da História brasileira, estamos resgatando a credibilidade sobre o conceito público de civilidade e dando o exemplo às novas gerações.
17	Ao darmos nossos votos, ficaram e ficarão registradas as responsabilidades de todos nesse período perante a História".
12	Suzana Lisboa, 44 anos, representante na comissão dos parentes de desaparecidos, confessa "imensa emoção" que já a fez chorar nas reuniões, por ver "restabelecida a verdade histórica que as famílias vêm denunciando há tanto tempo, mostrando que a versão oficial sobre supostos atropelamentos, tiroteios ou suicídios era mentirosa".
14	Responsabilidade - "Os governos da época tinham responsabilidade por aquela situação, já que a política oficial era essa, do aparato do Estado por trás dos órgãos de segurança.
11	Presidentes e ministros com exceção talvez de alguns sabiam o que estava acontecendo", completa.
9	As antigas e falsas versões continuam a fazer parte de documentos da área militar como nos relatórios sobre desaparecidos entregues pelos ministros da Marinha e Aeronáutica ao então ministro da Justiça Maurício Correa (governo Itamar Franco), alerta Suzana.
15	Por isso, ela sugeriu à comissão que, ao fim dos trabalhos seja publicado em livro o relatório final.
6	Até 9 de maio, a comissão, instalada no prédio Anexo II do Ministério da Justiça, estará recebendo novos pedidos de indenização de famílias de mortos e desaparecidos, cifras que variam de R\$ 100 mil a R\$ 150 mil.
8	Embora a comissão não tenha poder para responsabilizar individualmente os torturadores devido à Lei da Anistia, Suzana alerta que as famílias poderão utilizar futuramente as documentações obtidas e aprovadas para, por exemplo, "processar médicos legistas que deram laudos falsos para acobertar torturas".

Figura 5.7: Exemplo de aplicação do algoritmo sp_i^{wc} (caminhos mínimos) em texto-fonte do corpus TeMário. O texto-fonte aparece segmentado, e os números presentes no início de cada sentença indicam a ordem de prioridade dada pelo sumarizador sp_i^{wc} .

3, 10, 13	A Comissão Especial dos Desaparecidos Políticos está reescrevendo as versões sobre torturas, mortes durante a Revolução de 64, convicta de que os relatos oficiais da época estão cheios de mentiras.
18, 19	O general Osvaldo Pereira Gomes, representante das Forças Armadas, diz que nas reuniões atua como contraditório, procurando atenuar com isenção as paixões políticas.
2, 7, 17	O presidente da comissão, Miguel Reale Jr, destaca as arbitrariedades cometidas pelo regime militar, e tem a certeza de que o grupo está resgatando a verdade histórica e passando à posteridade o exato conceito de civilidade pública.
12	Suzana Lisboa, representante dos parentes dos desaparecidos, fala da sua emoção ao ser esclarecida a verdade que as famílias vinham buscando há tempo.
9, 15	E, dada a continuidade de versões oficiais falsas, sugeriu que este trabalho da comissão se concretize em um livro.
6	O trabalho da comissão inclui os pedidos de indenização para as famílias dos mortos e desaparecidos.

Figura 5.8: Resumo manual, retirado do corpus TeMário, construído para o texto-fonte da Figura 5.7. Os números ao lado de cada sentença são relacionados aos números dados às sentenças da Figura 5.7, e indicam quais sentenças do texto-fonte contêm as informações veiculadas por cada sentença do resumo manual.

Para as outras informações (opiniões de membros da comissão e de uma representante de familiares desaparecidos), parece haver uma maior maleabilidade quanto à sua inserção em um sumário. O extrato automático, apesar de não muito abrangente (deixa de incluir muitas informações contidas na segunda metade do texto-fonte), fornece uma boa idéia do conteúdo completo que procura sumarizar.

Porto Alegre - A Comissão Especial dos Desaparecidos Políticos, que terá sua quarta reunião na quinta-feira quando irá estudar 20 novos casos, está reescrevendo para a História do Brasil as páginas mais obscuras da ditadura militar de 64: as das torturas, mortes e dos desaparecidos.
O general Gomes se diz tranqüilo e isento para ajudar a reescrever, com seus votos, os anos de chumbo da história brasileira.
Por isso, também com tranqüilidade, o general Gomes apóia totalmente a segunda e futura etapa dos trabalhos da comissão na tentativa de localização dos corpos por “estar na lei de indenização dos desaparecidos.
Além disso, é uma questão humanitária, o direito de os familiares enterrarem os restos mortais dos seus entes queridos”.
O presidente da comissão, Miguel Reale Jr, 52 anos, destaca que “o regime militar autorizou e deu guarida a todas as violências, mesmo quando a versão oficial era mentirosa, até bisonha.

Figura 5.9: Extrato para o texto-fonte da Figura 5.7, gerado por sp_i^{wc} , com tamanho similar (em número de palavras) ao do resumo manual da Figura 5.8.

Já na Figura 5.10, encontra-se um texto-fonte segmentado do corpus DUC’2001, utilizado no segundo exemplo desta seção. A ordem de prioridade de seleção das sentenças na formação de um extrato é agora fornecida pelo sumarizador $r_i^{l,k}$, baseado nos d -anéis. A sentença com prioridade 1 é o *hub* na rede derivada desse texto (nó com maior k_i).

1	More than 3,000 passengers and crew members were evacuated early Wednesday from the Sovereign of the Seas, one of the world's largest cruise ships, after a fire broke out in a pantry.
2	One crewman was treated for smoke inhalation, but there were no other injuries in the fire, which broke out when the ship was moored in San Juan Harbor, Ports Authority spokesman David Rivera said.
10	"There was no panic," said passenger Tom Vento, 56, of Philadelphia.
14	"I was surprised that with so many people everyone was so calm.
15	At first, we thought it was a joke, but then we saw that it was serious."
12	The 14-deck, 800-foot luxury liner left Miami on Saturday for a seven-day voyage to La Badee, a private island near Haiti; Puerto Rico; and St. Thomas in the U.S. Virgin Islands.
3	Rich Steck, a spokesman for Miami-based Royal Caribbean Cruise Line, which owns the ship, said it would return to Miami on Wednesday night after a Coast Guard safety inspection rather than continue the cruise.
16	Passengers will be given a full refund, he said.
4	In Washington, the National Transportation Safety Board said a two-member team would meet the ship at Miami to investigate the fire.
5	The board has been concerned about the potential for accidents in the cruise ship industry and has held hearings around the nation.
6	Last year, it placed cruise ships on its "most wanted" list of safety improvements.
7	Steck said the fire, which started in a pantry between the fifth and seventh decks, apparently was caused by an electrical problem, but no details were available.
11	The fire broke out about 1 a.m. and spread to the 1,000-seat Follies Lounge, which was closed at the time.
13	It took about 4 1/2 hours to extinguish the blaze, the Coast Guard said.
8	The 2,318 passengers and most of the 818 crew members were evacuated to a nearby port terminal before they were allowed to return to the ship at daybreak, Steck said.
9	The 74,000-ton Sovereign of the Seas is one of the world's heaviest cruise ships and the largest in terms of its passenger capacity of 2,521, Steck said.

Figura 5.10: Exemplo de aplicação do algoritmo $r_i^{l,k}$ (d -anéis) em texto-fonte do corpus DUC'2001. O texto-fonte aparece segmentado, e os números presentes no início de cada sentença indicam a ordem de prioridade dada pelo sumarizador $r_i^{l,k}$.

Na sequência, sentenças pertencentes aos d -anéis próximos do *hub* são selecionadas, com a restrição de que possuam grau acima da média. Analogamente ao primeiro exemplo, a Figura 5.10 fornece uma noção geral da aplicação do algoritmo de sumarização, desconsiderando taxas de compressão. Dessa vez, o respectivo extrato manual, retirado do corpus DUC'2001 e apresentado na Figura 5.11, é tomado como referência. A comparação agora é mais simples, pois o extrato manual permite associação direta com as sentenças do texto-fonte, como pode ser visto na Figura 5.11, onde cada sentença está associada a uma única sentença do texto-fonte. Esse exemplo reflete o bom desempenho do sumarizador $r_i^{l,k}$ nos resultados obtidos com o corpus DUC'2001 (vide Seção 5.4.4), pois as sentenças do extrato manual apresentam, em sua maioria, alta prioridade quando selecionadas por $r_i^{l,k}$. Note também que, nesse caso, o Top-Baseline (um bom sumarizador, principalmente nos testes em inglês) selecionaria várias sentenças que não aparecem no extrato manual, como

as de prioridade 10, 14 e 15 (por sinal, sentenças que realmente não deveriam figurar em um extrato do tamanho do da Figura 5.11). Esse é, portanto, um exemplo que ajudou o sumarizador $r_i^{l,k}$ a superar o Top-Baseline em um dos experimentos de avaliação relatados anteriormente neste capítulo.

1	More than 3,000 passengers and crew members were evacuated early Wednesday from the Sovereign of the Seas, one of the world's largest cruise ships, after a fire broke out in a pantry.
2	One crewman was treated for smoke inhalation, but there were no other injuries in the fire, which broke out when the ship was moored in San Juan Harbor, Ports Authority spokesman David Rivera said.
12	The 14-deck, 800-foot luxury liner left Miami on Saturday for a seven-day voyage to La Badee, a private island near Haiti; Puerto Rico; and St. Thomas in the U.S. Virgin Islands.
3	Rich Steck, a spokesman for Miami-based Royal Caribbean Cruise Line, which owns the ship, said it would return to Miami on Wednesday night after a Coast Guard safety inspection rather than continue the cruise.
4	In Washington, the National Transportation Safety Board said a two-member team would meet the ship at Miami to investigate the fire.
5	The board has been concerned about the potential for accidents in the cruise ship industry and has held hearings around the nation.

Figura 5.11: Extrato manual, retirado do corpus DUC'2001, construído para o texto-fonte da Figura 5.10. Os número ao lado de cada sentença referem-se à prioridade dada pelo algoritmo $r_i^{l,k}$.

More than 3,000 passengers and crew members were evacuated early Wednesday from the Sovereign of the Seas, one of the world's largest cruise ships, after a fire broke out in a pantry.
One crewman was treated for smoke inhalation, but there were no other injuries in the fire, which broke out when the ship was moored in San Juan Harbor, Ports Authority spokesman David Rivera said.
Rich Steck, a spokesman for Miami-based Royal Caribbean Cruise Line, which owns the ship, said it would return to Miami on Wednesday night after a Coast Guard safety inspection rather than continue the cruise.
In Washington, the National Transportation Safety Board said a two-member team would meet the ship at Miami to investigate the fire.
The board has been concerned about the potential for accidents in the cruise ship industry and has held hearings around the nation.
Last year, it placed cruise ships on its "most wanted" list of safety improvements.
Steck said the fire, which started in a pantry between the fifth and seventh decks, apparently was caused by an electrical problem, but no details were available.

Figura 5.12: Extrato para o texto-fonte da Figura 5.10, gerado por $r_i^{l,k}$, com tamanho similar (em número de palavras) ao do extrato manual da Figura 5.11.

Foi gerado por $r_i^{l,k}$ um extrato com número de palavras aproximadamente igual ao do extrato manual (Figura 5.12). Nesse caso, a sentença de prioridade 12 não foi considerada (como no extrato manual), e duas outras sentenças foram adicionadas, dando ainda maior

cobertura ao extrato automático. Ao invés de informar a respeito do trajeto do cruzeiro marítimo (sentença de prioridade 12), o extrato automático inclui dados a respeito da crescente preocupação quanto à segurança em navios de cruzeiro e a respeito da possível causa do incêndio (problema elétrico). Ambos os extratos podem ser considerados satisfatórios, embora o extrato automático sofra uma penalização na avaliação automática por não ter incluído a sentença de prioridade 12.

Pelos exemplos dados nesta seção, percebe-se que mesmo a avaliação automática tem suas limitações. Um melhor cenário envolveria o uso de sumários de referência em grande quantidade, a fim de não penalizar sumarizadores que façam escolhas um pouco diferentes do que prescreve um único sumário de referência. Entretanto, deve-se lembrar que, apesar de em pequeno número, os sumários de referência aqui utilizados são, certamente, confiáveis. A disponibilidade atual de *corpus* para avaliação automática de sumários é maior do que a existente há alguns anos, e essa tendência de crescimento deve continuar, o que possibilitará avaliações ainda mais completas do que a realizada neste mestrado.

Por fim, no próximo capítulo, são tecidos alguns comentários finais sobre os métodos propostos e sobre os experimentos realizados. Além disso, são sugeridas algumas possíveis continuações deste trabalho.

Conclusões

Este projeto de mestrado segue a mesma linha de recentes pesquisas em lingüística computacional realizadas na USP-São Carlos¹, onde procura-se utilizar conceitos da área de Redes Complexas no processamento de textos. Nessas pesquisas, ao extrair parâmetros de textos representados por redes, relacionou-se a qualidade de redações de vestibular (Antiqueira et al., 2005, 2007), a qualidade de sumários (Pardo et al., 2006a,b), a tarefa de extração de terminologia (Antiqueira, 2005a,b) e o problema de identificação de autoria (Antiqueira et al., 2006) com propriedades associadas à conectividade de palavras em grafos. Essas propriedades são quantificadas por meio do uso de medidas que associam um número, por exemplo, a toda a rede, ou separadamente a cada vértice, e que refletem diversas características relacionadas à estrutura da rede. São diversos os exemplos dessas medidas. Existem as baseadas em distâncias (como o comprimento de caminhos mínimos), que ajudam a analisar a proximidade entre os vértices de uma rede. As medidas de coeficiente de aglomeração, por sua vez, permitem estudar a conectividade em torno de um vértice, ou seja, a conectividade entre seus vizinhos. Já o grau permite identificar os *hubs*, vértices que possuem um grande número de conexões. Essas medidas são tradicionalmente empregadas nos estudos em Redes Complexas, onde a disponibilidade de métricas para a análise de redes cresce continuamente. Neste trabalho de mestrado, tais métricas foram aplicadas na geração de extratos, visando integrar duas áreas aparentemente distantes:

¹Fruto da colaboração entre o Núcleo Interinstitucional de Lingüística Computacional, sediado no Instituto de Ciências Matemáticas e de Computação, e o Grupo de Pesquisa em Visão Cibernética, do Instituto de Física de São Carlos.

Redes Complexas e Sumarização Automática. O que permite essa interface é a representação de textos na forma de redes, que são posteriormente analisadas por meio de métricas como as supracitadas, as quais servem de parâmetros na escolha dos trechos de texto que devem formar um sumário (extrato). Determinadas questões devem ser respondidas, a fim de tornar frutífera a união desses dois campos de pesquisa. Dado um texto-fonte, como transformá-lo em uma rede, de maneira que facilite a interpretação das métricas? E como interpretar as métricas, à luz da sumarização?

Na presente pesquisa, optou-se por uma rede simples, que codifica um tipo de coesão lexical (repetição de substantivos) entre sentenças, as quais são representadas pelos vértices da rede. O intuito foi verificar qual o potencial da proposta, sem ainda utilizar recursos de PLN sofisticados, ou seja, o foco manteve-se mais nas métricas do que no aprimoramento das redes. Somente um pré-processamento ao nível lexical foi aplicado aos textos, a fim de refinar a criação das arestas, por meio da lematização e da eliminação de *stopwords*, tanto para textos em português (do Brasil) quanto em inglês. No caso da eliminação de *stopwords*, foram excluídas todas as palavras que não fossem substantivos, com o intuito de diminuir o grande número de arestas criadas quando as demais palavras são consideradas. Com a repetição de substantivos, possivelmente sentenças que tratam de assuntos relacionados são interligadas na rede. Supõe-se também que duas sentenças associadas na rede sejam complementares, ou seja, não redundantes, fato que prejudicaria a sumarização. Portanto, a presença de uma aresta entre dois nós tem efeito positivo quando uma rede é analisada, fato que influenciou as interpretações dadas às métricas utilizadas neste projeto (grau, coeficiente de aglomeração, caminhos mínimos, índice de localidade, índice de concordância e grau hierárquico). O tipo da métrica faz com que prioridade seja dada a determinados vértices na construção de um extrato, e servem para analisar a conectividade da rede de ângulos diferentes. Procurou-se também empregar outros conceitos na geração de extratos, além das referidas métricas. Determinados subgrafos foram utilizados (d -anéis, k -núcleos, w -cortes e comunidades) na criação de outros algoritmos de sumarização. Novamente, cada um desses subgrafos são produzidos levando-se em consideração diferentes características de uma rede, e são utilizados de maneiras distintas na construção de extratos. Inclusive, os métodos baseados nesses subgrafos foram convertidos em métricas sequenciais, que possibilitam associar uma pontuação a cada vértice. Note que a maior parte dessas métricas e subgrafos servem para destacar vértices ou grupos de vértices bem conectados (com exceção dos caminhos mínimos, do índice de concordância e dos d -anéis).

Todos esses conceitos (6 métricas, 4 subgrafos) deram origem a 26 sistemas de sumarização. Procurou-se avaliá-los automaticamente, por meio do uso das medidas ROUGE-1 e Precisão/Cobertura, contemplando as línguas inglesa e portuguesa. Sempre que possível, os resultados obtidos foram comparados com os de outros sistemas de sumarização extra-

tiva. O número de experimentos realizados mostra a abrangência da avaliação realizada, pois os resultados obtidos foram comparados com os de diversos outros sumarizadores e foram utilizados mais de 700 textos-fonte. Por meio de quatro experimentos de avaliação em textos jornalísticos, foi possível medir a informatividade dos extratos em diferentes taxas de compressão, em diferentes línguas e utilizando diferentes métricas de avaliação automática. Isso permitiu a identificação de alguns padrões ao analisar os resultados dos métodos aqui propostos, como a divisão dos sumarizadores entre dois grupos. Os métodos do Grupo-1, como os baseados nos d -anéis, mostraram-se melhores que os do Grupo-2, como os inspirados no grau hierárquico. Também foi possível perceber uma certa dificuldade em se aproximar dos melhores sistemas para a língua inglesa, enquanto que, para a língua portuguesa, resultados de maior destaque foram obtidos. Ainda devem ser melhor investigadas as diferenças de desempenho entre as duas línguas utilizadas. De maneira geral, os melhores métodos propostos baseiam-se nos d -anéis, nos k -núcleos, nos graus e nos caminhos mínimos.

São diversas as possibilidades de continuação deste trabalho. Uma delas refere-se à aplicação dos sumarizadores propostos em outros tipos de redes para textos. Uma pequena alteração no tipo de rede pode ser feita se for utilizado um *thesaurus*, ou seja, se palavras diferentes, mas de mesmo sentido, forem consideradas na definição das arestas. Outra mudança está relacionada ao incremento no peso de arestas consideradas mais importantes que as outras, tais como arestas que conectem sentenças com nomes próprios ou palavras-chave. Por outro lado, pode-se ainda manter o tipo de rede utilizada, para que se concentre nos algoritmos de extração de sentenças. Seguindo essa idéia, os diversos algoritmos propostos podem ser agrupados, dando origem a outros novos sumarizadores. Uma idéia está relacionada à criação de um esquema de votação, de maneira que as sentenças selecionadas por vários dos métodos baseados em redes complexas sejam escolhidas para formar o extrato. Outra possibilidade é o uso de algoritmos de aprendizado de máquina, o qual se basearia em uma tabela atributo-valor, onde cada métrica aqui utilizada seria transformada em um atributo numérico das sentenças. Por fim, uma última sugestão refere-se ao uso dos algoritmos PageRank e HITS nas redes utilizadas neste projeto. Em um tipo diferente de rede, esses algoritmos apresentaram bons resultados para a língua inglesa. A aplicação dos algoritmos PageRank e HITS pode indicar qual a direção que os algoritmos baseados em redes complexas devem tomar.

Grafos são extremamente flexíveis, utilizados não somente na representação de textos, mas também na representação de diversas outras estruturas não relacionadas à lingüística computacional (como a WWW). Se nos restringirmos apenas às línguas naturais, percebe-se que as possibilidades para uso de redes já são inúmeras. Com o avanço dos recursos disponibilizados pela comunidade de PLN, é possível incrementar uma rede com, por exemplo,

arestas relacionadas à estrutura discursiva de um texto, o que possibilitaria a construção de extratos mais coerentes. É possível, também, utilizar as árvores sintáticas geradas por um *parser*, de modo que os vértices representem orações, e não sentenças, o que possibilitaria uma maior versatilidade na seleção dos trechos que devem compor um extrato. Tudo isso sem considerar o uso de *thesauri* e *wordnets*, recursos que podem agregar mais conhecimento lingüístico ao processamento de textos. Muitos desses recursos podem ser combinados em uma única representação, já que é possível definir vários tipos de vértices ou de arestas em uma rede. Cabe ao pesquisador escolher, de maneira otimizada, quais informações devem ser consideradas em uma rede, visando sempre uma aplicação específica, como por exemplo, sumarização, tradução ou desambiguação. Os estudos em Redes Complexas entram na fase de análise das redes, pois fornecem uma ampla gama de técnicas e conceitos, como os aplicados neste mestrado, propícios à caracterização de tais estruturas. É visível, portanto, a diversidade de métodos que podem surgir quando representam-se textos na forma de redes. E, baseando-se em alguns dos resultados aqui relatados, espera-se que tais métodos resolvam de forma eficiente e robusta diversos problemas da área de PLN.

Referências Bibliográficas

- ABRAÇOS, J.; LOPES, G. P. Statistical methods for retrieving most significant paragraphs in newspaper articles. In: *Proceedings of the ACL/EACL Workshop on Intelligent Scalable Text Summarization*, Madrid, Spain, 1997, p. 51–57.
- ALBERT, R.; BARABÁSI, A. L. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, v. 74, p. 47–97, 2002.
- ALUISIO, S. M.; AIRES, R. V. *Etiquetação de um corpus e construção de um etiquetador de português*. Relatórios Técnicos do ICMC 107, Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo, São Carlos-SP, 18 p., 2000.
- ANTIQUÊIRA, L. O uso de redes complexas na elaboração de uma taxonomia para a área de Nanotecnologia. Projeto de Graduação I, Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo, 32 p., 2005a.
- ANTIQUÊIRA, L. Obtenção e associação de termos na construção de uma ontologia para a área de Nanotecnologia. Projeto de Graduação II, Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo, 40 p., 2005b.
- ANTIQUÊIRA, L.; NUNES, M. G. V.; OLIVEIRA JR., O. N.; COSTA, L. F. Modelando textos como redes complexas. In: *Anais do XXV Congresso da Sociedade Brasileira de Computação (III Workshop em Tecnologia da Informação e da Linguagem Humana - TIL)*, São Leopoldo-RS, Brasil, 2005, p. 2089–2098.
- ANTIQUÊIRA, L.; NUNES, M. G. V.; OLIVEIRA JR., O. N.; COSTA, L. F. Strong correlations between text quality and complex networks features. *Physica A*, v. 373, p. 811–820, physics/0504033.v2, 2007.
- ANTIQUÊIRA, L.; PARDO, T. A. S.; NUNES, M. G. V.; OLIVEIRA JR., O. N.; COSTA, L. F. Some issues on complex networks for author characterization. In: REZENDE, S. O.; DA SILVA FILHO, A. C. R., eds. *Fourth Workshop in Information and Human Language Technology (TIL'06) in the Proceedings of International Joint Conference, 10th Ibero-American Artificial Intelligence Conference, 18th Brazilian Artificial Intelligence Symposium, 9th Brazilian Neural Networks Symposium, IBERAMIA-SBIA-SBRN*, Ribeirão Preto, Brazil: ICMC-USP, 2006.

- BARABÁSI, A. L. *Linked: How everything is connected to everything else and what it means for business, science and everyday life*. Plume, 2003.
- BARABÁSI, A. L.; ALBERT, R. Emergence of scaling in random networks. *Science*, v. 286, p. 509–512, 1999.
- BARTHÉLEMY, M.; BARRAT, A.; PASTOR-SATORRAS, R.; VESPIGNANI, A. Characterization and modeling of weighted networks. *Physica A*, v. 346, p. 34–43, 2005.
- BARZILAY, R.; ELHADAD, M. Using lexical chains for text summarization. In: MANI, I.; MAYBURY, M. T., eds. *Advances in Automatic Text Summarization*, MIT Press, p. 111–121, 1999.
- BATAGELJ, V.; ZAVERNIK, M. Partitioning approach to visualization of large networks. In: KRATOCHVÍL, J., ed. *Proceedings of the Graph Drawing: 7th International Symposium (GD'99)*, Střirín Castle, Czech Republic: Springer-Verlag, 1999, p. 90–98 (LNCS, v.1731).
- BAXENDALE, P. B. Machine-made index for technical literature - an experiment. *IBM Journal of Research and Development*, v. 2, p. 354–365, 1958.
- BENBRAHIM, M.; AHMAD, K. *Computer-aided lexical cohesion analysis and text abridgment*. Computing Sciences Report CS-94-11, University of Surrey, 60 p., 1994.
- BOCCALETTI, S.; LATORA, V.; MORENO, Y.; CHAVEZ, M.; HWANG, D.-U. Complex networks: Structure and dynamics. *Physics Reports*, v. 424, n. 4-5, p. 175–308, 2006.
- BRUNN, M.; CHALI, Y.; DUFOUR, B. The University of Lethbridge text summarizer at DUC 2002. In: *Proceedings of the Document Understanding Conference (DUC)*, 2002.
- CASELLA, G.; BERGER, R. L. *Statistical inference*. Duxbury Press, 1990.
- CLAUSET, A.; NEWMAN, M. E. J.; MOORE, C. Finding community structure in very large networks. *Phys. Rev. E*, v. 70, p. 066111, 2004.
- COSTA, L. F. What's in a name? *Int. J. Mod. Phys. C*, v. 15, p. 371–379, 2004.
- COSTA, L. F.; DA ROCHA, L. E. C. A generalized approach to complex networks. *Eur. Phys. J. B*, v. 50, p. 237–242, cond-mat/0408076, 2006.
- COSTA, L. F.; KAISER, M.; HILGETAG, C. Beyond the average: detecting global singular nodes from local features in complex networks, physics/0607272, 2006a.
- COSTA, L. F.; RODRIGUES, F. A.; TRAVIESO, G.; VILLAS BOAS, P. R. Characterization of complex networks: A survey of measurements, cond-mat/0505185, 2006b.
- DE LUCCA, J. L.; NUNES, M. G. V. *Lematização versus stemming*. Relatórios Técnicos do ICMC 14, Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo, São Carlos-SP, 16 p., 2002.
- DOROGOVTSSEV, S. N.; MENDES, J. F. F. Evolution of networks. *Adv. Complex. Syst.*, v. 51, n. 4, p. 1079–1187, 2002.

- DOROW, B.; WIDDOWS, D.; LING, K.; ECKMANN, J. P.; SERGI, D.; MOSES, E. Using curvature and Markov clustering in graphs for lexical acquisition and word sense discrimination. In: *2nd Workshop organized by the MEANING Project (MEANING-2005)*, Trento, Italy, 2005.
- DUDA, R. O.; HART, P. E.; STORK, D. G. *Pattern classification*. Wiley-Interscience, 2000.
- EDMUNDSON, H. P. New methods in automatic abstracting. *Journal of the Association for Computing Machinery*, v. 16, n. 2, p. 264–285, 1969.
- ERDÖS, P.; RÉNYI, A. On random graphs I. *Publ. Math. Debrecen*, v. 6, p. 290–297, 1959.
- ERKAN, G.; RADEV, D. R. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, v. 22, p. 457–479, 2004.
- FALOUTSOS, M.; FALOUTSOS, P.; FALOUTSOS, C. On power-law relationships of the Internet topology. In: *Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication*, 1999, p. 251–262.
- FERRER I CANCHO, R.; CAPOCCI, A.; CALDARELLI, G. Spectral methods cluster words of the same class in a syntactic dependency network, cond-mat/0504165, 2005.
- FERRER I CANCHO, R.; SOLÉ, R. V. The small world of human language. *P. Roy. Soc. Lond. B Bio.*, v. 268, p. 2261, 2001.
- FERRER I CANCHO, R.; SOLÉ, R. V.; KÖHLER, R. Patterns in syntactic dependency networks. *Phys. Rev. E*, v. 69, p. 051915, 2004.
- HALL, M. A. Correlation-based feature selection for discrete and numeric class machine learning. In: *Proceedings of the 17th International Conference on Machine Learning*, 2000, p. 359–366.
- HARARY, F. *Graph theory*. Addison-Wesley, 1969.
- HEARST, M. A. TextTiling: segmenting text into multi-paragraph subtopic passages. *Comput. Linguist.*, v. 23, n. 1, p. 33–64, 1997.
- HIRAO, T.; SASAKI, Y.; ISOZAKI, H.; MAEDA, E. NTT’s text summarization system for DUC-2002. In: *Proceedings of the Document Understanding Conference (DUC)*, 2002.
- HOSMER, D. W.; LEMESHOW, S. *Applied logistic regression*. 2 ed. Wiley, 2000.
- KAISER, M.; HILGETAG, C. C. Edge vulnerability in neural and metabolic networks. *Biological Cybernetics*, v. 90, p. 311–317, 2004.
- KLEINBERG, J. M. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, v. 46, n. 5, p. 604–632, 1999.

- KOHONEN, T. The self-organizing map. *Proceedings of the IEEE*, v. 78, n. 9, p. 1464–1480, 1990.
- KOWALTOWSKI, T.; LUCCHESI, C. L.; STOLFI, J. *Finite automata and efficient lexicon implementation*. Relatório Técnico IC-98-2, DCC/UNICAMP, 1998.
- KUPIEC, J.; PEDERSEN, J.; CHEN, F. A trainable document summarizer. In: *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, USA: ACM Press, 1995, p. 68–73.
- LAROCCA NETO, J.; FREITAS, A. A.; KAESTNER, C. A. A. Automatic text summarization using a machine learning approach. In: BITTENCOURT, G.; RAMALHO, G. L., eds. *Proceedings of the 16th Brazilian Symposium on Artificial Intelligence (SBIA)*, Springer-Verlag, 2002, p. 205–215 (LNAI, v.2507).
- LAROCCA NETO, J.; SANTOS, A. D.; KAESTNER, A. A.; FREITAS, A. A. Generating text summaries through the relative importance of topics. In: MONARD, M. C.; SICHMAN, J. S., eds. *Proceedings of the International Joint Conference IBERAMIA-2000 (7th Ibero-American Conference on Artificial Intelligence) and SBIA-2000 (15th Brazilian Symposium on Artificial Intelligence)*, Atibaia, SP, Brazil: Springer-Verlag, 2000a, p. 300–309 (LNAI, v.1952).
- LAROCCA NETO, J.; SANTOS, A. D.; KAESTNER, C. A. A.; FREITAS, A. A. Document clustering and text summarization. In: MACKIN, N., ed. *Proceedings of the 4th International Conference Practical Applications of Knowledge Discovery and Data Mining (PADD-2000)*, London: The Practical Application Company, 2000b, p. 41–55.
- LEITE, D. S.; RINO, L. H. M. Selecting a feature set to summarize texts in Brazilian Portuguese. In: *Proceedings of the International Joint Conference IBERAMIA-SBIA 2006*, Springer-Verlag, 2006a, p. 462–471 (LNAI, v.4140).
- LEITE, D. S.; RINO, L. H. M. Uma comparação entre sistemas de sumarização automática extrativa. In: REZENDE, S. O.; DA SILVA FILHO, A. C. R., eds. *Fourth Workshop in Information and Human Language Technology (TIL'06 Poster Section) in the Proceedings of International Joint Conference, 10th Ibero-American Artificial Intelligence Conference, 18th Brazilian Artificial Intelligence Symposium, 9th Brazilian Neural Networks Symposium, IBERAMIA-SBIA-SBRN*, Ribeirão Preto, Brazil: ICMC-USP, 2006b.
- 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.
- LIN, C. Y.; HOVY, E. Automatic evaluation of summaries using n-gram co-occurrence statistics. In: *Proceedings of the 2003 Language Technology Conference (HLT-NAACL-2003)*, Edmonton, Canada, 2003.
- LUHN, H. P. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, v. 2, p. 159–165, 1958.

- LYMAN, P.; VARIAN, H. R. How much information? <http://www.sims.berkeley.edu/how-much-info-2003>, 2003.
- MANI, I. *Automatic summarization*. John Benjamins Publishing Co., 2001.
- MANI, I.; BLOEDORN, E. Summarizing similarities and differences among related documents. *Information Retrieval*, v. 1, n. 1-2, p. 35–67, 1999.
- MANI, I.; BLOEDORN, E.; GATES, B. Using cohesion and coherence models for text summarization. In: HOVY, E.; RADEV, D. R., eds. *Proceedings of the Spring Symposium on Intelligent Text Summarization (AAAI 98)*, Stanford, CA: AAAI Press, 1998, p. 69–76.
- MARGARIDO, P. R. A. Relatório Interno do NILC - Núcleo Interinstitucional de Linguística Computacional, 2007.
- MARTINS, C. B.; RINO, L. H. M. UNLSumm: Um sumarizador automático de textos UNL. In: *Anais do I Workshop de Teses e Dissertações em Inteligência Artificial (WTDIA)*, Porto de Galinhas-PE, Brasil, 2002.
- MÓDOLO, M. *SuPor: um ambiente para a exploração de métodos extrativos para a sumarização automática de textos em português*. Dissertação de mestrado, Universidade Federal de São Carlos, 2003.
- MIHALCEA, R. Language independent extractive summarization. In: *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, Ann Arbor-MI, United States: Association for Computational Linguistics, 2005, p. 49–52.
- MILGRAM, S. The small world problem. *Psychology Today*, v. 2, p. 60–67, 1967.
- MILLER, G. A. WordNet: a lexical database for English. *Commun. ACM*, v. 38, n. 11, p. 39–41, 1995.
- MITCHELL, T. M. *Machine learning*. WCB/McGraw-Hill, 1997.
- MOTTER, A. E.; MOURA, A. P. S.; LAI, Y. C.; DASGUPTA, P. Topology of the conceptual network of language. *Phys. Rev. E*, v. 65, p. 065102, 2002.
- NEWMAN, M. E. J. The structure and function of complex networks. *SIAM Rev.*, v. 45, p. 167–256, 2003.
- NUNES, M. G. V.; VIEIRA, F. M. C.; ZAVAGLIA, C.; SOSSOLOTE, C. R. C.; HERNANDEZ, J. A construção de um léxico para o português do Brasil: Lições aprendidas e perspectivas. In: *Anais do II Encontro para o Processamento Computacional do Português Escrito e Falado (PROPOR)*, 1996, p. 61–70.
- OVER, P. Introduction to DUC-2001: An intrinsic evaluation of generic news text summarization systems. http://www-nlpir.nist.gov/projects/duc/pubs/2001slides/pauls_slides/index.htm, 2001.

- OVER, P.; LIGGETT, W. Introduction to DUC: An intrinsic evaluation of generic news text summarization systems. <http://www-nlpir.nist.gov/projects/duc/pubs/2002slides/overview.02.pdf>, 2002.
- PAGE, L.; BRIN, S.; MOTWANI, R.; WINOGRAD, T. *The PageRank citation ranking: Bringing order to the web*. Relatório Técnico, Stanford Digital Library Technologies Project, 17 p., 1998.
- PAICE, C. D. The automatic generation of literature abstracts: an approach based on the identification of self-indicating phrases. In: *Proceedings of the 3rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Kent, UK: Butterworth & Co., 1981, p. 172–191.
- PARDO, T. A. S.; ANTIQUEIRA, L.; NUNES, M. G. V.; OLIVEIRA JR., O. N.; COSTA, L. F. Modeling and evaluating summaries using complex networks. In: *Proceedings of the 7th Workshop on Computational Processing of Written and Spoken Portuguese (PROPOR)*, Springer-Verlag, 2006a, p. 1–10 (LNAI, v.3960).
- PARDO, T. A. S.; ANTIQUEIRA, L.; NUNES, M. G. V.; OLIVEIRA JR., O. N.; COSTA, L. F. Using complex networks for language processing: The case of summary evaluation. In: *Proceedings of the International Conference on Communications, Circuits and Systems (ICCCAS'06) - Special Session on Complex Networks*, Gui Lin, China: UESTC Press, 2006b, p. 2678–2682.
- PARDO, T. A. S.; RINO, L. H. M. DMSumm: Um gerador automático de sumários. In: *Anais do I Workshop de Teses e Dissertações em Inteligência Artificial - WTDIA*, Porto de Galinhas-PE, Brasil, 2002.
- PARDO, T. A. S.; RINO, L. H. M. *TeMário: Um corpus para sumarização automática de textos*. Série de Relatórios do NILC NILC-TR-03-09, Núcleo Interinstitucional de Lingüística Computacional (NILC), São Carlos-SP, 11 p., 2003.
- PARDO, T. A. S.; RINO, L. H. M. *Descrição do GEI - Gerador de Extratos Ideais para o Português do Brasil*. Série de Relatórios do NILC NILC-TR-04-07, Núcleo Interinstitucional de Lingüística Computacional (NILC), São Carlos-SP, 8 p., 2004.
- PARDO, T. A. S.; RINO, L. H. M.; NUNES, M. G. V. GistSumm: A summarization tool based on a new extractive method. In: *Proceedings of the 6th Workshop on Computational Processing of Written and Spoken Portuguese (PROPOR)*, Springer-Verlag, 2003a, p. 210–218 (LNAI, v.2721).
- PARDO, T. A. S.; RINO, L. H. M.; NUNES, M. G. V. NeuralSumm: Uma abordagem conexcionista para a sumarização automática de textos. In: *Anais do IV Encontro Nacional de Inteligência Artificial - ENIA*, Campinas-SP, Brasil, 2003b, p. 1–10.
- PATHRIA, R. K. *Statistical mechanics*. Elsevier, 1996.
- QUINLAN, J. R. *C4.5: Programs for machine learning*. San Mateo, CA: Morgan Kaufmann, 1993.

- RABINER, L. R. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE*, v. 77, n. 2, p. 257–286, 1989.
- RATNAPARKHI, A. A maximum entropy part-of-speech tagger. In: *Proceedings of the Empirical Methods in Natural Language Processing Conference*, University of Pennsylvania, 1996.
- REYNAR, J. C.; RATNAPARKHI, A. A maximum entropy approach to identifying sentence boundaries. In: *Proceedings of the Fifth Conference on Applied Natural Language Processing*, Washington, D.C., 1997, p. 16–19.
- RINO, L. H. M.; MÓDOLO, M. SuPor: An environment for AS of texts in Brazilian Portuguese. In: *España for Natural Language Processing (EsTAL)*, Alicante, Spain, 2004, p. 419–430.
- RINO, L. H. M.; NUNES, M. G. V. *Sobre geração e sumarização de textos*. Notas Didáticas do ICMC 67, Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo, São Carlos-SP, 28 p., 2005.
- RINO, L. H. M.; PARDO, T. A. S. A sumarização automática de textos: Principais características e metodologias. In: *Anais do XXIII Congresso da Sociedade Brasileira de Computação - Volume VIII: III Jornada de Minicursos de Inteligência Artificial*, 2003, p. 203–245.
- RINO, L. H. M.; PARDO, T. A. S.; SILLA JR., C. N.; KAESTNER, C. A. A.; POMBO, M. A comparison of automatic summarizers of texts in Brazilian Portuguese. In: *Proceedings of the 17th Brazilian Symposium on Artificial Intelligence (SBIA)*, São Luis-MA, Brazil, 2004, p. 235–244.
- SALTON, G.; MCGILL, M. J. *Introduction to modern information retrieval*. New York: McGraw-Hill, 1983.
- SALTON, G.; SINGHAL, A.; MITRA, M.; BUCKLEY, C. Automatic text structuring and summarization. *Information Processing and Management*, v. 33, n. 2, p. 193 – 207, 1997.
- SCHLESINGER, J. D.; OKUROWSKI, M. E.; CONROY, J. M.; O’LEARY, D. P.; TAYLOR, A.; HOBBS, J.; WILSON, H. T. Understanding machine performance in the context of human performance for multi-document summarization. In: *Proceedings of the Document Understanding Conference (DUC)*, 2002.
- SIGMAN, M.; CECCHI, G. A. Global organization of the WordNet lexicon. *PNAS*, v. 99, n. 3, p. 1742–1747, 2002.
- SKOROCHOD’KO, E. F. Adaptive method of automatic abstracting and indexing. In: FREIMAN, C. V., ed. *Proceedings of the IFIP Congress 71*, 1971, p. 1179–1182.
- SPÄRCK JONES, K. Automatic summarizing: Factors and directions. In: MANI, I.; MAYBURY, M. T., eds. *Advances in Automatic Text Summarization*, cap. 1, MIT Press, p. 1–12, cmp-lg/9805011, 1999.

- VAN HALTEREN, H. A default first order family weight determination procedure for WPDV models. In: *Proceedings of the CoNLL-2000 - Association for Computational Linguistics*, 2000, p. 119–122.
- VAN HALTEREN, H. Writing style recognition and sentence extraction. In: *Proceedings of the Document Understanding Conference (DUC)*, 2002.
- VAPNIK, V. *The nature of statistical learning theory*. 2 ed. Springer, 2000.
- WATTS, D. J.; STROGATZ, S. H. Collective dynamics of ‘small-world’ networks. *Nature*, v. 393, p. 440–442, 1998.
- WITTEN, I. H.; FRANK, E. *Data mining: Practical machine learning tools and techniques*. 2 ed. Morgan Kaufmann, 2005.