UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

UM ESTUDO COMPARATIVO DE MODELOS BASEADOS EM ESTATÍSTICAS TEXTUAIS, GRAFOS E APRENDIZADO DE MÁQUINA PARA SUMARIZAÇÃO AUTOMÁTICA DE TEXTOS EM PORTUGUÊS

DANIEL SARAIVA LEITE

ORIENTADORA: PROFª. DRª. LUCIA HELENA MACHADO RINO

São Carlos - SP
Dezembro/2010

UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

UM ESTUDO COMPARATIVO DE MODELOS BASEADOS EM ESTATÍSTICAS TEXTUAIS, GRAFOS E APRENDIZADO DE MÁQUINA PARA SUMARIZAÇÃO AUTOMÁTICA DE TEXTOS EM PORTUGUÊS

DANIEL SARAIVA LEITE

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação, área de concentração Inteligência Artificial
Orientadora: Dra. Lúcia Helena Machado Rino

São Carlos - SP
Dezembro/2010

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

L533ec

Leite, Daniel Saraiva.
Um estudo comparativo de modelos baseados em estatísticas textuais, grafos e aprendizado de máquina para sumarização automática de textos em português / Daniel Saraiva Leite. -- São Carlos : UFSCar, 2011.
213 f.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2010.

1. Processamento da linguagem natural (Computação). 2. Sumarização automática. 3. Inteligência artificial. I. Título.

CDD: 006.35 (20ª)

Universidade Federal de São Carlos

Centro de Ciências Exatas e de Tecnologia

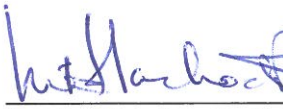
Programa de Pós-Graduação em Ciência da Computação

"Um Estudo Comparativo de Modelos Baseados em Estatísticas Textuais, Grafos e Aprendizado de Máquina para Sumarização Automática de Textos em Português"

DANIEL SARAIVA LEITE

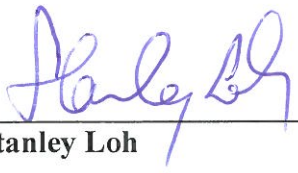
Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação

Membros da Banca:



Profa. Dra. Lucia Helena Machado Rino
(Orientadora - DC/UFSCar)

Prof. Dr. Thiago Alexandre Salgueiro Pardo
(ICMC/USP)



Prof. Dr. Stanley Loh
(UPEL)

São Carlos
Dezembro/2010

Aos meus pais, Sérgio e Nilva, e à Ariani.

AGRADECIMENTOS

A minha orientadora Lucia Rino pelos ensinamentos, incentivo ao mestrado e pela paciência.

Aos meus pais pelo incentivo e exemplo.

Ao amigo e colega de mestrado Igor Vitório Custódio, que sempre ajudou quando precisei.

Aos professores Thiago Pardo e Maria das Graças Volpe Nunes por estarem sempre dispostos a ajudar e pela colaboração no trabalho do artigo para o workshop da NAACL 2007.

A Marcelo Módolo pela ajuda com o SuPor.

A Lucas Antiqueira pelo trabalho conjunto realizado com as medidas de Redes Complexas e pelo fornecimento do cálculo dessas características para o corpus TeMário-2003.

Ao professor João Luis Rosa pelas contribuições na qualificação e pela disposição em responder dúvidas sobre o sistema SABio.

À professora Heloisa de Arruda Camargo pela ótima disciplina ministrada de Sistemas Nebulosas e pela sugestão do trabalho que foi a primeira versão do SuPor-2 Fuzzy.

Aos colegas pesquisadores do Lalic e do NILC, pelas ajudas ao longo deste trabalho.

À Rada Mihalcea pela prestatividade ao responder dúvidas sobre o TextRank e sua avaliação.

Ao professor Stanley Loh pelo aceite do convite para a banca e contribuições.

Aos professores e funcionários do DC/UFSCar e do Programa de Pós-Graduação em Ciência da Computação.

À UFSCar e aos professores dessa universidade pela formação desde a graduação,

À CAPES, FAPESP e CNPq pelo auxílio financeiro.

Tudo deveria se tornar o mais simples possível, mas não simplificado.

(Albert Einstein)

RESUMO

A tarefa de Sumarização Automática de textos tem sido de grande importância dentro da área de Processamento de Linguagem Natural devido à necessidade de se processar gigantescos volumes de informação disponibilizados nos diversos meios de comunicação. Assim, mecanismos em larga escala para sintetizar e facilitar o acesso a essas informações são de extrema importância. Esses mecanismos visam à preservação do conteúdo mais relevante e com pouca ou nenhuma intervenção humana. Partindo do sumarizador extrativo SuPor e contemplando o Português, este trabalho de mestrado visou explorar variadas características de sumarização pela utilização de métodos computacionais baseados em estatísticas textuais, grafos e aprendizado de máquina. Esta exploração consistiu de uma extensão significativa do SuPor, pela definição de novos modelos baseados nessas três abordagens de forma individual ou híbrida. Por serem originários desse sistema, manteve-se a relação com seu nome, o que resultou na denominação genérica *SuPor-2*. Os diversos modelos propostos foram, então, comparados entre si em diversos experimentos, avaliando-se intrínseca e automaticamente a informatividade dos extratos produzidos. Foram realizadas também comparações com outros sistemas conhecidos para o Português. Os resultados obtidos evidenciam uma melhora expressiva de algumas variações do SuPor-2 em relação aos demais sumarizadores extrativos existentes para o Português. Os sistemas que se evidenciaram superiores podem ser disponibilizados no futuro para utilização geral por usuários comuns ou ainda para utilização como ferramentas em outras tarefas do Processamento de Língua Natural ou em áreas relacionadas. A portabilidade para outras línguas é possível com a substituição dos recursos dependentes de língua, como léxico, etiquetadores morfossintáticos e *stoplist* Os modelos supervisionados foram treinados com textos jornalísticos até o momento. O treino para outros gêneros pode ser feito pelos usuários interessados através dos próprios sistemas desenvolvidos.

Palavras-chave: SUMARIZAÇÃO AUTOMÁTICA EXTRATIVA, SUMARIZAÇÃO AUTOMÁTICA BASEADA EM GRAFOS, SUMARIZAÇÃO AUTOMÁTICA BASEADA EM ESTATÍSTICAS TEXTUAIS, SUMARIZAÇÃO AUTOMÁTICA BASEADA EM APRENDIZADO DE MÁQUINA, SUMARIZAÇÃO AUTOMÁTICA BASEADA EM TÉCNICAS HÍBRIDAS, SUMARIZAÇÃO AUTOMÁTICA DE TEXTOS, PROCESSAMENTO DE LÍNGUA NATURAL, INTELIGÊNCIA ARTIFICIAL

ABSTRACT

Automatic text summarization has been of great interest in Natural Language Processing due to the need of processing a huge amount of information in short time, which is usually delivered through distinct media. Thus, large-scale methods are of utmost importance for synthesizing and making access to information simpler. They aim at preserving relevant content of the sources with little or no human intervention. Building upon the extractive summarizer SuPor and focusing on texts in Portuguese, this MsC work aimed at exploring varied features for automatic summarization. Computational methods especially driven towards textual statistics, graphs and machine learning have been explored. A meaningful extension of the SuPor system has resulted from applying such methods and new summarization models have thus been delineated. These are based either on each of the three methodologies in isolation, or are hybrid. In this dissertation, they are generically named after the original SuPor as SuPor-2. All of them have been assessed by comparing them with each other or with other, well-known, automatic summarizers for texts in Portuguese. The intrinsic evaluation tasks have been carried out entirely automatically, aiming at the informativeness of the outputs, i.e., the automatic extracts. They have also been compared with other well-known automatic summarizers for Portuguese. SuPor-2 results show a meaningful improvement of some SuPor-2 variations. The most promising models may thus be made available in the future, for generic use. They may also be embedded as tools for varied Natural Language Processing purposes. They may even be useful for other related tasks, such as linguistic studies. Portability to other languages is possible by replacing the resources that are language-dependent, namely, lexicons, part-of-speech taggers and stop words lists. Models that are supervised have been so far trained on news corpora. In spite of that, training for other genres may be carried out by interested users using the very same interfaces supplied by the systems.

Keywords: EXTRACTIVE AUTOMATIC SUMMARIZATION, GRAPH-BASED AUTOMATIC SUMMARIZATION, AUTOMATIC SUMMARIZATION BASED UPON STATISTICS, MACHINE LEARNING APPROACH FOR AUTOMATIC SUMMARIZATION, HYBRID METHODS FOR AUTOMATIC SUMMARIZATION, AUTOMATIC SUMMARIZATION, NATURAL LANGUAGE PROCESSING, ARTIFICIAL INTELLIGENCE

ÍNDICE

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

AM - Aprendizado de Máquina

ARFF - *Attribute Relation File Format*

CFS - *Correlation Feature Selection*

DUC - *Document Understanding Conference*

EA - Extrato Automático

ER - Extrato de Referência

IG - *Information Gain*

IT - Importância dos Tópicos

PLN - *Processamento de Língua Natural*

QA - *Question Answering*

RC - Redes Complexas

RNA - Redes Neurais Artificiais

ROUGE - *Recall-Oriented Understudy for Gisting Evaluation*

SA - Sumarização Automática

SCU - *Summary Content Unit*

SOM - *Self-Organizing Map*

SU - *Symetrical Uncertainty*

SVM - *Support Vector Machines*

TAC - *Text Analysis Conference*

TC - Taxa de Compressão

TF-IDF - *Term Frequency – Inverse Document Frequency*

TF-IPF - *Term Frequency – Inverse Paragraph Frequency*

TF-ISF - *Term Frequency – Inverse Sentence Frequency*

TG - Teoria dos Grafos

WEKA - *Waikato Environment for Knowledge Engineering*

WWW - *World Wide Web*

Capítulo 1

INTRODUÇÃO

Neste capítulo é feita uma definição introdutória da área de Sumarização Automática e são apresentados os problemas abordados, as hipóteses consideradas e os objetivos deste trabalho.

De modo geral, a Sumarização Automática (SA) busca produzir uma versão reduzida de um texto, geralmente pela seleção ou generalização de seu conteúdo informativo mais relevante (Spärck Jones 1999).

A tarefa de SA vem ganhando importância dentro da área de Processamento de Linguagem Natural devido ao gigantesco volume de informação disponibilizado nos diversos meios de comunicação. Um estudo do final de 2009 da Universidade de Califórnia (Bohn e Short 2009) apontou que a população americana gasta em média 12h por dia consumindo informações veiculadas na TV, Internet, rádio, telefone, jogos ou em meios impressos. Embora o estudo aponte que o consumo de informação veiculada em meio impresso venha caindo, estando por volta de 5% do tempo total gasto, o consumo de informações escritas na Internet é o triplo e vem crescendo. Essa explosão da era da informação tem motivado a busca por mecanismos em larga escala para sintetizar e facilitar o acesso a essas informações, com a preservação do conteúdo mais relevante e com pouca ou nenhuma intervenção humana.

Conforme propõe Spärck Jones (1999), a SA é baseada num modelo de processamento de três fases: (1) a *interpretação* do texto-fonte para criar uma representação conceitual do mesmo; (2) a *transformação* da representação do texto-fonte para uma representação conceitual do sumário; (3) a *geração* do texto do sumário a partir de sua representação conceitual.

Do processo de sumarização, pode-se originar um extrato ou um *abstract* (Sparck Jones 1997; Mani 2001). Um extrato corresponde ao texto produzido diretamente pela extração de segmentos inteiros desse texto, justapostos na mesma ordem original. Usualmente, a construção dos extratos é feita pela justaposição de sentenças. Já um *abstract* envolve a reescrita do texto e, portanto, consiste de um novo texto, em geral com vocabulário e estrutura distintos do texto-fonte. A geração de *abstracts* remete a uma abordagem profunda ou fundamental de SA e ao modo como a tarefa é feita manualmente.

Na geração de *abstracts*, como geralmente utilizam-se paráfrases, generalizações e especializações, o poder de síntese do sumário produzido é teoricamente maior. Além disso, o fato de haver a reescrita em oposição à simples justaposição de segmentos proporciona geralmente uma legibilidade melhor que a dos *extratos*.

Já na geração de extratos, os sumários produzidos podem apresentar tipicamente problemas de coesão e clareza referencial[1], que é a propriedade de um texto permitir ao leitor identificar a quem ou que um pronome ou sintagma nominal está se referindo. Um problema comum de clareza referencial, por exemplo, é a presença de anáforas não resolvidas no texto porque seus referentes não são endofóricos, isto é, não estão explicitados anteriormente no texto-fonte. Aqui não se contemplam as referenciações exofóricas, isto é, as menções a elementos de uma realidade exterior, não explicitadas no texto-fonte. Anáforas que remetam a esse tipo de referentes não deveriam representar um problema para a SA, se considerado que o referente já não está explicitado no próprio texto-fonte.

Como exemplo de anáforas problemáticas, considere o trecho do texto-fonte mostrado na Figura 1-1, do corpus TeMário-2003 (Pardo e Rino 2003). Na Figura 1-2 é mostrado um extrato hipotético para esse texto, onde se nota que a ideia principal da notícia não é transmitida com clareza. A expressão “Semelhanças”, por si, não permite identificar que o texto argumenta que os discursos dos candidatos à presidência citados, FHC e Lula, são muito semelhantes. Há ainda a expressão “ambos” no extrato sem o referente explícito, prejudicando também a interpretação do texto.

[1] Definição apresentada na DUC 2005 (http://duc.nist.gov/duc2005/) em seu questionário de avaliação qualitativa. Esse questionário é retomado na Seção 2.4.1.

A leitura atenta das entrevistas de Fernando Henrique Cardoso (candidato pela coligação do PSDB) e Luiz Inácio Lula da Silva (da aliança liderada pelo PT) à Folha revela muito mais coincidências do que divergências. Semelhanças que começam nas primeiras medidas que cada um deles anuncia como prioritárias para o início de governo.

Ambos mencionam a reforma tributária como ponto de partida, o que, de resto, segue uma lógica inescapável: o Estado brasileiro está em evidente estado falimentar e não há governo que possa fazer o que quer que seja se, antes, não conseguir reorganizar racionalmente as suas fontes de recursos.

Até a fórmula para a implementação dessa reforma indispensável, a negociação com a sociedade, é semelhante entre os dois candidatos que lideram todas as pesquisas de intenção de voto.

Figura 1-1 - Trecho do texto-fonte op94ag14-a.txt

Semelhanças que começam nas primeiras medidas que cada um deles anuncia como prioritárias para o início de governo. Ambos mencionam a reforma tributária como ponto de partida, o que, de resto, segue uma lógica inescapável: o Estado brasileiro está em evidente estado falimentar e não há governo que possa fazer o que quer que seja se, antes, não conseguir reorganizar racionalmente as suas fontes de recursos.

Figura 1-2 - Extrato exemplo do texto-fonte op94ag14-a.txt

Já o possível *abstract* da Figura 1-3, construído manualmente a partir de trechos do *abstract* manual[2] que acompanha o TeMário-2003, apresenta uma

[2] São os sumários produzidos por pessoas. Na área de Sumarização Automática, é comum chamar tais sumários de sumários manuais, em oposição a sumários automáticos

cobertura muito maior da informação do texto-fonte e sem a introdução de problemas de legibilidade.

> Pelas entrevistas dos dois candidatos à presidência --- Fernando Henrique Cardoso (PSDB) e Luís Inácio Lula da Silva (PT) conclui-se que há mais convergências do que divergências entre eles. Ambos mencionam a reforma tributária como ponto de partida para o próprio governo, coerentes, aliás, com a condição básica para a sobrevivência do país. E os dois pretendem fazê-la, negociando com a sociedade.

Figura 1-3 - Abstract manual do texto op94ag14-a.txt

Embora a legibilidade e o poder de síntese da informação dos *abstracts* sejam em geral superiores aos extratos, a geração de *abstracts* é muito mais custosa que a geração de extratos e pode demandar recursos linguísticos e computacionais mais avançados e nem sempre disponíveis, tais como analisadores sintáticos, discursivos e semânticos. Em contraste, a geração de extratos constitui, atualmente, um grande apelo, do ponto de vista de usabilidade, praticidade e abrangência: sumarizadores extrativos podem ser independentes de língua natural ou domínio e gênero textual. Além disso, as técnicas e métodos de Inteligência Artificial, como o Aprendizado de Máquina (AM), são mais facilmente aplicáveis em sumarizadores extrativos. Por conta disso, a SA extrativa tem sido o foco das pesquisas acadêmicas envolvendo abordagens totalmente automáticas (Mani 2001; Spärck Jones 2007).

Um dos sistemas de SA para o português do Brasil é o SuPor (Módolo 2003) e segue também a abordagem extrativa. Esse sistema permite que os usuários livremente escolham características das sentenças a serem usadas como fatores de decisão na escolha das sentenças. Através de um modelo estatístico, as características escolhidas são combinadas e ponderadas, permitindo a pontuação da cada sentença de acordo com sua relevância.

O bom desempenho do SuPor frente a outros sumarizadores para o Português, reportado em Rino et al. (2004), serviu de motivação inicial para continuidade dessa linha de pesquisa extrativa. Assim, objetivou-se neste trabalho a construção de modelos de sumarização que superem o desempenho do SuPor e que também tenham uma potencialidade de aplicação mais ampla. Vários desses

modelos mantém fortes vínculos com as características principais do SuPor e por isso mantiveram a denominação SuPor-2 em seus nomes. Eles serão descritos no Capítulo 5.

De modo geral, na construção dos modelos propostos, considerou-se dois problemas que permeiam o uso de um sistema como o SuPor e a qualquer sistema extrativo tipicamente:

Problema 1. Quais características utilizar?

Este problema consiste em como selecionar conjuntos adequados de características, usualmente das palavras ou sentenças, para se definir o modelo de sumarização. O conjunto de características pode ser influenciado pelo gênero linguístico ou domínio. Note-se que, inclusive, que utilizar conjuntos maiores de características não levará necessariamente aos melhores resultados. No trabalho de Kupiec et al. feito num corpus de 188 artigos técnicos e científicos, verificou-se que um subconjunto de características leva ao melhores resultados. Das 5 características disponíveis no sistema, apenas três delas (localização, presença de palavras indicativas e tamanho) foram utilizadas. Essa mesma variação no desempenho do sumarizador de acordo com o conjunto de características também é observada no SuPor (Módolo 2003). No caso do SuPor, existem ao todo 11 características possíveis.

Segundo Mani (2001), o processo de se desenvolver boas características de sumarização é uma arte.

Problema 2. Como combinar e ponderar as características escolhidas?

Uma vez escolhidas as características, resta ainda a questão de como combiná-las e ponderá-las. Da mesma forma que no problema anterior, este problema é geralmente dependente de gênero ou domínio. A questão está em como determinar os pesos do modelo de saliência. Isto é, o modelo que irá atribuir um peso a cada característica e julgar se o segmento textual é relevante ou não. A complexidade do problema também irá crescer conforme o número de características aumentar. No caso do SuPor, um modelo estatístico Bayesiano é utilizado, mas outros modelos são possíveis.

Considerando-se esses dois problemas, partiu-se da premissa de que é possível combinar e ponderar automaticamente características de diversas naturezas para refinar a SA de textos em Português do Brasil, conforme já foi mostrado, por exemplo, por Módolo (2003) e Kupiec et al. (1995). Buscou-se verificar, então, as seguintes hipóteses:

Hipótese 1. A combinação de características de SA diversas pode permitir a consideração de diferentes fatores ou pontos de vista na análise da relevância das sentenças e, assim, levar a melhores extratos.

Hipótese 2. A pré-seleção automática de características pode configurar de forma adequada o sumarizador para um determinado corpus e levar aos melhores resultados. Ou seja, a pré-seleção automática de características pode ser uma forma de tratar o Problema 1 explicado acima.

Hipótese 3. A forma de combinação e ponderação das características tem influência significativa na qualidade dos sumários. Quanto maior a capacidade do modelo de ponderação em construir bons *rankings* de sentenças, melhor será o desempenho do sumarizador.

Para verificação dessas hipóteses, adotou-se a seguinte metodologia:

- A definição de modelos matemáticos e computacionais para a SA, sua prototipação, avaliação e comparação com outros modelos;

- Explorar em cada modelo de sumarização definido:
  - Conjuntos de características de naturezas diferentes;
  - Abordagens diferentes de ponderação e consequente combinação das características.

Na construção dos modelos propostos, três abordagens principais foram adotadas: a utilização de estatísticas textuais, o uso de aprendizado de máquina (AM) e o emprego de métodos e medidas relacionados à Teoria dos Grafos (TG).

O restante desta dissertação está organizado da seguinte forma: No Capítulo 2 são apresentados os principais conceitos e metodologias utilizados na avaliação de sumários automáticos[3]. Esse capítulo é referência tanto para as metodologias de avaliação utilizadas neste trabalho quanto para as citadas no Capítulo 3, onde é feita uma revisão de alguns sistemas de sumarização já propostos, bem como trabalhos recentes e correlatos da área. Ainda no Capítulo 3, distinguem-se as três grandes abordagens exploradas neste trabalho: estatísticas textuais, grafos e aprendizado de máquina. O Capítulo 4 dedica-se a um aprofundamento na descrição dos principais métodos de aprendizado de máquina, discorrendo sobre seu uso para SA. O Capítulo 5 descreve os modelos de SA desenvolvidos neste trabalho. O Capítulo 6 é focado na descrição e discussão dos experimentos de avaliação conduzidos, tanto de cada modelo individualmente quanto de sua comparação com sistemas de outros autores. Por fim, no Capítulo 7, apresentam-se as principais conclusões, contribuições e limitações deste trabalho, assim como possíveis desdobramentos e continuações.

[3] Neste texto, por se considerar a abordagem extrativa, adota-se também o termo "extrato automático"

Capítulo 2

AVALIAÇÃO DE SUMÁRIOS

Neste capítulo são apresentados os principais aspectos envolvidos na avaliação de sistemas de sumarização. As metodologias de avaliação aqui apresentadas são utilizadas nos capítulos subsequentes.

2.1 Conceitos Iniciais

Segundo Mani (2001) e Spärck Jones (2007), a avaliação é uma etapa de fundamental importância na Sumarização Automática. Uma variedade de bases de comparação pode ser utilizada para avaliar o desempenho dos sumarizadores. Os sumários podem ser julgados em relação ao texto-fonte, comparados com sumários de referência ou *gold-standards*, comparados com sumários produzidos por outros sistemas ou *baselines*.

Mani (2001) destaca as seguintes dificuldades principais na avaliação:

- Pode ser necessário utilizar trabalho manual para julgar o resultado dos sumarizadores, encarecendo a avaliação;
- A dificuldade do que se pode definir como um bom sumário manualmente, já que pode haver discordância entre os juízes, principalmente em avaliações intrínseca, definidas na Seção 2.2. Isso se deve, por exemplo, ao fato de um texto estar sujeito a diferentes interpretações com base em pressupostos ou expectativas que um leitor traz, previamente;

- A avaliação é feita para uma taxa de compressão[4] previamente estipulada. Nem sempre os sumários de referência, quando necessários para comparação, têm tamanho compatível ao dos sumários automáticos, o que pode gerar diferenças significativas de conteúdo e, assim, a impossibilidade de se efetuar uma comparação consistente;
- Sumários podem ser avaliados de diversas formas, se considerados os seus variados propósitos. Por exemplo, um sumário pode ser muito útil como indicativo do conteúdo de um texto, mas muito ruim se considerado um substituto dele.

A seguir, apresentam-se os tipos principais de avaliação que podem ser feitos.

2.2 Tipos de Avaliação

Assim como é distinguido na área de PLN em geral, a avaliação pode ser *intrínseca* ou *extrínseca* (Mani 2001). Na intrínseca, o sumarizador é avaliado de acordo com a qualidade direta dos textos produzidos. Na extrínseca, é mensurado o quanto os sumários são úteis para alguma outra tarefa que os utiliza, como por exemplo QA (*Question Answering*) ou Categorização de Textos. Na primeira tarefa, pode-se medir o quão bem um sumário construído a partir de um conjunto de documentos responde uma dada pergunta. Na segunda tarefa, por exemplo, pode-se avaliar a utilidade dos sumários como substitutos dos textos-fonte na determinação de suas categorias.

Caso o sistema seja avaliado observando-se apenas suas entradas e suas saídas, a avaliação é dita *black-box*. A avaliação será do tipo *glass-box* quando se observar também o que ocorre nos módulos e estados internos do sumarizador, e não apenas em seu funcionamento global.

[4] A definição matemática da taxa de compressão é dada na Seção 3.1.2 deste texto.

Neste trabalho, é dada ênfase nas avaliações intrínsecas. Nas seções seguintes, são descritas metodologias objetivas e subjetivas para esse tipo de avaliação.

2.3 Metodologias Objetivas de Avaliação Intrínseca

As avaliações objetivas podem ser geralmente feitas por medidas automáticas, comparando-se os sumários automáticos com sumários manuais, sendo estes os dados de referência e, por isso, usualmente denominados *gold-standards.* Para isso, calculam-se medidas numéricas que reproduzam essa correspondência com os dados de referência. Na área de Sumarização Automática, foram propostas inúmeras medidas, cada uma com sua forma de cálculo particular, para se avaliar sumários automáticos. Aqui destacamos somente as mais diretamente relacionadas às avaliações feitas para os modelos de SA propostos neste trabalho.

2.3.1 Medidas de Precisão, Cobertura e F-Measure

As medidas de Precisão (*P*), Cobertura (*C*) e *F-Measure* (*F*) foram inicialmente propostas para avaliar sistemas de Recuperação da Informação (Salton e Buckley 1988). Nesse caso, elas são utilizadas para avaliar o quão satisfatórias são as respostas recuperadas por um sistema de recuperação da informação. A Precisão (*P*) representa a quantidade de documentos relevantes para o usuário dentre os documentos recuperados na busca. A Cobertura (*C*) representa a quantidade de documentos relevantes recuperados dentre os documentos relevantes existentes na base de dados. Já a F-Measure (*F*) representa a média harmônica ente *P* e *C.* A F-Measure é útil por combinar numa única medida tanto a Precisão quanto a Cobertura.

Devido à tradição da área de Recuperação de Informação, essas medidas foram adotadas também para a avaliação de sumários automáticos (Mani 2001). Para isso, elas foram adaptadas para, em vez de se considerarem documentos completos recuperados, considerarem-se apenas sentenças apontadas pelo

sumarizador como relevantes para incluir no sumário em construção. A utilização de unidades sentenciais representa uma correspondência mais direta com a identificação de documentos importantes no modelo de Salton e Buckley (1988). Entretanto, nada impediria que outras unidades intratextuais fossem utilizadas como unidades elementares, no cálculo das medidas *P, C e F*

Quando é feita a comparação do extrato automático (*EA*) com um sumário ou extrato de referência (ER), as medidas podem ser assim definidas: *P* representa o número de sentenças do extrato automático contidas no de referência dividido pelo número sentenças do extrato automático; *C* representa o número de sentenças do extrato automático contidas no de referência dividido pelo número sentenças do extrato de referência; *F* é a média harmônica entre *P* e *C*.

$$P = \frac{|EA \cap ER|}{|EA|} \quad [1]$$

$$C = \frac{|EA \cap ER|}{|ER|} \quad [2]$$

$$F = 2\frac{P.C}{P + C} \quad [3]$$

A construção dos extratos de referência (ou extratos ideais) necessários para esse tipo de avaliação pode ser feita de forma automática a partir de sumários manuais. Uma das ferramentas para esse fim é o GEI – Gerador de Extratos Ideais (Pardo e Rino 2004). O GEI é capaz de produzir o extrato ideal correspondente, isto é, o sumário formado pelas sentenças do texto-fonte que mais se assemelham a cada uma das sentenças do sumário manual. Nesse processo, pode haver casos em que a sentença do texto-fonte que mais se assemelhe a uma determinada sentença do sumário manual já tenha sido associada a outra sentença do sumário manual. Quando isso ocorre, o GEI busca pela próxima sentença do texto-fonte com maior similaridade com a sentença do sumário manual em questão. Se essa outra sentença também já tiver sido selecionada, busca-se outra, e assim por diante. Pelo fato de se selecionar uma sentença do texto-fonte para cada sentença do sumário manual, o extrato ideal conterá o mesmo número de sentenças do sumário manual. Entretanto, a taxa de compressão dos extratos ideais (em número de palavras)

poderá acabar ficando bem diferente do sumário manual, em função das diferenças no tamanho das sentenças, para mais ou para menos.

2.3.2 Medida de Informatividade da Ferramenta ROUGE

Nos últimos anos têm sido muito utilizadas as medidas de avaliação presentes no pacote ROUGE[5] (*Recall-Oriented Understudy for Gisting Evaluation*) para calcular a informatividade dos sumários automáticos. Segundo Lin e Hovy (2003), as medidas ROUGE têm alta correlação com as avaliações humanas. As medidas ROUGE são utilizadas desde a DUC 2003[6] (Document Understanding Conferences), concurso de SA que passou a se chamar TAC[7] (Text Analysis Conference) em 2008.

Há varias formas de se calcular as medidas ROUGE e elas são dependentes do número de n-gramas que se considera para verificar a informatividade de extratos ou sumários automáticos em relação aos seus correspondentes sumários de referência, produzidos manualmente. A premissa das métricas é que a coocorrência de n-gramas entre um sumário manual e um sumário automático reflete sua informatividade.

As medidas calculadas pela ROUGE são denotadas genericamente por ROUGE-N, onde o parâmetro N indica o número de n-gramas considerado. Assim, ROUGE-1 indica a utilização de unigramas, ROUGE-2 indica a utilização de bigramas, etc. De forma geral, ROUGE-N é uma medida do tipo Cobertura, embora a ferramenta permita o cálculo da Precisão e F-measure. O cálculo da cobertura com base em n-gramas, é dado pela fórmula abaixo:

$$ROUGE-N = \frac{\sum_{S \in \{Sumários de Referência\}} \sum_{n-grama \in S} P(n-grama)}{\sum_{S \in \{Sumários de Referência\}} \sum_{n-grama \in S} 1} \qquad [4]$$

[5] Disponível em http://berouge.com/default.aspx (Junho/2010)
[6] http://www-nlpir.nist.gov/projects/duc/index.html (Junho/2010)
[7] http://www.nist.gov/tac (Junho/2010)

em que:

S denota cada sentença dos sumários de referência;

$P(n-grama)$ é uma função que apresenta o valor 1 quando o n-grama existir no sumário avaliado ou 0 caso contrário.

É importante frisar que embora o cálculo da métrica seja automático na ROUGE, a necessidade de uso de sumários manuais para comparação continua demandando uma equipe de profissionais especializados para sua elaboração.

2.3.3 Comparações objetivas entre sistemas de SA

As comparações objetivas entre sumarizadores têm sido feitas geralmente através de medidas automáticas, como as métricas de Precisão, Cobertura e F-measure (Seção 2.3.1) e métricas disponíveis na ferramenta ROUGE (Seção 2.3.2).

Em geral, tomam-se como referência as medidas médias de cada sumarizador obtidas para um corpus de avaliação. A questão que surge na comparação entre as medidas é se as médias são estatisticamente diferentes. Em outras palavras, se os resultados têm significância estatística. De forma geral, quanto maior o número de textos utilizados para o cálculo das medidas médias, tanto menores serão as diferenças necessárias para que os resultados sejam significativos.

Considerando essa análise estatística, o Apêndice C descreve o uso de possíveis métodos e ferramentas para a análise de significância estatística na comparação entre pares de sumarizadores. Particularmente, neste trabalho foi adotado o teste t-Student ou t-teste. Para mais sumarizadores outras abordagens são apresentadas em Densar (2006).

2.4 Metodologias Subjetivas de Avaliação Intrínseca

De modo geral, as metodologias subjetivas de avaliação buscam julgar os sumários de acordo com critérios qualitativos, como legibilidade, qualidade de

conteúdo ou utilidade. Embora a avaliação de informatividade possa ser feita de forma automática, com o uso da ROUGE, por exemplo, a avaliação manual costuma capturar melhor a sobreposição entre a informação expressa nos sumários manual e automático. Isso se deve ao fato de os sumários poderem expressar os mesmos conceitos com o uso de vocabulário completamente distinto, por meio do emprego de sinônimos, generalizações ou especializações. Como a avaliação automática geralmente se apoia na sobreposição de palavras, ela pode ser falha nesse sentido.

A seguir, descrevem-se três formas de avaliação subjetivas que têm sido utilizadas nas DUCs e TACs.

2.4.1 Questionário de Avaliação Qualitativa das DUCs e TACs

Nas DUCs e TACs tem sido utilizado o questionário mostrado na Figura 2-1 para avaliação manual da qualidade dos extratos. Esse questionário é preenchido por um único juiz. Como se pode ver, as notas de cada quesito devem variar no intervalo de 1 (muito ruim) a 5 (muito bom). As características a avaliar envolvem vários níveis da produção textual: o sintático (gramaticalidade); o semântico, de conteúdo (redundância) ou clareza referencial; o discursivo, de foco ou coerência; e o estrutural, que irá remeter aos níveis anteriores em maior ou menor grau. De um modo geral, esses parâmetros, juntos, servem para julgar a textualidade dos resultados automáticos.

Além desse questionário focado na textualidade, uma medida geral de utilidade do sumário tem sido proposta nas DUCs e TACs. Ela é denominada *Overall Responsiveness* e abrange tanto os quesitos de conteúdo quanto os de textualidade. Ela também pode variar de 1 (muito ruim) a 5 (muito bom).

1. **Gramaticalidade**
O sumário não pode apresentar erros com relação ao emprego incorreto de maiúsculas e minúsculas, formatação interna do sistema ou obviamente sentenças não-gramaticais que tornam o texto de difícil leitura.
   1. Muito ruim
   2. Ruim
   3. Aceitável
   4. Bom
   5. Muito bom

2. **Redundância**
Não deve existir repetição desnecessária no sumário. Essa repetição desnecessária pode aparecer como sentenças inteiras repetidas, fatos repetidos ou a utilização um sintagma nominal maior (ex. "O presidente") quando um pronome simples seria suficiente (ex. "Ele").
   1. Muito ruim
   2. Ruim
   3. Aceitável
   4. Bom
   5. Muito bom

3. **Clareza Referencial**
Deve ser facilmente identificável a quem ou ao que os pronomes e sintagmas nominais se referem no texto.[8]
   1. Muito ruim
   2. Ruim
   3. Aceitável
   4. Bom
   5. Muito bom

4. **Foco**
O sumário deve ter foco, contendo apenas sentenças que contenham informação relacionada ao restante do sumário.
   1. Muito ruim
   2. Ruim
   3. Aceitável
   4. Bom
   5. Muito bom

5. **Estrutura e Coerência**
O sumário deve estar bem-estruturado e bem-organizado. Ele deve ser construído sentença a sentença de modo a formar um conjunto coerente de informações sobre um assunto.
   1. Muito ruim
   2. Ruim
   3. Aceitável
   4. Bom
   5. Muito bom

Figura 2-1 – Questionário de avaliação subjetiva proposto nas DUCs e TACs

[8] A propriedade de clareza referencial já foi citada no Capítulo 1, onde se mencionou que os extratos podem possuir com frequência problemas relacionados a essa propriedade, como a presença de anáforas endofóricas não resolvidas.

2.4.2 O Método da Cobertura de Unidades Elementares (Basic Elements)

O método da Cobertura das Unidades Elementares ou *Basic Elements* (Tratz e Hovy 2008) foi utilizado até a DUC 2005 e trata da avaliação de conteúdo. Basicamente, ele propõe a divisão dos sumários manuais em pequenas unidades elementares de conteúdo. Essa divisão tem o propósito de facilitar a verificação da cobertura das informações pelos juízes humanos.

A vantagem da utilização dessas unidades é que elas permitem considerar expressões compostas por múltiplas palavras como uma única entidade. Por exemplo, enquanto na abordagem de coocorrência de unigramas da medida ROUGE-1, as palavras que compõem o sintagma "Estados Unidos da América" podem ser consideradas isoladamente, na abordagem de Cobertura de Unidades Elementares todas indicam uma mesma unidade. Além disso, por esse método os sintagmas "Estados Unidos da América", "Estados Unidos" e "EUA" são considerados unidades elementares sobrepostas, pois apontam o mesmo conceito.

O processo completo de julgamento é o seguinte:

- O sumário de referência é dividido em unidades elementares de conteúdo, que correspondem grosso modo às unidades elementares do discurso. O processo de anotação é manual[9];
- Um juiz humano analisa o sumário automático e encontra todas as unidades que cobrem pelo menos algumas das informações das unidades elementares do sumário manual;
- Para cada unidade elementar de conteúdo do sumário manual, o juiz então analisa se o sumário automático expressou entre 0, 20%, 40%, 60%, 80% ou 100% da informação.

Obtida a pontuação do sumário, seu escore final é normalizado entre 0 e 1 e representa, assim, a cobertura das unidades de conteúdo do sumário de referência.

[9] Estudos (Tratz e Hovy 2008) têm sido feitos para tentar obter automaticamente as unidades elementares. Entretanto, a determinação manual tem sido mais frequente.

2.4.3 O Método da Pirâmide

O método da Pirâmide (Nenkova et al. 2007) busca avaliar o conteúdo dos sumários automáticos considerando que as pessoas, quando sumarizam os mesmos textos-fonte, podem selecionar informações diferentes na composição de seus sumários manuais. A metodologia considera a frequência com que as informações coocorrem entre os sumários manuais. Ou seja, é mais indicada quando existem múltiplos sumários de referência.

Assim como no método de Cobertura das Unidades Elementares, no método da Pirâmide anotam-se manualmente as chamadas SCUs (*Summary Content Units*) dos sumários de referência. Essas unidades são fragmentos do texto com algum significado.

A importância de uma SCU está ligada à frequência com que a informação expressa nela é encontrada no conjunto de sumários. Quando maior a frequência, mais importante é considerada a SCU. Um conjunto de SCUs geralmente forma uma pirâmide. Na base, existe um número grande de SCUs com peso 1 ou 2, enquanto no ápice um conjunto bem menor e com pesos bem maiores e consequentemente mais importantes.

De forma semelhante ao método anterior, os sumários automáticos são confrontados com as SCUs. Toda vez que a informação expressa na SCU for coberta pelo sumário, ele recebe o valor *n*, que é o peso da SCU. Caso contrário, recebe 0. O escore final do sumário é razão entre a soma dos pesos das SCUs cobertas e a soma dos pesos de um sumário ótimo, construído a partir da seleção das SCUs com maiores pesos (do ápice).

Como citado, o Método da Pirâmide tem a robustez de considerar que vários sumários de referência podem expressar informações diferentes e com vocabulário diferente. Entretanto, a principal dificuldade do método é a anotação manual e subjetiva das SCUs.

O exemplo a seguir é fornecido pelos autores (Nenkova et al. 2007, p. 3). Os textos foram obtidos a partir de sumários referenciais da DUC 2003 e traduzidos por nós para o Português a título de clareza. Os trechos foram identificados por uma letra, indicando o sumário de procedência, e por um número indicando a posição da sentença nesse sumário.

A1. Em 1998, dois líbios indiciados em 1991 pelo atentado em Lockerbie ainda estavam na Líbia.

B1. Dois líbios foram indiciados em 1991 por ter derrubado um jumbo da PanAm sobre Lockerbie na Escócia, em 1988.

C1. Dois líbios, acusados pelos EUA e pelo Reino Unido de explodir um jato nova-iorquino da Pan Am sobre Lockerbie na Escócia em 1988, matando 270 pessoas, estavam refugiados na Líbia que alegou que eles não teriam um julgamento justo nos EUA ou no Reino Unido.

D2. Dois suspeitos líbios foram indiciados em 1991.

A anotação inicia-se com a identificação de trechos similares, grifados no exemplo. A partir disso, faz-se uma análise mais profunda, que poderá levar a identificação de trechos mais fortemente relacionados. Nesse exemplo, duas SCUs foram identificadas a título de exemplo. Cada SCU recebe um peso conforme o número de sumários que a possuem.

SCU1 (w = 4): dois líbios foram oficialmente acusados pelo atentado em Lockerbie.

A1 [dois líbios] [indiciados]
B1 [dois líbios foram indiciados]
C1 [dois líbios] [acusados]
D2 [dois suspeitos líbios foram indiciados]

SCU2 (w = 3): o indiciamento dos suspeitos do atentado em Lockerbie foi em 1991

A1 [em 1991]
B1 [em 1991]
D2 [em 1991]

Outras SCUs, embora de pesos menores, podem ainda ser identificadas nesses trechos. Uma vez identificadas as SCUs, elas podem ser dispostas numa pirâmide. Nos níveis superiores da pirâmide devem estar dispostas as SCUs de maior peso. Isso permitirá identificar os sumários ótimos, que deverão ser construídos dando preferência às sentenças do topo e ir descendo até a base conforme a taxa de compressão permitir. A Figura 2-2 ilustra dois sumários ótimos (de seis possíveis) com tamanho de 4 SCUs.

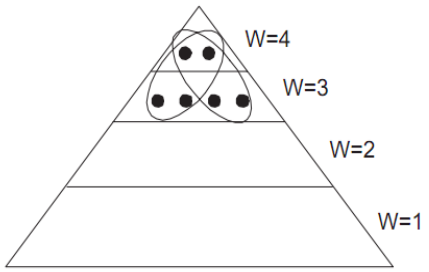


Figura 2-2 – Exemplo do Método da Pirâmide

O escore final atribuído a cada sumário avaliado é a razão entre a soma dos pesos da SCU desse sumário com a soma dos pesos de um sumário ótimo de mesmo tamanho.

A revisão feita neste capítulo é retomada nos capítulos subsequentes: no Capítulo 5, onde é feita a avaliação dos modelos propostos neste trabalho em uma série de experimentos, e no próximo capítulo, onde se reporta o desempenho de uma série de sumarizadores já existentes.

Capítulo 3

ABORDAGENS DE SA EXTRATIVA

Neste capítulo são apresentados os modelos clássicos de Sumarização Automática, alguns sumarizadores existentes para o português do Brasil e trabalhos recentes. Ao final do capítulo, é feita um síntese onde se distinguem as três grandes abordagens exploradas neste trabalho: estatísticas textuais, grafos e aprendizado de máquina.

3.1 Modelos Clássicos de Sumarização Automática Extrativa

Conforme já citado no Capítulo 1, a qualidade dos textos produzidos por modelos extrativos é em geral questionável, sendo difícil controlar ou modelar processos de decisão que garantam sua textualidade, considerando as questões de coesão e coerência (Halliday e Hasan 1976). Muitas vezes sumarizadores automáticos extrativos produzem extratos informativos, mas com textualidade ruim.

Por conta disso, buscam-se maneiras de se contemplar na SA Extrativa maior conhecimento linguístico. Como exemplo, pode-se citar desde o pré-processamento textual, que é dependente da língua (p.ex., *stemming* ou *tagging*), até o uso de técnicas mais robustas como o encadeamento lexical explorado por Barzilay e Elhadad (1999), descrito ainda neste capítulo. Essas técnicas visam suprir a ausência, na abordagem essencialmente empírica, dos processos de abstração e de fusão, sendo o primeiro responsável por qualquer reformulação de material relevante do texto-fonte e o segundo, por sua condensação. Esses processos são comumente associados ao tratamento da coerência do texto final (Sparck Jones e Galliers 1996) e são viabilizados quando se processam as decisões de reestruturação do sumário; neste caso, remetendo à modelagem fundamental.

No que segue, apresentam-se alguns dos modelos principais da área de Sumarização Automática extrativa, que estão bastante relacionados ao presente trabalho. Sempre que aplicável, apontam-se as vantagens ou desvantagens dos modelos apresentados.

3.1.1 Modelo de Frequência das Palavras

O modelo mais tradicional de SA surgiu no final da década de 50, particularmente com o trabalho de Luhn (1958). Nesse trabalho, Luhn verificou que a distribuição de palavras relevantes num texto estava associada a uma lei de Zipf (Zipf 1949; Gelbukh e Sidorov 2001). Através da análise dessa distribuição, Luhn propôs um método para julgar a relevância de uma sentença a partir da análise das frequências das palavras dessa sentença no texto. De modo geral, a hipótese do modelo é de que as sentenças mais importantes são as que possuem as palavras com maior poder de resolução, isto é, palavras que estejam relacionadas às informações mais relevantes do texto. Estas palavras são as que se concentram numa região estabelecida por dois cortes na curva de distribuição de frequências. A Figura 3-1 ilustra essa distribuição com a função função *E*, em forma de sino, que representa o poder de resolução das palavras.

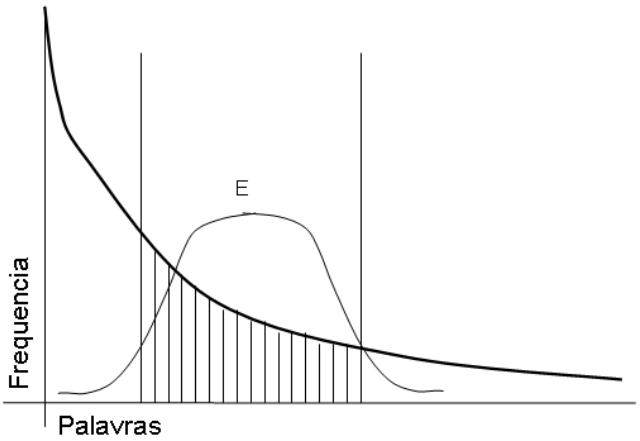


Figura 3-1 – Distribuição de Luhn

Usualmente, o corte superior de frequências é feito removendo-se palavras conhecidas que apresentam esse comportamento, as chamadas *stopwords*.

Principais vantagens: construir um extrato a partir de apenas informações superficiais do texto; baixo custo computacional.

Principais desvantagens: considerar apenas a frequência das palavras como determinante da relevância das sentenças.

3.1.2 Modelo de Combinação de Características

A hipótese básica dos processos extrativos é de que é possível se gerar um resumo a partir de um ranking de sentenças. Ou seja, uma vez avaliadas as sentenças, selecionam-se de forma ordenada para compor o extrato as sentenças com os maiores índices. A seleção é feita até que se atinja o tamanho desejado do texto, determinado por sua taxa de compressão (TC), definida a seguir:

$$TC = \frac{TamanhoExtrato}{TamanhoTextoOriginal} \qquad [5]$$

Essa fórmula para a TC indica o percentual do tamanho dos sumários automáticos produzidos em relação texto-fonte, ao contrário de indicar o quanto o texto-fonte deve ser comprimido (que seria o conceito usual associado a esse termo). Segue, portanto, a definição de Mani (2001). O tamanho dos textos é usualmente medido pelo número de palavras.

Partindo dessa hipótese, Edmundson (1969) propôs o arcabouço geral que vem sendo seguido até os dias atuais. Para Edmundson, a saliência ou relevância de uma sentença pode ser associada a características superficiais dessa sentença, como sua localização no texto, tamanho e a presença de certas palavras indicativas, que podem tanto indicar tanto sua importância como sua dispensabilidade. De forma geral, o modelo pondera e combina as características consideradas relevantes no processo de julgamento para se determinar um número indicativo de sua saliência. Em sua forma mais simples, o modelo é linear e apresentado na fórmula seguinte:

$$Saliência \quad (s) = w_1 \times C_1(s) + ... + w_n \times C_n(s) \qquad [6]$$

em que:
s é a sentença;
C_i (s) é o valor de uma característica i para a sentença;
w_i é o peso atribuído à característica.

A grande questão na utilização do modelo de Edmundson é como deve ser feita a atribuição dos pesos às características para combiná-las, justamente o Problema 2 apresentado na introdução deste trabalho. Esses pesos poderão variar, por exemplo, conforme o gênero linguístico considerado. Em textos jornalísticos, por exemplo, é usual que as informações principais concentrem-se no início do texto (Mani 2001). Por conta disso, a característica de localização possivelmente terá um peso maior.

No trabalho de Edmundson, os pesos foram ajustados empiricamente com base na comparação com extratos gerados manualmente, para um corpus de textos científicos. Todo o trabalho deveria ser refeito caso um novo gênero ou corpus muito diferente fosse utilizado.

Principais vantagens: possibilita a combinação de múltiplas características para sumarização.

Principais desvantagens: a determinação da função de combinação das características e dos pesos atribuídos a elas é manual.

3.1.3 Utilização de Aprendizado de Máquina

Em 1995, com a abordagem pioneira de Kupiec et al. (1995), foi introduzido o uso de aprendizado de máquina supervisionado[10] para a tarefa de SA. Esse trabalho propôs a utilização do classificador Naïve-Bayes (e.g., Mitchel 1997) e características (*features*) binárias (*True* indicando a presença e *False* a ausência) que refletem propriedades superficiais do texto. Essa abordagem se mostrou promissora, por produzir extratos mais informativos que as antigas abordagens não-supervisionadas. Além disso, a utilização de corpora de treino surgiu como mecanismo para combinação das características, que é um dos maiores desafios no arcabouço de Edmundson.

[10] O Aprendizado de Máquina Supervisionado é aquele que utiliza em sua fase de treino pares problema/solução usualmente feito por humanos. O Capítulo 4 apresenta uma descrição mais detalhada da utilização de Aprendizado de Máquina para SA.

O classificador proposto é o Naïve-Bayes para determinar, para cada sentença de um texto-fonte, a probabilidade de ela estar incluída em seu sumário. Espera-se que quanto maior o valor dessa probabilidade, maior a relevância da sentença para a composição do extrato.

São cinco as características utilizadas pelo classificador: presença de palavras relevantes, presença de nomes próprios, presença de termos indicativos, tamanho da sentença e posição da sentença. Como as características são binárias, seu cômputo é feito estabelecendo-se um valor de corte. Por exemplo, se o número de palavras relevantes na sentença for maior que 5, então a *característica* assume *True* como valor; caso contrário, *False*.

Para o cálculo da relevância de uma sentença, a fórmula usada, apresentada a seguir, é derivada do teorema de Bayes e indica a probabilidade de a sentença *s* ser incluída no extrato, dado um conjunto de valores booleanos para as *k* características da sentença.

$$P((s \in E) \mid C_1, C_2, ..., C_k) = \frac{\prod_{j=1}^{k} P(C_j \mid s \in E) \times P(s \in E)}{\prod_{j=1}^{k} P(C_j)} \qquad [7]$$

O termo C_j representa o valor da característica *j*. $P(C_j \mid s \in E)$ é a probabilidade do valor C_j ocorrer nas sentenças pertencentes aos extratos do corpus de treinamento para a *característica j*. $P(s \in E)$ é a probabilidade a priori de *s* estar incluída no extrato. Essa probabilidade é dependente da taxa de compressão (TC), dada pela fórmula [5]. Ou seja, $P(s \in E)$ é constante para cada texto a ser sumarizado. Finalmente, $P(C_j)$ é a probabilidade do valor *Cj* ocorrer em todo o corpus de treinamento para a característica *j*.

Para cada novo texto a ser sumarizado, calculam-se os valores das *características* e obtém-se uma tupla $C_1, C_2, ..., C_k$ para cada sentença. Feito isso, aplica-se o classificador Naïve-Bayes para cada tupla e ordenam-se as probabilidades em ordem decrescente. As sentenças que comporão o extrato são, então, selecionadas de acordo com as maiores probabilidades. Seu número será proporcional à taxa de compressão desejada.

Desde a introdução da abordagem de Kupiec et al., vários outros classificadores e combinação de características tem sido empregados, sempre buscando refinar o modelo de seleção.

No Apêndice A é apresentado um exemplo completo de utilização do modelo.

Principais vantagens: possibilita a combinação de múltiplas características para sumarização e o ajuste da função de combinação com base num processo treinado usando aprendizado de máquina.

Principais desvantagens: não resolve a questão de quais características devem ser consideradas no modelo; limitações dos modelos de aprendizado de máquina ou hipóteses violadas nesses modelos, como a assunção de independência das probabilidades condicionais no modelo Naïve-Bayes (Mitchel 1997), podem introduzir distorções nos resultados obtidos.

3.1.4 Mapa de Relacionamentos

A proposta de Salton et al. (1997) é construir uma representação da coesão no texto por meio de um grafo que contém as ligações entre os parágrafos. Palavras iguais em parágrafos diferentes são consideradas uma ligação entre eles. Parágrafos conectados a vários outros são considerados salientes no texto, pois devem conter tópicos discutidos em vários outros parágrafos. A Figura 3-2 ilustra o chamado Mapa de Relacionamentos construído pelo método de Salton et al.

De forma geral, a construção do grafo que contém a ligação dos parágrafos do texto é o principal passo desse método. Os parágrafos do texto que comporão um extrato são selecionados percorrendo-se esse grafo por três modos distintos. A unidade mínima extraída é o parágrafo, porque seus autores consideram que o uso dessa unidade, em vez de segmentos de granularidade menor, faz com que problemas de legibilidade e coerência sejam amenizados.

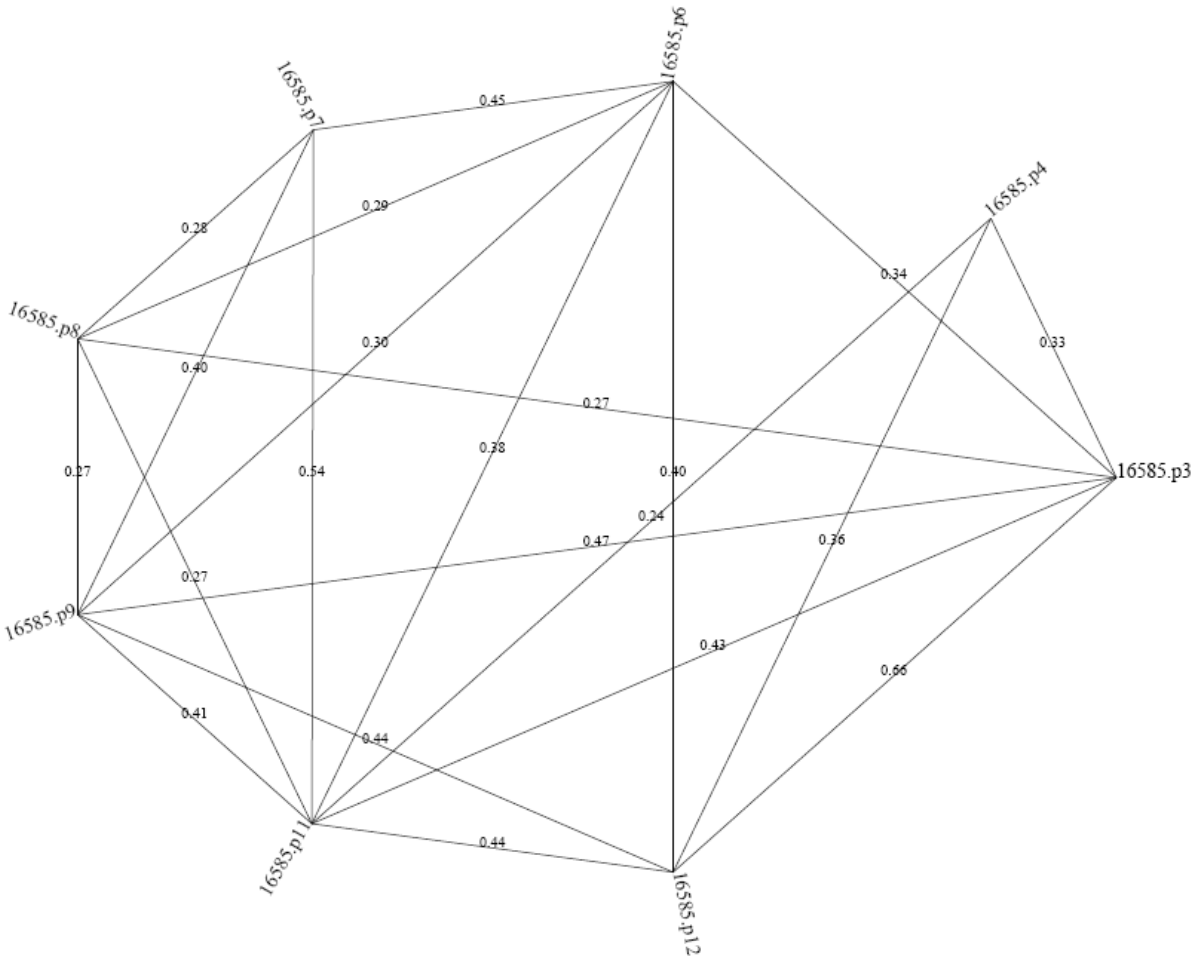


Figura 3-2 – Exemplo de Mapa de Relacionamentos (Salton et al.), para uma estrutura de parágrafos altamente ligada

Os passos para geração de extratos pelo método são:

1) *Pré-processamento do texto-fonte:*
   a. Remoção das *stopwords*;
   b. *Stemming* ou geração de quadrigramas[11].

2) *Construção do mapa de relacionamentos:*
   a. Cálculo dos vetores TF-IPF[12] (*Term frequency – Inverse paragraph frequency*) para cada parágrafo do texto;

[11] São sequências sobrepostas de quatro letras de cada palavra. Para a palavra *estufa*, por exemplo, seriam: _est / estu / stuf / tufa / ufa_ /

[12] Trata-se de uma adaptação medida TF-IDF, substituindo a noção de documento pela noção de parágrafo.

b. Cálculo da similaridade entre os parágrafos. Essa similaridade é determinada pela aplicação de uma medida semelhante à dos co-senos aos vetores TF-IPF.
c. Construção do mapa de relacionamentos, considerando as ligações entre parágrafos que tenham similaridade acima de um valor mínimo.

3) *Seleção dos parágrafos: o* extrato é obtido percorrendo-se o grafo através de três caminhos:

- **Caminho 1** (denso): são selecionados os parágrafos que contém o maior número de ligações até que a taxa de compressão seja atingida.
- **Caminho 2** (profundo): primeiro é selecionado o parágrafo com o maior número de ligações (mais denso). A partir do parágrafo mais denso, é selecionado aquele que tem o maior número de ligações com ele. A partir desse segundo parágrafo, é selecionado o terceiro que tem mais ligações com este último, e assim sucessivamente, até atingir a taxa de compressão escolhida.
- **Caminho 3** (segmentado): este caminho procura selecionar diferentes parágrafos de cada tópico do texto-fonte. Os tópicos são produzidos através de uma simplificação do mapa, proposta pelos autores do método. Usando essa divisão em tópicos, é selecionado pelo menos um parágrafo de cada tópico, sendo que outros parágrafos podem ser selecionados dos tópicos com maior número de parágrafos, até se obter a taxa de compressão desejada.

Como se vê, esses caminhos são baseados em três hipóteses distintas, para a construção dos extratos. No caminho denso, supõe-se que a seleção dos parágrafos mais densos leve a uma boa cobertura dos conceitos principais do texto. No entanto, considerar simplesmente os parágrafos mais densos não garante coerência. Já o caminho profundo privilegia a coerência, mas não cobre todos os conceitos principais, ou seja, perde em informatividade. O caminho segmentado justamente tenta balancear informatividade e coerência, utilizando a divisão do texto em tópicos.

Principais vantagens: busca contemplar num modelo extrativo informações relacionadas à estrutura coesiva do texto, por meio da similaridade lexical.

Principais desvantagens: a similaridade lexical por si só pode não ser suficiente para construir um modelo da estrutura coesiva no texto; existem 3 caminhos para seleção dos parágrafos finais e não há uma regra geral de como utilizá-los conjuntamente; a unidade mínima de extração é o parágrafo.

3.1.5 Utilização de Cadeias Lexicais

Cadeias lexicais são sequências de palavras semanticamente relacionadas entre si pelos mecanismos de repetição, sinonímia/antonímia, hiperonímia/hiponímia ou holonímia/metonímia. Elas são importantes na medida em que estão relacionadas à estrutura coesiva do texto. A figura a seguir mostra um exemplo de cadeia, traduzido de Barzilay e Elhadad (1999, p. 2):

O Sr. Kenny é a pessoa que inventou a máquina anestésica, que utiliza um microcomputador para controlar o bombeamento de anestésicos no sangue. Tais máquinas não são novas. Mas, seu dispositivo utiliza dois microcomputadores para obter um monitoramento mais preciso do bombeamento do anestésico no paciente.

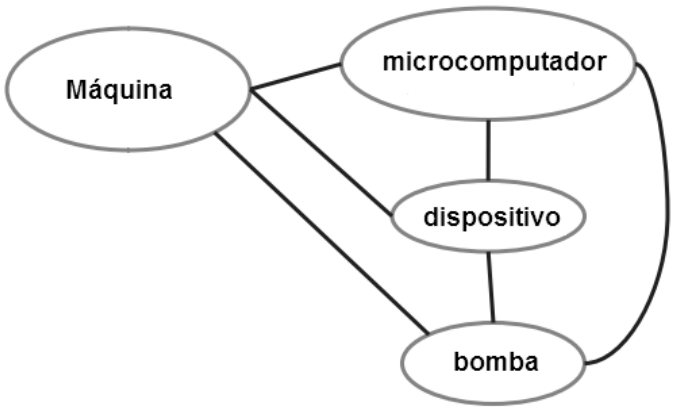


Figura 3-3 – Exemplo de Cadeais Lexicais

A partir da obtenção das cadeias, a ideia do método é a de que as cadeias fortes num texto representam seus conceitos importantes, daí a utilidade para a sumarização automática. Sentenças mais relacionadas a essas cadeias devem, portanto, ser consideradas na geração do extrato.

O número de relações em uma cadeia lexical, e seus respectivos pesos, são utilizados para que as cadeias mais promissoras sejam selecionadas para cada segmento. Esses segmentos são delimitados pelo algoritmo TextTiling, o qual segmenta um texto em grupos coerentes de sentenças (Hearst 1993). As cadeias dos diferentes segmentos são unidas quando possuem um termo em comum (de mesmo sentido), o que dá origem a uma rede de relações semânticas entre os termos do texto-fonte.

Para selecionar as sentenças para compor o extrato, os autores propuseram três heurísticas (H1, H2 e H3):

- **H1**: seleção de toda sentença *s* do texto-fonte baseada em cada membro *m* da toda cadeia lexical forte do texto. No caso, *s* é a sentença que contém a primeira ocorrência do membro *m*;
- **H2**: a seleção é feita de maneira similar à heurística anterior, com a diferença que são considera somente os membros julgados como representativos da cadeia lexical. Um membro é considerado representativo se sua frequência na cadeia for maior que a média da frequência de todos os membros;
- **H3**: novamente, a heurística é similar à anterior, com a diferença de que considera a estrutura de tópicos do texto, selecionando as cadeias representativas de cada tópico do texto.

Barzilay e Elhadad avaliaram seu método em um experimento utilizando sumários construídos manualmente, e obtiveram melhores resultados de Precisão e Cobertura do que a ferramenta *AutoResumo* do Microsoft Word.

Principais vantagens: busca contemplar num modelo extrativo informações relacionadas à estrutura coesiva do texto; utilização de relações semânticas.

Principais desvantagens: dificuldade de implementação do modelo para todas as línguas devido a necessidade de uma ontologia do tipo WordNet (Miller, Beckwith et al. 1990) para determinação das relações semânticas; existem 3 heurísticas para seleção das sentenças finais e não há uma regra geral de como utilizá-las conjuntamente.

3.1.6 Importância dos Tópicos

O método proposto por Larocca Neto et al. (2000) objetiva a identificação dos tópicos principais do texto e a consideração da importância do tópico na seleção das sentenças. Para isso, o texto é dividido em tópicos que são avaliados segundo suas importâncias. A relevância de uma sentença para a composição do extrato será, então, proporcional à importância do tópico em que ela figura e também proporcional à sua similaridade com relação a esse mesmo tópico.

Os passos básicos descritos pelo método são:

2) Pré-processamento do texto-fonte:
    a. Remoção das *stopwords*;
    b. *Stemming* ou geração de quadrigramas.
3) Divisão do texto em tópicos: uma versão modificada do algoritmo TextTiling, proposta pelos autores do método, é aplicada ao texto para sua divisão em tópicos (tiles).
4) Cálculo da força dos tópicos: a força do tópico é definida como a soma da média dos vetores TF-ISF[13] (Term frequency – Inverse sentence frequency) de suas sentenças. Os valores obtidos para todos os tópicos são, então, normalizados no intervalo [0,1].
5) Cálculo do número de sentenças de cada tópico: é calculado, com base na importância dos tópicos, um número proporcional de sentenças a extrair de cada tópico.
6) Seleção das sentenças: para cada tópico, selecionam-se as sentenças que têm maior similaridade com o centróide do tópico. O centróide do tópico é o vetor resultante da média dos vetores TF-ISF das sentenças desse tópico. Para o cômputo da similaridade da sentença com o centróide é usada a clássica medida de similaridade dos co-senos (Salton e Buckley 1988).

[13] Trata-se de uma adaptação da clássica medida TF-IDF (*Term frequency – Inverse document frequency*), de Salton & Buckley (1988), usada no campo da recuperação da informação. A adaptação consiste em substituir a noção de documento pela noção de sentença.

Principais vantagens: considera a estrutura de tópicos do texto, podendo ser útil na construção de extratos que abordam não somente o tópico principal do texto.

Principais desvantagens: depende de uma boa divisão do texto em tópicos.

3.1.7 TextRank

O TextRank (Mihalcea 2005) é baseado no mesmo algoritmo usado pelo Google™ para julgar a relevância das páginas da WWW, o PageRank (Brin e Page 1998). Este considera que a importância de uma página é proporcional ao número de recomendações que existem na WWW para ela, as quais são, basicamente, os *links* que apontam para a página. O PageRank constrói um grafo em que os vértices são as páginas e as arestas são suas recomendações. É capaz de percorrer o grafo para definir a importância das páginas, que são consideradas em seu processo de busca de informações.

Seguindo a mesma ideia, o TextRank constrói um grafo similar, em que os vértices são sentenças e as arestas são denotadas pelo grau de similaridade entre pares de sentenças. Em vez dos *links* entre páginas, as "recomendações" entre sentenças são medidas por seu grau de similaridade, calculado de acordo com a seguinte equação:

$$Sim(S_i, S_j) = \frac{|\{w_k \mid w_k \in S_i \wedge w_k \in S_j\}|}{\log(|S_i|) + \log(|S_j|)} \quad [8]$$

em que:

S_i e S_j são sentenças

w_k é um termo comum entre as sentenças.

Uma vez construído o grafo, a importância da sentença é calculada por meio do algoritmo de passeio aleatório (*random walk*) mostrado na Equação 9.

$TR(V_i)$ sinaliza a importância da sentença representada pelo vértice V_i, $Sim(S_i, S_j)$ *é* a similaridade lexical entre as sentenças *Si* e *Sj*, d é um parâmetro de calibração no intervalo [0;1], e *N* é o número de sentenças do texto. O parâmetro *d* representa no modelo a probabilidade de migração de um vértice a outro. Ele é a

componente responsável pelo passeio aleatório. A implementação de Mihalcea (2005) utiliza o valor 0,85.

$$TR(V_i) = (1-d) + d \times \sum_{j=0}^{N-1} \left(TR(V_j) \times \frac{Sim(S_i, S_j)}{\sum_{k=0}^{N-1} Sim(S_j, S_k)} \right) \quad [9]$$

A ideia intuitiva dessa fórmula está em considerar que a importância de uma sentença ($TR(V_i)$) está associada ao quão similar essa sentença é com sentenças que possuam alto $TR(V_j)$. Em outras palavras, "sentenças relevantes" devem ser "recomendadas" por outras sentenças relevantes. Mede-se aqui o nível de recomendação pela similaridade lexical entre as sentenças.

Como a medida é recursiva, os valores iniciais são definidos aleatoriamente no intervalo (0;1). Depois de sucessivas iterações, os valores tendem a convergir. Calculadas as importâncias das sentenças, elas são ordenadas em ordem decrescente e selecionadas até que seja atingida a taxa de compressão.

Mihalcea (2005) também propôs algumas variações no método original, a saber:

a) A utilização de grafos dirigidos, conforme a ordem em que as sentenças aparecem no texto. Nesse caso, o percurso pode ser feito de duas formas: *backward* (considerando as sentenças predecessoras) ou *forward* (considerando as sucessoras)

b) A utilização do método HITS (Kleinberg 1999), que trabalha de forma semelhante ao PageRank. Porém, em vez de agregar em uma única medida os *links* predecessores e sucessores, produz duas medidas independentes chamadas respectivamente de *authority* e *hubs*.

Principais vantagens: modelo fundamentado pela Teoria dos Grafos; fácil implementação e independente de língua; considera no cômputo da importância de uma sentença também a importância das sentenças que a recomendam.

Principais desvantagens: não há garantias formais de convergência para textos; não considera nenhum tipo de pré-processamento linguístico, nem mesmo *stemming* ou remoção de *stopwords* na proposta original dos autores.

3.2 Sumarizadores para o Português do Brasil

Apresentam-se nesta seção alguns dos sumarizadores extrativos principais para o Português do Brasil.

3.2.1 GistSumm

O GistSumm (Pardo et al. 2003) é um sumarizador que busca identificar a ideia central (*gist*) do texto-fonte utilizando técnicas estatísticas. Ele utiliza o método das palavras-chave ou da métrica TF-ISF, a critério do usuário, para escolher a sentença mais bem pontuada que será, então, tomada como a *gist sentence*[14]. A partir da inclusão dessa sentença no extrato, outras sentenças são escolhidas conforme suas pontuações, 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 de extratos. No primeiro caso, foi utilizado um corpus 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.

Melhorias recentes no sistema (Balage et al. 2006) visaram diminuir a restrição da premissa de que texto possui apenas uma única *gist sentence*. Para isso, esses autores propuseram a delimitação da estrutura textual do texto, identificando suas seções. Por exemplo, um texto científico pode possuir as seções “Introdução”, “Metodologia”, “Desenvolvimento” e “Conclusões”. Uma vez identificadas as seções, procede-se a sumarização como se cada seção fosse um texto individual. Pela avaliação dos autores, essas melhorias são promissoras.

Principais vantagens: simples do ponto de vista computacional; tende a favorecer a coerência ao determinar a *gist sentence* e escolher as sentenças relacionadas a ela.

[14] A expressão “gist sentence” refere-se à sentença que mais representa a ideia central do texto-fonte.

Principais desvantagens: parte da hipótese de que o texto possui uma única sentença que contém a ideia principal, na proposta original; a versão modificada ainda não é capaz de lidar com textos que possuem vários tópicos, mas que não apresentam uma divisão clara em seções[15].

3.2.2 ClassSumm

O ClassSumm (Larocca Neto et al. 2002) é baseado na proposta de Kupiec et al. (1995). A grande diferença é que os autores utilizam 12 características superficiais em vez de apenas 5. Além disso, exploraram a utilização do classificador baseado em árvore de decisão, o C4.5 (Quinlan 1993).

Nos experimentos reportados o algoritmo Naïve-Bayes foi superior ao algoritmo C4.5. Além disso, o desempenho do modelo foi superior ao método considerado *baseline*, o qual seleciona as primeiras sentenças do documento a ser sumarizado.

Principais vantagens: explora um número maior de características que o modelo de Kupiec et al. (1995).

Principais desvantagens: as mesmas do modelo de Kupiec et al. (1995).

3.2.3 NeuralSumm

O NeuralSumm (Pardo 2003) utiliza uma rede neural do tipo SOM (*Self-Organizing Map*), proposta por Kohonen (1990), para classificar as sentenças do texto a ser sumarizado, com base em um conjunto de características pré-selecionado.

Essa rede neural 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. A prioridade

[15] Isso poderia ser viabilizado com a utilização de um algoritmo detector de tópicos, como o TextTiling, já utilizado no método proposto por Larocca et al. (2000).

na seleção das sentenças é primeiro selecionar as essenciais, em seguida as complementares e em último caso as supérfluas.

Os autores utilizaram um conjunto de 8 características para SA, entre elas 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 do Brasil, anotado por juízes 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.

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 outros trabalhos. Entretanto, quando comparado a outros seis sumarizadores (Rino et al. 2004), o NeuralSumm obteve um desempenho inferior a outros sumarizadores supervisionados e não-supervisionados para o Português.

Principais vantagens: utilização de um modelo conexionista que teve menores taxas de erro que os algoritmos Naïve-Bayes e C4.5 em algumas avaliações.

Principais desvantagens: diferente da abordagem de Kupiec et al. (1995), o modelo exige a anotação manual de cada sentença nas 3 categorias estabelecidas para treino. No modelo de Kupiec et al., isso não é necessário.

3.2.4 SuPor

O SuPor[16] (Módolo 2003) foi apresentado inicialmente no Capítulo 1 devido a esse sistema ter servido de motivação inicial para este trabalho. Como já citado, na comparação com outros seis sumarizadores extrativos para o português (Rino et al. 2004) foi considerado o melhor, atingindo a *F-Measure* de 42,8%.

[16] O nome “SuPor” vem de “SUmarização para o PORtuguês”

As abordagens utilizadas pelo SuPor envolvem a utilização de aprendizado de máquina com o uso de um classificador Naïve-Bayes seguindo o modelo de Kupiec et al. (1995). Por se tratar de um ambiente de SA, o SuPor permite ao usuário a escolha das características que serão utilizadas para a SA dos textos

Essas características empregadas são definidas com base em diversos métodos clássicos de SA e outros fatores relevantes, descritos na Seção 3.1. Os métodos utilizados são o método de Cadeias Lexicais (Barzilay e Elhadad 1999), o cômputo da frequência das palavras (Luhn 1958) e a localização das sentenças (Edmundson 1969). Outras características consideradas incluem o comprimento das sentenças, a ocorrência de substantivos próprios (Kupiec et al., 1995), a representação da coesão do texto por meio de um Mapa de Relacionamentos entre parágrafos (Salton, Singhal et al. 1997) e a identificação de sentenças indicativas dos tópicos principais do texto (Larocca Neto, Santos et al. 2000). As características usadas pelo SuPor, assim como no modelo proposto por Kupiec et al. (1995), são todas binárias. Ou seja, se um método recomenda uma sentença para compor o extrato, o valor da característica associada ao método será *True*; caso contrário, *False*. A Tabela 3-1 resume os métodos utilizados e as características associadas no SuPor.

Tabela 3-1 – Métodos e características associadas no SuPor

Método	Condição para a Característica assumir *True*
Frequência das Palavras	Escore (soma das frequências de cada palavra) deve ser maior que a média
Tamanho da sentença	Número de palavras da sentença deve ser maior que cinco
Posição	Sentença deve figurar nos parágrafos iniciais e finais do texto, ou ser uma sentença inicial ou final de qualquer parágrafo
Nomes próprios	Escore (soma das frequências de cada nome) deve ser maior que o mínimo
Cadeias lexicais	A sentença deve ser recomendada por pelo menos uma das heurísticas do método
Importância dos tópicos	A sentença deve ser saliente no tópico em que se encontra
Mapa de relacionamentos	A sentença deve ser recomendada por pelo menos um dos três percursos no mapa

Além das próprias características, o usuário pode escolher também algumas opções de pré-processamento relacionadas aos métodos utilizados. A Tabela 3-2 mostra todos os métodos de processamento mais as opções de pré-processamento que cada método admite, definindo as características utilizadas. O processamento

das características tamanho da sentença, posição e nomes próprios não possui opções de pré-processamento, pois, devido à natural simplicidade, não requer tratamento linguístico mais sofisticado do texto.

Tabela 3-2 - Opções de processamento e opções de pré-processamento do SuPor

Método	Opções de Pré-processamento	Influência
Tamanho da sentença	-	-
Posição	-	-
Frequência das Palavras	Quadrigramas	Realiza o cômputo das medidas de frequência baseando-se nos quadrigramas de uma palavra
	Radicais (*Stemming*)	Realiza o cômputo das medidas de frequência baseando-se nos radicais das palavras[17]
Nomes Próprios	-	-
Cadeias lexicais	TextTiling	Utiliza os tiles fornecidos pelo algoritmo como tópicos do texto
	Parágrafos	Utiliza os parágrafos como tópicos
Importância dos Tópicos	Quadrigramas	Realiza o cômputo das medidas de frequência baseando-se nos quadrigramas de uma palavra
	Radicais (*Stemming*)	Realiza o cômputo das medidas de frequência baseando-se nos radicais das palavras
Mapa de Relacionamentos	Quadrigramas	Realiza o cômputo das medidas de frequência e similaridade entre sentenças baseando-se nos quadrigramas de uma palavra
	Radicais (*Stemming*)	Realiza o cômputo das medidas de frequência e similaridade entre sentenças baseando-se nos quadrigramas de uma palavra

Na implementação do autor do sistema, se uma opção de pré-processamento é escolhida, ela será utilizada por todos os métodos e características selecionados que dependam também dessa opção. Por exemplo, escolhido o pré-processamento por quadrigramas, os métodos Mapa dos Relacionamentos, Importância dos Tópicos e Palavras mais frequentes irão utilizar, obrigatoriamente, esse pré-processamento se forem também selecionados pelo usuário. Dessa forma, o número total de possíveis configurações para a SA no SuPor é de 348, ou seja, o ambiente permite, via customização realizada pelo usuário, gerar 348 sumarizadores automáticos distintos.

[17] Obtidos através de um *stemmer*.

Como citado anteriormente, vários foram os sumarizadores desenvolvidos neste trabalho a partir SuPor original. O primeiro deles foi chamado de “SuPor-2” e é descrito na Seção 5.1.

Principais vantagens: combinação de diferentes abordagens para SA, incluindo a utilização de estatísticas textuais, grafos e aprendizado de Máquina.

Principais desvantagens: a escolha das características e opções de pré-processamento é complicada e exige um usuário especialista de domínio. De forma geral, essa questão é bastante relacionada ao Problema 1 enunciado na introdução do presente trabalho.

3.2.5 SABio

O SABio (Orrú et al. 2006) utiliza uma abordagem conexionista para a SA de textos por meio de um algoritmo para treino da rede neural inspirado biologicamente, chamado GeneRec. Assim como faz o conhecido algoritmo BackPropagation (Haykin 1999), o GeneRec determina os pesos mais adequados para a rede numa etapa de treino.

O sistema também é similar ao NeuralSumm com relação às características empregadas. Sete das oito características do NeuralSumm são empregadas. Uma diferença é que o SABio considera seis níveis de importância para uma sentença, em vez de apenas três como o NeuralSumm: Nenhuma, Pequena, Pequena-Média, Média, Média-Grande, Grande.

No treino do sistema realizado pelos autores, foram utilizados dois terços do corpus TeMário-2003 (Pardo e Rino 2003). A determinação das saídas atribuídas a cada sentença, isto é, do nível de importância dentro das seis opções disponível foi feita de modo empírico. Como relatam os autores, a atribuição foi feita comparando-se a sentença com o extrato ideal por meio de método similar ao utilizado pelo sistema GistSumm.

Para avaliação do sistema, os autores replicaram o experimento de Rino et al. (2004), que aponta o SuPor como o mais bem classificado dentre outros sete sistemas. Nessa avaliação, o SaBio ficou logo abaixo do ClassSumm, segundo mais bem classificado.

Principais vantagens: a utilização de uma abordagem inspirada biologicamente pode em teoria realizar a tarefa de modo como ela é feita manualmente.

Principais desvantagens: a determinação das saídas associadas a cada sentença foi feita utilizando-se um método que pode introduzir ruídos no resultado final.

3.2.6 Sumarizadores Baseados em Medidas de Redes Complexas

Sistemas complexos modelados como grafos são conhecidos como Redes Complexas (RC) e têm influenciado várias áreas atualmente, conforme apontam Albert e Barabási (2002). Textos podem ser representados por meio de grafos de diversas maneiras, tal como fazem os já citados métodos TextRank e Mapa dos Relacionamentos.

Antiqueira (2007) propôs 26 medidas numéricas baseadas em redes complexas para a SA extrativa. Cada medida foi usada individualmente gerando 26 modelos diferentes de SA. O texto é representado exatamente como no método TextRank. Os grupos de medidas utilizadas são descritos brevemente a seguir, acompanhados de ilustrações do que representam extraídas de Antiqueira (2009). Maiores detalhes sobre o cálculo podem ser obtido em Antiqueira (2009):

- ***Degree*** (Costa, Rodrigues et al. 2006). Calcula o número de arestas associadas a um nó. A medida visa sinalizar o quão conectado um nó é em relação a seus vizinhos. Quanto maior o grau de conexão do nó que representa a sentença, mais importante ela é para a SA.

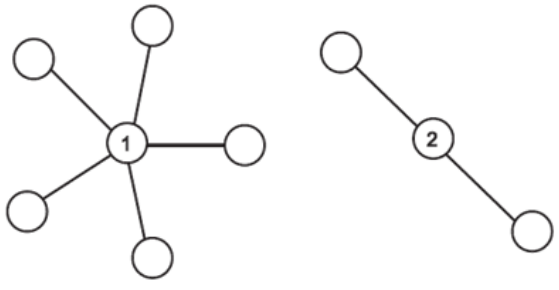


Figura 3-4 – Exemplo medida degree

- ***Clustering Coefficient*** (Watts e Strogatz 1998). Quantifica o quão próximo um nó e seus vizinhos estão de se tornarem um clique, um tipo de agrupamento da Teoria dos Grafos em que cada nó está conectado aos demais.

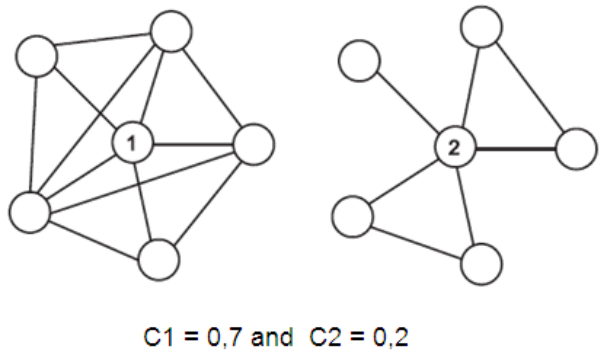


Figura 3-5 - Exemplo medida Clustering Coefficient

- ***Minimal Paths*** (Costa, Rodrigues et al. 2006). Para cada nó, todos os caminhos mínimos são calculados para os demais. A média da distância desses caminhos para cada outro nó é então calculada. Essa média sinaliza diretamente a relevância da sentença para a SA. Isso se relaciona ao fato de que as sentenças mais próximas às demais podem conter a ideia principal do texto.

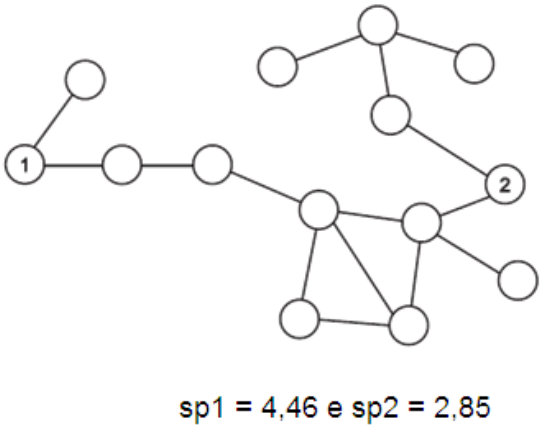


Figura 3-6 - Exemplo medida Minimal Paths

- ***Locality Index*** (Costa, Kaiser et al. 2006). Também considera o agrupamento de nós, mas leva em conta também as conexões fora do agrupamento.

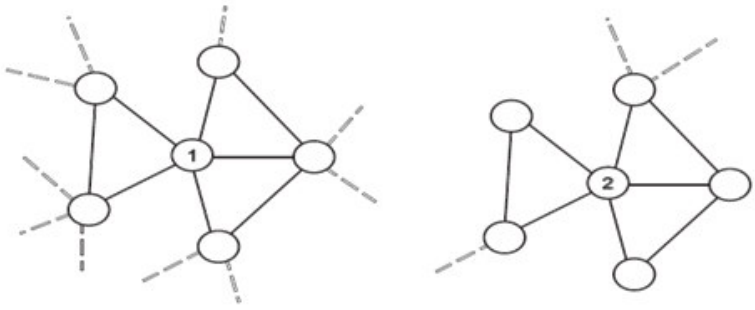


Figura 3-7 - Exemplo medida Locality Índex

- ***Matching Index*** (Costa, Rodrigues et al. 2006). Ao medir a força da conexão entre dois nós, a medida visa selecionar sentenças que cobrem diferentes grupos de informação ou tópicos do texto.

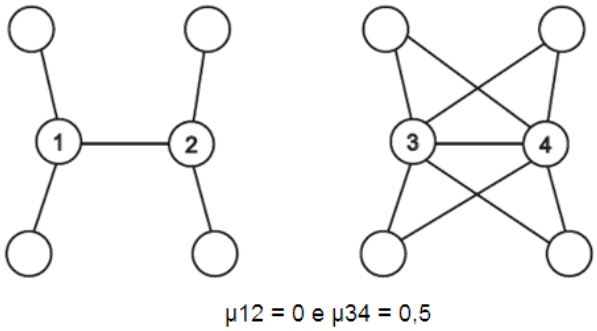


Figura 3-8 - Exemplo medida Matching Índex

- ***Dilation*** (Costa e da Rocha 2006). Expressa a importância dos nós utilizando relações hierárquicas dentro do grafo. Anéis de nós são traçados ao redor do nó focalizado: anéis de ordem 1 capturam os nós ligados por até uma aresta; anéis de ordem 2, capturam os que estão ligados por até duas arestas, etc. Os anéis visam capturar a conectividade entre nós na vizinhança que estão mais longe do nó focalizado. O grau hierárquico do nível *h* é definido então como o número de nós entre esse nível e o nível do próximo anel. Atribuí-se

maior poder de sumarização às sentenças que possuem maior grau hierárquico.

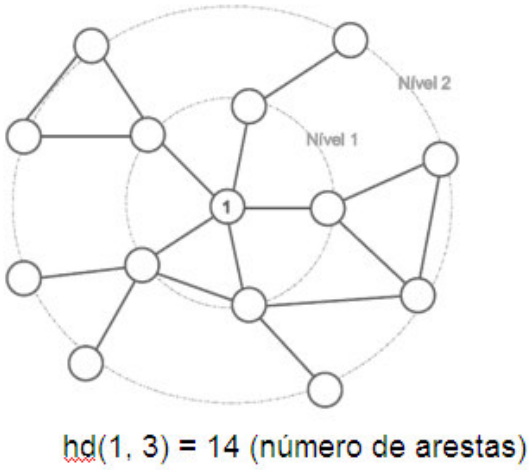


Figura 3-9 - Exemplo medida Dilation

- ***Hubs*** (Antiqueira et al. 2009). Também considera a operação de *dilation*, mas dá preferência aos nós com maior número de conexões no grafo, os chamados *hubs*.

- ***K-cores*** (Batagelj e Zaversnik 1999). Um *k-core* é um subgrafo cujos nós possuem o grau mínimo de *k*. Para a SA, o subgrafo mais relevante é o que possui o maior *k*.

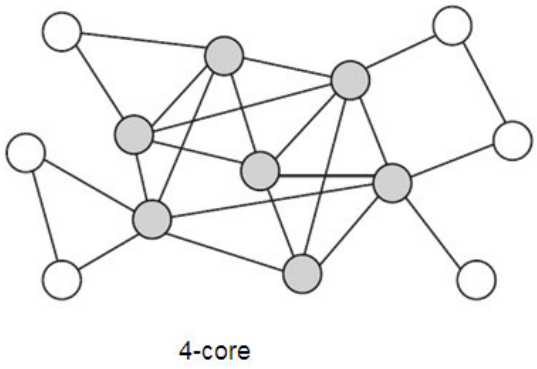


Figura 3-10 - Exemplo medida K-cores com K = 4

- ***W-cuts*** (Antiqueira et al. 2009). Objetiva encontrar grupos coesos de sentenças considerando a interconexão de nós com alto valor para a medida *degree*.

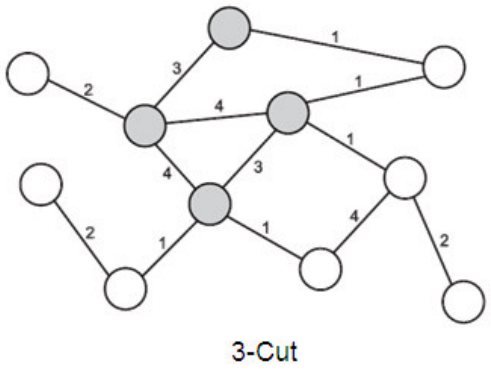


Figura 3-11 - Exemplo medida W-cuts com W = 3

- ***Communities*** (Clauset et al. 2004). Uma comunidade é um grupo de nós que são muito conectados entre si. Nós em diferentes comunidades não são significantemente conectados. Assim, as comunidades sinalizam a densidade de conexões no grafo. No caso da SA, elas podem indicar a estrutura de tópicos do texto. Dentro da mesma comunidade, sentenças com maior *degree* possuem preferência.

De acordo com as definições acima, algumas dessas medidas relacionam-se a alguns métodos já citados. Por exemplo, ao considerar a medida *degree* e *minimal paths* existe uma relação com o Caminho 1 (denso) do método Mapa de Relacionamentos. A medida *locality index* ao Caminho 2 (profundo). As medidas *matching índex* e *dilation* parecem relacionarem-se ao Caminho 3 (segmentado), enquanto hubs parece relacionar-se tanto ao Caminho 2 quanto ao 3. Entretanto, estudos adicionais são necessários para confirmar essas proposições.

Uma correspondência mais clara é vista nos três últimos grupos de medidas: *K-cores* não necessariamente obtém coesão na medida em que se relaciona a caminhos densos. *W-cuts* busca contornar o problema de forma semelhante ao Caminho 2 (profundo) do Mapa de Relacionamentos. *Communities* abrange diferentes tópicos do texto, relacionando-se ao Caminho 3 (segmentado) do Mapa de Relacionamentos e ao Método de Importância dos Tópicos.

Em um trabalho mais recente, Antiqueira et al. (2009) propuseram a combinação dessas medidas por meio de uma estratégia de votação, em que cada medida contribui para a geração do ranking das sentenças. Essa versão do sumarizador foi chamada de "CN-Voting". Pelos resultados reportados pelos autores, apresentou desempenho inferior ao SuPor-2 — proposto neste trabalho[18] — e um pouco acima do SuPor original. Entretanto, os autores não reportam significância estatística nas diferenças.

Principais vantagens: a utilização de diversas medidas do grafo pode ser útil para capturar diferentes características das sentenças ou mesmo capturar aspectos ligados ao nível do discurso.

Principais desvantagens: no trabalho de Antiqueira (2007) não é proposta uma metodologia para a combinação dessas características e, assim, a definição de um único modelo de SA; a estratégia de combinação apresentada no CN-Voting (Antiqueira et al. 2007) pode ser muito simplista ao deixar de ponderar a relevância das características ou medidas.

3.3 Trabalhos Recentes

Apresentam-se a seguir alguns trabalhos recentes na área e que se relacionam ao nosso.

3.3.1 Utilização de Aprendizado de Máquina

3.3.1.1 Sistema IIIT

Pingali et al. (2007) atingiram a melhor medida pelo método da Pirâmide[19] na DUC de 2007, com seu sistema IIIT. Eles focaram na chamada tarefa *Update*

[18] Ver Seção 5.1.
[19] O método da Pirâmide foi descrito na Seção 2.4.3.

Summarization, que consiste em construir um sumário sob a assunção de que o leitor já leu um conjunto de textos sobre o tópico.

Os autores propuseram uma medida que leva em conta dois fatores: a importância individual da sentença e sua relação com o tópico. Para o fator de importância individual, foi utilizado um modelo Naïve-Bayes com o propósito de relacionar os conjuntos de palavras do texto candidato com as palavras dos textos já lidos.

Principais vantagens: a associação da relevância por meio de uma componente individual, não relacionada ao tópico, e a outra componente vinculada ao tópico; embora não detalhado pelos autores, é realizado um procedimento para diminuir a redundância no texto final produzido.

Principais desvantagens: abordagem é mais adequada para o tipo de tarefa específico da DUC.

3.3.1.2 Utilização de Support Vector Machines (SVM)

Li et al. (2007) propuseram um sistema que utiliza o método de aprendizado de máquina conhecido como SVM - Support Vector Machines (Vapnik 1998) para a tarefa *Update Summarization*. Ao todo, seis características foram utilizadas, incluindo uma característica baseada na similaridade medida pela WordNet entre a sentença e o tópico. O sistema foi o quinto mais bem colocado e foi avaliado pelas métricas ROUGE-2 e ROUGE-SU4 (Lin e Hovy 2003).

Principais vantagens: SVMs são considerados como um dos métodos estado-da-arte da área de Aprendizado de Máquina (Vapnik 1995; Witten e Frank 2005).

Principais desvantagens: os autores relatam que poderia ter sido utilizado um processo melhor de geração dos dados de treino e que também características mais elaboradas deveriam ser utilizadas.

3.3.1.3 Utilização de Support Vector Machines (SVM) e Treino Dirigido pela ROUGE

Galanis e Malakasiotis (2008) propuseram um sumarizador baseado no modelo de regressão por Support Vector Machines, de forma semelhante ao trabalho de Li et al. (2007). A tarefa foi a mesma, *Update Summarization*, na TAC de 2008. A diferença principal é que utilizaram um processo diferente de geração dos dados de treino, utilizando como variável de saída a média entre as medidas ROUGE-2 e ROUGE-SU4. Ou seja, o modelo proposto busca determinar sentenças que tem uma boa média entre essas medidas.

Pela avaliação realizada, o sistema foi o quinto colocado quando julgado pelas medidas ROUGE-2 e ROUGE-SU4, atingindo os índices de 0,113 e 0,165 respectivamente. Esses índices ficam acima do sumarizador de Li et al.

Principais vantagens: mesmas do modelo de Li et al. (2007) e a utilização de métricas específicas da área no processo de treino.

Principais desvantagens: os autores relatam que não atingiram bom desempenho em avaliações humanas por não empregaram nenhum mecanismo de reescrita e tratamento de redundância.

3.3.1.4 Utilização de Support Vector Machines (SVM) e Seleção Automática de Características

Schilder et al. (2008) utilizaram 8 características e também o método de regressão baseado em Support Vector Machines, de forma semelhante a Li et al. (2007). Porém, utilizaram um procedimento de seleção automática de características, que leva em conta as correlações entre cada característica e a saída desejada Efron et al. (2004). Porém, esse método de seleção não leva em conta a redundância entre as características.

O sistema foi o quarto colocado na avaliação pelo Método da Pirâmide, em relação a outros 35 sistemas.

Principais vantagens: mesmas do modelo de Li et al. (2007) e a utilização de seleção automática de características.

Principais desvantagens: o método de seleção de características poderia levar em conta a redundância entre elas.

3.3.1.5 Utilização de grandes conjuntos de características

Wong et al. (2008) propuseram um sumarizador que utiliza 15 características numa abordagem envolvendo aprendizado de máquina.[20] Eles argumentam que considerar a importância das sentenças por apenas um ponto de vista, ou característica, não é efetivo. Assim, sugerem a utilização de muitas características combinadas por meio de aprendizado de máquina. De forma semelhante ao sistema SuPor (Módolo 2003), eles utilizaram características superficiais e métodos completos mapeados como características. Por exemplo, o método TextRank (Mihalcea 2005) foi utilizado como característica.

Os autores utilizaram para comparação os algoritmos SVM e Naïve-Bayes. Também utilizaram um processo conhecido como *Co-training* que busca suprir a falta de dados rotulados (exemplos manuais) combinando dados rotulados e não rotulados para treinar os dois classificados simultaneamente.

Através das métricas ROUGE-1, ROUGE-2 e ROUGE-L os autores compararam suas abordagens entre si (SVM isolado, Naïve-Bayes isolado e *co-training*) com corpora das DUCs de 2001 e 2007. Eles verificaram que o método de *co-training* traz resultados em geral melhores quando há poucos dados rotulados disponíveis.

Principais vantagens: utilização de um grande conjunto de características; utilização de método para contornar o problema da utilização de corpus de treino pouco expressivo.

Principais desvantagens: devido à utilização de muitas características, o sistema proposto provavelmente é bastante sensível a quais conjuntos de características são utilizadas, assim como o sistema SuPor.

[20] Pelo nosso conhecimento trata-se do trabalho que havia utilizado o maior número de características até o momento.

3.3.2 Sumarização utilizando lógica nebulosa

Kiani-B e Akbarzadeh-T (2006) propuseram um sumarizador que utiliza uma abordagem híbrida genético-nebulosa para determinar as sentenças mais relevantes para inclusão no texto. Através de um processo de treino não-supervisionado o algoritmo proposto maximiza uma série de funções de *fitness*, entre elas a presença de palavras no extrato que também estão no título do texto original.

São utilizadas seis características como variáveis nebulosas, a saber:

- Número de palavras da sentença que estão no título;
- Se a sentença é a primeira a figurar no parágrafo;
- Se a sentença é a última a figurar no parágrafo;
- O número de palavras da sentença (tamanho);
- O número de radicais da sentença, obtidos removendo-se *stopwords* e realizando-se o processo de *stemming*;
- O número de palavras importantes tais como “most”, “very”, etc.

Para avaliação dos resultados, os autores utilizaram um corpus de textos jornalísticos de tópicos variados e de tamanho não especificado. As medidas focadas foram Precisão, Cobertura e *F-Measure* entre as sentenças extraídas e as que constavam no extrato manual. Os autores relatam desempenho superior ao sumarizador comercial Copernic e à ferramenta *AutoResumo* do Microsoft Word. Em *F-Measure*, o sistema obteve 0,752 contra 0,62 e 0,26 do sumarizador Copernic e da ferramenta *AutoResumo*, respectivamente.

Principais vantagens: a utilização de lógica nebulosa pode ser interessante à tarefa de SA já que diferenciar uma sentença boa de uma ruim pode ser considerado um pouco nebuloso até mesmo na tarefa manual.

Principais desvantagens: avaliação não focou em medidas mais atuais para sumarização automática; características utilizadas são apenas superficiais.

3.4 Síntese e Comparação das Abordagens para SA

Três grandes abordagens podem ser identificadas como principais na construção dos sistemas de SA descritos nas seções anteriores:

I. **O Uso de Estatísticas Textuais**

Compreendem medidas extraídas diretamente a partir da estrutura superficial do texto, como a distribuição das frequências de palavras (vide o Modelo de Luhn, Seção 3.1.1), posição e tamanho das sentenças (Seção 3.1.2). A grande vantagem da utilização dessas medidas está justamente no fácil cômputo e também no fato de demandarem pouco ou nenhum processamento linguístico. Geralmente, demandam apenas pré-processamentos simples como a remoção de *stopwords* e a uniformização dos termos usando *stemming* ou quadrigramas. Em geral, as estatísticas textuais são independentes de língua.

Por outro lado, podem ser dependentes de gênero. Por exemplo, a característica de posição da sentença pode ser mais importante num gênero jornalístico que num gênero científico. Outro ponto negativo é que os modelos baseados somente nesses tipos de medidas desconsideram importantes relações linguísticas entre os termos e sentenças, como relações coesivas e de coerência.

II. **O Uso de Medidas de Grafos**

A representação em grafo de informações textuais permite recuperar mais claramente as relações entre elas, particularmente as relações que podem indicar segmentos mais relevantes para a SA. Diferente das medidas estatísticas simples, a representação em grafo permite mais facilmente a recuperação dos elos coesivos e de coerência no texto. Entretanto, as próprias medidas extraídas do grafo podem favorecer certas características em detrimento de outras. Por exemplo, na Seção 3.1.4 foi descrito o método de *Mapa Relacionamentos*, que é uma abordagem baseada em grafos. Por esse método, o caminho *denso* pode levar a geração de sumários informativos, mas com pouca coerência. O caminho *profundo* favorece a coerência em detrimento da abrangência. Já o caminho *segmentado*, busca tanto a abrangência de informações quanto a coerência. As mesmas questões surgem com a utilização de medidas de redes complexas (Seção 3.2.6).

III. **O Uso de Aprendizado de Máquina**

A utilização de aprendizado de máquina é recorrente em vários modelos apresentados (e.g., Kupiec et al., 1995; Módolo, 2003; Kiani-B e Akbarzadeh-T, 2006; Wong et al., 2008). Ela pode ser útil na combinação e ponderação de características para SA. O trabalho que é necessário nessa abordagem é geralmente o treino dos modelos. O fato de alguns métodos de aprendizado de máquina possuírem premissas fortes e não atendidas e o fato de muitas vezes os modelos de aprendizado serem prejudicados por características irrelevantes ou redundantes de SA são problemas que devem ser considerados na utilização dessa abordagem.

O Capítulo 4 deste trabalho apresenta uma descrição mais elaborada dos modelos de aprendizado de máquina mais atuais e a questões que envolvem seu uso para SA. O detalhamento nesse capítulo deve-se a diversidade dos métodos explorados neste trabalho.

Como exposto, cada uma das abordagens tem aspectos positivos e negativos com relação a sua utilização. Neste trabalho, buscou-se combiná-las para construção dos modelos de SA que serão descritos no Capítulo 5.

Capítulo 4

MÉTODOS BASEADOS EM APRENDIZADO DE MÁQUINA

Este capítulo apresenta os fundamentos teóricos e uma descrição dos principais métodos de aprendizado de máquina explorados neste trabalho.

4.1 Introdução à Mineração de Dados e à Classificação

A Mineração de Dados se concentra essencialmente na busca de padrões potencialmente úteis. A Classificação é uma das tarefas mais comuns na Mineração de Dados. Seu objetivo é determinar o valor de uma variável categórica (discreta, portanto) para um exemplar de um conjunto de dados. Essa variável discreta que se deseja determinar é chamada de classe ou ainda rótulo. O número de categorias da variável é arbitrário. Por exemplo, uma classe "Risco", que indica o risco de concessão de empréstimo a uma pessoa, poderia assumir uma das seguintes categorias: baixo, médio ou alto.

De modo geral, os algoritmos para classificação trabalham da seguinte forma: é feita a análise de um subconjunto de dados já rotulado (chamado nesse caso de conjunto de treino) presente na base de dados e se deseja determinar a classe para registros não-rotulados (sem a informação de qual classe pertence). Os registros são compostos de vários campos (chamados de características) que devem ser os mesmos para os registros já rotulados e para os que ainda não possuem rótulos.

Como exemplo, suponha um sistema inteligente que busca determinar uma categoria para o risco dos tomadores de empréstimos de uma empresa financeira. Essa empresa já contém em sua base de dados um grande conjunto de registros de

clientes e a informação sobre a categoria de risco em que eles se enquadram. Um exemplo desse conjunto de registros é mostrado na Tabela 4-1. Considere que a empresa obteve esse conjunto a partir de dados históricos de seus clientes (características Idade, Rendimento Anual, Sexo e Ocupação) e determinou a classe "Risco" a partir de estudos realizados com base em análises de seus movimentos financeiros por vários anos (em geral, a análise é realizada pelos analistas de risco). Suponha, ainda, que o problema da empresa seja classificar o risco de um novo cliente. A tarefa, nesse caso, é justamente de classificação.

Tabela 4-1 – Conjunto de Treino do Exemplo de Risco de Crédito

Idade	Rendimento Anual	Sexo	Ocupação	Risco
21	50.000	F	Engenheiro	Médio
45	20.000	M	Desempregado	Alto
60	8.000	M	Aposentado	Baixo
35	49.000	F	Engenheiro	Baixo

Para resolver o problema, o primeiro passo é configurar um classificador de interesse, o qual resulta do treino para obtenção de um modelo. O algoritmo aplicado irá modelar as relações entre as características de cada cliente e sua classe. Feito isso, será possível classificar o risco do novo cliente. Nesse caso, a entrada para o classificador é constituída das características do novo cliente (Idade, Rendimento Anual, Sexo e Ocupação) e a resposta obtida será uma categoria de risco: baixa, média ou alta, no caso.

A tarefa de classificação pode ser usada para gerar modelos de classificação de sentenças, visando à escolha de unidades textuais para compor os extratos de interesse em nosso trabalho, tema da próxima seção.

4.1.1 Classificação para Sumarização Automática

Conforme apresentado na Seção 3.1.3, a tarefa de julgar se uma sentença é importante ou não pode ser tratada por meio de métodos de classificação. Essa abordagem, como já citado anteriormente, foi inicialmente proposta por Kupiec et al. (1995).

Remetendo ao *Problema 2,* sobre *Como combinar e ponderar as características escolhidas?* (veja introdução deste trabalho)*,* busca-se, a partir da análise das características de uma sentença, determinar se ela é relevante ou não para a formação do extrato.

Formalmente, seja $X = C_1 \times,...,\times C_n$ um espaço vetorial sobre o conjunto de características $C_1,...,C_n$. Um classificador é um mapeamento $\hat{s} : X \rightarrow U$, onde *U* é um conjunto de rótulos ou classes de saída.

Frequentemente, para SA modela-se o problema como uma tarefa de classificação binária, indicando se a sentença deve estar presente (*True*) ou não (*False*) no extrato. Há, portanto, duas classes que uma sentença pode assumir:

U = {True, False}

Entretanto, diferentemente da maioria dos problemas tradicionais abordados pela Classificação, é de interesse na SA determinar um escore que indique o quão relevante a sentença é. Em outras palavras, a classe de interesse é sempre a "True" e um ranking das sentenças com base nessa essa classe é desejado. Formalmente, o classificador deve fornecer o mapeamento $\hat{r} : X \rightarrow \Re$, atribuindo um escore de relevância $\hat{r}(x)$ para cada sentença representada pelo conjunto de características *X,* definido acima.

A maioria dos métodos de classificação permite a extração de um escore que indique a aderência a uma classe particular. Os métodos apresentados a seguir são os métodos que seguem esse modelo, de interesse para este trabalho.

4.1.2 Métodos Bayesianos

O classificador Naïve-Bayes foi o primeiro a ser utilizado para a SA, conforme a Seção 3.1.3. Para determinação da relevância de uma sentença, é utilizada a probabilidade dessa sentença pertencer a classe *True*, ou seja, a probabilidade de ela pertencer ao extrato.

O classificador é denominado ingênuo (*naïve*) por assumir que as características são condicionalmente independentes. Isto é, assume que a informação de uma característica não é informativa sobre nenhuma outra. Entretanto, há estudos (e.g., Zhang e Su 2004) que indicam bom desempenho do

classificador Naïve-Bayes para geração de rankings, característica de interesse para SA.

A utilização tradicional do classificador Naïve-Bayes para características numéricas assume que as características têm distribuição normal. Assim, a distribuição é geralmente representada por uma média μ e por um desvio-padrão σ.

A variação introduzida no método Flexible-Bayes (John e Langley 1995) é a utilização da técnica conhecida como Kernel Density Estimation que aproxima a real função de distribuição de probabilidades por uma soma de gaussianas centralizadas ao longo de cada valor da característica. Segundo os autores, os resultados obtidos com Flexible-Bayes são no mínimo melhores do que aqueles obtidos a partir de uma simples distribuição gaussiana. A figura seguinte mostra o princípio do método:

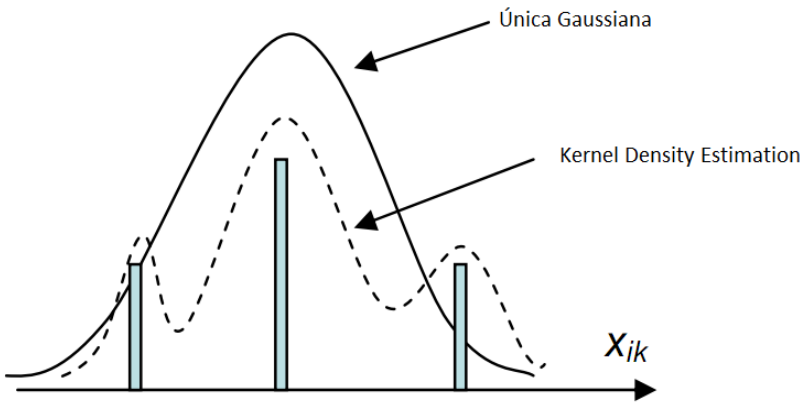


Figura 4-1 – Princípio do Flexible-Bayes para tratamento de características numéricas

Um exemplo completo de aplicação do método Bayesiano é apresentado no Apêndice A.

Principais vantagens: a associação da relevância da sentença a uma probabilidade é algo natural; apesar da simplicidade, os métodos Bayesianos são bastante usados e possuem bom desempenho na área do Processamento de Línguas Naturais.

Principais desvantagens: a premissa de independência condicional das características.

4.1.3 C4.5

O C4.5 (Quinlan 1993) é um algoritmo bastante popular para a criação de árvores de decisão. A árvore de decisão chega a sua decisão pela execução de uma sequência de testes. Cada nó interno da árvore corresponde a um teste do valor de uma das características, e os ramos deste nó são identificados com os possíveis valores do teste. Cada nó folha da árvore determina a classe se a folha for atingida.

A Figura 4-2 seguinte mostra um exemplo de árvore de decisão para um modelo de SA com duas características. A primeira característica considerada é a de posição da sentença no texto, que pode estar situada no início, no meio ou no fim. A segunda é a de frequência média das palavras da sentença no texto, que foi distinguida em três categorias: alta, média ou baixa.

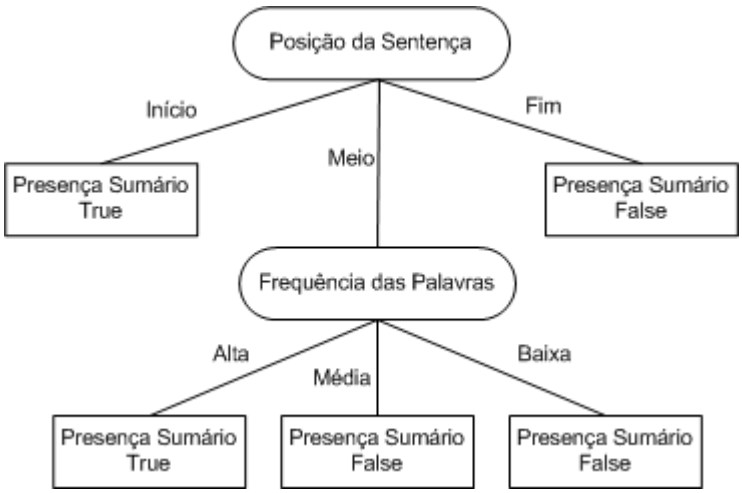


Figura 4-2 – Exemplo de árvore de decisão para um modelo de SA com duas características

Para classificar uma sentença, basta começar pela raiz (no caso do exemplo, posição da sentença), seguindo cada nó de decisão de acordo com a característica da sentença até que uma folha seja alcançada. Quando uma folha é alcançada, é atribuída ao exemplo a classe com maior frequência nessa folha.

Já para construir o ranking de sentenças por relevância, deve-se calcular a frequência relativa da classe *True* (presente no extrato) dentro das folhas da árvore, considerando todos os exemplos do conjunto de treino que são classificados pela respectiva folha. Um problema pode ocorrer se não existir nenhuma sentença da classe *True* que for classificada pela folha. Nesse caso, a folha torna-se inútil.

Alguns estudos (e.g., Zadrozny e Elkan 2001) apontam também que o C4.5 pode não produzir bons rankings através desse procedimento, devido a:

- Alto *bias*: Árvores de decisão tendem a deixar as folhas homogêneas. Assim, as frequências variam geralmente nos extremos 0 ou 1.
- Alta variância: quando o número de instâncias da base de treino associadas a uma folha for pequeno, as frequências observadas não são estatisticamente significantes.

Para manipulação de características numéricas, o método de Discretização Supervisionada (Fayyad e Irani 1993) é utilizado. Esse método visa dividir em intervalos adequados os valores das características, tornando-os categóricos.

Principais vantagens: a representação dos conhecimentos por meio de árvores é intuitiva.

Principais desvantagens: a obtenção de um escore para determinar a relevância da sentença é difícil e pode não ser precisa.

4.1.4 Support Vector Machines

SVM ou Support Vector Machines (Vapnik 1998) determinam uma fronteira de decisão entre duas classes mapeando os exemplos de treino (sentenças rotuladas) num espaço dimensional maior e determinando o hiperplano ótimo de separação nesse espaço.

Para o cálculo da relevância de uma sentença, assume-se que ela é igual à distância Euclidiana entre o hiperplano e o vetor de características da sentença. Quanto mais próxima estiver a sentença do lado das sentenças da classe *True* (presente no extrato), mais relevante ela será considerada.

Esse tipo de modelo tem sido utilizado em alguns sistemas com bons resultados nas últimas DUCs e TACs, conforme exposto na Seção 3.3.1.

Principais vantagens: é considerando um método robusto. Implementações mais recentes consideram aspectos de não-linearidade.

Principais desvantagens: a obtenção de um escore para determinar a relevância da sentença pode não ser calibrada em função de sua forma de cálculo.

4.1.5 Regressão Logística

Trata-se de uma variação da regressão tradicional (Witten e Frank 2005). Ela é usada quando a variável dependente da regressão é binária. No caso, *True* para presente no extrato ou *False* para não presente no extrato). O modelo do classificador é dado na equação seguinte: p é a probabilidade da classe *True*, X_1, …, X_n são os valores das características consideradas e β_0, β_1,…, β_n são os coeficientes de regressão obtidos por meio de treino. O resultado logarítmico é então transformado numa probabilidade por meio de uma função logística. As probabilidades obtidas indicam a relevância das sentenças.

$$\log\left(\frac{p}{1-p}\right) = \beta_o + \beta_1 X_1 + ... + \beta_n X_n \qquad \textbf{[10]}$$

Principais vantagens: é reportado na literatura que o método gera boas estimativas de probabilidade, sendo adequado a tarefa de ranking, assim como os métodos Bayesianos.

Principais desvantagens: assume relação linear entre características e a relevância da sentença.

4.1.6 Redes Neurais Artificiais

As Redes Neurais Artificiais ou RNA (Haykin 1999) são sistemas computacionais estruturados para realizar cômputos por meio de ligações entre suas unidades elementares de processamentos, chamados de neurônios ou simplesmente nós. A inspiração original para essa técnica advém do exame das estruturas do cérebro, em particular do exame de neurônios.

A Figura 4-3 a seguir ilustra uma possível rede neural de exemplo para SA, para um modelo de SA com duas características, como já foi feito no exemplo da

Seção 4.1.3. A rede recebe como entrada o valor de duas características numéricas da sentença — a posição relativa no texto e a frequência das palavras das sentença — e produz duas saídas binárias. O valor da primeira saída será 1 se a sentença for relevante ou será 0 se for irrelevante. Para a segunda saída, vale o contrário. Os nós interdiários da rede são responsáveis por realizar processamentos intermediários a partir dos valores da camada de entrada e fornecer a entrada para os nós de saída, que irão combinar todos os valores e produzir a saída.

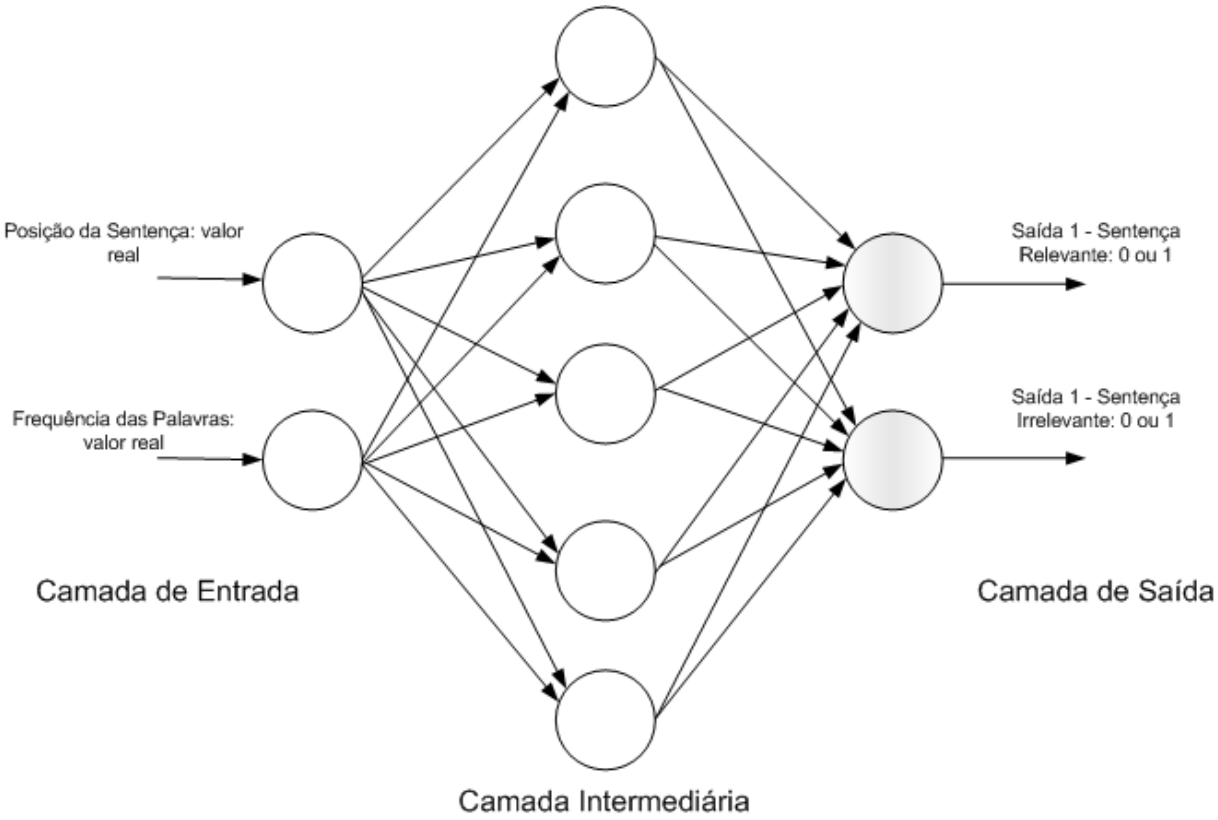


Figura 4-3 – Representação de uma possível rede neural para duas características

Nos neurônios da camada de saída, o resultado da combinação dos valores das camadas intermediárias é dado como entrada para uma função que irá determinar a saída binária da rede, chamada de função de ativação. Normalmente, é utilizada uma função sigmoide, ilustrada na Figura 4-4.

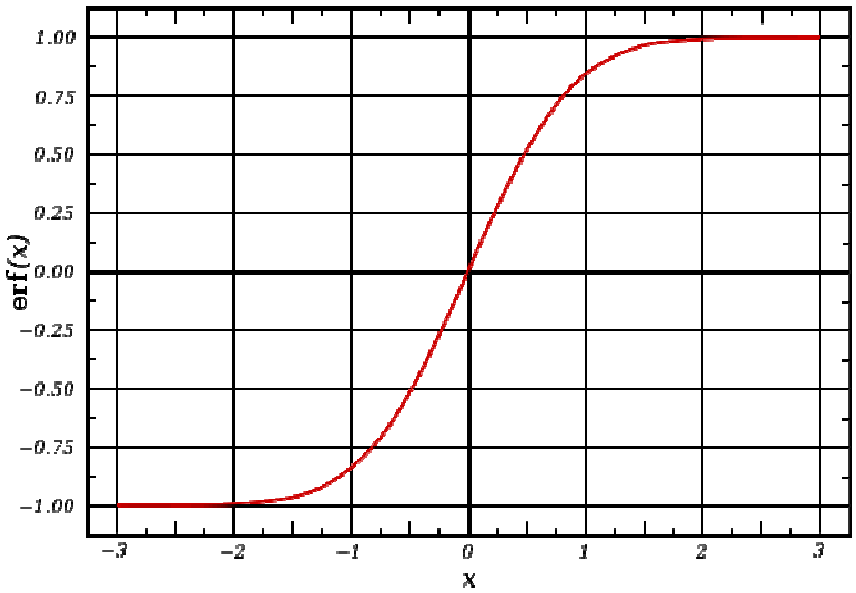


Figura 4-4 – Função sigmóide

Em vez de se utilizar apenas dois neurônios de saída, níveis intermediários de relevância das sentenças poderiam ser modelados com a utilização de mais neurônios na camada de saída, como é o caso dos já citados[21] NeuralSumm (Pardo 2003), que modela três níveis de relevância, e do SABIo (Orrú et al. 2006), que modela seis níveis de importância. Em ambos os casos, a saída da rede é discreta e acaba restringindo os níveis de relevância que podem ser atribuídos às sentenças. Além disso, para que a rede possa modelar uma saída mais abrangente que a binária (relevante ou não), os dados de treino precisam ser ajustados ou complementados manualmente. Usualmente, os dados de treino que se dispõem possuem uma saída também binária, indicando ou não a presença daquela sentença em um extrato de referência manual.

Principais vantagens: Pode tratar problemas não-lineares.

Principais desvantagens: Como a saída da rede é determinada pelas saídas dadas pela função de ativação (usualmente binária), pode ser complicado distinguir adequadamente a relevância das sentenças e construir um ranking por relevância.

[21] Vide Seção 3.2

4.2 Seleção Automática de Características

É conhecido na área de aprendizado de máquina e Mineração de Dados que os algoritmos de classificação podem ter seu desempenho prejudicado devido à presença de características irrelevantes (e.g.,Hall 2000; Witten e Frank 2005). O mesmo é observado na utilização de algoritmos de aprendizado de máquina para a SA. No Capítulo 1 foi citado, por exemplo, que a utilização de apenas 3 características (em vez de 5) trouxe os melhores resultados no trabalho de Kupiec et al. (1995). De fato, o *Problema 1* lá apresentado, sobre *Quais características utilizar?,* pode ser abordado de forma automática por meio de técnicas de seleção automática de características[22].

A seleção automática de características permite identificar subconjuntos promissores de características de modo a evitar que o modelo de aprendizado de máquina utilize todas as características disponíveis. No caso da Sumarização Automática Extrativa baseada em aprendizado de máquina, busca-se com isso determinar um subconjunto de características que maximize a capacidade do sistema de classificar corretamente os casos. Em outras palavras, de maximizar a determinação de sentenças relevantes para composição do extrato.

Além de esse processo poder conduzir a melhores extratos, do ponto de vista computacional é mais eficiente, pois o modelo será baseado num número menor de características.

Há duas abordagens principais para se tratar o problema de seleção de características: *Filter* e *Wrapper*. A abordagem *Filter* opera independente do algoritmo de classificação considerado e filtra as características antes do processo de classificação. Métodos baseados na abordagem *Filter* costumam selecionar as características com base em medidas estatísticas ou em medidas baseadas na teoria da informação (Shannon 1948). O resultado dos métodos pode ser tanto um ranking das características por ordem de relevância ou um subconjunto recomendado de características. Existe ainda a técnica de análise dos componentes principais (e.g., Haykin 1999) que também se considera que segue a abordagem *Filter*, por operar independente do algoritmo de classificação. Essa técnica busca

[22] A tarefa também é conhecida como "Seleção de Atributos" ou "Attribute Selection" e "Feature Selection" em inglês.

reduzir o número de características a um pequeno número de combinações lineares das características originais. Já a abordagem *Wrapper*, opera em conjunto com o algoritmo de classificação, por meio de *cross-validation*[23]. Ou seja, avalia-se diretamente o desempenho do classificador para cada subconjunto de características considerado.

4.2.1 Métodos Wrapper para Seleção Automática de Características em SA

A abordagem *Wrapper* opera da seguinte forma:

a) Selecionam-se potenciais subconjuntos de características por meio de um processo de busca heurística.
b) Para cada subconjunto, deve-se proceder sua avaliação da seguinte forma:
    i. Construir (treinar) o modelo de aprendizado de máquina com as características consideradas;
    ii. Avaliar o desempenho do modelo, usualmente por meio de critérios como taxa de acerto ou *F-Measure.*

Por conta desse processo, a crítica em geral feita aos métodos *Wrapper* é com relação ao seu alto custo computacional. Além desse custo, existem algumas complicações adicionais quando se considera aplicar essa abordagem na Sumarização Automática extrativa baseada em aprendizado de máquina:

a) As avaliações do modelo não são feitas por medidas de taxas de acerto ou *F-Measure* do classificador, mas sim por medidas que indicam, por exemplo, a informatividade dos extratos produzidos. As implementações existentes, por exemplo do WEKA (Witten e Frank 2005), não têm condições de tratar essa particularidade;
b) Por conta dos subconjuntos de características serem avaliados com base nos extratos produzidos, a avaliação desses subconjuntos fica condicionada à taxa de compressão que foi adotada. Caso uma outra taxa fosse considerada,

[23] A Seção 6.1.5 apresenta um resumo do funcionamento dessa técnica e seus objetivos.

não necessariamente o conjunto de características selecionado seria o mesmo.

Principais vantagens: A seleção das características é enviesada de modo a produzir os melhores resultados com o classificador escolhido.

Principais desvantagens: Alto custo computacional e uso de medidas não específicas de SA para avaliação da performance do classificador.

4.2.2 Métodos Filter - Medidas Baseadas na Teoria da Informação

Na década de 1940, Claude Shannon estava interessado em maximizar a quantidade de informação que poderia ser transmitida por um canal de comunicação, particularmente um canal imperfeito, com ruídos. Isso o levou a definir matematicamente o conceito de informação e como quantificá-la. A solução encontrada foi definir informação como uma redução da incerteza e, dessa forma, quantificar a informação quantificando a incerteza (Shannon 1948).

Suponha um dispositivo que transmita três símbolos — A, B ou C — através de um canal para um receptor. Enquanto o receptor espera pelo símbolo, há a incerteza sobre o símbolo que será recebido. Assim que o símbolo enviado chega ao receptor e é reconhecido, a incerteza diminui e recebe-se alguma quantidade de informação. A questão é, então, como medir a incerteza para se determinar a essa quantidade de informação recebida. A maneira mais simples seria considerar que se tem uma “incerteza de três símbolos”, para o exemplo.

Suponha, agora, um outro dispositivo que envia simultaneamente ao dispositivo anterior os símbolos 1 e 2 ao mesmo receptor. As possibilidades na recepção serão A1, A2, B1, B2, C1 ou C2. Ou seja, uma “incerteza de seis símbolos”. Essa não é maneira usual de se pensar, quer em relação ao conteúdo informativo, quer em relação à incerteza, pois esses conceitos são, intuitivamente, aditivos, não multiplicativos. Usando-se a função logarítmica essa propriedade aditiva pode ser conseguida. Considere que a incerteza para o primeiro dispositivo seja $log(3)$ e para o segundo $log(2)$, então a incerteza para os dois dispositivos será igual a $log(6) = log(3) + log(2)$. A base do logaritmo determina a unidade. Se a base dois for usada, o resultado será em bits. Assim, no envio de dois símbolos, tem-se uma incerteza de 1 bit.

A fórmula apresentada no parágrafo anterior para incerteza é $log_2(N)$, em que N é o número de símbolos. Ou, equivalentemente, $-log_2(P)$, pois $1/N$ é a probabilidade P de cada símbolo ser recebido. Ela não serve para casos em que as probabilidades de cada símbolo são diferentes. Suponha que haja quatro símbolos possíveis, mas dois deles nunca sejam recebidos. A incerteza será de apenas 1 bit, não de 2 bit.

Shannon estendeu essas ideias para símbolos com diferentes probabilidades e definiu a grandeza conhecida como entropia (H), que mede a incerteza de uma mensagem codificada num alfabeto com distribuição de probabilidades $P(X)$, como segue:

$$H(X) = -\sum_i P(x_i).\log_2(P(x_i)) \qquad [11]$$

em que:

x_i é um dos símbolos do alfabeto.

Note que a entropia é uma medida positiva, pois o sinal negativo é anulado pelo logaritmo negativo, já que as probabilidades são sempre menores ou igual a 1.

Considere, como exemplo, o lançamento de uma moeda. As faces observadas correspondem aos símbolos que serão recebidos. Se a probabilidade de ocorrer cara for a mesma de ocorrer coroa, a entropia será:

$$H(X) = -P(X = cara).\log_2(P(X = cara)) - P(X = coroa).\log_2(P(X = coroa))$$

$$H(X) = -\frac{1}{2}\times(-1) - \frac{1}{2}(-1) = 1 \text{ bit}$$

Suponha agora que a moeda seja viciada e que a probabilidade de ocorrer cara é de 75%. Então, a entropia será:

$$H(X) = -\frac{3}{4}\times\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\times\log_2\left(\frac{1}{4}\right) \approx 0{,}8 \text{ bit}$$

Ou seja, a entropia para o lançamento de uma moeda viciada é menor, já que a incerteza sobre qual face se observará é menor. Esse comportamento pode ser mais bem observado no gráfico da Figura 3, que mostra a função de entropia para o

caso de a variável *X* assumir apenas dois valores (símbolos). Observe que quando um dos símbolos tem probabilidade máxima (100%), a entropia é zero, pois não há incerteza. A entropia atinge seu valor máximo quando a probabilidade dos dois símbolos é a mesma (50%). Ou seja, quanto mais equiprováveis forem os símbolos, maior será a entropia, pois maior será a incerteza. Observe ainda que quando os símbolos são equiprováveis, a fórmula para a entropia se reduz à inicialmente apresentada — *log(N),* sendo *N* o número de símbolos.

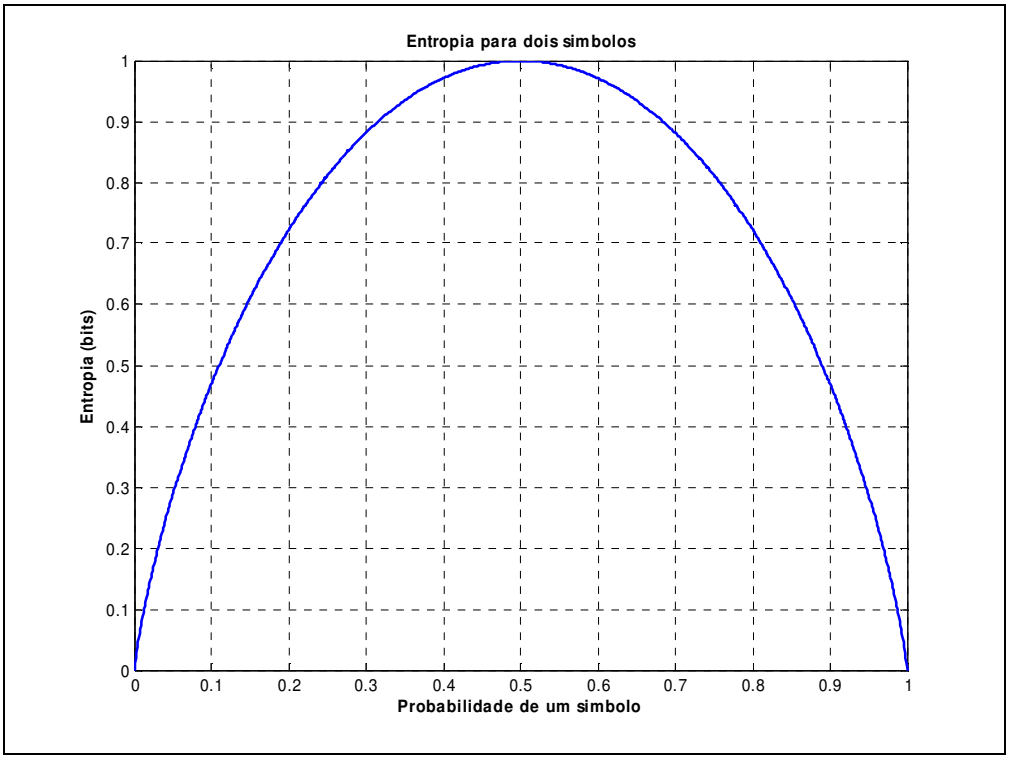


Figura 4-5 - Gráfico da Entropia de dois símbolos

Uma outra visão para a entropia é o número médio de bits necessário para transmitir uma sequência de símbolos com distribuição de probabilidades *P(X)*. Uma mensagem cujos símbolos forem equiprováveis (distribuição uniforme) necessitará de um número maior de bits para representá-la, pois cada símbolo terá o mesmo número de bits. Se a distribuição não for uniforme (entropia menor), os símbolos mais frequentes (ou prováveis) serão representados por um número menor de bits e o número médio de bits para codificar a mensagem diminuirá.

Considere, como exemplo, uma mensagem cujos símbolos pertençam ao alfabeto A, B, C e D. As frequências relativas de cada símbolo na mensagem são dadas por 4, 2, 1 e 1, respectivamente. As probabilidades, então, de cada símbolo

ser enviado são 1/2, 1/4, 1/8 e 1/8, respectivamente. Ignorando o fato de as frequências serem conhecidas, a mensagem poderia ser codificada da forma a seguir, que usa, em média, dois bits por símbolo.

A = 00
B = 01
C = 10
D = 11

Essa codificação usa sempre dois bits para cada símbolo. Shannon (1948) mostrou que sempre existe um código mínimo que pode representar a mensagem com o número médio de bits por símbolo dado pela entropia. Para o exemplo, o cálculo da entropia, mostrado a seguir, revela que esse código irá utilizar em média 1,75 bit por símbolo.

$$H(X) = -\frac{1}{2}\times\log_2\left(\frac{1}{2}\right) - \frac{1}{4}\times\log_2\left(\frac{1}{4}\right) - \frac{1}{8}\times\log_2\left(\frac{1}{8}\right) - \frac{1}{8}\times\log_2\left(\frac{1}{8}\right) = 1.75 \text{ bit}$$

Uma dessas codificações é a seguinte:

A = 0
B = 10
C = 110
D = 111

Observe que o número médio de bits ($\overline{N}$) para cada símbolo enviado nessa codificação é exatamente 1.75, pois, utilizando a média ponderada pela frequência dos símbolos, tem-se:

$$\overline{N} = \frac{4\times1 + 2\times2 + 1\times3 + 1\times3}{8} = 1.75 \text{ bits/símbolo}$$

Segundo Mani (2001) o número médio necessário de bits para se transmitir mensagens em inglês é de aproximadamente 2 bits/caractere.

Definida a entropia, Shannon formalizou a taxa de informação (*R*) com a seguinte expressão:

$$R = H_{antes} - H_{depois} \quad [12]$$

Como exemplo, considere a seguinte situação (Shannon 1948). Um emissor pode enviar dois símbolos (0 ou 1) equiprováveis por um canal com ruído. A probabilidade de se receber o símbolo 1 quando, na verdade, for enviado um 0 é de 0.01 e a probabilidade de um 0 ser recebido quando um 1 é enviado é também 0.01. No envio de um símbolo, a incerteza *antes* (H_{antes}) é 1 (vide Figura 4-5) e a incerteza *depois* (H_{depois}) será:

$$H_{depois} = -0.99 \times \log_2(0.99) - 0.01 \times \log_2(0.01) = 0.081 \text{ bit}$$

A quantidade real de informação recebida é dada, então, por:

$$R = H_{antes} - H_{depois} = 1 - 0.081 = 0.919 \text{ bit}$$

Observe que se o canal for perfeito (sem ruídos), a entropia depois será zero, pois não haverá incertezas. A informação, então, se reduz a $R = H_{antes}$.

4.2.2.1 Métodos Filter - Information Gain

A relação apresentada por Shannon para quantificar a informação ($R = H_{antes} - H_{depois}$) pode ser adaptada para quantificar a relevância das características em um problema de aprendizado de máquina. Intuitivamente, características mais relevantes serão aquelas mais informativas. A questão é descobrir quais os contextos que devem ser adotados para se calcular as entropias *antes* e *depois*. Uma possibilidade é considerar a entropia *antes* como sendo a da classe (*X*), e a *depois* como sendo a entropia da classe depois que se conheça o

valor da característica (*Y*) cuja informatividade se deseja medir. Isso pode ser descrito por:

$$R = H_{antes} - H_{depois} = H(X) - H(X \mid Y) \qquad [13]$$

Essa relação indica justamente a redução na incerteza (ganho de informação) de *X* pelo conhecimento de *Y*. Ou seja, quanta informação que a característica *Y* traz para a determinação da classe *X*. A entropia *depois* $H(X \mid Y)$ representa esse conceito de entropia de uma variável *X* após a observação do valor de outra variável *Y*, que Shannon definiu como *entropia condicional*, apresentada a seguir:

$$H(X \mid Y) = -\sum_j \left(P(y_j) \sum_i P(x_i \mid y_j) \log_2 (P(x_i \mid y_j)) \right) \qquad [14]$$

$P(y_j)$ é a probabilidade *a priori* para todos os valores de *Y* e $P(x_i \mid y_j)$ a probabilidade de *X* dados os valores de *Y*.

Na Teoria da informação e na Estatística, de modo geral, essa medida é conhecida como *informação mútua*. Em mineração de dados, o termo empregado é *information gain* (*IG*) onde é bastante utilizada para se gerar rankings por pesos de relevância das características. Quinlan (1993) também a utiliza como critério para escolher a característica que deve ser usada para particionar recursivamente os exemplos de uma árvore de decisão. Características com maior *IG* são escolhidas.

No Apêndice B é apresentado um exemplo de cálculo da medida, no contexto de Sumarização Automática extrativa baseada em aprendizado de máquina.

Principais vantagens: Medida bastante usada e popular no campo da Estatística.

Principais desvantagens: Por se tratar de um método *filter*, o resultado é um ranking de características em vez do conjunto recomendado. Isso implica em se estabelecer um número de corte de características.

4.2.2.2 Métodos Filter - Correlation Feature Selection (CFS)

Como Hall (2000) aponta, a medida IG (*information gain*) tende a favorecer as características com maior diversidade de valores. Ou seja, os IGs dessas características tendem a ser mais altos, mesmo que elas não sejam tão informativas. Além disso, para se comparar de maneira adequada valores *IG* de duas ou mais características, a medida deveria ser normalizada no intervalo [0,1]. Para contornar esse desequilíbrio da IG, Hall sugere a medida S*ymetrical Uncertainty* (*SU*), definida por:

$$SU(X,Y) = 2\frac{IG(X,Y)}{H(X)+H(Y)} = 2\frac{H(X)-H(X\mid Y)}{H(X)+H(Y)} \qquad [15]$$

Essa medida, como as outras já apresentadas, pode ser usada para avaliar a relevância de uma dada característica. Entretanto, a simples escolha das características mais relevantes não garante que se consiga o melhor conjunto de características, ou mesmo um que se aproxime do melhor. Primeiramente porque não se sabe quantas características devem ser escolhidas e também por não se considerar o grau de redundância entre as características, problema que pode ter bastante influência sobre o desempenho dos algoritmos de aprendizado de máquina (Witten e Frank 2005).

Hall (2000) propõe uma heurística para se escolher um subconjunto de características com base na medida *SU*. A premissa é de que um bom subconjunto de características tem características altamente correlacionadas com a classe, mas pouco correlacionadas entre si. Ou seja, sua heurística busca conjuntos relevantes com baixo grau de redundância entre suas características. A fórmula a seguir formaliza a heurística, em que A_i e A_j são características e C é a classe:

$$Mérito_S = \frac{\sum_j SU(A_j,C)}{\sqrt{\sum_i \sum_j SU(A_{i,}A_j)}} \qquad [16]$$

Teoricamente, o melhor conjunto seria obtido calculando-se a medida para todos os possíveis conjuntos, o que é inviável, pois o número é exponencial em relação ao número de características (2^n). Como exemplo, a Figura 4-6 ilustra o

espaço de busca para de determinação do melhor subconjunto de características para um sumarizador que utiliza como características: tamanho das sentenças, frequência das palavras, posição da sentença e a presença de nomes próprios. Por dessa complexidade, Hall (2000) sugere a utilização de uma heurística de busca para localizar o subconjunto, como o método *best-first*.

O algoritmo começa a procura a partir de um subconjunto vazio de características e gera os possíveis conjuntos com apenas uma característica. O subconjunto de maior mérito é então expandido da mesma maneira. Se o resultado piorou, o algoritmo volta ao estado anterior. O critério de parada é a ocorrência consecutiva de cinco expansões sem melhoras.

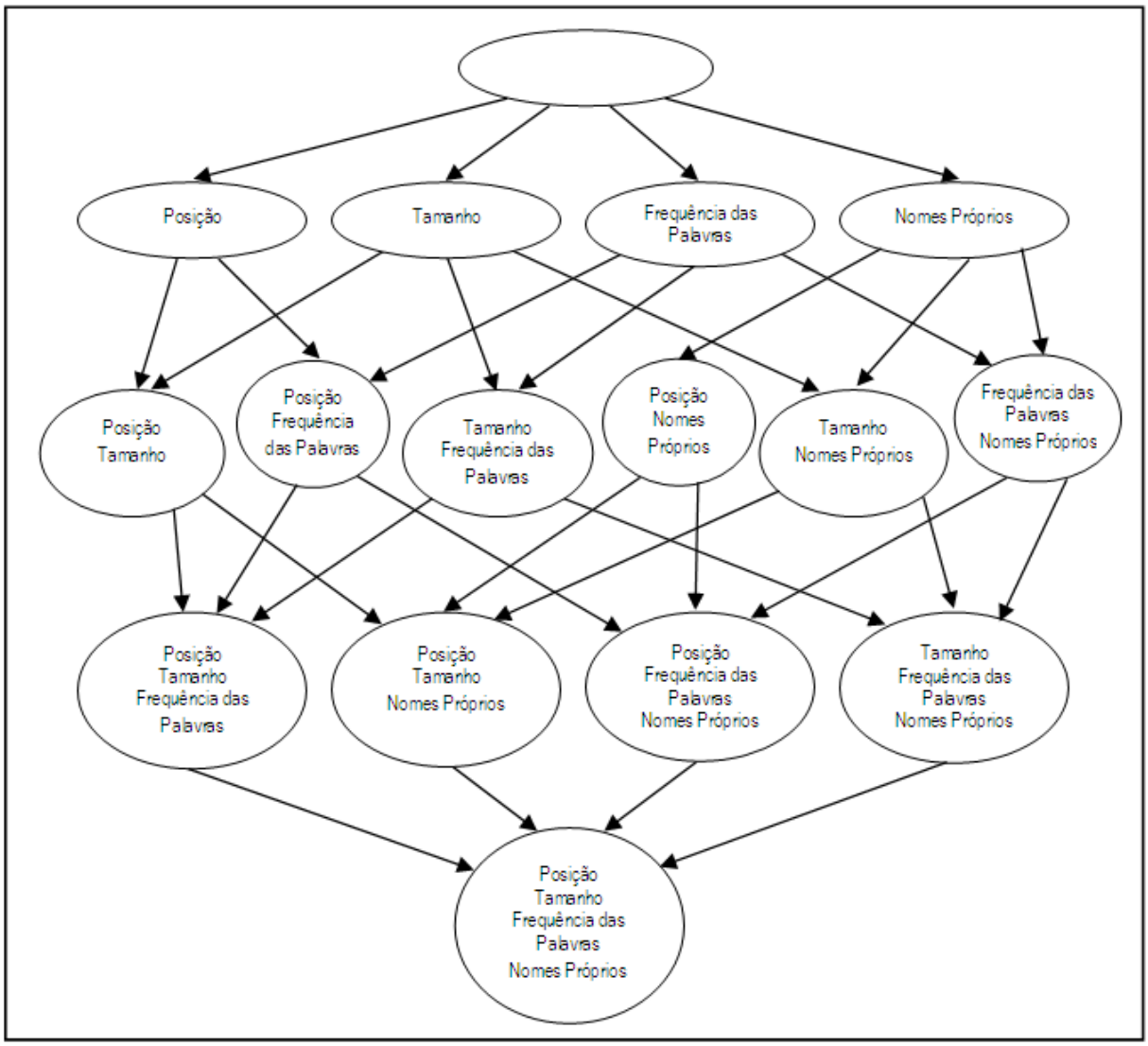


Figura 4-6 – Exemplo de espaço de busca para o melhor subconjunto de características

Principais vantagens: O resultado produzido pelo método é o conjunto recomendado de características. O método foca tanto em redundância quanto relevância das características.

Principais desvantagens: Custo computacional maior, embora viável se utilizadas heurísticas de busca.

4.2.3 Métodos Filter - Medidas Estatísticas

Como Biesiada et al. (2005) apontam, medidas estatísticas de dependência entre variáveis aleatórias podem ser uma alternativa às medidas baseadas na Teoria da Informação, para a seleção de características. Uma dessas medidas é a estatística do qui-quadrado (χ^2), bastante eficiente para se avaliar a associação existente entre variáveis (ou valores) discretas. Seu princípio é comparar as diferenças entre as frequências observadas e esperadas. Para a seleção de características, calcula-se a medida χ^2 para a característica que se deseja avaliar com respeito à classe do problema, conforme a seguinte fórmula:

$$\chi^2 = \sum_{i=1}^{m}\left(\sum_{j=1}^{k}\frac{\left(A_{ij}-E_{ij}\right)^2}{E_{ij}}\right) \qquad [17]$$

em que:

m é número de valores que a característica pode assumir;

k é número de valores que a classe pode assumir;

A_{ij} é a frequência observada. No caso, o número de instâncias da característica que assumem o i-ésimo valor possível para essa característica e com o j-ésimo valor possível para a classe;

E_{ij} é a frequência esperada, definida por $E_{ij} = \frac{R_i \times C_j}{N}$;

R_i é número de instâncias que possuem o i-ésimo valor possível para a característica.

C_j é número de instâncias com o j-ésimo valor possível para a classe;

N é o número total de instâncias (tuplas).

Quanto maior o valor da medida, maior será a dependência entre a característica e a classe.

No Apêndice B é apresentado um exemplo de cálculo da medida, no contexto de Sumarização Automática Extrativa baseada em aprendizado de máquina.

Principais vantagens: Medida bastante usada e popular no campo da Estatística.

Principais desvantagens: Por se tratar de um método *filter*, o resultado é um ranking e características em vez do conjunto recomendado. Isso implica em se estabelecer um número de corte de características.

4.2.4 Métodos Filter - Análise dos Componentes Principais

A análise dos Componentes Principais (e.g., Haykin 1999; Witten e Frank 2005) é um dos métodos estatísticos mais usados na análise de dados multivariados. O objetivo principal da análise de componentes principais é a obtenção de um pequeno número de combinações lineares de um conjunto de variáveis, chamados de componentes principais, que retenham o máximo possível da informação contida nas variáveis originais. Frequentemente, um pequeno número de componentes pode ser usado no lugar das variáveis originais, nas análises de regressões, análises de agrupamentos etc.

Mais formalmente, dado um conjunto D com n instâncias e p atributos ou características ($x_1, x_2,..., x_p$), uma transformação linear para um novo conjunto de características $z_1, z_2,..., z_p$ pode ser calculada como:

$$z_1 = a_{11} x_1 + a_{21} x_2 + \ldots + a_{p1} x_p$$
$$z_2 = a_{12} x_1 + a_{22} x_2 + \ldots + a_{p2} x_p$$
$$\ldots$$
$$z_p = a_{1p} x_1 + a_{2p} x_2 + \ldots + a_{pp} x_p$$

A determinação dos componentes envolve o cálculo dos autovalores da matriz de covariâncias dos dados. Para isso, assume-se que todas as características são numéricas ou tenham sido transformadas de forma adequada em números.

Os componentes são extraídos na ordem do mais explicativo para o menos explicativo. Teoricamente o número de componentes é sempre igual ao número de variáveis. Entretanto, alguns poucos componentes são responsáveis por grande parte da explicação total.

Diferente dos demais métodos das seções anteriores, que apenas avaliam características, a análise dos componentes principais gera novas características mais informativas e que são combinações das variáveis originais.

Principais vantagens: técnica bastante usada e popular no campo da Estatística; os componentes principais correspondem a características não correlacionadas e, portanto, não redundantes.

Principais desvantagens: a utilização de novas características que são combinações das características originais pode levar a perda do aspecto intuitivo e do significado linguístico presente nelas; a combinação linear pode não fazer sentido para características que sejam discretas; a complexidade do método é cúbica com relação ao número de características, podendo levar a um alto custo computacional para modelos de sumarização com grandes números de características.

4.3 Sistemas Nebulosos

A lógica nebulosa, ou lógica *fuzzy* (e.g., Klir e Yuan 1995), é uma extensão da lógica booleana que admite valores lógicos intermediários entre o falso e o verdadeiro. Tradicionalmente, uma proposição lógica tem dois extremos: ou

"completamente verdadeiro" ou "completamente falso". Entretanto, na lógica nebulosa, uma proposição varia em grau de verdade de 0 a 1, o que leva a ser parcialmente verdadeira e parcialmente falsa.

A lógica nebulosa é baseada na teoria dos conjuntos nebulosos que por sua vez estende a teoria dos conjuntos tradicionais[24] utilizando o conceito de grau de verdade. De forma geral, um conjunto é definido por uma função chamada de função característica *μ(x)*, que declara quais elementos de *x* são membros do conjunto e quais não são. Enquanto a função característica dos conjuntos tradicionais é discreta (0 ou 1), a função característica dos conjuntos nebulosos pode assumir qualquer valor real entre 0 e 1, indicando o grau de pertinência desse elemento ao conjunto.

Exemplo 1: Um homem de 1,80 m e um homem de 1,75 m podem ambos pertencerem ao conjunto "alto", embora o homem de 1,80 metro tenha um grau de pertinência maior a este conjunto. Um possível função característica (ou função de pertinência) para esse conjunto "alto" poderia ser:

$$\mu_A(x) = \begin{cases} 1, x > 1{,}75 \\ 0, x < 1{,}6 \\ \dfrac{x - 1{,}6}{0{,}15}, 1{,}6 \le x \le 1{,}75 \end{cases}$$

E uma representação gráfica é dada na Figura 4-7. Veja que o grau de pertinência ao conjunto "alto" para indivíduos com menos de 1,60m é 0. A partir dessa estatutura, o nível de pertinência começa a aumentar, até a estatura de 1,80m. Acima dessa estatura, todos os indivíduos tem o grau máximo de pertinência ao conjunto.

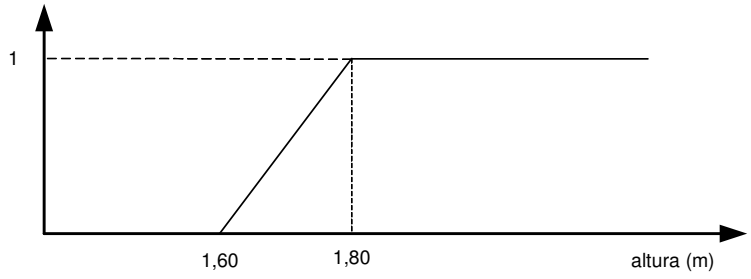


Figura 4-7 – Representação gráfica do conjunto "alto"

[24] Nos trabalhos sobre lógica nebulosa, os conjuntos tradicionais são referidos usualmente pelo termo *crisp* do inglês.

Exemplo 2: Considere a figura seguinte que mostra possíveis conjuntos fuzzy para a variável temperatura: baixíssima, baixa, média, alta ou altíssima. Perceba que, por exemplo, que uma tempetura de 15ºC pode tanto ter um grau de pertinência no conjunto "baixa" quanto no "média". Além disso, diferente do Exemplo 1, que representava o conjunto por meio de um triângulo, aqui a representação é por funções em forma de sino. A escolha de um determinado formato deve ser norteada pela compatibilidade do formato com o conceito que se deseja representar.

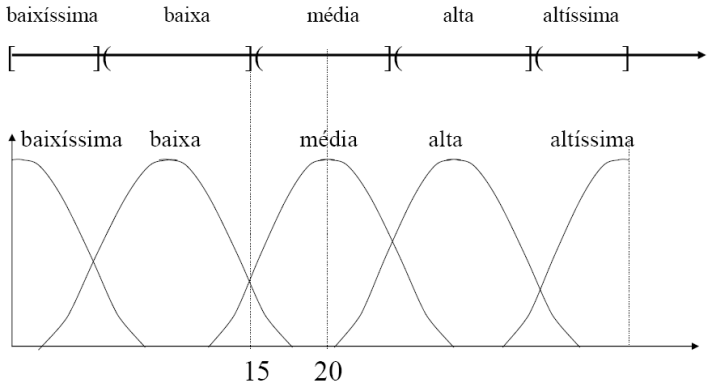


Figura 4-8 – Exemplo conjuntos nebulosos para a variável tempetura

4.3.1 Sistemas de Inferência Nebulosos

Os artifícios da teoria dos conjuntos nebulosos permitem que estados indeterminados possam ser modelados. Desse modo, é possível avaliar conceitos não-quantificáveis precisamente, como por exemplo a veracidade de um argumento: *corretíssimo*, *correto*, *incoerente*, *falso*, *totalmente errôneo*, etc. Cada um desses conceitos poderia ser modelado por meio de um conjunto nebuloso.

Partindo dessa premissa, sistemas de inferência nebulosos têm sido aplicados em sistemas de controle, sistemas especialistas e sistemas de visão computacional. Esses sistemas utilizam um processo de inferência nebuloso, que é o processo de se mapear uma dada entrada numa saída com base na teoria de conjuntos nebulosos. Esse mapeamento provê uma base a partir do qual decisões podem ser tomadas ou padrões discernidos.

A maioria dos sistemas de inferência nebulosos é baseada em regras. A inferência nebulosa baseada em regras interpreta as variáveis de entrada e, com

base em algum conjunto de regras, designa valores para a saída. O processo de inferência nebuloso envolve, então, funções matemáticas que definem o grau de pertinência de um elemento a um conjunto nebuloso, operadores nebulosos (*E* corresponde ao operador de mínimo e *OU* corresponde ao operador de máximo) e regras *SE-ENTÃO*, em que a parte *SE* da regra é chamada de premissa e a parte *ENTÃO* é chamada de consequente. Na inferência nebulosa, as premissas podem ser "ativadas" com um certo grau de verdade. Consequentemente, as saídas ou implicações possuem um grau de verdade associado.

4.3.2 Classificador Nebuloso

Um tipo particular de sistema de inferência nebuloso é o classificador nebuloso. Um classificador nebuloso possui quatro componentes: um processador de entrada (ou *fuzzificador*), um conjunto de regras linguísticas, um método de inferência nebuloso e um processador de saída (ou *defuzzificador*), gerando um número real como saída.

As regras de classificação nebulosas são definidas da seguinte forma:

SE C é Condição ENTÃO classe

em que:

C_k é o vetor de características dos exemplos (ou sentenças) C_{jk} $\{C_{1k},..., C_{nk}\}$,

Condição é o conjunto de funções características A_{ij} $\{A_{i1},..., A_{in}\}$

Classe é u_i $\{u_1,..., u_M\}$, com $i = 1,...M$ o número de classes,

$j = 1,... ,n$ o número de características

$k = 1,...,N$ as os exemplos a classificar.

O procedimento geral para classificação é o seguinte:

1) Calcular o grau de compatibilidade do exemplo (C_k) com o antecedente de cada regra R_i :

$$\omega(R_i, C_k) = \min(A_{i1}(C_{1k})),..., A_{in}(C_{nk}))) \quad [18]$$

2) Para cada classe u_i atribui-se um escore usualmente pela soma dos graus de compatibilidade do exemplo com as regras dessa classe;

3) A classe do exemplo será a classe que obtiver o maior escore conforme o passo 2.

A grande questão quanto ao uso dos sistemas de inferência nebulosos é a construção da base de regras e a modelagem dos conjuntos nebulosos envolvidos. A primeira forma é a construção manual, com o auxílio de um especialista de domínio. Uma segunda possibilidade é a utilização de métodos heurísticos, como o método de Wang e Mendel (1998), que geras regras que espelham os conjuntos nebulosos com maior grau de compatibilidade com os exemplos existentes em uma base de treinamento. Mais recentemente, abordagens de construção automática têm sido desenvolvidas, com o uso de algoritmos genéticos (Eberhart e Shi 2007). Seguindo essa linha Kiani-B e Akbarzadeh-T (2006) utilizaram, de forma pioneira pelo nosso conhecimento, um sistema nebuloso para modelar o processo de escolha de sentenças, conforme descrito na Seção 3.3.2.

Principais vantagens: A decisão de se uma sentença é relevante ou não pode ser considerada um processo nebuloso: uma sentença pode ser relevante em determinadas condições e com um certo grau de verdade. Este comportamento tem sido tratado usualmente por meio da teoria de probabilidade (por exemplo, com o classificador Naïve-Bayes). Entretanto, a utilização de sistemas de inferência nebulosos pode ser uma alternativa viável.

Principais desvantagens: A grande dificuldade está na construção dos conjuntos de regras nebulosos e na modelagem das funções características dos conjuntos nebulosos. Isto é, a modelagem das funções de pertinência dos elementos aos conjuntos.

4.4 Síntese e Comparação das Abordagens Envolvendo Aprendizado de Máquina

Os diversos métodos da área de aprendizado de máquina apresentados neste capítulo podem ser aplicados na área de Sumarização Automática com o objetivo de tratar os dois problemas fundamentais desta pesquisa, já apresentados no Capítulo 1:

Problema 1. Quais características utilizar?
Problema 2. Como combinar e ponderar as características escolhidas?

Nos primeiros trabalhos de SA, esses problemas eram tratados de forma manual, com modelos construídos e calibrados por engenheiros de conhecimento e especialistas de domínio. Como já sugerido anteriormente, a grande vantagem dos métodos baseados em aprendizado de máquina é justamente o fato de serem automáticos e permitirem a construção de modelos mais escaláveis e mais facilmente adaptáveis a outros domínios ou gêneros linguísticos. Embora um trabalho envolvendo engenheiros de conhecimento e especialistas de domínio possa produzir resultados melhores, geralmente todo o trabalho deve ser refeito para aplicação em um novo domínio no geral. Com os métodos automáticos, geralmente basta-se executá-los novamente e retreinar o sistema.

Por outro lado, métodos automáticos são baseados em modelos matemáticos que geralmente desconsideram a semântica dos dados. Como eles apenas buscam otimizar as taxas de acerto para um conjunto de dados, não se observa necessariamente razões linguísticas para seus parâmetros e para os pesos que atribuem às características, por exemplo. Além disso, podem muitas vezes produzir modelos de inferência ou decisão pouco intuitivos para os humanos.

Outro problema da utilização dos métodos baseados em aprendizado de máquina é que geralmente é necessário definir premissas sobre a natureza e distribuição dos dados para que se possa utilizá-los. Muitas vezes não há como se garantir que os dados provenientes de textos livres tenham alguma propriedade ou distribuição particular. Na maioria das vezes, as premissas são simplesmente relaxadas.

Além disso, os próprios métodos geram modelos, que naturalmente tem limitações. Nesse sentido, os métodos de seleção automática de características podem ser encarados como complementares aos classificadores e métodos de inferência, pois visam pré-processar e tratar os dados de forma a gerar um melhor desempenho do modelo como um todo.

Considerando os métodos apresentados nesta seção, a Tabela 4-2 sintetiza as utilidades de cada um, suas principais vantagens e desvantagens.

No próximo capítulo, parte-se para a descrição dos modelos de SA propostos neste trabalho, que além da utilização do aprendizado de máquina, fizeram uso de estatísticas textuais e medidas de grafos.

Tabela 4-2 – Comparação entre os métodos de aprendizado de máquina

Problema	Método	Vantagens	Desvantagens
1	**Wrapper**	As características determinadas são as que produzirão os melhores resultados com o classificador escolhido.	Custo computacional e uso de medidas não específicas no caso de SA
	Análise dos Componentes Principais	Bastante utilizada na Estatística em problemas com muitas variáveis.	Possibilidade de introdução de novas variáveis, sem necessariamente significado linguístico; alto custo computacional.
	Information Gain	Medida popular	Necessário fixar número características
	Qui-quadrado	Medida popular	Necessário fixar número características
	CFS	Gera um conjunto recomendado de características relevantes e pouco redundantes	Custo computacional maior, embora viável se utilizadas heurísticas de busca.
2	**Naïve-Bayes**	Bom desempenho geral nas tarefas de PLN. Modelagem por meio de probabilidades.	Premissa independência características
	C4.5	Representação por árvores	Geração dos rankings é difícil e pode não ser precisa.
	SVM	Método robusto e que pode tratar não-linearidades	Rankings podem não ser calibrados
	Regressão Logística	Adequado a tarefa de ranking	Assume relação linear entre características e a relevância da sentença
	Redes Neurais	Pode tratar problemas não-lineares	Pode ser complicado configurar o modelo para geração de rankings de sentenças

Capítulo 5

DESENVOLVIMENTO DOS MODELOS DE SA BASEADOS EM GRAFOS, ESTATÍSTICAS TEXTUAIS E APRENDIZADO DE MÁQUINA

Nesta seção apresentam-se os modelos de Sumarização Automática que foram desenvolvidos com base nas três metodologias exploradas neste trabalho, empregando-as de forma individual ou em conjunto.

Os modelos de SA desenvolvidos neste trabalho foram construídos a partir da combinação ou da utilização individual das três grandes abordagens exploradas neste trabalho: estatísticas textuais, grafos e aprendizado de máquina. Ao todo, foram 29 modelos de SA propostos, sintetizados na Tabela 5-1:

Tabela 5-1 – Modelos desenvolvidos

Modelos	Seção do texto
1. SuPor-2	Seção 5.1
2. TextRank+StemmingStopRem	Seção 5.2
3. TextRank+Thesaurus	Seção 5.3
4-27. Modelos baseados em redes complexas e características do SuPor-2 combinadas	Seção 5.4
28. SuPor-2 modificado (mesmas características do SuPor-2 fuzzy) 29. SuPor-2 Fuzzy	Seção 5.4

Na Seção 5.1 descreve-se o primeiro sistema desenvolvido, o SuPor-2, construído a partir do SuPor, já citado na introdução deste trabalho.

Na Seção 5.2, descrevem-se os dois modelos construídos com base na abordagem de Grafos, seguindo o algoritmo TextRank, já apresentado na Seção 3.1.7.

Na Seção 5.3 são descritos os modelos de SA baseados na combinação por meio do Aprendizado de Máquina de características diversas para SA, incluindo estatísticas textuais e métricas de grafos.

Na Seção 5.4, é descrito um modelo de SA que buscou verificar os potenciais benefícios da utilização de aprendizado de máquina baseada da Teoria de Conjuntos Nebulosos.

Ao final do capítulo, é feita uma síntese dos modelos construídos, identificando-se suas características principais e as abordagens exploradas em cada um.

5.1 Desenvolvimento do SuPor-2

Como citado no Capítulo 1 e na Seção 3.2.4, o SuPor serviu de motivação inicial para utilização das abordagens focadas neste trabalho devido a seu bom desempenho quando comparado a outros seis sumarizadores para o Português do Brasil (Rino et al. 2004). Assim, descreve-se nesta seção sua primeira evolução, o SuPor-2 (Leite e Rino 2006). Neste sumarizador, buscou-se eliminar a necessidade de um especialista humano que o SuPor impõe para determinar as características mais adequadas Para isso, foram desenvolvidas as seguintes propostas:

Proposta 1 - Aperfeiçoamento nas características utilizadas pelo esquema de aprendizado de máquina (Seção 5.1.1);

Proposta 2 - Pré-seleção das características mais promissoras (Seção 5.1.2);

Proposta 3 - Foi proposta também uma mudança na arquitetura do sistema (Seção 5.1.3), com a substituição do próprio módulo de aprendizado pelo WEKA (Witten e Frank 2005), um ambiente para mineração de dados. A utilização do WEKA no modelo do SuPor-2 é justificada devido à facilidade oferecida para aplicar e avaliar diferentes estratégias de aprendizado. Essa avaliação inicial permitiu

determinar a melhor configuração do sistema no experimento descrito na Seção 5.1.4.

5.1.1 Proposta 1 – Aperfeiçoamento das Características

A primeira modificação feita no SuPor para combinar e ponderar melhor as características foi alterar sua representação binária. Por esse modelo de característica, a única informação resultante do método ou medida é uma indicação ou não para aquela sentença compor o sumário. Ou seja, a representação binária não informa o grau em que a indicação é feita. O uso de características mais significativas, que incorporem mais informação para o modelo de aprendizado de máquina, sugere a possibilidade de se tomar decisões de extração mais expressivas.

A abordagem que se utilizou para se especificar características mais significativas que as do SuPor considerou, basicamente:

1) A utilização dos dados numéricos que os métodos de sumarização usam para julgar a relevância das sentenças, quando possível (casos a-d listados a seguir);
2) A utilização de informações que refletem a forma como cada sentença foi selecionada, no caso de métodos que não utilizam números para julgar a relevância das sentenças (casos e-g listados a seguir).

No que segue, são especificadas, para cada método, as mudanças realizadas nas características:

a) **Alteração na Característica associada ao Método de Frequência das Palavras**

O método de Frequência das Palavras utiliza um dado numérico para julgar a relevância das sentenças — o somatório das frequências de suas palavras em todo o texto. No SuPor original, a saída binária da característica associada era obtida verificando se o somatório das frequências das palavras da sentença ultrapassa (valor *True*) ou não (valor *False*) um determinado valor de corte. Já no SuPor-2, a

característica (*C*) foi definida pelo próprio somatório das frequências das palavras da sentença.

Para evitar treinamento tendencioso, optou-se por relativizar o valor dessa característica em relação ao texto-fonte, normalizando-se seu valor no intervalo [0,1]. A normalização consistiu, simplesmente, em dividir-se o valor da característica de cada sentença pelo maior valor obtido em todo o texto, tanto na fase de treinamento quanto na de extração, conforme fórmula a seguir:

$$CaracterísticaNormalizada(i) = \frac{C(i)}{Máx(C)} \qquad [19]$$

em que *i* é o número da sentença do texto processado.

b) **Alteração no Uso da Característica de Tamanho da Sentença**

Similarmente ao que foi feito para o método de frequência das palavras, o valor dessa característica foi definido como o número de palavras da sentença. Foi adotado, também, o mesmo procedimento de normalização da Equação [19].

c) **Alteração no Uso da Característica de Nomes Próprios**

Novamente, a mudança feita foi similar àquela feita para o método de frequência das palavras. Considerou-se a soma das frequências dos nomes próprios da sentença em todo o texto como o valor da característica. Da mesma forma, adotou-se o mesmo procedimento de normalização, conforme Equação [19].

d) **Alteração no Uso do Método de Importância dos Tópicos**

Pelo método de *Importância dos Tópicos*, a relevância de uma sentença está associada a dois valores numéricos no intervalo [0,1]:

- A importância do tópico (*T*) relacionado à sentença (*S*): *Sim (T, S)*
- A similaridade da sentença (*S*) com o centróide do tópico (*C*): *Sim (C, S)*

Quanto maiores forem a importância do tópico relacionado à sentença e sua similaridade com o centróide do tópico, maior será a relevância dessa sentença para a composição do extrato.

$$C_{IT} = 2\frac{Sim(T,S) \times Sim(C,S)}{Sim(T,S) + Sim(C,S)} \quad [20]$$

Para preservar a noção de que as duas medidas são importantes, definiu-se o valor da característica de Importância dos Tópicos C_{IT} como a média harmônica entre elas, conforme Equação [20]. Note que não é necessário normalizar o resultado, pois as duas medidas já estão no intervalo [0,1].

e) **Alteração no Uso da Característica de Posição**

A característica de Posição, que não utiliza dados numéricos, permaneceu categórica. No entanto, para torná-la mais informativa, detalhou-se também, para cada categoria, a posição da sentença em seu parágrafo e a posição deste no próprio texto. Essa mudança é sintetizada na tabela a seguir:

Tabela 5-2 – Característica de Posição no SuPor-2

Rótulo	Posição do parágrafo	Posição da sentença no parágrafo
II	Início do texto	Início do parágrafo
IM	Início do texto	Meio do parágrafo
IF	Início do texto	Final do parágrafo
MI	Meio do texto	Início do parágrafo
MM	Meio do texto	Meio do parágrafo
MF	Meio do texto	Final do parágrafo
FI	Final do texto	Início do parágrafo
FM	Final do texto	Meio do parágrafo
FF	Final do texto	Final do parágrafo

Para a especificação dessa característica, cada sentença é rotulada de acordo com um dos códigos da primeira coluna da tabela anterior, conforme a posição de seu parágrafo e sua posição no próprio parágrafo. De maneira semelhante à versão original do SuPor, são considerados parágrafos iniciais somente aqueles que figuram entre os 10% do início do texto; são considerados finais somente os que estão nos 5% parágrafos do fim do texto.

f) **Alteração no Uso do Método de Cadeias Lexicais**

Como o método não faz uso de dados numéricos para julgar a relevância das sentenças, considerou-se:

- se alguma heurística do método recomendou a sentença;
- quais heurísticas recomendaram a sentença.

Assim, a característica pode assumir oito rótulos distintos, conforme mostra a tabela a seguir.

Tabela 5-3 – Característica associada ao método de Cadeias Lexicais no SuPor-2

Rótulo	**Significado**
False	nenhuma heurística recomendou a sentença
H1	apenas a heurística 1 (*primeira ocorrência*) recomendou a sentença
H2	apenas a heurística 2 (*membro representativo*) recomendou a sentença
H3	apenas a heurística 3 (*concentração no tópico*) recomendou a sentença
H1+H2	as heurísticas 1 e 2 recomendaram a sentença
H1+H3	as heurísticas 1 e 3 recomendaram a sentença
H2+H3	as heurísticas 2 e 3 recomendaram a sentença
H1+H2+H3	as heurísticas 1, 2 e 3 recomendaram a sentença

g) **Alteração no Uso do Método de Mapa de Relacionamentos**

De modo similar ao método das cadeias lexicais, esse método também não utiliza dados numéricos para julgar a relevância das sentenças, já que elas são selecionadas por três heurísticas que determinam o modo de percurso no mapa de relacionamentos entre os parágrafos. Assim, também foi considerado um conjunto de rótulos linguísticos para determinar o valor da característica para cada sentença, conforme a Tabela 5-4.

Tabela 5-4 – Característica associada ao Mapa de Relacionamentos no SuPor-2

Rótulo	Significado
False	nenhum caminho selecionou a sentença
C1	apenas o caminho 1 (*profundo*) selecionou a sentença
C2	apenas o caminho 2 (*segmentado*) selecionou a sentença
C3	apenas o caminho 3 (*denso*) selecionou a sentença
C1+C2	os caminhos 1 e 2 selecionaram a sentença
C1+C3	os caminhos 1 e 3 selecionaram a sentença
C2+C3	os caminhos 2 e 3 selecionaram a sentença
C1+C2+C3	os caminhos 1, 2 e 3 selecionaram a sentença

Considerando-se as alterações descritas nos tópico anteriores, a tabela seguinte resume as características utilizadas no SuPor-2. Essa tabela apresenta os mesmos métodos do SuPor original (Tabela 3-1). A diferença básica existente é a utilização no SuPor-2 de características com domínio mais abrangente. No SuPor-2 as características podem assumir valores numéricos e multinomiais, em vez de apenas valores binários.

Tabela 5-5 – Quadro-resumo de características exploradas no SuPor-2

Característica	Nome	Domínio
C1	Cadeias Lexicais	{'False', 'H1', 'H2', 'H3', 'H1H2', 'H1H3', 'H2H3', 'H1H2H3'}.
C2		
C3	Tamanho da Sentença	[0, 1]
C4	Nomes Próprios	[0,1]
C5	Posição da Sentença	{'II','IM','IF','MI','MM','MF','FI','FM','FF'}
C6	Frequência das Palavras	[0, 1]
C7		
C8	Mapa de Relacionamento	{'False,'C1','C2','C3', 'C1C2', 'C1C3', 'C2C3', 'C1C2C3'}.
C9		
C10	Importância dos Tópicos	[0, 1]
C11		

A justificativa para adotar o mesmo número de características do SuPor original (11) é consoante com a Hipótese 1 deste trabalho, no sentido que a utilização de características diversas é benéfica para a SA. Entretanto, no SuPor original a representação binária das características pode prejudicar esse beneficio e por isso as características foram alteradas conforme foi descrito.

5.1.2 Proposta 2 – Seleção Automática de Características

Seguindo a Hipótese 2 deste trabalho, que sugere a pré-seleção de características, incorporou-se também no SuPor esse processo. É justamente essa

modificação que foi feita para evitar a necessidade de um especialista humano selecionar as características mais adequadas manualmente.

A partir da análise feita sobre as abordagens de seleção automática de características (Seção 4.2), optou-se pelo uso no SuPor-2 do método CFS (Hall 2000). O motivo principal foi o fato de o método indicar um subconjunto de características em vez de gerar um ranking. No segundo caso, seria necessário estabelecer um número de corte no número de características. Ou seja, caso fosse utilizada uma abordagem como o cálculo da medida *Information Gain,* haveria ainda a necessidade de um especialista definir qual seria o número adequado de corte para um determinado corpus. Além disso, o autor relata boa performance do método CFS quando utilizado com os principais classificadores.

.

5.1.3 Proposta 3 – Utilização do WEKA na Arquitetura do SuPor-2

No SuPor-2, todo o modelo de aprendizado de máquina foi acoplado ao WEKA. Assim, o método de classificação no SuPor-2 é configurável para qualquer classificador disponível no WEKA. O padrão do sistema é o método Flexible-Bayes descrito na Seção 4.1.2. A utilização desse classificador em vez do Naïve-Bayes tradicional é devido à presença de características numéricas.

O modelo de características modificado, definido nas seções anteriores, foi representado no WEKA por arquivos do tipo *ARFF* [25] (Witten e Frank 2005). No WEKA, arquivos *ARFF* podem ser usados tanto na fase de treino quanto na de classificação (extração). Um arquivo *ARFF* de treino contém todas as sentenças do corpus de treino representadas na forma de tuplas. Cada tupla contém os valores para as 11 *características* definidas no modelo do SuPor-2 mais um valor booleano, a classe no contexto de Aprendizado de Máquina, que indica se aquela sentença pertence ao extrato ideal de seu texto. No caso de arquivos *ARFF* de classificação, a diferença é que os valores para a classe estão ausentes, justamente porque se deseja determinar quais sentenças irão compor o extrato, e as tuplas se referem a apenas um texto.

A Figura 5-1 mostra um exemplo de arquivo *ARFF* de treino, com somente algumas tuplas (as características consideradas são indicadas pelo rótulo @attribute

[25] Attribute Relation File Format.

e as tuplas, delimitadas pelo rótulo @data). O último atributo definido é justamente a classe, que foi chamada de *PresencaSumario*. Note que há 11 características, e não 7, que é número de métodos de SA utilizados pelo SuPor. Como já citado anteriormente, isso ocorre porque alguns métodos contribuem com 2 características, diferenciadas pela forma como o pré-processamento é feito (indicado após o caractere "_"). Observe ainda que os valores das tuplas (indicados pela sequência de valores @data) seguem a ordem em que os atributos (*características*) são definidos.

```
@relation ObtencaoSumario

@attribute CadeiasLexicais_Paragrafos { FALSE, H1, H2, H3, H1+H2, H1+H3,H2+H3, H1+H2+H3 }
@attribute CadeiasLexicais_TextTiling { FALSE, H1, H2, H3, H1+H2, H1+H3,H2+H3, H1+H2+H3 }

@attribute TamanhoSentencas real

@attribute NomesProprios real

@attribute PosicaoParagrafos {II, IM, IF, MI, MM, MF, FI, FM, FF }

@attribute FrequenciaPalavras_Radicais real
@attribute FrequenciaPalavras_Quadrigramas real

@attribute MapaRelacionamentos_Radicais { FALSE, C1, C2, C3, C1+C2, C1+C3,C2+C3, C1+C2+C3 }
@attribute MapaRelacionamentos_Quadrigramas { FALSE, C1, C2, C3, C1+C2, C1+C3,C2+C3, C1+C2+C3
}

@attribute ImportanciaTopicos_Radicais real
@attribute ImportanciaTopicos_Quadrigramas real

@attributePresencaSumario {TRUE,FALSE}

@data

FALSE,H1+H2+H3,0.6327,1,II,0.6275,0.352,C1+C2+C3,FALSE,0.788,0.6186,0.6331,TRUE
H1+H2,FALSE,0.9388,1,IF,0.6667,0.748,C1+C2+C3,FALSE,0.9068,0.8955,0.7348,TRUE
FALSE,H1+H2+H3,0.6939,0,MI,0.3529,0.644,FALSE,FALSE,0.9088,0.8175,0.3203,TRUE
FALSE,FALSE,0.5102,1,MF,0.2157,0.124,FALSE,FALSE,0.7242,0.6814,0.4452,TRUE
FALSE,FALSE,1,0,MI,0.6471,1,FALSE,FALSE,1,1,0.745,FALSE
FALSE,FALSE,0.4286,1,MF,0.0784,0.364,FALSE,FALSE,0.6156,0.5413,0.4993,FALSE
H3,FALSE,0.0204,0,MI,0,0.112,FALSE,FALSE,0.1056,0.1059,0.0755,TRUE
FALSE,FALSE,0.3878,0,MF,0.2745,0.412,FALSE,FALSE,0.5644,0.5686,0.332,TRUE
FALSE,FALSE,0.5102,0,MI,0.6863,0.652,FALSE,C1,0.6036,0.6015,0.586,FALSE
FALSE,FALSE,0.5306,1,MM,0.3922,0.364,FALSE,C1,0.5653,0.5866,0.6054,FALSE
FALSE,FALSE,0.2041,0,MF,0.1176,0.084,FALSE,C1,0.4119,0.4294,0.2539,FALSE
```

Figura 5-1 – Arquivo ARFF de Treino do SuPor-2

Com relação à arquitetura resultante, o SuPor-2 possui dois módulos, assim como o SuPor original: de treinamento e extração.

No módulo de treinamento do SuPor-2, as mudanças feitas consistiram, basicamente, na geração de arquivos *ARFF* de treinamento, como ilustrado na Figura 5-2. A saída desse módulo é um arquivo com os parâmetros do classificador. Se for usado o Flexible-Bayes (padrão do sistema), por exemplo, esse arquivo contém as probabilidades usadas pelo classificador na etapa de extração.

No módulo de extração (Figura 5-3), as entradas são o texto-fonte e a taxa de compressão desejada pelo usuário. A saída é o extrato gerado.

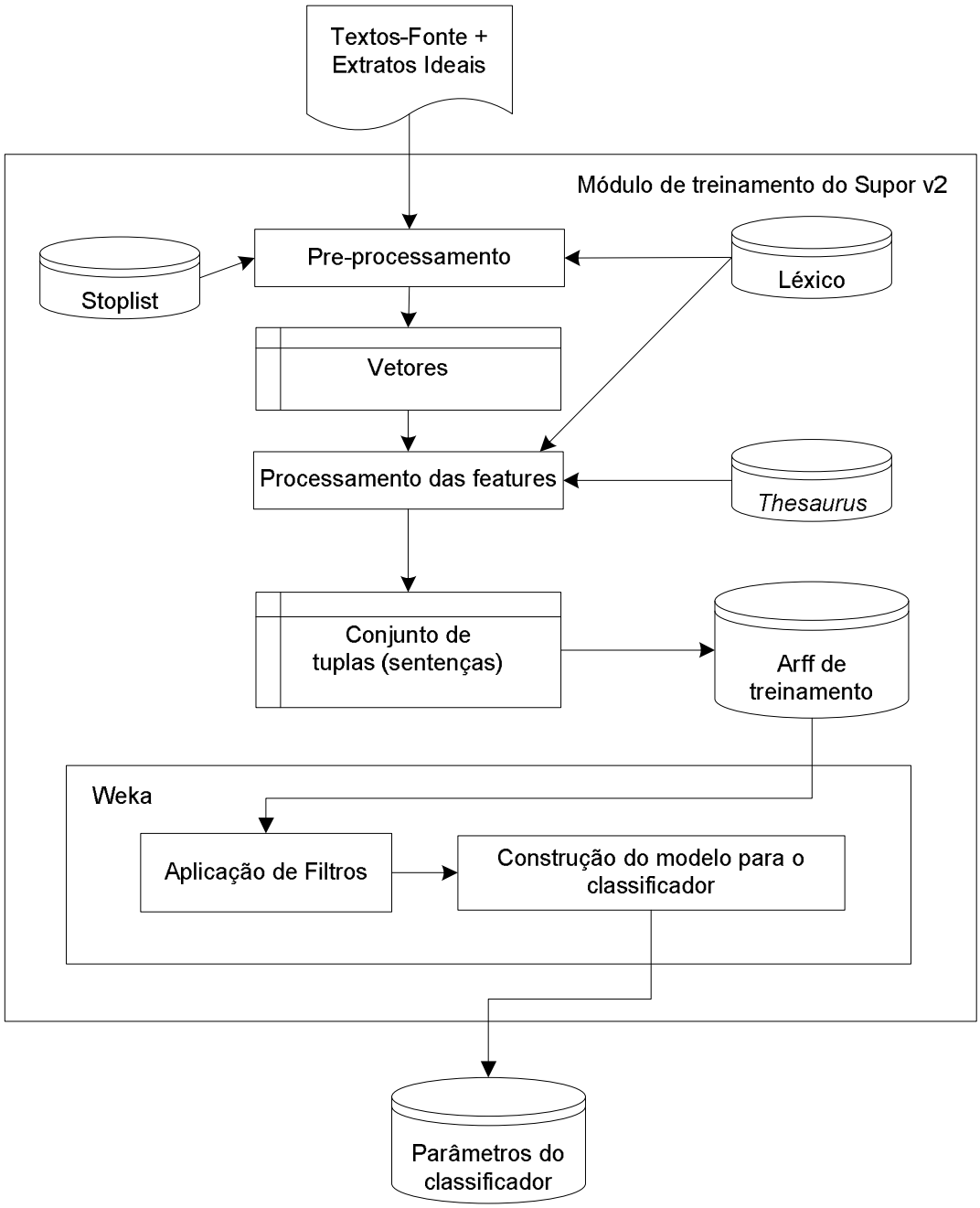


Figura 5-2 – Arquitetura do módulo de treinamento do SuPor-2

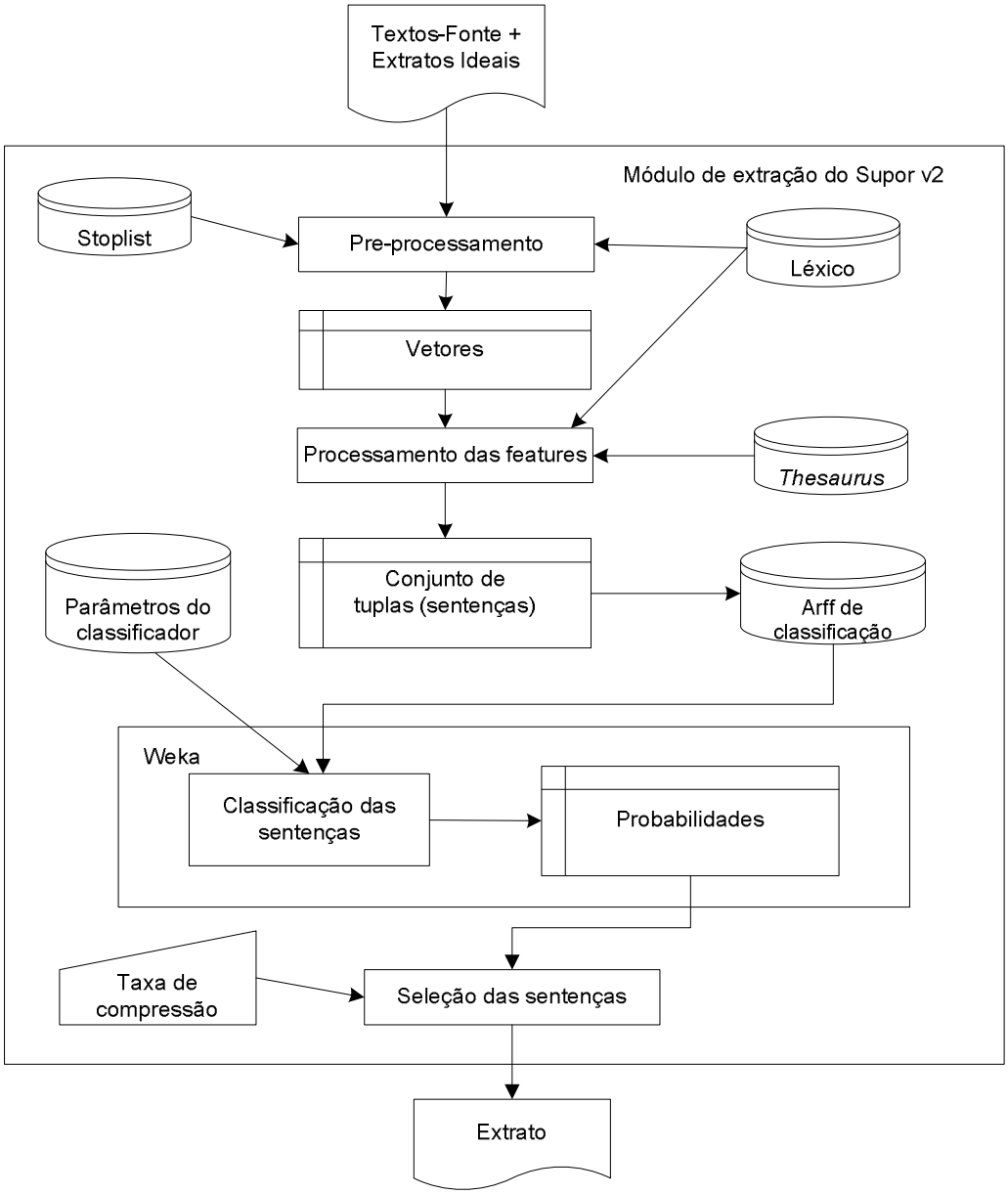


Figura 5-3 – Módulo de Extração do SuPor-2

A próxima seção descreve a maneira como se configurou empiricamente o modelo de aprendizado de máquina no SuPor-2. O acoplamento ao WEKA facilitou essa configuração, permitindo a avaliação de diferentes estratégias.

5.1.4 Determinação da melhor configuração para o SuPor-2

Através dos modelos disponíveis no WEKA, determinou-se a melhor configuração do SuPor-2 com relação ao seu modelo de aprendizado de máquina, isto é, seu modelo de classificação (Leite e Rino 2006). Foram selecionados para o

experimento os dois modelos Bayesianos (Naïve-Bayes tradicional e Flexible-Bayes) e também o classificador C4.5, bastante tradicional na área de Aprendizado de Máquina e cuja utilização foi sugerida em Módolo (2003). Ambos os classificadores são descritos na Seção 4.1. Cabe lembrar que o modelo original do SuPor é o Naïve-Bayes tradicional, que discretiza as características numéricas de modo permitir o cálculo das probabilidades. Alem disso, avaliou-se o uso da seleção automática de características por meio do método CFS (Seção 4.2.2.2).

O experimento foi feito de forma automática, gerando-se extratos com taxa de compressão de 30% para o corpus TeMário-2003 (Pardo e Rino 2003). Esse tamanho de extrato é compatível com os sumários manuais que acompanham o TeMário-2003.

Para evitar a necessidade de corpora separados para treino e teste das estratégias, foi utilizada a técnica *10-fold cross-validation* Por esta técnica, o conjunto de dados é separado em 10 subconjuntos disjuntos e em cada uma de 10 fases há um conjunto de dados de treino obtido concatenando 9 dos subconjuntos e um conjunto de dados de validação que usa o restante do subconjunto; o processo é repetido 10 vezes, permutando de forma circular os subconjuntos.

As medidas escolhidas para avaliação dos modelos foram as medidas de Precisão, Cobertura e *F-Measure*, descritas na Seção 2.3.1. Essas medidas foram escolhidas devido ao fácil cômputo, permitindo a avaliação de diferentes estratégias no WEKA. As estratégias e os resultados de cada uma são mostrados na Tabela 5-6. A configuração determinada como a melhor para o SuPor-2 foi destacada nesta tabela.

Tabela 5-6 – Avaliação de diferentes estratégias de aprendizado de máquina para o SuPor-2

Classificador	Emprego do CFS	Cobertura (%)	Precisão (%)	F-measure (%)
Flexible-Bayes	Não	43.9	47.4	45.6
	Sim	42.8	46.6	44.6
Naïve-Bayes tradicional	Não	42.2	45.8	43.8
	Sim	42.0	45.9	43.9
C4.5	Não	37.7	40.6	39.1
	Sim	40.2	43.8	41.9

Assim como na versão original do SuPor, o classificador Bayesiano obteve melhor desempenho. Particularmente, a variação Flexible-Bayes teve melhor

desempenho, provavelmente devido ao fato de tratar melhor características numéricas, que estão presentes no SuPor-2 (vide Seção 5.1).

Com relação ao desempenho bem inferior do C4.5, provavelmente é devido ao fato do modelo não ser muitas vezes um bom gerador de rankings (vide Seção 4.1.3).

Considerando o emprego do método CFS, ele não trouxe melhora de resultados ao ser combinado com os modelos Bayesianos. Quando é utilizado em conjunto com o C4.5, há apenas uma discreta melhora. Tais resultados sugerem que dado o fato de as características do SuPor-2 terem sua representação e poder de expressão melhorados em relação ao SuPor original (Seção 5.1.1), um conjunto mais estável de características pode ter sido obtido. Em outras palavras, o SuPor-2 pode ser menos dependente da escolha de características que o SuPor original. Para corroborar essa hipótese, comparou-se através de medidas de avaliação de características (Seção 4.2.2) a relevância das características do SuPor e do SuPor-2. O gráfico a seguir mostra essa comparação, onde é possível perceber um considerável aumento na relevância das características do SuPor-2 em relação ao SuPor.

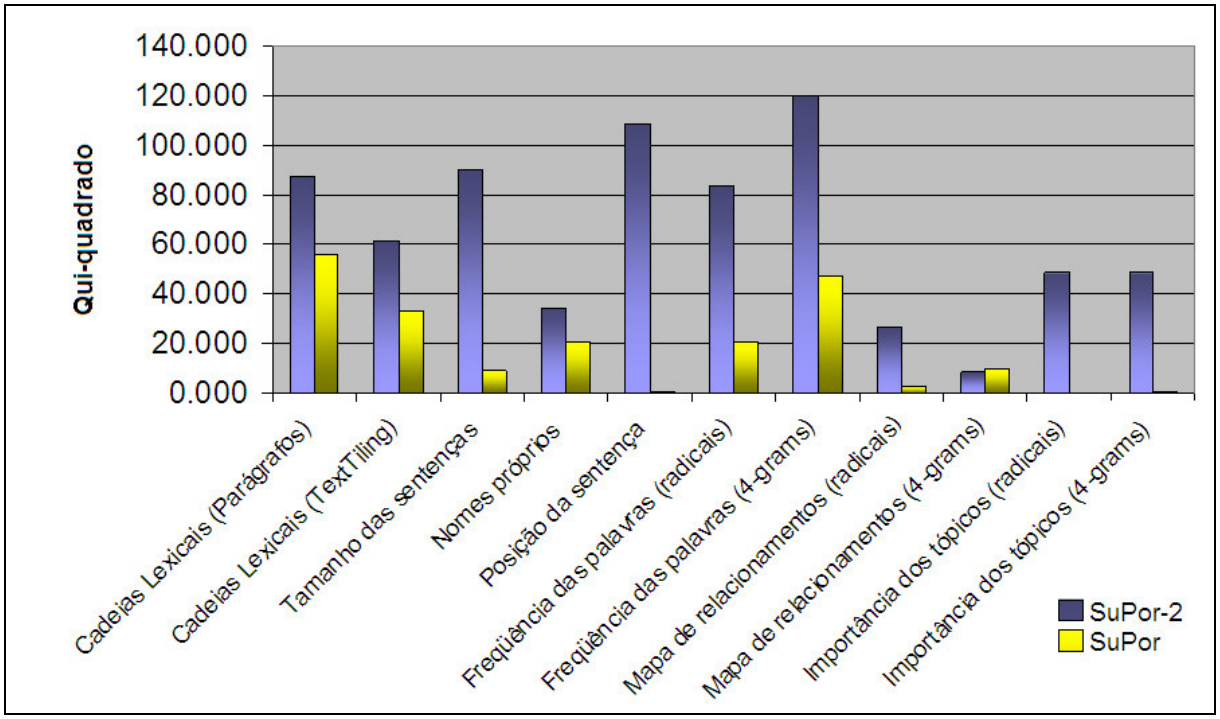


Figura 5-4 – Comparação da relevância das características do SuPor e SuPor-2 pela estatística qui-quadrado

5.2 Desenvolvimento de Modelos Baseados no TextRank

Considerando a abordagem baseada na Teoria dos Grafos, que este trabalho propõe explorar, um dos métodos recentes para SA baseado nessa teoria é o TextRank, já descrito na Seção 3.1.7. Seu principal ponto-forte é a portabilidade para diferentes línguas e domínios, inclusive já tendo sido aplicado ao Português pela autora do método (Mihalcea 2005).

Foram dois interesses principais que motivaram a utilização do modelo do TextRank neste trabalho:

a) Trata-se de um método de ponta da Teoria dos Grafos e que não é baseado em aprendizado de máquina. Assim, existe uma única característica sendo associada à relevância de uma sentença, em contraste com o já apresentado SuPor-2, onde existem várias. A investigação desse modelo poderia permitir a refutação da Hipótese 1 deste trabalho, de que a combinação de características é benéfica para a SA.

b) Apesar de alguns dos resultados obtidos com o método terem sido considerados bons pelos autores, o TextRank apresenta como desvantagem o fato de não considerar nenhum tipo de pré-processamento ou processamento linguístico. Neste trabalho, vislumbrou-se sua melhoria com a produção de duas variações do método com a incorporação de conhecimento linguístico. O foco das duas variações, descritas a seguir, foi modificar a maneira como o método realiza o cômputo da similaridade entre as sentenças. O processo geral de percurso no grafo não foi alterado, adotando-se o modelo original de grafos não-dirigidos.

5.2.1 Modelo TextRank+Stem+StopwordsRem

Ao se aplicar a Equação [8], apresentada na Seção 3.1.7, para cálculo das similaridades, somente casamentos exatos entre os termos são considerados. Uma vez que no Português existem muitas variações morfológicas e terminações flexionais para a maioria das palavras, casamentos semanticamente válidos podem ser ignorados.

Para se contornar o problema, foi utilizado neste modelo um *stemmer* para o Português do Brasil (Junior, Imamura et al. 2001), baseado no algoritmo de Porter (Porter 1980).

Foram também removidas *stopwords* do Português.

5.2.2 Modelo TextRank+Thesaurus

A segunda variação proposta do TextRank envolve o uso de um thesaurus para o Português do Brasil (Dias-da-Silva, Moraes et al. 2000), contemplando substantivos. A hipótese aqui é que a similaridade semântica das palavras envolvidas é também importante para melhorar a informatividade dos extratos produzidos pelo método.

Palavra: **acomodação** =
S={acomodação, instalação, } A={} &
S={acomodação, ajeitação, ajeitamento, arranjo, arrumação} A={} &
S={acomodação, aposento, cômodo, quarto } A={} &
S={acomodação, emprego, ocupação } A={} &
S={acomodação, comodismo, conformismo, desambição } A={inconformismo } &
S={acomodação, adaptação, adequação, apropriação } A={inadaptação, inadequação } &
S={acomodação, adaptação, adequação, ajustamento, conformação } A={desajuste, inadaptação }

Figura 5-5 – Exemplo de estrutura do thesaurus utilizado no modelo TextRank+Thesaurus

Embora a simples utilização do thesaurus não envolva profundas mudanças no método, existem algumas questões de projeto a serem consideradas:

a) Devem ser considerados apenas sinônimos ou antônimos também em adição a repetição lexical?

b) Como tratar a polissemia das palavras e desambiguar os sentidos?

c) Uma vez determinadas as relações semânticas, como elas devem ser ponderadas? Simplesmente considerá-las como de igual importância pode não necessariamente levar aos melhores resultados.

Considerando a questão (a), as relações de sinonímia, antonímia e repetição lexical foram todas levadas em conta, como sugerido por alguns autores (e.g., Barzilay e Elhadad 1999). Quanto a (b), decidiu-se não tratar o problema devido a falta de um processo efetivo de desambiguação para o Português do Brasil. Já quanto a (c), foram adotados os mesmos pesos propostos por Barzilay e Elhadad (1999) no método de Cadeias Lexicais. Para repetição e sinonímia, assume-se o peso 10. Para relações de antonímia, utiliza-se o peso 7.

5.3 Desenvolvimento de Modelos Baseados em Características de Redes Complexas e do SuPor-2 Combinadas

Seguindo-se a abordagem baseada em Grafos, outro trabalho que se considerou interessante para a SA foi o de Antiqueira et al. (2007), pois propõe um conjunto grande de medidas para SA baseadas na teoria de Redes Complexas, apresentadas na Seção 3.2.6. As Redes Complexas vem sendo cada vez mais estudadas devido à popularização das redes sociais na WWW.

A grande desvantagem dos modelos propostos por Antiqueira et al. (2007) como já citado na Seção 3.2.6, é que as medidas utilizadas são sempre exploradas isoladamente e não de forma combinada. Já na estratégia de votação chamada de CN-Voting (Antiqueira et al. 2007), a forma de combinação empregada não leva em conta as diferenças de relevâncias das características. Ou seja, não é feita a ponderação das características para sua combinação.

De forma a explorar de forma conjunta as medidas de redes complexas, buscou-se combinar de forma automática essas medidas por meio do aprendizado de máquina, utilizando-as como características para a SA. Diferente dos modelos baseados no TextRank (Seção 5.2), buscou-se aqui confirmar a Hipótese 1, de que a combinação de características é benéfica, em vez de refutá-la.

Seguiu-se a linha de trabalhos recentes como o de Wong et al. (2008), descrito na Seção 3.3.1.5, que argumentam que a utilização de conjuntos de características diversos é interessante para a SA. Além das próprias características

de redes complexas, explorou-se a união desse conjunto com características clássicas de SA e já exploradas no SuPor-2.

5.3.1 Características Exploradas e Análise de Relevância

O conjunto total de características exploradas é mostrado na Tabela 5-7. A primeira coluna indica o sistema ou abordagem teórica que originou a característica: SuPor-2 ou Redes Complexas (RC). Na terceira coluna, indica-se entre parênteses a eventual variação na característica. No caso das características do SuPor-2, essas variações são detalhadas na Tabela 3-2. No caso das características de redes, elas são descritas em Antiqueira (2007).

Dado o número total de características exploradas (37) ser grande, realizou-se uma avaliação estatística preliminar de quais características são mais promissoras. Essa avaliação buscou verificar se era possível verificar algum padrão na relevância das características e, assim, determinar manualmente, os subconjuntos mais adequados a explorar. A análise utilizou as medidas de avaliação individual descritas na Seção 4.2.2. Os resultados são mostrados no gráfico da Figura 5-6.

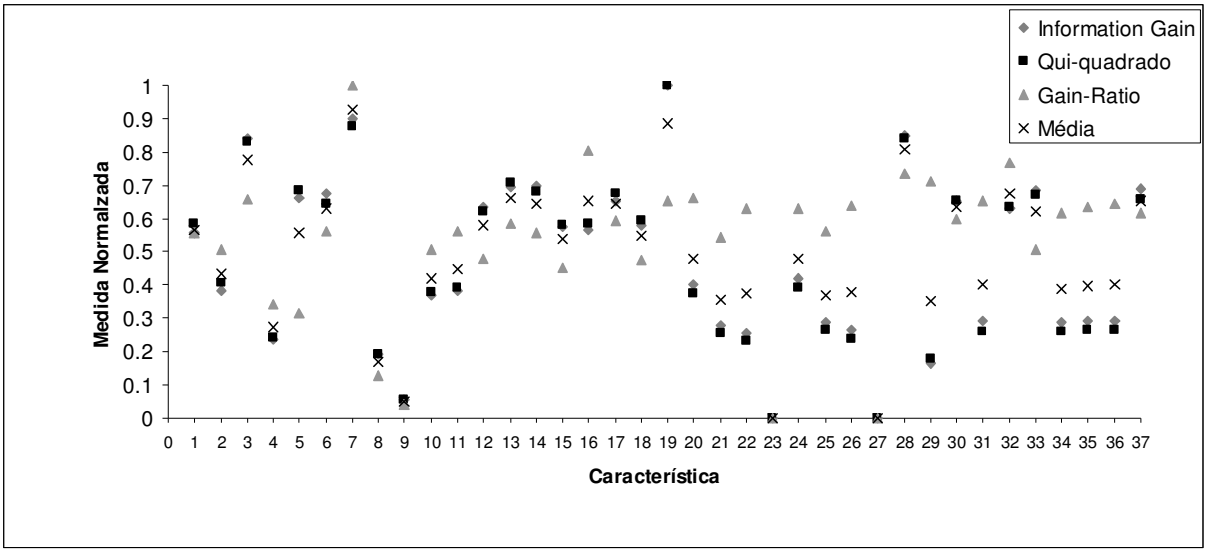


Figura 5-6 - Comparação da relevância das características

Tabela 5-7 – Características exploradas nos modelos baseados em características de RC e SuPor-2 combinadas

Fonte da Característica	Número	Nome da Característica
SuPor-2	1	Cadeias Lexicais (Divisão em dos tópicos pelos parágrafos)
SuPor-2	2	Cadeias Lexicais (Divisão em tópicos usando TextTiling)
SuPor-2	3	Tamanho da sentença
SuPor-2	4	Nomes próprios
SuPor-2	5	Posição
SuPor-2	6	Frequência das Palavras (pré-processamento usando *stemming*)
SuPor-2	7	Frequência das Palavras (pré-processamento usando quadrigramas)
SuPor-2	8	Mapa de Relacionamentos (pré-processamento usando *stemming*)
SuPor-2	9	Mapa de Relacionamento (pré-processamento usando quadrigramas)
SuPor-2	10	Importância dos Tópicos (pré-processamento usando *stemming*)
SuPor-2	11	Importância dos Tópicos (pré-processamento usando quadrigramas)
RC	12	Degree
RC	13	Degree (variação com ponderação)
RC	14	Clustering Coefficient
RC	15	Clustering Coefficient (variação com ponderação)
RC	16	Minimal Paths
RC	17	Minimal Paths (variação pesos complementares)
RC	18	Minimal Paths (variação pesos inversos)
RC	19	Locality Index
RC	20	Locality Index (modificada)
RC	21	Matching Index
RC	22	Dilation (nível 2)
RC	23	Dilation (nível 2, acumulativa)
RC	24	Dilation (nível 3)
RC	25	Dilation (nível 3, acumulativa)
RC	26	Dilation (nível 2, ponderada)
RC	27	Dilation (nível 2, ponderada, acumulativa)
RC	28	Dilation (nível 3, ponderada)
RC	29	Dilation (nível 3, ponderada, acumulativa)
RC	30	Hubs (ordenado pelo grau)
RC	31	Hubs (ordenado pela localidade)
RC	32	Hubs (ordenado pela localidade e com corte de grau)
RC	33	K-Cores (ordernado pela localidade)
RC	34	K-Cores (ordernado pelo grau)
RC	35	W-Cuts (ordernado pela localidade)
RC	36	W-Cuts (ordenado pelo grau)
RC	37	Communities

Como pode ser observado, em alguns casos há muita divergência entre os valores apontados pelas medidas na Figura 5-6. Não se pode dizer a princípio que as características de redes complexas são melhores que as do SuPor-2 e vice-versa.

Considerando a média, a melhor característica foi a 7 (Frequência das Palavras, com pré-processamento por quadrigramas). As duas piores foram as características 23 e 27 (*Dilation*). O resultado dessa avaliação em forma tabular é fornecido no Apêndice D.

A partir desse estudo, optou-se por aplicar o mesmo processo de seleção automática de características já empregado no SuPor-2 (Seção 5.1), o CFS. A utilização do CFS é consoante com a Hipótese 2 deste trabalho, de que a pré-seleção automática de características pode ser benéfica para a SA.

5.3.2 Os Modelos que Combinam Características do SuPor-2 e RC

Combinando-se os conjuntos de características (SuPor-2 ou RC), quatro diferentes classificadores e o uso ou não do método CFS, foram desenvolvidos os modelos listados na Tabela 5-8. O símbolo de união indica que o modelo utiliza as 37 características da Tabela 5-7, sendo 11 do SuPor-2 e 26 de RCs.

A escolha dos classificadores considerados levou em conta a análise feita na Seção 4.1. A utilização de vários classificadores pretendeu aprofundar e investigar melhor a Hipótese 3 proposta neste trabalho, relacionada a forma de combinação e ponderação das características.

Tabela 5-8 – Modelos com características de RC e SuPor-2 combinadas

Origem do Conjunto de Características	Classificador	Uso do CFS
RC	Flexible-Bayes	Não
RC	Flexible-Bayes	Sim
RC	C4.5	Não
RC	C4.5	Sim
RC	Regressão Logística	Não
RC	Regressão Logística	Sim
RC	SVM	Não
RC	SVM	Sim
SuPor-2[26]	Flexible-Bayes	Não
SuPor-2	Flexible-Bayes	Sim
SuPor-2	C4.5	Não
SuPor-2	C4.5	Sim
SuPor-2	Regressão Logística	Não
SuPor-2	Regressão Logística	Sim
SuPor-2	SVM	Não
SuPor-2	SVM	Sim
SuPor-2 ∪ RC	Flexible-Bayes	Não
SuPor-2 ∪ RC	Flexible-Bayes	Sim
SuPor-2 ∪ RC	C4.5	Não
SuPor-2 ∪ RC	C4.5	Sim
SuPor-2 ∪ RC	Regressão Logística	Não
SuPor-2 ∪ RC	Regressão Logística	Sim
SuPor-2 ∪ RC	SVM	Não
SuPor-2 ∪ RC	SVM	Sim

Na próxima seção, mostra-se a arquitetura geral desses modelos, que também utilizaram o WEKA para o aprendizado de máquina.

5.3.3 Arquitetura dos Modelos e Utilização do WEKA

A implementação dos modelos seguiu a mesma arquitetura utilizada no SuPor-2, com a utilização dos classificadores e do algoritmo CFS disponíveis no WEKA. Não houve reimplementação das medidas propostas por Antiqueira (2007). Todo o cômputo das características foi realizado pelos sistemas de Antiqueira (2007).

A Figura 5-7 seguinte ilustra o arquivo *ARFF* de treino usado, considerando o conjunto total de características SuPor-2 ∪ RC.

[26] O modelo dessa linha equivale a melhor configuração do SuPor-2 obtida na Seção 5.1.4.

```
@RELATION ObtencaoSumario
@attribute CadeiasLexicais_Paragrafos { FALSE, H1, H2, H3, H1+H2, H1+H3,H2+H3, H1+H2+H3 }
@attribute CadeiasLexicais_TextTiling { FALSE, H1, H2, H3, H1+H2, H1+H3,H2+H3, H1+H2+H3 }
@attribute TamanhoSentencas real
@attribute NomesProprios real
@attribute PosicaoParagrafos {II, IM, IF, MI, MM, MF, FI, FM, FF }
@attribute FrequenciaPalavras_Radicais real
@attribute FrequenciaPalavras_Quadrigramas real
@attribute MapaRelacionamentos_Radicais { FALSE, C1, C2, C3, C1+C2, C1+C3,C2+C3, C1+C2+C3 }
@attribute MapaRelacionamentos_Quadrigramas
                            { FALSE, C1, C2, C3, C1+C2, C1+C3,C2+C3, C1+C2+C3 }
@attribute ImportanciaTopicos_Radicais real
@attribute ImportanciaTopicos_Quadrigramas real
@ATTRIBUTE k-CORE-DEG NUMERIC
@ATTRIBUTE k-CORE-SEQ NUMERIC
@ATTRIBUTE CLUST-COEFF NUMERIC
@ATTRIBUTE COMMUNITIES NUMERIC
@ATTRIBUTE DILATIONS-DEG NUMERIC
@ATTRIBUTE DILATIONS-MOD NUMERIC
@ATTRIBUTE DILATIONS-SEQ NUMERIC
@ATTRIBUTE HUBS-E NUMERIC
@ATTRIBUTE HUBS-HRQ-E-2-CUMUL NUMERIC
@ATTRIBUTE HUBS-HRQ-E-2-NOCUMUL NUMERIC
@ATTRIBUTE HUBS-HRQ-E-3-CUMUL NUMERIC
@ATTRIBUTE HUBS-HRQ-E-3-NOCUMUL NUMERIC
@ATTRIBUTE HUBS-HRQ-W-2-CUMUL NUMERIC
@ATTRIBUTE HUBS-HRQ-W-2-NOCUMUL NUMERIC
@ATTRIBUTE HUBS-HRQ-W-3-CUMUL NUMERIC
@ATTRIBUTE HUBS-HRQ-W-3-NOCUMUL NUMERIC
@ATTRIBUTE HUBS-W NUMERIC
@ATTRIBUTE LOCALITY-INDEX-MOD NUMERIC
@ATTRIBUTE LOCALITY-INDEX NUMERIC
@ATTRIBUTE MATCHING-INDEX NUMERIC
@ATTRIBUTE PRUNING-DEG NUMERIC
@ATTRIBUTE PRUNING-SEQ NUMERIC
@ATTRIBUTE SP-1 NUMERIC
@ATTRIBUTE SP-compl NUMERIC
@ATTRIBUTE SP-inv NUMERIC
@ATTRIBUTE W-CLUST-COEFF NUMERIC
@ATTRIBUTE PresencaSumario {True,False}

@DATA
H1+H2+H3,  H1+H2,  0.5088,  0.5714,  II,  0.3725,  0.797,  C1+C2+C3,  C1+C2+C3,  0.7683,  0.7615,
0.1111111,  0.0370370,  0.5514706,  0.2962963,  0.1111111,  0.0740741,  0.0740741,  0.9444444,
0.7215190,  0.5882353,  0.7037037,  0,  0.73,  0.5454545,  0.7156863,  0,  0.8620690,  0.3703704,
0.9344692, 0.1481481, 0.1481481, 0.1111111, 0.4861111, 0.4651163, 0.4565538, 0.6125, True
...
```

Figura 5-7 – Exemplo de arquivo ARFF de treino considerando características do SuPor-2 e de Redes Complexas

5.4 Desenvolvimento de Modelos com Base em Ranking Nebuloso

Conforme exposto na Seção 4.3, sistemas nebulosos podem ser aplicados para se modelar o processo de decisão de relevância das sentenças. Em outras palavras, podem ser usados para combinar e ponderar características e, assim, permitir o aprofundamento da Hipótese 3 deste trabalho seguindo essa abordagem.

Pelo nosso conhecimento, são poucos os trabalhos que exploram para sumarização a utilização da teoria de conjuntos nebulosos. Uma exceção é o sistema de Kiani-B e Akbarzadeh-T (2006), já apresentado na Seção 3.3.2. Para tornar viável a inferência nebulosa, Kiani-B e Akbarzadeh-T construíram de forma automática a base de regras nebulosas, por meio de algoritmos genéticos.

Uma desvantagem apontada do modelo de Kiani-B e Akbarzadeh-T (2006) é que o processo de aprendizado, conduzido por meio de algoritmos genéticos, não utiliza medidas específicas e atuais de SA como função de *fitness*, que é responsável no conceito de algoritmos genéticos por indicar o quão boa uma solução candidata do problema é. Em outras palavras, é a função responsável por avaliar a aptidão do cromossomo[27] que codifica uma possível solução candidata.

O modelo de sumarização nebuloso desenvolvido neste trabalho tomou como base o trabalho de Kiani-B e Akbarzadeh-T (2006). Porém, também utilizou um artifício que vem sendo usado nas últimas TACs, vide o sistema de Galanis e Malakasiotis (2008), que é dirigir o processo de treino ou refinamento do modelo pela otimização de métricas específicas de SA, como as medidas ROUGE. Em outras palavras, procura-se ajustar o modelo para que ele produza sumários bem avaliados pela métrica em foco. Nesse caso, parte-se do princípio que se a medida de avaliação de sumários for adequada, o sumarizador deverá produzir bons textos. Se a medida de avaliação não for adequada, o sumário produzido poderá estar enviesado para características que não necessariamente são esperadas de um bom sumário.

A estratégia de desenvolvimento do modelo foi partir do trabalho já desenvolvido no SuPor-2, buscando-se alterar seu módulo de classificação por um classificador baseado em regras nebulosas, mas se mantendo, quando possível,

[27] No contexto do Aprendizado Evolutivo, um cromossomo é a representação em código de um indivíduo, que é uma possível solução para o problema.

suas principais características como sumarizador. O modelo de SA desenvolvido através desse sistema nebuloso foi chamado de SuPor-2 Fuzzy. A seguir, descrevem-se as principais adaptações necessárias no SuPor-2, as principais decisões de projeto e suas motivações.

5.4.1 Adaptações no Conjunto de Características

Grande parte dos métodos de classificação nebulosos trabalham usualmente com dados numéricos, não multinomiais. Isso porque características multinomiais são em essência discretas, em contraste com os números, que permitem a representação de incertezas (e.g., Klir e Yuan 1995).

Neste trabalho, optou-se, então, por seguir essa tendência, manipulando-se somente características numéricas. Como o conjunto original de características do SuPor-2 contém características discretas, foram realizadas as seguintes adaptações nas características abaixo:

a) **Posição da Sentença**. A representação categórica da posição de uma sentença dentro do parágrafo e do parágrafo dentro do texto foi substituída por duas características numéricas. A primeira indica a posição relativa da sentença dentro do parágrafo. A segunda, a posição relativa do parágrafo considerando o texto todo.

b) **Cadeias Lexicais**. Foram desconsideradas as duas características que sinalizam a presença de cadeias lexicais devido a natureza do método indicar apenas qual heurística seleciona a sentença, um dado essencialmente discreto, sem fornecer uma informação que possa ser modelada de forma natural por meio de um conjunto nebuloso.

c) **Importância dos Tópicos**. Em vez de se considerar, como na versão original, uma característica que indica a média harmônica entre a importância do tópico em que a sentença ocorre e a similaridade da sentença para o centróide do tópico, dividiu-se essa informação em duas características novas: a primeira que indica a importância do tópico e a segunda a similaridade da sentença em relação ao tópico em que ela figura.

d) **Mapa dos Relacionamentos.** Como as informações produzidas pelas características correspondentes a esse método são essencialmente categóricas (caminhos), substituíram-se essas características por uma nova, baseada num método que também utiliza a representação de um texto por grafos, mas que produz um valor numérico como saída. O método escolhido foi o TextRank, descrito na Seção 3.1.7 e já explorado nos modelos da Seção 5.2. A implementação escolhida utilizou *stemming* e remoção de *stopwords*, sendo equivalente ao modelo TextRank+Stem+StopwordsRem.

A tabela a seguir resume as mudanças efetuadas nas características do SuPor-2. Observe-se que foi mantido o mesmo número de 11 características, considerando-se as variações de pré-processamento indicadas entre parênteses.

Tabela 5-9 – Características utilizadas no modelo nebuloso

Característica	Descrição	Domínio
C1	Tamanho da Sentença	[0, 1]
C2	Nomes Próprios	[0, 1]
C3	Posição Relativa da Sentença no Parágrafo	[0, 1]
C4	Posição Relativa do Parágrafo no Texto	[0, 1]
C5	TextRank	[0, 1]
C6	Frequência das Palavras (Stemming)	[0, 1]
C7	Frequência das Palavras (Quadrigramas)	[0, 1]
C8	Similaridade da Sentença com Centróide do Tópico (Stemming)	[0, 1]
C9	Similaridade da Sentença com Centróide do Tópico (Quadrigramas)	[0, 1]
C10	Importância do Tópico (Stemming)	[0, 1]
C11	Importância do Tópico (Quadrigramas)	[0, 1]

5.4.2 Base de Conhecimento Nebulosa

Como já citado na Seção 4.3, o projeto de um sistema nebuloso passa geralmente pela construção de uma base de conhecimento nebulosa, que consiste de dois componentes:

a) Os conjuntos nebulosos que devem ser modelos para se representar as variáveis (em nosso caso, características) relevantes;

b) O conjunto de regras nebulosas que irá guiar o processo de inferência nebuloso.

5.4.2.1 Modelagem dos Conjuntos Nebulosos

Optou-se pela definição manual dos conjuntos nebulosos seguindo uma abordagem tradicional de representação por triângulos (Eberhart e Shi 2007). Três triângulos uniformes representam 3 conjuntos nebulosos possíveis (Baixo, Médio ou Alto) para cada característica. Ou seja, para cada características de sumarização existem três "conceitos" que poderão ser atribuídos: Baixo, Médio ou Alto.

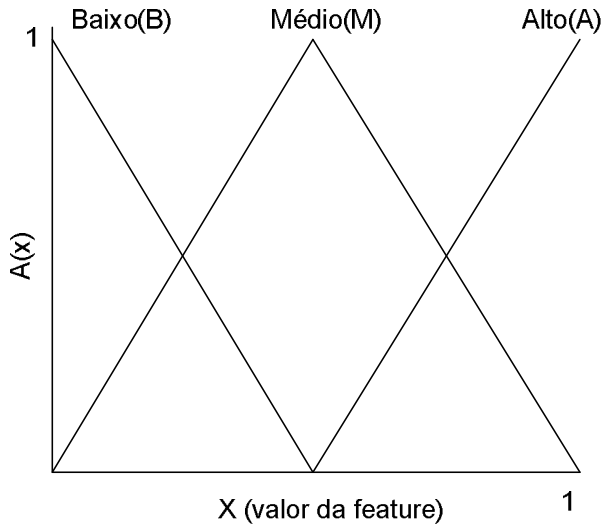


Figura 5-8 – Representação dos conjuntos nebulosos

O grau como as sentenças estão relacionadas aos conceitos Baixo, Médio ou Alto de cada características vai ser atribuído em função do valor que a sentença apresenta para a característica. As equações seguintes mostram as funções de pertinência para cada conceito:

$$Baixo(x) = \begin{cases} 1 - \dfrac{x}{0.5}, x \le 0.5 \\ 0, x > 0.5 \end{cases}$$

$$Médio(x) = 1 - \frac{|\, x - 0.5 \,|}{0.5}$$

$$Alto = \begin{cases} \dfrac{x - 0.5}{0.5}, x \geq 0.5 \\ 0, x < 0.5 \end{cases}$$

5.4.2.2 Geração do Conjunto de Regras

Uma abordagem comum no projeto de sistemas nebulosos baseados em regras é através da geração manual por especialistas de domínio. Isto pode ser bastante complicado, especialmente quando o número de variáveis for grande e o problema complexo. Outra abordagem é por meio da utilização de algoritmos genéticos, como adotado por Kiani-B e Akbarzadeh-T (2006). Essa última abordagem vem se tornando mais popular devido aos bons resultados proporcionados e por não demandar o trabalho manual de definição das regras (Eberhart e Shi 2007).

Neste trabalho, optou-se por seguir uma abordagem totalmente automática na geração de regras. Foi utilizado, para isso, um algoritmo genético seguindo a chamada abordagem de Pittsburgh (Eberhart e Shi 2007), que representa uma base de regras completa num cromossomo. Isto é, cada cromossomo representa uma base de regras completa. A contrapartida da abordagem de Pittsburgh é a de Michigan, que representa por meio de cada cromossomo uma única regra. Ela foi deixada de lado neste trabalho por se entender que tem pior desempenho computacional quando utilizada com funções de *fitness* complexas, como as métricas de SA.

O tamanho da população foi definido em 200 cromossomos. A população inicial foi gerada a partir da combinação de 2000 regras aleatórias e 2000 regras geradas pelo método de (Wang e Mendel 1992).

5.4.2.3 Codificação do Cromossomo

Pela abordagem de Pittsburgh, cada cromossomo representa um conjunto de regras completo. Por decisão de projeto, cada cromossomo possui 100 regras. Portanto, o tamanho do vetor cromossômico tem 1200 posições. Esse número é obtido multiplicando-se o número regras codificadas, que é 100, por 12 (11 posições para o antecedente e 1 para a classe).

A representação dos conjuntos foi feita a partir de números inteiros, da seguinte forma:

- Para o antecedente:
  - 0 indica que a característica não aparece no antecedente ("*dont-care*");
  - 1 a 3 representam os conjuntos Baixo, Médio e Alto, respectivamente.
- Para a classe (consequente):
  - 0 indica a classe *False* (*PresencaSumario = False)*
  - 1 indica a classe *True* (*PresencaSumario = True)*

A figura a seguir ilustra um segmento parcial de cromossomo que codifica a regra: *Se C1 é Alto e C2 é Médio e C3 é Baixo então a classe é True.*

3	2	1	0	0	0	0	0	0	0	0	1	...

Figura 5-9 – Exemplo de Representação de um Cromossomo

5.4.2.4 A Função de Fitness e o Método de Seleção

Foi adotada a medida ROUGE-2 como função de *fitness*. A escolha da ROUGE deve-se ao fato de a ferramenta ser bastante usada em trabalhos acadêmicos e competições de SA. Além disso, seus autores relatam boa correlação com as avaliações humanas, conforme já citado na Seção 2.3.2.

Para se avaliar um cromossomo, isto é, um conjunto de regras nebulosas, as regras codificadas são utilizadas em conjunto com o classificador nebuloso descrito na Seção 5.4.3. Após a geração dos extratos, o valor de aptidão (*fitness*) da base de regras é a medida ROUGE-2 dos extratos produzidos por essa base.

Uma vez avaliados os cromossomos, a probabilidade de esse cromossomo permanecer na próxima geração será proporcional ao valor da função de aptidão (ROUGE). Esse método é conhecido como "roleta", uma vez que cada cromossomo ocupa um espaço na roleta proporcional ao seu valor de aptidão, conforme Figura

5-10. Cada vez que “a roleta vira”, indivíduos com maior valor de aptidão têm chance de ser selecionados mais vezes. Isso tende a facilitar a convergência uma vez que haverá dominância dos indivíduos mais aptos desde as primeiras gerações.

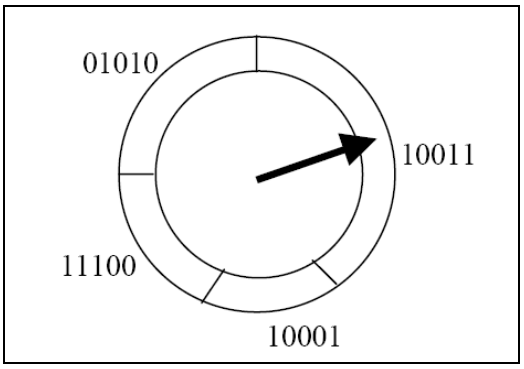


Figura 5-10 – Método da roleta

5.4.2.5 Cruzamento e Mutação

Ao longo do processo evolucionário do algoritmo genético, operadores de cruzamento e mutação são utilizados com a finalidade de se derivar novos indivíduos. A taxa de cruzamento foi definida em 90% e a de mutação em 1%. Os operadores são descritos a seguir:

a) **Cruzamento.** Utilizou-se o cruzamento simples (e.g.,Herrera, Lozano et al. 1993) em que o ponto de cruzamento foi escolhido aleatoriamente entre as codificações de cada regra. Por este método, a partir do ponto escolhido são trocas as informações genéticas dos pais, conforme figura:

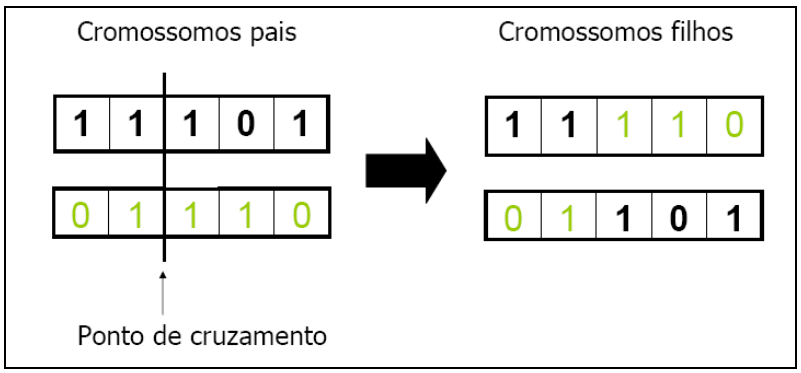


Figura 5-11 – Método de cruzamento genético

b) **Mutação.** Utilizou-se uma mutação não-uniforme (e.g.,Herrera, Lozano et al. 1993) com no máximo 10% das posições do cromossomo sendo alteradas. Isto é, determina-se aleatoriamente o número máximo de alterações que o cromossomo irá sofrer, de 0 até 120 (10% de 1200). Cada posição selecionada, então, para mutação é atualizada conforme as seguintes equações:

$$c_k' = c_k + \Delta(0, LS(c_k) - c_k) \text{ se um número aleatório uniforme for 0}$$

$$c_k' = c_k + \Delta(0, c_k) \text{ se um número aleatório uniforme for 1}$$

em que:

c_k é o valor atual do gene *k;*

c_k' é o novo valor para o gene *k*;

$LS(c_k)$ é 3 se c_k for um gene que codifica uma *característica*. Se codificar uma classe é 1;

$\Delta(a,b)$ é um número aleatório uniforme *a* e *b*.

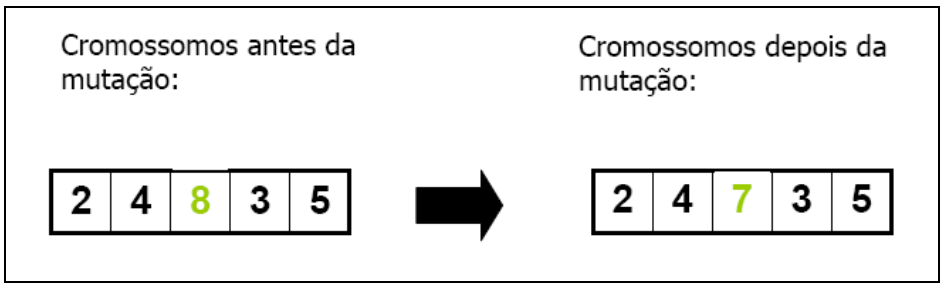


Figura 5-12 – Método de mutação genética

5.4.2.6 Algoritmo Genético

O trecho de pseudocódigo seguinte mostra o fluxo geral do processo de geração da base de regras ao longo do processo evolucionário. *P(t)* indica a população corrente na geração *t*. O número máximo de iterações foi definido como 1000 gerações.

```
      Início
      t=0;
      Iniciar P (t) a partir do Método de Wang-Mendel e de regras
Aleatórias;
      Avaliar P(t);
      Enquanto (t < MAXIMO_ITERACOES) Faça
       Início
          t=t+1;
          Seleção por roleta de P(t) a partir P(t-1);
          Aplicar Cruzamento em P(t);
          Aplicar Mutação em P(t);
          Avaliar P(t) sumarizando-se os textos do corpus de treinamento,
para cada base de regras, e avaliando na ferramenta ROUGE;
       Fim
      Fim
```

5.4.3 Modelo de Classificação Nebuloso

O modelo de classificação de sentenças nebuloso é utilizado em conjunto com as regras obtidas através do algoritmo genético para se gerar o ranking de sentenças por relevância. O modelo proposto para determinar a importância (*I*) de uma sentença *s* e é baseado na seguinte equação:

$$I(s) = \sum_{i|\exists R_i \to True} \omega(R_i, s) - \sum_{i|\exists R_i \to False} \omega(R_i, s) \qquad [21]$$

A fórmula considera a contribuição de cada regra (*R*) no cálculo. Regras que indicam que a sentença *s* não deve figurar no extrato ($R_i \rightarrow False$) tem viés negativo. A função ω indica a força de ativação de regras dadas as características da sentença *s*. Em outra palavras, mede o grau de compatibilidade entre o antecedente da regra e o vetor de características de *s*. Isto é feito calculando a função mínimo entre o conjunto calculado a partir do grau de pertinência de cada característica da sentença com o antecedente da regra (Klir e Yuan 1995).

Como exemplo, considerando-se os conjuntos nebulosos adotados, mostra-se o cálculo da importância das sentenças:

Exemplo 1: uma única regra

a) Regra nebulosa: suponha que tenhamos a regra abaixo

Se *C1 é Médio e C2 é Médio e C3 é Médio e C4 é Médio e F5 é Alto e F6 é Médio e F7 é Médio e F8 é Médio s F9 é Médio e F10 é Alto e F11 é Alto* então a classe é *True*.

(b) vetor de características de *s,* ou seja, o resultado do cômputo das 11 características de sumarização:

[C1=0.5, C2=0.5, C3=0.5, C4=0.5, C5=0.6, C6=0.5, C7=0.5, C8=0.5, C9=0.5, C10=1, C11=1]

Através das funções de pertinência triangulares, calcula-se o grau de pertinência aos conjuntos:

Médio(0.5) = 1.0, *Alto*(0.6) = 0.2 e *Alto* (1.0) = 1.0

A função mínimo sinaliza o grau de ativação da regra pela sentença *s*, indicando sua importância.

$\omega = min\ (1, 1, 1, 1, 0.2, 1, 1, 1, 1, 1, 1) = 0.2$

Como aqui consideramos apenas uma regra, a importância será o próprio grau de compatibilidade:

$I(s) = 0.2$

Exemplo 2: duas regras

a) Regras nebulosas: suponha que tenhamos as regras abaixo

R1: Se C1 é Médio e C5 é Alto então a classe é True
R2: Se C1 é Baixo e C5 é Baixo então a classe é False

(b) vetor de características de *s,* ou seja, o resultado do cômputo das 11 características de sumarização:

[C1=0.4, C2=0.6, C3=0.5, C4=0.5, C5=0.6, C6=0.5, C7=0.5, C8=0.5, C9=0.5, C10=1, C11=1]

Através das funções de pertinência triangulares, calcula-se o grau de pertinência aos conjuntos, conforme ilustra a Figura 5-13:

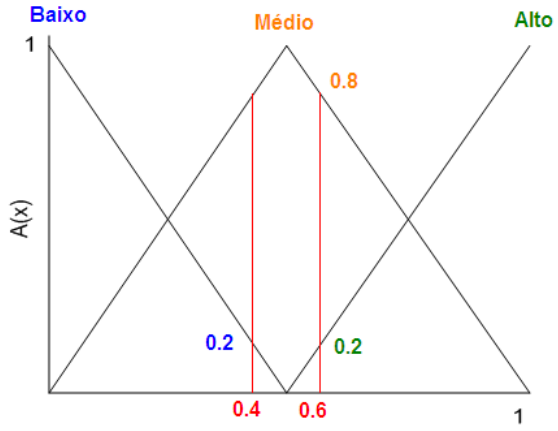


Figura 5-13 – Cálculo dos graus de pertinência para o exemplo 2

A função mínimo sinaliza o grau de ativação de cada regra pela sentença *s*:

$$\omega(R1,s) = min\ (0.8, 0.2) = 0.2$$
$$\omega(R2,s) = min\ (0, 0) = 0$$

Como aqui consideramos apenas uma regra, a importância será a diferença entre o grau de compatibilidade com a regra que recomenda a sentença e a regra que não recomenda a sentença:

$$I(s) = \omega(R1,s) - \omega(R2,s) = 0.2 - 0 = 0.2$$

5.4.4 Arquitetura do SuPor-2 Fuzzy

Assim como o SuPor-2, o SuPor-2 Fuzzy tem uma fase inicial de treino que tem o propósito construir uma base de conhecimento nebulosa a partir de um corpus de treino. A Figura 5-14 ilustra essa fase, desde a geração da base de dados até a geração da base de regras final.

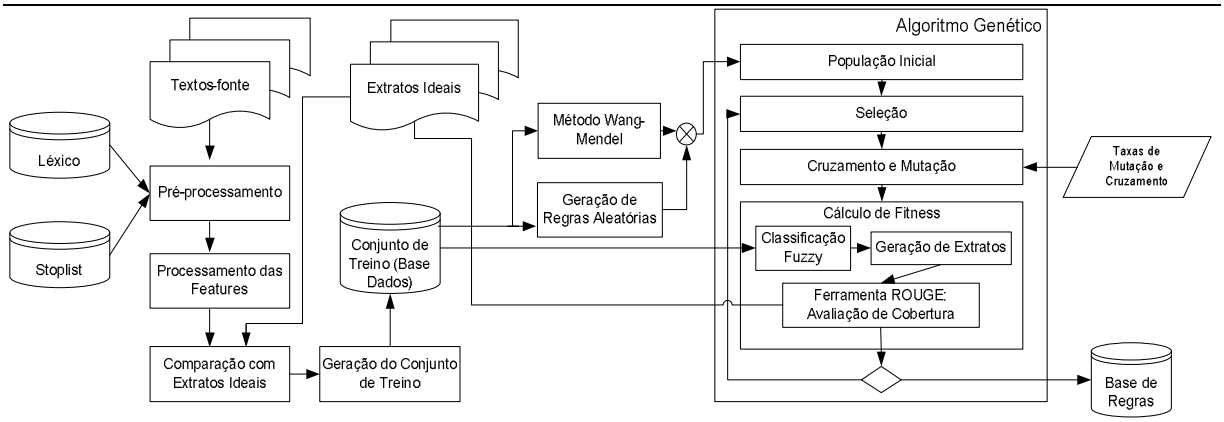


Figura 5-14 – Fase de Treino do SuPor-2 Fuzzy

Para armazenagem da base de dados, isto é, o conjunto de textos processados com suas características, adotou-se o mesmo formato *ARFF* do WEKA, adotado também no SuPor-2. Como exemplo, a figura a seguir representa um *ARFF* de treino real do SuPor-2 Fuzzy:

```
@relation ObtencaoSumario
@attribute PosicaoParagrafo real
@attribute PosicaoSentenca real
@attribute TamanhoSentencas real
@attribute NomesProprios real
@attribute TextRank real
@attribute FrequenciaPalavras_Radicais real
@attribute FrequenciaPalavras_Quadrigramas real
@attribute ImportanciaTopico_Radicais real
@attribute SimilaridadeCentroide_Radicais real
@attribute ImportanciaTopico_Quadrigramas real
@attribute SimilaridadeCentroide_Quadrigramas real
@attribute PresencaSumario
@data
0 , 0 , 0.3171 , 0 , 0.6134 , 0.4091 , 0.5541 , 1 , 0.414 , 1 , 0.4815 , FALSE
0 , 0.0294 , 0.2439 , 0 , 0.4805 , 0.4091 , 0.4069 , 1 , 0.3821 , 1 , 0.4941 , FALSE
0 , 0.0588 , 0.3902 , 0 , 0.8496 , 0.8409 , 0.7965 , 1 , 0.4569 , 1 , 0.6148 , TRUE
0.0769 , 0.0882 , 0.561 , 0 , 0.4935 , 0.5682 , 0.6364 , 1 , 0.6261 , 1 , 0.8595 , FALSE
0.0769 , 0.1176 , 0.5854 , 0 , 0.4727 , 0.2727 , 0.2208 , 1 , 0.6222 , 1 , 0.6961 , TRUE
0.1538 , 0.1471 , 0.3659 , 0 , 0.8631 , 0.9318 , 0.8788 , 1 , 0.5231 , 1 , 0.5816 , FALSE
0.1538 , 0.1765 , 0.2683 , 0 , 0.4318 , 0.4545 , 0.0606 , 1 , 0.423 , 1 , 0.5145 , FALSE
```

Figura 5-15 – Arquivo ARFF de Treino do SuPor-2 Fuzzy

Já na fase de geração de extratos, conforme figura seguinte, o fluxo é basicamente o mesmo do SuPor-2, distinguindo-se o modo como as sentenças são classificadas e as características utilizadas.

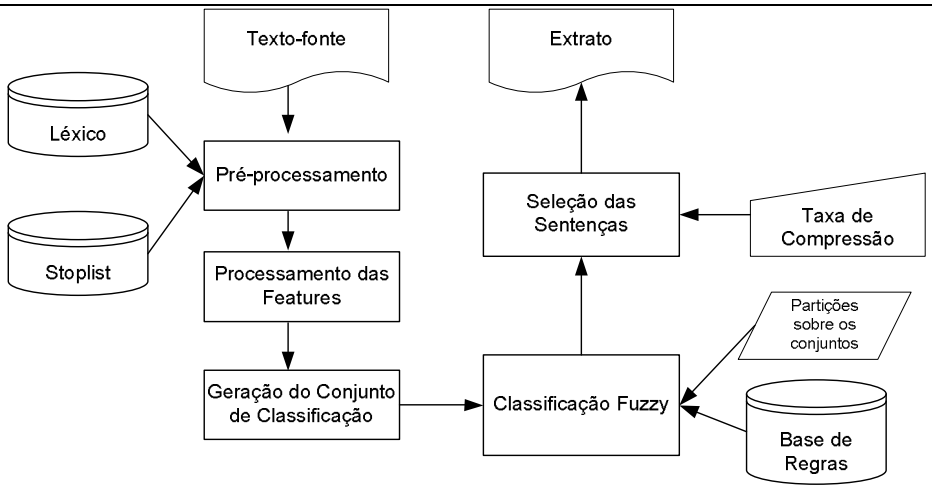


Figura 5-16 – Fase de geração de extratos do SuPor-2 Fuzzy

5.5 Síntese dos Modelos Desenvolvidos

Neste capítulo foram descritos os modelos e protótipos de SA desenvolvidos com base nas três grandes abordagens exploradas:

I. Estatísticas textuais;
II. Métodos baseados em grafos;
III. Métodos baseados em aprendizado de máquina.

Cada protótipo desenvolvido buscou combinar abordagens e características distintas de modo a tratar os dois problemas principais focados neste trabalho, apresentados no Capítulo 1:

Problema 1. Quais características utilizar?
Problema 2. Como combinar e ponderar as características escolhidas?

As duas primeiras abordagens — de uso de Estatísticas Textuais (I) e de uso de Grafos (II) — concentram-se no Problema 1. Através delas, derivam-se medidas, chamadas aqui de características, com o objetivo de considerar diversos fatores na escolha das sentenças, como por exemplo a posição das sentenças e a presença de elos coesivos indicados por medidas de grafos.

Já a abordagem – de uso de Aprendizado de Máquina (III)– foca no Problema 2. Assim, é utilizada para ponderar e combinar características, construindo o modelo de decisão responsável por julgar a relevância das sentenças.

A Tabela 5-10 mostra os protótipos desenvolvidos, as abordagens exploradas em cada um deles (I, II ou III) e a seção deste texto onde o modelo é descrito. O protótipo 28 foi construído apenas como *baseline* para as avaliações que serão descritas no Capítulo 6. Esse protótipo utiliza as mesmas características do SuPor-2 Fuzzy e com o mesmo classificador do SuPor-2 original.

Tabela 5-10 – Protótipos desenvolvidos e abordagens exploradas

Protótipo	Abordagens Exploradas	Seção
1. SuPor-2	I, II, III	5.2
2. TextRank+StemmingStopRem 3. TextRank+Thesaurus	II	5.3
4-27. Protótipos baseados em redes complexas e características do SuPor-2 combinadas	I, II, III	5.4
28. SuPor-2 modificado (mesmas caract. do SuPor-2 fuzzy) 29. SuPor-2 Fuzzy	I, II, III	5.5

A Tabela 5-11 classifica os 29 protótipos quanto ao tipo de modelo de ranking de sentenças (1ª coluna) e a utilização ou não de pré-seleção de características (2ª coluna). O tipo de ranking é um dos fatores chave que o Problema 2 apresenta. Nesse trabalho, a exploração de vários modelos de ranking foi feita de modo a permitir a investigação da Hipótese 3, de que a forma de combinação e ponderação de características tem influência significativa na qualidade dos sumários. Já a variação da pré-seleção de características, teve o objetivo de investigar a Hipótese 2 deste trabalho.

Tabela 5-11 – Protótipos desenvolvidos e modelos de ranking de sentenças

Protótipo	Modelos de Ranking	Pré-seleção de características
1. SuPor-2	Bayes e variações C4.5 também possíveis	Sim, método CFS
2. TextRank+StemmingStopRem 3. TextRank+Thesaurus	-	-
4-27. Protótipos baseados em redes complexas e características do SuPor-2 combinadas	Flexible-Bayes C4.5 Regressão SVM	Sim, método CFS
28. SuPor-2 modificado (mesmas caract. Do SuPor-2 fuzzy) 29. SuPor-2 Fuzzy	Classificação Nebulosa	Não

A Tabela 5-12 relaciona os protótipos e o número máximo de características que cada um pode empregar. O número indicado de características é máximo, pois o modelo pode empregar um método de pré-seleção de características, conforme apontado na Tabela 5-11. A utilização de conjuntos de características com tamanho e origem bem diferentes teve o objetivo de investigar a Hipótese 1 deste trabalho, que propõe que a utilização de conjuntos de características diversas pode permitir a consideração de diferentes fatores ou pontos de vista na análise da relevância das sentenças e, assim, levar a melhores extratos. Nos casos extremos, foram explorados modelos de SA com uma única característica até modelos de SA com 37 características.

Tabela 5-12 – Protótipos desenvolvidos e número máximo possível de características

Protótipo	Número máximo de características
1. SuPor-2	11
2. TextRank+StemmingStopRem	1
3. TextRank+Thesaurus	1
4-27. Protótipos baseados em redes complexas e características do SuPor-2 combinadas	37 (26 + 11)
28. SuPor-2 modificado (mesmas caract. do SuPor-2 fuzzy) 29. SuPor-2 Fuzzy	11

O próximo capítulo apresenta as propostas de avaliação dos 29 protótipos e sua correspondente análise.

Capítulo 6

EXPERIMENTOS DE AVALIAÇÃO

Nesta seção descrevem-se e discutem-se os experimentos de avaliação automática dos modelos desenvolvidos.

6.1 Arcabouço Geral dos Experimentos Realizados

Neste trabalho foram feitas avaliações buscando-se comparar entre si os diversos modelos propostos e também com outros sumarizadores existentes. Devido ao número de modelos considerados, o foco das comparações foi objetivo e não subjetivo.

Para isso, foram conduzidas avaliações intrínsecas automáticas. Nessas avaliações, os extratos produzidos pelos modelos foram comparados com sumários de referência manuais (*gold-standards)*. Segundo Spärck-Jones (2007) esse tipo de avaliação é interessante, pois considera de forma implícita o propósito dos sumários. Se os sumários manuais forem construídos visando propósitos particulares, a comparação com os sumários automáticos tenderá a levar em conta a adequação a esses propósitos. Além disso, Spärck-Jones (2007) também apontam que esse tipo de avaliação tem sido a mais utilizada e mais operacionalmente viável nos últimos anos.

As avaliações automáticas conduzidas foram exclusivamente *black-box*[28], sem se considerar estados internos dos algoritmos de aprendizado de máquina empregados. Uma avaliação *glass-box* automática seria possível, por exemplo, se

[28] Ver Seção 2.2 sobre avaliações glass-box e black-box.

fossem utilizados mecanismos automáticos de geração de regras a partir dos classificadores produzidos (e.g., Witten e Frank 2005). Essas regras poderiam permitir, em teoria, o estudo das fronteiras de decisão delineadas pelos classificadores empregados. Entretanto, devido ao número de modelos avaliados ser grande, essa análise adicional traria uma complexidade maior e, por isso, não foi focada tendo em vista os objetivos principais deste trabalho.

Sempre que possível, a configuração dos experimentos foi ajustada de modo a permitir a comparação com resultados publicados por outros autores. Por isso, existe variação de configuração dos experimentos conforme o subconjunto de modelos avaliados, considerando corpora, taxa de compressão, etc. A Seção 6.1.1 descreve a divisão que foi seguida nos experimentos.

6.1.1 Proposta de Experimentos

Os experimentos foram conduzidos em 5 fases, de forma a avaliar primeiramente subconjuntos de modelos para posteriormente realizar uma avaliação final com os modelos mais promissores. A divisão é como segue:

Experimento 1 (Seção 6.2): trata da avaliação do SuPor-2, descrito na Seção 5.1, e em sua comparação com outros sistemas, incluindo o próprio SuPor original. Nesta fase, verificou-se a efetividade das modificações introduzidas no SuPor-2 em relação ao SuPor original. Além disso, os resultados obtidos também originaram evidências para verificação da Hipótese 3 deste trabalho, que propõe que a forma de combinação e ponderação das características tem influência significativa na qualidade dos sumários.

Experimento 2 (Seção 6.3): trata da avaliação dos modelos construídos com base no método TextRank (Seção 5.2). Os resultados produzidos nesta fase também serviram como evidência para verificação da Hipótese 1 proposta neste trabalho, de que a combinação de características distintas é benéfica para a SA;

Experimento 3 (Seção 6.4): trata da avaliação dos modelos que combinam características de redes complexas com características do SuPor-2, descritos na Seção 5.3. Os resultados desta fase serviram como base também para a verificação das três hipóteses propostas neste trabalho;

Experimento 4 (Seção 6.5): trata da avaliação do modelo de SA baseado em ranking nebuloso de sentenças, descrito na Seção 5.4. Os resultados obtidos nesta etapa serviram como evidência na verificação da Hipótese 3 deste trabalho;

Experimento 5 (Seção 6.6): esta última etapa objetivou determinar o melhor modelo de SA, a partir do conjunto dos modelos mais promissores indicados nos experimentos anteriores.

6.1.2 Corpora Utilizados

Os corpora utilizados para avaliação dos modelos desenvolvidos foram todos de textos jornalísticos em Português do Brasil e incluem tanto os textos originais quanto os sumários de referência construídos manualmente. São eles:

a) **TeMário-2003** (Pardo e Rino 2003). Compreende um conjunto de 100 textos jornalísticos extraídos da Folha de São Paulo e do Jornal do Brasil. Um sumário manual construído por especialista humano acompanha cada texto-fonte, com tamanho entre 25% e 30%, em relação ao tamanho do texto original. O corpus também possui extratos ideais de referência, construídos automaticamente a partir dos sumários manuais pela ferramenta GEI (Pardo e Rino 2004). Ao todo, o corpus possui 2940 instâncias (sentenças) rotuladas.

b) **Summ-it** (Collovini et al. 2007). Compreende um conjunto de 50 textos de divulgação científica, com tamanho variando de 127 a 654 palavras. Cada texto também é acompanhado de seu respectivo sumário de referência construído manualmente. O corpus também possui extratos ideais de referência, construídos automaticamente a partir dos sumários manuais pela ferramenta GEI (Pardo e Rino 2004). Ao todo, o corpus possui 851 instâncias (sentenças) rotuladas.

c) **TeMário-2006** (Maziero et al. 2007). Construído nos mesmos moldes do TeMário original, é, na verdade, um complemento daquele, porém agora com 151 textos extraídos do jornal on-line Folha de São Paulo, de diversos cadernos. Ao todo, o corpus possui 9027 instâncias (sentenças) rotuladas.

6.1.3 Taxa de Compressão

A definição de taxa de compressão utilizada nos experimentos foi apresentada na Seção 3.1.2. As taxas foram determinadas para serem compatíveis com o tamanho dos sumários de referência utilizados ou compatíveis com as utilizadas em experimentos conduzidos por outros autores, permitindo a comparação de resultados.

Há duas formas de se calcular a taxa de compressão: baseando-se no número de sentenças ou no número de palavras. Neste trabalho, foram utilizadas as duas formas. O critério adotado foi indicado em cada etapa de avaliação descrita neste capítulo. O critério preferencial utilizado foi o de cálculo com base no número de palavras. O cálculo com base no número de sentenças foi utilizado apenas para permitir a comparação com resultados publicados por outros autores.

6.1.4 Métricas de Avaliação

Foram utilizadas tanto as métricas automáticas de Precisão, Cobertura e *F-measure* quanto as métricas calculadas pela ferramenta ROUGE (vide Seção 2.3), com 95% de confiança. Ambos os tipos de métrica focam na avaliação da informatividade dos extratos produzidos em comparação com os sumários manuais. A preferência neste trabalho foi a utilização da ROUGE devido ao fato de ela ser a mais utilizada em avaliações automáticas (Spärck Jones 2007), além de ter sido reportado pelos autores da ferramenta que suas medidas apresentam boa correlação com as avaliações humanas, conforme já citado anteriormente.

6.1.5 Avaliação dos Modelos Treinados

No caso dos modelos que exigem treinamento, a avaliação foi feita geralmente utilizando a técnica de *N-fold cross-validation*, já utilizada na Seção 5.1.4. A utilização dessa técnica permite intercambiar dados de treino e teste, evitando a necessidade de corpora de treino e testes separados.

O funcionamento é como segue: o conjunto de dados é separado em *N* subconjuntos disjuntos e em cada uma de *N* fases há um conjunto de dados de treino obtido concatenando $N - 1$ dos subconjuntos e um conjunto de dados de

validação que usa o restante do subconjunto; o processo é repetido *N* vezes, permutando de forma circular os subconjuntos. O *N* adotado foi indicado em cada avaliação conduzida.

6.2 Experimento 1 - Avaliação do SuPor-2

A avaliação do SuPor-2 buscou compará-lo com o desempenho de outros sete sumarizadores para o Português do Brasil, inclusive o próprio SuPor original, já avaliados numa avaliação conjunta conduzida por 2 grupos de pesquisadores (Rino et al. 2004). O experimento reproduzido aqui também é descrito em Leite e Rino (2006).

O corpus adotado foi o mesmo utilizado no experimento replicado, o TeMário-2003. A taxa de compressão utilizada foi a mesma, definida em 30% do número de sentenças.

A configuração utilizada do SuPor-2 foi a determinada como a mais promissora na Seção 5.1.4. Já a configuração do SuPor utilizada foi determinada manualmente como a melhor dentre as 348 possíveis.

A Tabela 6-1 apresenta os resultados. Os sistemas *From-top* e *Random Order* são *baselines* da avaliação. O primeiro seleciona as primeiras sentenças do texto e o segundo seleciona sentenças aleatórias. Os demais são descritos na Seção 3.2.

Cabe ressaltar que os resultados para os sistemas de outros autores não foram recalculados. Além disso, como se pretendeu, aqui, reproduzir integralmente o experimento realizado em 2004, não houve testes estatísticos de significância porque esses resultados não estavam disponíveis para os demais sistemas.

Tabela 6-1 – Comparação do SuPor-2 com outros sumarizadores

Sistema	Precisão (%)	Cobertura (%)	F-measure (%)
SuPor-2	47.4	43.9	45.6
SuPor	44.9	40.8	42.8
ClassSumm	45.6	39.7	42.4
From-top	42.9	32.6	37.0
TF-ISF-Summ	39.6	34.3	36.8
GistSumm	49.9	25.6	33.8
NeuralSumm	36.0	29.5	32.4
Random order	34.0	28.5	31.0

Os resultados obtidos para a *F-measure* mostram um desempenho superior de aproximadamente 3 pontos percentuais do SuPor-2 em relação ao SuPor. Analisando-se as variações presentes na tabela entre os demais sistemas, considera-se esse avanço expressivo. Os resultados sugerem ainda que considerar características não-binárias, como faz o SuPor-2, pode ser benéfico para a SA.

É interessante notar também que os três sistemas mais bem classificados utilizam métodos Bayesianos, o que parece confirmar seu bom desempenho para a tarefa de ranking, conforme já sugerido na Seção 4.1.2. Justamente esse ponto ajuda a corroborar a Hipótese 3 deste trabalho, de que a forma de combinação e ponderação das características tem influência para a SA.

6.3 Experimento 2 - Avaliação dos Modelos Baseados no TextRank

A avaliação dos modelos baseados no TextRank foi feita de modo a permitir a comparação com os resultados publicados por Mihalcea (2005), adotando-se metodologia similar à anterior, isto é, de reprodução fiel dos dados de teste, taxa de compressão e medidas de avaliação. Assim, adotou-se o corpus TeMário 2003, a taxas de compressão de 25% à 30% em número de palavras e a métrica ROUGE-1. O experimento reproduzido nesta etapa também é descrito em Leite et al. (2007). Os resultados para as versões de Mihalcea (2005) não foram recalculados.

Os resultados são mostrados na Tabela 6-2. As linhas sombreadas representam os sistemas desenvolvidos neste trabalho. Foi incluído na comparação o SuPor-2. Para seu treino, utilizou-se a técnica *10-fold cross-validation*, já que se dispunha de apenas um corpus. As variações do TextRank propostas por Mihalcea (2005) também estão na tabela: uso das medidas HITS e uso de grafos dirigidos com modo de percurso *backward* ou *forward*. O sistema *Baseline* é um sumarizador construído por Mihalcea (2005) para fins de comparação, que seleciona as sentenças na ordem em que aparecem. Testes estatísticos de significância não foram feitos nessa etapa devido à indisponibilidade de resultados por texto dos demais sistemas.

Tabela 6-2 – Resultados da avaliação dos modelos baseados no método TextRank

Sistema	**ROUGE-1**
SuPor-2	0.5839
TextRank+Thesaurus	0.5603
TextRank+Stem+StopwordsRem	0.5426
TextRank (PageRank -backward)	0.5121
TextRank (HIT hub - forward)	0.5002
TextRank (HITS authority -backward)	0.5002
Baseline – From-Top	0.4963
TextRank (PageRank -undirected)	0.4939
TextRank (HITS authority -forward)	0.4834
TextRank (HIT hub - backward)	0.4834
TextRank (HITS authority - undirected)	0.4814
TextRank (HIT hub - undirected)	0.4814
TextRank (PageRank - forward)	0.4574

Como pode ser observado, as duas variações propostas obtiveram resultados superiores às versões sem nenhum pré-processamento linguístico, representando um aumento de quase 6% na versão TextRank+Stem+StopwordsRem e aproximadamente 9% com o TextRank+Thesaurus.

Tais resultados sugerem que a utilização de métodos baseados em grafos sem nenhum tipo de pré-processamento linguístico, mesmo que simples, pode não produzir extratos informativos o suficiente. O fato de o *Baseline* superar a maioria das versões do TextRank corrobora essa hipótese.

Quanto ao SuPor-2, ele superou todos os demais sistemas, possivelmente por incorporar várias características numa abordagem de Aprendizado de Máquina. Esse desempenho superior corrobora a Hipótese 1, de que a combinação de características diversas, como o SuPor-2 emprega, é benéfica para a SA.

Entretanto, há de se considerar o contraste na comparação, já que o SuPor-2 é supervisionado e apresenta um custo computacional maior que os modelos baseados no método TextRank. É interessante notar que o sistema TextRank+Thesaurus está relativamente perto do SuPor-2, considerando que o esforço de construção do TextRank+Thesaurus consistiu em apenas incorporar pré-processamentos simples no TextRank original.

6.4 Experimento 3 - Avaliação dos Modelos Baseados na Combinação de Características do SuPor-2 e Redes Complexas

Os 24 modelos com características da área de Redes Complexas (RC) e do SuPor-2 combinadas (Tabela 5-8) foram comparados entre si nesta etapa. O experimento desta etapa também foi reportado e discutido em Leite e Rino (2008).

As características de redes complexas foram calculadas para o corpus TeMário-2003 por Lucas Antiqueira (2007). Como esses dados foram previamente processados e os sistemas de redes não se encontram disponíveis, adotou-se, para a avaliação combinada, o mesmo corpus, tomando-se o cuidado de evitar o problema de seu tamanho pequeno com a técnica de validação cruzada (*10-fold cross-validation*). A métrica adotada foi também a mesma proposta por Antiqueira — ROUGE-1 — permitindo assim a comparação de resultados sem que se reprocessassem as medidas de redes complexas. A taxa de compressão foi de 30% (em palavras).

A Tabela 6-3 mostra os resultados obtidos, indicando qual o classificador utilizado em cada modelo, assim como o uso (ou ausência de uso) da pré-seleção de características por meio do CFS. Em cada modelo, é indicado também o conjunto base de características utilizado: características de redes complexas (RC) apenas, características do SuPor-2 apenas ou a união de ambos os conjuntos de características (SuPor-2 $\cup$ RC). O resultado completo da aplicação do CFS, indicando os subconjuntos recomendados de características é fornecido no Apêndice F.

O melhor modelo aqui obtido foi chamado de *SuPor-2 LogReg.* Como mostrado, os classificadores Regressão Logística e Flexible-Bayes apresentaram desempenho superior à maioria dos modelos que utilizam SVM ou C4.5. Novamente, isso corrobora a ideia de que classificadores probabilísticos são mais adequados à tarefa de ranking (vide seções 0 e 4.1.5) e também a Hipótese 3 deste trabalho, que defende que o modelo de combinação e ponderação de características tem influência importante para a SA.

Com relação ao uso do CFS, relacionado à Hipótese 2 de que a pré-seleção automática de características é útil, fica claro que seu benefício varia de acordo com o conjunto de características. Por exemplo, quando a união dos conjuntos de

características foi utilizada, o uso do CFS melhorou os resultados, exceto para o modelo M21 que obteve o mesmo resultado. Considerando o conjunto de características RC, todos os modelos obtiveram resultados iguais ou piores utilizando o CFS. Esses resultados sugerem que, quando considerados conjuntos individuais de características (SuPor-2 ou RC), o uso do CFS não traz ganhos. Entretanto, quando conjuntos maiores são explorados, o uso do CFS pode ser benéfico.

Tabela 6-3 – Comparação dos modelos com características do SuPor-2 e Redes Complexas combinadas

Modelo	Conjunto de Característica	Classificador	Uso do CFS	ROUGE-1
M16	SuPor-2	Regressão Logística	Não	0.5316
M17	SuPor-2	Regressão Logística	Sim	0.5288
M13	SuPor-2	Flexible-Bayes	Sim	0.5284
M21	SuPor-2 ∪ RC	Flexible-Bayes	Sim	0.5278
M25	SuPor-2 ∪ RC	Regressão Logística	Sim	0.5270
M23	SuPor-2 ∪ RC	C4.5	Sim	0.5253
M20	SuPor-2 ∪ RC	Flexible-Bayes	Não	0.5249
M15	SuPor-2	C4.5	Sim	0.5238
M4	RC	Flexible-Bayes	Não	0.5237
M5	RC	Flexible-Bayes	Sim	0.5236
M8	RC	Regressão Logística	Não	0.5230
M24	SuPor-2 ∪ RC	Regressão Logística	Não	0.5228
M12	SuPor-2	Flexible-Bayes	Não	0.5227
M14	SuPor-2	C4.5	Não	0.5212
M9	RC	Regressão Logística	Sim	0.5188
M22	SuPor-2 ∪ RC	C4.5	Não	0.5184
M6	RC	C4.5	Não	0.5167
M18	SuPor-2	SVM	Não	0.5158
M19	SuPor-2	SVM	Sim	0.5158
M26	SuPor-2 ∪ RC	SVM	Não	0.5158
M27	SuPor-2 ∪ RC	SVM	Sim	0.5158
M7	RC	C4.5	Sim	0.5157
M10	RC	SVM	Não	0.5032
M11	RC	SVM	Sim	0.5032

Cabe-se ressaltar que a utilização e a combinação de características de RC por meio do aprendizado de máquina representa um avanço sobre a utilização de medidas individuais, propostas por Antiqueira (2007). A melhor medida individual proposta por esse autor— C13 da Tabela 5-7 — obteve o índice ROUGE-1 de 0.5020 (Antiqueira 2007), enquanto a combinação de medidas de redes complexas por meio do aprendizado de máquina obteve o índice 0.5237. Essa é mais uma evidência a favor da Hipótese 1 deste trabalho, de que a combinação de características deve ser explorada para a SA.

A Tabela 6-4 mostra a análise de significância estatística dos resultados obtidos pelo teste *t-student* com nível de significância de 5%. Os pares significativos encontram-se hachurados (p-valores menores ou iguais a 0,05).

De forma geral, não se garante significância para as diferenças encontradas nos sistemas mais bem classificados. Só há significância quando são consideradas as diferenças entre os sistemas do topo da Tabela 6-3 e os situados mais abaixo. A utilização de um corpus maior e a disponibilidade de mais sumários de referência poderiam ser úteis na verificação da significância estatística desses resultados.

Tabela 6-4 – Análise de significância estatística do Experimento 3 – p-valores

Mod.	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20	M21	M22	M23	M24	M25	M26	M27
M4	-	0,98	0,16	0,10	0,87	0,21	**0,00**	**0,00**	0,81	0,28	0,62	0,99	0,08	0,26	0,25	0,25	0,74	0,25	0,36	0,77	0,83	0,39	0,25	0,25
M5	0,98	-	0,15	0,11	0,85	0,17	**0,00**	**0,00**	0,82	0,33	0,63	1,00	0,11	0,32	0,26	0,26	0,76	0,29	0,37	0,80	0,83	0,44	0,26	0,26
M6	0,16	0,15	-	0,78	0,19	0,69	**0,04**	**0,04**	0,30	**0,02**	0,42	0,20	**0,01**	**0,03**	0,88	0,88	0,12	**0,03**	0,83	0,11	0,27	**0,05**	0,88	0,88
M7	0,10	0,11	0,78	-	0,15	0,54	**0,04**	**0,04**	0,15	**0,01**	0,33	0,15	**0,00**	**0,01**	0,99	0,99	0,07	**0,01**	0,68	0,07	0,15	**0,02**	0,99	0,99
M8	0,87	0,85	0,19	0,15	-	0,24	**0,00**	**0,00**	0,94	0,24	0,69	0,89	0,08	0,26	0,29	0,29	0,66	0,24	0,40	0,68	0,95	0,37	0,29	0,29
M9	0,21	0,17	0,69	0,54	0,24	-	**0,03**	**0,03**	0,43	**0,04**	0,70	0,40	**0,01**	0,05	0,66	0,66	0,08	**0,01**	0,90	0,25	0,41	0,06	0,66	0,66
M10	**0,00**	**0,00**	**0,04**	**0,04**	**0,00**	**0,03**	-	-	**0,01**	**0,00**	**0,00**	**0,00**	**0,00**	**0,00**	**0,00**	**0,00**	**0,00**	**0,00**	**0,03**	**0,00**	**0,00**	**0,00**	**0,00**	**0,00**
M11	**0,00**	**0,00**	**0,04**	**0,04**	**0,00**	**0,03**	-	-	**0,01**	**0,00**	**0,00**	**0,00**	**0,00**	**0,00**	**0,00**	**0,00**	**0,00**	**0,00**	**0,03**	**0,00**	**0,00**	**0,00**	**0,00**	**0,00**
M12	0,81	0,82	0,30	0,15	0,94	0,43	**0,01**	**0,01**	-	0,09	0,76	0,82	**0,03**	0,09	0,26	0,26	0,60	0,17	0,40	0,62	0,97	0,25	0,26	0,26
M13	0,28	0,33	**0,02**	**0,01**	0,24	**0,04**	**0,00**	**0,00**	0,09	-	0,09	0,32	0,38	0,90	**0,04**	**0,04**	0,36	0,87	0,07	0,49	0,22	0,70	**0,04**	**0,04**
M14	0,62	0,63	0,42	0,33	0,69	0,70	**0,00**	**0,00**	0,76	0,09	-	0,45	**0,02**	0,10	0,36	0,36	0,50	0,18	0,62	0,27	0,75	0,17	0,36	0,36
M15	0,99	1,00	0,20	0,15	0,89	0,40	**0,00**	**0,00**	0,82	0,32	0,45	-	0,07	0,28	0,15	0,15	0,85	0,43	0,33	0,66	0,84	0,45	0,15	0,15
M16	0,08	0,11	**0,01**	**0,00**	0,08	**0,01**	**0,00**	**0,00**	**0,03**	0,38	**0,02**	0,07	-	0,39	**0,01**	**0,01**	0,12	0,35	**0,01**	0,11	**0,01**	0,18	**0,01**	**0,01**
M17	0,26	0,32	**0,03**	**0,01**	0,26	0,05	**0,00**	**0,00**	0,09	0,90	0,10	0,28	0,39	-	**0,03**	**0,03**	0,36	0,80	0,06	0,38	0,12	0,59	**0,03**	**0,03**
M18	0,25	0,26	0,88	0,99	0,29	0,66	**0,00**	**0,00**	0,26	**0,04**	0,36	0,15	**0,01**	**0,03**	-	-	0,19	0,07	0,71	0,11	0,25	0,07	-	-
M19	0,25	0,26	0,88	0,99	0,29	0,66	**0,00**	**0,00**	0,26	**0,04**	0,36	0,15	**0,01**	**0,03**	-	-	0,19	0,07	0,71	0,11	0,25	0,07	-	-
M20	0,74	0,76	0,12	0,07	0,66	0,08	**0,00**	**0,00**	0,60	0,36	0,50	0,85	0,12	0,36	0,19	0,19	-	0,27	0,27	0,94	0,63	0,54	0,19	0,19
M21	0,25	0,29	**0,03**	**0,01**	0,24	**0,01**	**0,00**	**0,00**	0,17	0,87	0,18	0,43	0,35	0,80	0,07	0,07	0,27	-	0,11	0,59	0,22	0,84	0,07	0,07
M22	0,36	0,37	0,83	0,68	0,40	0,90	**0,03**	**0,03**	0,40	0,07	0,62	0,33	**0,01**	0,06	0,71	0,71	0,27	0,11	-	0,22	0,39	0,11	0,71	0,71
M23	0,77	0,80	0,11	0,07	0,68	0,25	**0,00**	**0,00**	0,62	0,49	0,27	0,66	0,11	0,38	0,11	0,11	0,94	0,59	0,22	-	0,60	0,63	0,11	0,11
M24	0,83	0,83	0,27	0,15	0,95	0,41	**0,00**	**0,00**	0,97	0,22	0,75	0,84	**0,01**	0,12	0,25	0,25	0,63	0,22	0,39	0,60	-	0,21	0,25	0,25
M25	0,39	0,44	**0,05**	**0,02**	0,37	0,06	**0,00**	**0,00**	0,25	0,70	0,17	0,45	0,18	0,59	0,07	0,07	0,54	0,84	0,11	0,63	0,21	-	0,07	0,07
M26	0,25	0,26	0,88	0,99	0,29	0,66	**0,00**	**0,00**	0,26	**0,04**	0,36	0,15	**0,01**	**0,03**	-	-	0,19	0,07	0,71	0,11	0,25	0,07	-	-
M27	0,25	0,26	0,88	0,99	0,29	0,66	**0,00**	**0,00**	0,26	**0,04**	0,36	0,15	**0,01**	**0,03**	-	-	0,19	0,07	0,71	0,11	0,25	0,07	-	-

6.5 Experimento 4 - Avaliação dos Modelos Baseados em Regras Nebulosas

O objetivo principal desta etapa foi verificar se o desempenho de um modelo de ranking nebuloso de sentenças poderia ser superior ao modelo Bayesiano do SuPor-2. Essa etapa de avaliação também foi descrita em Leite e Rino (2009).

Como o conjunto de características do SuPor-2 Fuzzy é diferente do SuPor-2, a simples comparação não seria justa e não permitiria comparar apenas os modelos de ranking. Por conta disso, conforme citado na Seção 5.5, foi construído o SuPor-2

Modificado, adotando o mesmo modelo de ranking do SuPor-2, porém com as características do SuPor-2 Fuzzy.

Dada a complexidade da fase de treino do SuPor-2 Fuzzy, que envolve algoritmos genéticos, decidiu-se não utilizar o processo de *cross-validation* e utilizar dois corpora: o TeMário-2003 para treino e o Summ-it para avaliação.

O processo de evolução do algoritmo genético levou 168h até convergência da função de *fitness* da melhor base de regras. A Figura 6-1 mostra o gráfico da evolução.

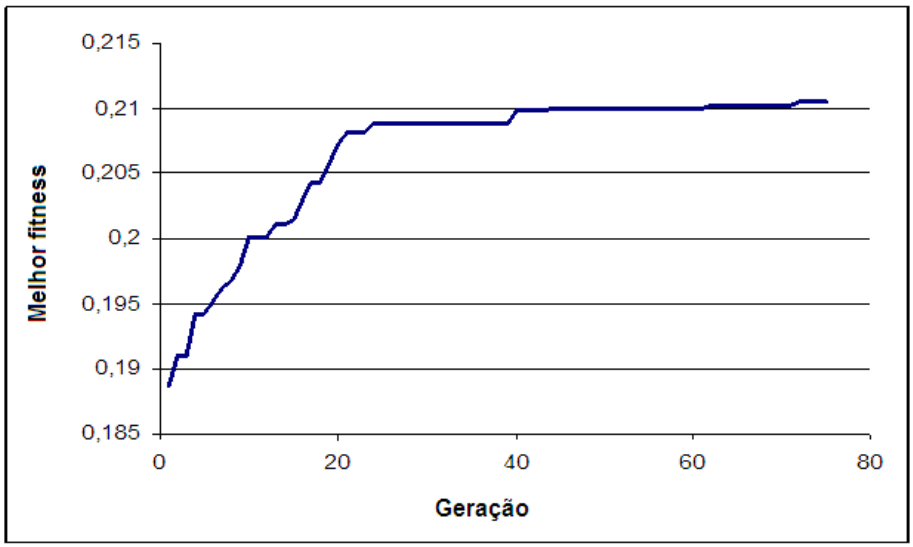


Figura 6-1 – Evolução da etapa de treino do SuPor-2 Fuzzy

A taxa de compressão foi definida em 30% no número de palavras. Foram utilizadas na avaliação tanto a métrica ROUGE-1 quanto a ROUGE-2, as mais comuns nas últimas DUCs/TACs.

Tabela 6-5 – Avaliação do SuPor-2 Fuzzy

Modelo	Sistema	ROUGE-1	ROUGE-2
M28	SuPor-2 Fuzzy	0.74583	0.73859
M29	SuPor-2 modificado	0.73205	0.72323

Os resultados mostram um desempenho superior do SuPor-2 Fuzzy em relação ao classificador Bayesiano considerando tanto a medida ROUGE-1 quanto ROUGE-2. Ressalta-se que o tempo de treino do SuPor-2 Fuzzy é muito maior que os modelos Bayesianos. Entretanto, o processo é conduzido uma única vez.

Cabe lembrar também que o SuPor-2 Fuzzy teve seu treino dirigido pela ROUGE e foi avaliado pela mesma medida. Ou seja, assumindo a premissa de que a ROUGE seja uma medida que diferencie bem sumários informativos de não-informativos, o SuPor-2 Fuzzy foi projetado de modo a produzir sumários com índices altos da medida. Embora a medida de treino e avaliação seja a mesma, em nenhum momento utilizou-se qualquer dado de teste para treino. Em outras palavras, não existe viés que torne a comparação injusta.

O teste *t-student* apresentou p-valores de 0.12014 para a ROUGE-1 e 0.09466 para ROUGE-2. Com 95% de confiança não há diferenças significativas. Entretanto, considerando que os p-valores são baixos, as diferenças estão bem perto de serem significativas.

Novamente, o fato de um modelo com ranking nebuloso superar um Bayesiano corrobora a Hipótese 3, de que o modelo de combinação e ponderação de características influencia significativamente o desempenho dos modelos de SA.

6.6 Experimento 5 - Avaliação Global de Todos os Modelos

Os melhores modelos de SA indicados nas etapas anteriores de avaliação foram selecionados para uma avaliação conjunta utilizando um único corpus e a mesma taxa de compressão (a usual de 30% em número de palavras). Os modelos selecionados são apresentados na Tabela 6-6, em que cita-se também os recursos dependentes de língua utilizados em cada um.

Nesta avaliação global, o corpus utilizado foi o TeMário-2006 (151 textos), que não havia sido utilizado anteriormente e possui o maior número de textos, porém é de mesmo gênero dos anteriores, o que torna o treinamento ou classificação das diversas versões consistente com os dados de teste. Neste caso, particularmente, usou-se para treino dos sistemas o corpus TeMário-2003 inteiro (100 textos). O motivo da escolha de um corpus separado para treino em vez da utilização do processo de *cross-validation* foi evitar o treino mais custoso do SuPor-2 Fuzzy, demandando mais tempo usando *cross-validation*. Cabe lembrar que não há sobreposição de textos entre o TeMário-2003 e o TeMário-2006.

Tabela 6-6 – Comparativo das características dos modelos propostos

Modelo	Nome	Número de Características	Modelo de Combinação de Características	Recursos Dependentes de Língua Natural
M1	SuPor-2	11	Flexible-Bayes	*Stoplist*[29] Léxico *Stemmer* Etiquetador morfossintático *Thesaurus*
M2	TextRank+StemmingStopRem	1	Nenhum	*Stoplist* *Stemmer*
M3	TextRank+Thesaurus	1	Nenhum	*Stoplist* *Stemmer* *Thesaurus*
M16	SuPor-2 LogReg	11	Regressão Logística	Stoplist *Léxico* *Stemmer* Etiquetador morfossintático *Thesaurus*
M29	SuPor-2 Fuzzy	11	Classificador Nebuloso	*Stoplist* *Stemmer*

Os extratos gerados tiveram a informatividade medida pela ROUGE, através de suas medidas ROUGE-1 e ROUGE-2. As tabelas seguintes mostram os resultados obtidos e análise de significância das diferenças para cada par de modelos pelo teste *t-student*. O Apêndice E mostra o resultado da avaliação para cada texto do corpus, dados que foram usados no teste *t-student.*

Tabela 6-7 – Resultados da avaliação conjunta dos modelos de SA sobre o TeMário-2006

Modelo	Sistema	ROUGE-1	ROUGE-2
M1	SuPor-2	0,65398	0,31670
M16	SuPor-2 LogReg	0,64069	0,14888
M29	SuPor-2 Fuzzy	0,62460	0,13151
M3	TextRank+Thesaurus	0.60056	0.25449
M2	TextRank+StemmingStopRem	0.57856	0,25183

Tabela 6-8 – P-valores do teste t-student para a medida ROUGE-1

p-valores	M1	M2	M3	M16	M29
M1		0.00%	0.00%	0.00%	0.00%
M2	0.00%		0.02%	0.00%	0.00%
M3	0.00%	0.02%		0.00%	0.00%
M16	0.00%	0.00%	0.00%		0.00%
M29	0.00%	0.00%	0.00%	0.00%	

[29] Trata-se de uma lista de *stopwords*, palavras muito comuns e consideradas irrelevantes.

Tabela 6-9 – P-valores do teste t-student para a medida ROUGE-2

p-valores	M1	M2	M3	M16	M29
M1		0.00%	0.00%	74.47%	0.00%
M2	0.00%		63.82%	0.00%	0.00%
M3	0.00%	63.82%		0.00%	0.00%
M16	74.47%	0.00%	0.00%		0.00%
M29	0.00%	0.00%	0.00%	0.00%	

Como pode ser observado, o SuPor-2 apresentou diferenças significantes para os demais quando avaliado pela medida ROUGE-1. Quando a ROUGE-2 é utilizada, o teste *t-student* não indica significância apenas na comparação com a versão SuPor-2 LogReg, que utiliza Regressão Logística em vez do Flexible-Bayes. Houve, portanto, uma inversão na posição desses modelos em relação à avaliação conduzida na Seção 6.4, onde a Regressão Logística havia superado todos os demais classificadores. Entretanto, naquela avaliação o corpus de teste era menor (TeMário-2003), as diferenças obtidas entre os sistemas menores e nenhuma significância estatística foi verificada.

Com relação aos demais modelos, o ranking obtido confirma os resultados anteriores. No caso do SuPor-2 Fuzzy, seu desempenho superior em relação aos modelos baseados no TextRank é muito provavelmente explicado pelo fato desse sistema ser supervisionado e por fazer uso de combinação de características. Com relação a seu desempenho inferior ao SuPor-2 e ao SuPor-2 LogReg, a diferença está no fato de o conjunto de características empregadas no modelo nebuloso ser uma simplificação do modelo original do SuPor-2, conforme exposto na Seção 5.4.1.

A seguir, mostram-se exemplos de sumários gerados para o corpus Temário-2006 do melhor sistema identificado, o SuPor-2. Exemplos de sumários gerados para outros modelos são dados no Apêndice G.

EXEMPLO 1 (br94de25-11.txt) – texto de opinião

- **Texto-fonte br94de25-11.txt**

O efeito Gutenberg

MARCELO LEITE

Um leitor qualificado e perspicaz sugeriu-me outro dia uma pergunta difícil: como se comportará a Folha em relação ao governo Fernando Henrique Cardoso? No maniqueísmo inerente ao jornalismo, só haveria uma alternativa: ou amor ou ódio.

A questão é pertinente, dada a notória proximidade do jornal com o presidente eleito. Até setembro de 1992, FHC mantinha uma coluna semanal na pág. 1-2, publicada às quintas-feiras. Tal relação de colaboração só foi interrompida porque o senador peessedebista se tornou chanceler de Itamar Franco.

(Segundo praxe da Folha, um colunista não pode simultaneamente ocupar ou candidatar-se a cargo no Executivo. Nesta condição, sua coluna correria o risco de transformar-se em tribuna para defesa de um interesse privado –a reputação como governante.)

Fernando Henrique não foi o único tucano a ocupar esse espaço, conhecido na Redação como coluna vertical. Depois de amanhã deverá ser publicado o último texto do futuro ministro do Planejamento José Serra.

Pertinente, a questão não é porém nova. O próprio retrospecto das colunas de ombudsman aponta para uma simpatia espúria:

Durante a campanha eleitoral, minha antecessora apontou fernandohenriquismo do jornal;

Ao estrear, emiti a opinião de que este e outros diários tinham mesmo henricado;

A 30 de outubro, na coluna Lua-de-mel na Europa, critiquei a condescendência da Folha com o presidente eleito.

Quando FHC enfim se lançou ao primeiro ato de governo, montar seu ministério, temi pelo pior.

No episódio da escolha de Pedro Malan para ministro da Fazenda, intencionalmente vazada para repórteres, os jornais evidenciaram sua tibieza. Com arrogância, FHC desqualificou as manchetes de 1º de dezembro, dizendo que era um ministério Gutenberg (referência a Johannes Gutenberg, que inventou a imprensa de tipos móveis no século 15).

Ficou por isso mesmo. Em outras épocas, a Folha teria posto a boca no trombone, denunciando a tentativa de manipulação.

FHC seguiu a seu modo a receita do seu sucessor na Fazenda, aquele premiado com a embaixada em Roma pela ajuda ao candidato: esconder o que é ruim (as pressões para indicar Serra no lugar de Malan) e faturar o que é bom (a imagem favorável de Malan).

No último dia 14, perguntei em minha crítica interna da edição –documento distribuído diariamente na Redação– se o termo loteamento também não se aplicaria às negociações em curso, em especial as tratativas com o fisiológico PMDB. Afinal, eram um tanto semelhantes às entabuladas por Itamar dois anos antes e desancadas pelo jornal.

Dois dias depois, uma chamada na capa do jornal anunciava: FHC cede a pressão e loteia ministério. Ao elogiar a iniciativa crítica, no entanto, fiz uma ressalva:

"Faltou mencionar um ponto importante, na análise das 'pressões': FHC teria condições, sem contemplar PMDB, de fazer a reforma constitucional (ou pelo menos fiscal) exigida por todos, inclusive esta Folha?

É uma espécie de outro lado –neste caso, da questão. A necessidade de criticar o emprego de métodos políticos atrasados, como a distribuição de cargos, não desobriga de outra, a de eventualmente reconhecer que pode não haver outra moeda no mercado para negociar a estabilidade.

A palavra-chave do comportamento que a Folha deve observar frente ao governo –qualquer governo– é equilíbrio. Sem simpatia nem rancor.

O perigo das relações estremecidas, como no caso FHC-Folha, são as hiper-reações resultantes de encontrões fortuitos.

Foi o que sucedeu com o sociólogo Luciano Martins, amigo de FHC e organizador de um convescote acadêmico em Brasília. Na véspera do seminário, ele tinha dado entrevista à Folha e falado da crise do Estado-Nação, publicada sob título Acabou o Estado nacional, diz tucano.

Era um exagero, mas confesso que nem me chamou a atenção. Por vaidade, ou cioso das diferenciações que matizam o pensamento, seu ofício, Martins chiou.

Em carta ao Painel do Leitor, expôs suas divergências e levou troco imediato, na forma de uma atordoante Nota da Redação:

"Por serem resumos extremamente condensados, os títulos jornalísticos quase nunca comportam filigranas como esta que tanto preocupa o missivista. Para Luciano Martins, o conceito de Estado nacional não acabou, mas está em crise. E

daí? A imprensa deve melhorar seus títulos, não há dúvida. Mas os intelectuais agora transformados em aprendizes de políticos ajudariam muito se começassem a falar de maneira categórica ou, pelo menos, clara."

O reflexo desse destempero pôde ser visto pelo público no próprio Painel do Leitor, 11 dias depois: quatro cartas de protesto, nenhuma de apoio ao jornal, nenhuma nova nota justificando ou se desculpando pela anterior.

Os leitores estão certos. Se o jornal acha que intelectuais não têm nada de importante ou compreensível para dizer, não deveria insistir em entrevistá-los. Se entrevista, tem de cobrar clareza durante a conversa; depois, só lhe resta ser fiel ao que dizem.

Atritos como esse são exceção. No geral, a relação entre tucanos e repórteres é afável. Sua melhor expressão é o "off", um acordo entre fonte e jornalista para manter a primeira no anonimato.

Na última terça-feira, o colunista Luís Nassif levantou questões pertinentes sobre o abuso dessa modalidade de investigação. Seu alvo eram as muitas reportagens abusivamente atribuídas à famosa equipe econômica.

Aproveitei a deixa para anotar que a distorção afetava grande parte, talvez a maior, do noticiário sobre o governo Itamar Cardoso. No caso deste jornal, sem que as reportagens respeitassem norma do "Novo Manual da Redação", que manda identificar o "off" com a expressão "a Folha apurou".

Foi o caso, entre outros, da notícia sobre a escolha de Malan para a Fazenda (ironizada e depois confirmada). E também da indicação de Bresser para o Itamaraty (manchetada e depois revista).

Trata-se de uma distorção, sim. Embora a prática jornalística brasileira sugira o "off" como ferramenta básica de repórteres, ele contraria o direito à informação. Deve ser encarado como exceção, e nunca oferecido pelo próprio repórter, muito menos aceito, se o confidente não tiver motivos sólidos para manter-se em sigilo.

Não me parece que o anúncio a conta-gotas do "ministério possível" de FHC, todo ele em "off", se enquadre nessa exigência.

A identificação da fonte é crucial para a credibilidade de uma informação. O jornalista que admite a exceção não pode esconder do leitor que se trata de um "off", pelo simples fato de que o interesse no anonimato pode comprometer aquilo que se revela.

Afinal, não foi para esconder informações que Gutenberg inventou a imprensa.

O ombudsman estará de folga até o dia 2. Se você tiver alguma reclamação, deixe recado na secretária eletrônica ou mande fax. Na volta, respondo.

- **Extrato SuPor-2 br94de25-11.txt**

[1] Um leitor qualificado e perspicaz sugeriu-me outro dia uma pergunta difícil: como se comportará a Folha em relação ao governo Fernando Henrique Cardoso?

[2] A questão é pertinente, dada a notória proximidade do jornal com o presidente eleito.

[3] Até setembro de 1992, FHC mantinha uma coluna semanal na pág 1-2, publicada às quintas-feiras.

[4] Tal relação de colaboração só foi interrompida porque o senador peessedebista se tornou chanceler de Itamar Franco.

[5] Fernando Henrique não foi o único tucano a ocupar esse espaço, conhecido na Redação como coluna vertical.

[6] A 30 de outubro, na coluna Lua-de-mel na Europa, critiquei a condescendência da Folha com o presidente eleito.

[7] Quando FHC enfim se lançou ao primeiro ato de governo, montar seu ministério, temi pelo pior.

[8] No episódio da escolha de Pedro Malan para ministro da Fazenda, intencionalmente vazada para repórteres, os jornais evidenciaram sua tibieza.

[9] Com arrogância, FHC desqualificou as manchetes de 1º de dezembro, dizendo que era um ministério Gutenberg (referência a Johannes Gutenberg, que inventou a imprensa de tipos móveis no século 15).

[10] No último dia 14, perguntei em minha crítica interna da edição – documento distribuído diariamente na Redação – se o termo loteamento também não se aplicaria às negociações em curso, em especial as tratativas com o fisiológico PMDB.

[11] "Faltou mencionar um ponto importante, na análise das 'pressões': FHC teria condições, sem contemplar PMDB, de fazer a reforma constitucional (ou pelo menos fiscal) exigida por todos, inclusive esta Folha?”

[12] A necessidade de criticar o emprego de métodos políticos atrasados, como a distribuição de cargos, não desobriga de outra, a de eventualmente reconhecer que pode não haver outra moeda no mercado para negociar a estabilidade.

[13] A palavra-chave do comportamento que a Folha deve observar frente ao governo – qualquer governo – é equilíbrio.

[14] O perigo das relações estremecidas, como no caso FHC-Folha, são as hiper-reações resultantes de encontrões fortuitos.

[15] Se o jornal acha que intelectuais não têm nada de importante ou compreensível para dizer, não deveria insistir em entrevistá-los.

[16] No caso deste jornal, sem que as reportagens respeitassem norma do "Novo Manual da Redação", que manda identificar o "off" com a expressão "a Folha apurou".

[17] Embora a prática jornalística brasileira sugira o "off" como ferramenta básica de repórteres, ele contraria o direito à informação.

- **Avaliação do Extrato**

De forma geral, este sumário permite a identificação de que se trata de um texto de opinião, discutindo eventuais complacências do jornal do autor, a Folha, e o presidente FHC.

As primeiras 9 sentenças do texto apresentam de forma conexa uma ideia concisa sobre o que o autor quer transmitir. Um problema ocorre após a sentença [9], pois falta a informação importante de que “Ficou por isso mesmo. Em outras épocas, a Folha teria posto a boca no trombone, denunciando a tentativa de manipulação”. Ou seja, sem essa informação o exemplo que consta no extrato não contribui muito para a argumentação do autor.

Um outro problema ocorre na sentença [11], que fica sem sentido sem a antecessora do texto-original: “Dois dias depois, uma chamada na capa do jornal anunciava: FHC cede a pressão e loteia ministério. Ao elogiar a iniciativa crítica, no entanto, fiz uma ressalva:”.

Já na sentença [14], a informação ali expressa fica também desnecessária e sem sentido, uma vez que o exemplo que se sucede no texto-fonte, sobre o caso de Luciano Martins, foi suprimido. O mesmo ocorre nas sentenças que se seguem no

extrato. Na sentença [15], por exemplo, acaba ficando a expressão “intelectuais” sem o referente.

EXEMPLO 2 (td94ab03-08.txt) – texto de divulgação

- **Texto-fonte td94ab03-08.txt**

Maquetes crescem com mercado imobiliário

Trabalhos são vendidos por até US$ 15 mil e podem ser também usados em projetos educacionais e artísticos

CLÁUDIA RIBEIRO MESQUITA

Free-lance para a Folha

Maquetes são como "bolas de cristal" que antecipam, de forma tridimensional, edifícios, parques, usinas, cenários, projetos educacionais e culturais e os mais variados tipos de produtos. Seu grande filão é o mercado imobiliário, que, quando aquecido –como está ocorrendo este ano–, agita freneticamente os artesãos das oficinas.

Maquetes de prédios e conjuntos residenciais respondem por mais de 80% dos pedidos –e são as mais bem pagas. Uma maquete simples, de um prédio de 20 andares, por exemplo, pode custar entre US$ 4.000 e US$ 7.000.

Outros modelos mais complexos chegam a valer o dobro, como uma maquete do projeto de um conjunto residencial em Campinas (interior de São Paulo), o Bougamville, encomendada à Kenji Maquetes por US$ 15 mil.

O objetivo desse tipo de modelo é promocional, para auxiliar na venda dos imóveis. "A função da maquete, nesse caso, é elucidar ao leigo o que foi projetado em duas dimensões e instigá-lo", diz Kenji Furuyama, 61. Há 32 anos no mercado, Kenji conta com 15 empregados e fatura por mês cerca de US$ 30 mil, 20% dos quais computados como lucro.

Segundo ele, as despesas com mão-de-obra ficam em quase 70%. "Meus funcionários recebem um salário e uma comissão de 30% em cada trabalho que executam", diz.

Kenji começou a trabalhar com maquetes aos 19 anos na Kevel, uma das poucas maquetarias de São Paulo no ano de 1954. Por ali passaram também dois

outros maquetistas da cidade, Adhemir Fogassa e Achiles Maimoni. Os três aprenderam o ofício na prática.

"A formação de um maquetista é aleatória", afirma o professor Júlio Katinsky, do departamento de história da Faculdade de Arquitetura e Urbanismo da USP. Segundo ele, muitos começaram com aeromodelismo e, hoje, são profissionais bem remunerados.

São poucos os que se dedicam a essas miniaturas. Em São Paulo, de acordo com os maquetistas, deve haver cerca de 60 profissionais. Quem está no ramo não reclama. Um autônomo, em um bom mês, pode faturar até US$ 6.000.

Mário Segall, 39, abriu seu escritório em 93. Segundo ele, o investimento para montar a oficina ficou em torno de US$ 10 mil. Alguns equipamentos foram trazidos de Londres. Sua capacidade de produção é de quatro a cinco maquetes por mês. Em meses de pico, Segall afirma que fatura cerca de US$ 6.000, e seu lucro beira os 25%. "As maquetes, em Londres, são respeitadas como parte do projeto", conta. "Aqui, nem tanto."

Roberto Cardoso, arquiteto recém-formado, começou a fazer maquetes para os projetos de faculdade e, hoje, a maior parte de sua renda vem delas. Segundo ele, dá para lucrar, em média, US$ 1.000 por mês. Mas em 92, por exemplo, ele e mais um grupo de maquetistas receberam, cada um, US$ 5.000 por 45 dias de trabalho para a produção de uma maquete do projeto de despoluição do rio Tietê, apresentada na Eco 92.

- **Extrato SuPor-2 td94ab03-08.txt**

[1] Maquetes crescem com mercado imobiliário

[2] Trabalhos são vendidos por até US$ 15 mil e podem ser também usados em projetos educacionais e artísticos

[3] Maquetes são como "bolas de cristal" que antecipam, de forma tridimensional, edifícios, parques, usinas, cenários, projetos educacionais e culturais e os mais variados tipos de produtos.

[4] Outros modelos mais complexos chegam a valer o dobro, como uma maquete do projeto de um conjunto residencial em Campinas (interior de São Paulo), o Bougamville, encomendada à Kenji Maquetes por US$ 15 mil.

[5] "A função da maquete, nesse caso, é elucidar ao leigo o que foi projetado em duas dimensões e instigá-lo ", diz Kenji Furuyama, 61.

[6] Kenji começou a trabalhar com maquetes aos 19 anos na Kevel, uma das poucas maquetarias de São Paulo no ano de 1954.

[7] "A formação de um maquetista é aleatória", afirma o professor Júlio Katinsky, do departamento de história da Faculdade de Arquitetura e Urbanismo da USP.

[8] Roberto Cardoso, arquiteto recém-formado, começou a fazer maquetes para os projetos de faculdade e, hoje, a maior parte de sua renda vem delas.

[9] Mas em 92, por exemplo, ele e mais um grupo de maquetistas receberam, cada um, US$ 5.000 por 45 dias de trabalho para a produção de uma maquete do projeto de despoluição do rio Tietê, apresentada na Eco 92.

- **Avaliação do Extrato**

Trata-se de um bom extrato que indica de forma consistente o conteúdo da noticia e cobre as principais informações. Uma ressalva pode ser feita com relação à textualidade na sentença [9]. O sintagma “mas” indica contraste, porém a informação ali expressa não contrasta com a informação da sentença [8], apenas complementa-a. O que ocorreu foi que a sentença intermediária (“Segundo ele, dá para lucrar, em média, US$ 1.000 por mês.”) foi suprimida, causado esse problema.

EXEMPLO 3 (di94mr20-20.txt) – texto sobre política

- **Texto-fonte di94mr20-20.txt**

Os custos sociais do liberalismo suicida

Países centrais tomam consciência da gravidade dos problemas gerados por uma política liberal irresponsável

MARIA DA CONCEIÇÃO TAVARES

Especial para a Folha

Finalmente, políticos e intelectuais dos países centrais começam a se dar conta da gravidade dos problemas sociais e econômicos gerados por mais de uma década de um liberalismo irresponsável, dogmático e anárquico.

Esta tardia tomada de consciência se manifesta no encontro de cúpula dos ministros do Trabalho dos países centrais em Detroit (o Job Summit) e em recentes declarações de renomados e respeitáveis economistas conservadores. Pela primeira vez, o G-7 se reúne para discutir o problema do desemprego em massa nos países desenvolvidos, que não pára de crescer, lançando uma parcela cada vez maior da população na marginalidade.

Intimamente ligada a este processo está a questão da deslocalização, onde setores e até comunidades inteiras são destruídas, pois suas indústrias deixaram de ser competitivas num ambiente de globalização financeira e abertura comercial indiscriminada.

A combinação de taxas de desemprego crescente com a decadência econômica de regiões onde ocorre a deslocalização gera um quadro social terrível, cujas consequências são bem conhecidas.

Não falam, é claro, os liberais como os da revista "The Economist", que ainda no número da semana passada repetem a ladainha de que o problema do desemprego é resultado da rigidez do mercado de trabalho dos países desenvolvidos, em particular os europeus. A solução, como sempre, seria aumentar a "flexibilidade" do mercado de trabalho, com a retirada do seguro-desemprego e demais empecilhos ao livre jogo das forças de mercado.

Em outras palavras, o problema do desemprego viria do fato de que as economias centrais, no que diz respeito ao mercado de trabalho, são liberais de menos e a solução seria mais liberalismo.

Depois de anos de crescente "flexibilização" do mercado de trabalho, acompanhado de grande aumento e não de diminuição do desemprego, é natural que os governos e até alguns liberais de renome comecem a desconfiar que a solução para os males sociais causados pelo liberalismo irresponsável não seja mais liberalismo.

Em um artigo recente, o professor Maurice Allais, que recebeu o Prêmio Nobel de Economia em 1988 por suas contribuições à teoria neoclássica (teoria de

onde a fé liberal busca obter credibilidade "científica"), faz um ataque frontal à aplicação, nas condições contemporâneas, da doutrina das vantagens comparativas.

Segundo ele, esta "só é aplicável sob condicões altamente restritivas, particularmente se as taxas de câmbio correspondem ao equilíbrio das balanças comerciais e se as vantagens comparativas são permanentes, o que em geral não é o caso". Allais, talvez por vício profissional, ou sentimento de impotência ante a realidade, se esqueceu de mencionar a necessidade da hipótese de pleno emprego.

Na maioria dos casos, o resultado da política liberal foi uma enorme destruição de empregos locais, em troca de uma pequena redução no preço do produto para o consumidor e um grande custo fiscal para a sociedade toda, sobretudo para os próprios consumidores que mativeram-se empregados.

Os custos sociais estão hoje em evidência em toda parte. Um relatório recente da OIT prevê para o final da década taxas de desemprego em torno de 30% para os países desenvolvidos. Esta situação e a falta de perspectiva para os mais jovens cria um caldo de cultura propício à marginalidade e aos movimentos de extrema direita, visíveis em toda a Europa.

Frente a esta situação de catástrofe social, o ex-liberal Maurice Allais recomenda o fechamento comercial do mercado comum europeu, através do controle quantitativo de importações dos países extra-comunitários. No caso de a CEE não adotar francamente uma política de bloco, frontalmente contrária às regras do Gatt, recomenda que a França o faça sozinha. Na verdade, apesar da retórica liberal, é esta a prática corrente nos Estados Unidos e no Japão em matéria de comércio de mercadorias que ameaçam suas indústrias.

De outro lado, renascem também as propostas utópicas onde há os que, como Ricardo Petrella –em recente artigo no "Le Monde Diplomatique"–, esperam que a ONU no seu próximo encontro de cúpula sobre a questão social, a ser realizado em Copenhague em 1995, estabeleça as bases para uma nova ordem econômica e financeira mundial!

Independentemente do caráter conservador ou utópico e da viabilidade técnica ou política de quaisquer destas propostas, é um consolo saber que as pessoas estão reaprendendo que a solução para o problema do desemprego, resultante da modernização conservadora e dos excessos do liberalismo, não pode ser simplesmente mais liberalismo.

Enquanto isso, chega ao Brasil lady Margaret Thatcher, símbolo do que há de pior no liberalismo socialmente irresponsável e é aplaudida de pé pela nata do empresariado brasileiro.

As classes produtoras brasileiras não tomam juízo. Pagam US$ 100 mil para ouvir um show requentado da pseudo-rainha de um ex-império, cuja indústria entrou em decadência há 100 anos. Enquanto isto, sabotam, em nome do "livre mercado", mais um plano de estabilização, apesar de supostamente apoiarem o ministro como candidato.

Melhor fossem em caravana a Washington (e não a Nova York) verificar "in locu" as duas caras do consenso na capital do império. Na verdade, o que deviam escutar e estudar são os planos de reestruturação da indústria e a reforma do sistema de saúde, privado e público, que o governo dos Estados Unidos está aplicando para melhorar a situação interna do seu país.

Não deveriam impressionar-se tanto com as receitas e pressões do FMI e do secretário do Tesouro norte-americano sobre o Brasil e muito menos deslumbrar-se com a performance de uma atriz coadjuvante. Se prestassem atenção ao que está ocorrendo com as mudanças na economia norte-americana, ficariam surpresos, por exemplo, com o grau de estatização do novo programa de telecomunicações.

Talvez aprendessem também que o aumento de produtividade sistêmica é incompatível com o sucateamento do Estado e não implica, do lado empresarial, simplesmente aumentar o desemprego e subir os preços.

Finalmente, concluiriam que o governo americano não está baixando os impostos nem desregulando sua economia, mas regulando-a mais intensamente do que nunca, para enfrentar a concorrência dos países asiáticos e do Japão.

Ao mesmo tempo, o "Consenso de Washington" pretende obter da América Latina um déficit comercial, através de uma sobrevalorização da nossa moeda, o que permitiria aos Estados Unidos reequilibrar a curto prazo suas contas externas.

Isto significa que o Brasil, o último país a resistir ao novo ajuste, que é o oposto do de 1982/83, deve submeter-se à dolarização e promover a toque de caixa e no segredo dos gabinetes a reforma constitucional, no capítulo da ordem econômica, numa direção supostamente liberal, o que sustentaria novo ciclo de endividamento.

Mas seria pedir demais às classes produtoras brasileiras, interessadas apenas no botim imediato, que tomassem consciência do seu destino e do destino

da nação. Provincianos e deslumbrados pela mídia, parecem não saber o que acontece no mundo e são incapazes de pensamento estratégico.

Continuam viciados numa ideologia liberal suicida, preocupados apenas com os seus desejos incontidos de ganância especulativa e patrimonial, que vão custar ao governo, este ano, mais de US$ 10 bilhões em juros internos. Somando os juros da dívida externa (cuja negociação ainda não terminou), o próprio FMI estima em 5,7% do PIB (mais de US$ 22 bilhões) a conta global de juros, uma cifra inacreditável, cuidadosamente oculta pela equipe econômica, e superior ao impacto fiscal ocorrido no auge da crise da dívida externa!

É por isso que o "ajuste fiscal" nunca termina e que o processo de privatização é uma farsa sinistra.

Na verdade, como disse recentemente Clovis Rossi nesta Folha, estamos precisando mesmo é de uma "ruptura democrática" que exponha o nosso empresariado aos ventos da negociação e da verdadeira produtividade e que termine de vez com o seu caráter de parasitas financeiros.

O saneamento do Estado e o cuidado com o povo, seguramente não cabem a eles e sim ao avanço da consciência e do desejo de cidadania do próprio povo, particularmente na escolha de seus representantes no Congresso e dos futuros governos da nação.

MARIA DA CONCEIÇÃO TAVARES, 63, é economista, professora emérita da Universidade Federal do Rio de Janeiro (UFRJ) e professora associada da Universidade de Campinas (Unicamp).

- **Extrato SuPor-2 di94mr20-20.txt**

[1] Finalmente, políticos e intelectuais dos países centrais começam a se dar conta da gravidade dos problemas sociais e econômicos gerados por mais de uma década de um liberalismo irresponsável, dogmático e anárquico.

[2] Esta tardia tomada de consciência se manifesta no encontro de cúpula dos ministros do Trabalho dos países centrais em Detroit (o Job Summit) e em recentes declarações de renomados e respeitáveis economistas conservadores.

[3] Pela primeira vez, o G-7 se reúne para discutir o problema do desemprego em massa nos países desenvolvidos, que não pára de crescer, lançando uma parcela cada vez maior da população na marginalidade.

[4] Intimamente ligada a este processo está a questão da deslocalização, onde setores e até comunidades inteiras são destruídas, pois suas indústrias deixaram de ser competitivas num ambiente de globalização financeira e abertura comercial indiscriminada.

[5] Não falam, é claro, os liberais como os da revista "The Economist", que ainda no número da semana passada repetem a ladainha de que o problema do desemprego é resultado da rigidez do mercado de trabalho dos países desenvolvidos, em particular os europeus.

[6] Em outras palavras, o problema do desemprego viria do fato de que as economias centrais, no que diz respeito ao mercado de trabalho, são liberais de menos e a solução seria mais liberalismo.

[7] Depois de anos de crescente "flexibilização" do mercado de trabalho, acompanhado de grande aumento e não de diminuição do desemprego, é natural que os governos e até alguns liberais de renome comecem a desconfiar que a solução para os males sociais causados pelo liberalismo irresponsável não seja mais liberalismo.

[8] Em um artigo recente, o professor Maurice Allais, que recebeu o Prêmio Nobel de Economia em 1988 por suas contribuições à teoria neoclássica (teoria de onde a fé liberal busca obter credibilidade" científica "), faz um ataque frontal à aplicação, nas condições contemporâneas, da doutrina das vantagens comparativas.

[9] Na maioria dos casos, o resultado da política liberal foi uma enorme destruição de empregos locais, em troca de uma pequena redução no preço do produto para o consumidor e um grande custo fiscal para a sociedade toda, sobretudo para os próprios consumidores que mativeram-se empregados.

[10] Frente a esta situação de catástrofe social, o ex-liberal Maurice Allais recomenda o fechamento comercial do mercado comum europeu, através do controle quantitativo de importações dos países extra-comunitários.

[11] De outro lado, renascem também as propostas utópicas onde há os que, como Ricardo Petrella – em recente artigo no "Le Monde Diplomatique" –, esperam que a ONU no seu próximo encontro de cúpula sobre a questão social, a ser

realizado em Copenhague em 1995, estabeleça as bases para uma nova ordem econômica e financeira mundial!

[12] Independentemente do caráter conservador ou utópico e da viabilidade técnica ou política de quaisquer destas propostas, é um consolo saber que as pessoas estão reaprendendo que a solução para o problema do desemprego, resultante da modernização conservadora e dos excessos do liberalismo, não pode ser simplesmente mais liberalismo.

[13] Finalmente, concluiriam que o governo americano não está baixando os impostos nem desregulando sua economia, mas regulando-a mais intensamente do que nunca, para enfrentar a concorrência dos países asiáticos e do Japão.

- **Avaliação do Extrato**

O texto-fonte deste extrato versa sobre política e apresenta vários conceitos implícitos, mesclando a notícia (encontro da cúpula dos ministros de trabalho) com a opinião da autora. Sendo assim, o próprio texto-fonte não é um texto de simples leitura. Essa propriedade naturalmente foi refletida no extrato.

De forma geral, o extrato permite identificar muito bem a parte que versa sobre a notícia e também os principais argumentos da autora. Todo o trecho que versa sobre a crítica ao empresariado brasileiro foi suprimido. Entretanto, a sentença [13] se refere ao empresariado brasileiro e acabou ficando sem sentido.

Capítulo 7

CONSIDERAÇÕES FINAIS

Nesta seção apresentam-se as conclusões principais deste trabalho, suas contribuições principais limitações e possíveis desdobramentos.

7.1 Principais Conclusões

Esta dissertação apresentou modelos de SA baseados em aprendizado de máquina, na Teoria dos Grafos e em diversas estatísticas textuais para o Português do Brasil. Esses modelos buscaram combinar e selecionar características diversas de SA, de forma a explorar diferentes aspectos do texto e, assim, levar à geração de sumários mais informativos e úteis.

Retomando-se as hipóteses formuladas no capítulo inicial desta dissertação:

Hipótese 1. A combinação de características de SA diversas pode permitir a consideração de diferentes fatores ou pontos de vista na análise da relevância das sentenças e, assim, levar a melhores extratos.

Hipótese 2. A pré-seleção automática de características pode configurar de forma adequada o sumarizador para um determinado corpus e levar aos melhores resultados. Ou seja, a pré-seleção automática de características pode ser uma forma de tratar o Problema 1 (*quais características utilizar para a SA?)* apresentado no Capitulo 1.

Hipótese 3. A forma de combinação e ponderação das características tem influência significativa na qualidade dos sumários. Quando maior a capacidade do modelo de ponderação em construir bons *rankings* de sentenças, melhor será o desempenho do sumarizador.

Conclui-se com relação a Hipótese 1 que não necessariamente quanto maior o número de características empregadas, melhor o desempenho do sumarizador. Isso foi verificado, por exemplo, quando se combinaram as características do SuPor-2 com as de redes complexas (Seção 5.3): o resultado obtido foi pior do que o obtido com a utilização simplesmente do conjunto de características do SuPor-2.

Entretanto, a combinação de características diferentes pode sim gerar bons modelos de SA. Por exemplo, ao se considerar a combinação das características de redes complexas. No trabalho de Antiqueira (2007), cada medida proposta originava uma máquina distinta de SA. Com o modelo de combinação explorado na Seção 5.3, o desempenho foi superado. Outra evidência foi o desempenho superior de todos os modelos com características combinadas em relação aos modelos de uma única característica baseada no método TextRank, conforme mostrado na Seção 6.6. Ou seja, deve existir um balanço ideal no número de características do sumarizador.

Com relação a Hipótese 2, a utilização de métodos de seleção automática de características, como o método CFS, é factível para a SA baseada em aprendizado de máquina, assim como é para vários problemas em Mineração de Dados. Entretanto, considerando o CFS, sua utilização parece compensar apenas para conjuntos maiores de características. Os resultados obtidos não permitem sugerir nenhum número crítico de características, conforme Seção 6.4.

Sobre a Hipótese 3, verificou-se sua validade. Assim, a escolha do modelo de ranking, isto é, do algoritmo de aprendizado de máquina, tem influência considerável nos modelos de SA supervisionados. Como as evidências teóricas e referências acadêmicas sobre esse ponto sugerem, classificadores probabilísticos ou que de certam forma modelam a incerteza ou imprecisão tem desempenho melhor para SA em relação aos demais. Neste trabalho, isso foi verificado pelos resultados obtidos com os modelos de ranking Bayesiano, por regressão logística ou nebuloso.

Partindo dos resultados obtidos através dos experimentos realizados, considera-se também relevante as formulações adicionais sugeridas a seguir:

a) A utilização de características numéricas e multinomiais, em oposição a características simplesmente binárias, tem em geral influência positiva na SA baseada em aprendizado de máquina. Este ponto é sugerido pelo fato do SuPor-2 ter desempenho relativamente superior ao SuPor, mantido o mesmo classificador, conforme resultados da Seção 6.2.

b) A utilização de métodos e características baseadas em grafos é interessante devido ao fato desses métodos serem de fácil cômputo e mesmo sendo independentes de língua, em geral, poderem explorar diferentes aspectos coesivos do texto. Entretanto, os resultados tendem a melhorar significantemente se for incorporado algum pré-processamento linguístico, como foi evidenciado pela construção dos modelos baseados no TextRank.

c) De forma geral, das três abordagens exploradas neste trabalho (estatísticas textuais, grafos e aprendizado de máquina) não é possível dizer qual é mais promissora ou útil, até porque os propósitos dessas abordagens são distintos e defende-se aqui sua combinação. As estatísticas textuais e as medidas de grafos podem ambas fornecer características úteis para SA. No estudo descrito na Seção 5.3.1, de avaliação de relevância das características, a melhor delas foi a Frequência de Palavras, estatística textual explorada desde 1958 para SA por Luhn. Entretanto, o escore foi seguido de perto por uma medida de redes complexas, baseada em grafos. O melhor modelo de características foi justamente o do SuPor-2 que combina tanto medidas estatísticas quanto métodos baseados em grafos. Com relação ao aprendizado de máquina, seu propósito é claro de construir um modelo de ranking de sentenças sem demandar o ajuste manual do modelo por um especialista, gargalo do modelo de Edmundson (1969).

7.2 Contribuições

A primeira contribuição apontada é a de que vários modelos de SA propostos e desenvolvidos neste trabalho superaram outros sumarizadores para o Português do Brasil, conforme apontando em (1) e (2) a seguir:

1) Considerando-se o SuPor-2, todos os experimentos conduzidos neste trabalho (ver Capítulo 6) indicaram-no como o que produz extratos mais

informativos, incluindo a própria replicação do experimento de Rino et al. (2004) que o comparou a outros oito sistemas. Além disso, o SuPor-2 também superou a utilização individual e combinada de medidas de redes complexas, como foi mostrado na Seção 6.4 e também apontado por Antiqueira et al. (2009). Entretanto, como reportado aqui e em Antiqueira et al., os testes não indicaram significância estatística. A condução de uma avaliação em um corpus maior poderia ser útil nessa comprovação.

Apresenta-se na Tabela 7-1 um quadro listando os sistemas com os quais o SuPor-2 foi comparado. A tabela mostra os métodos de combinação e ponderação explorados em cada um, o número e tipo de características empregadas, uma indicação estimada do custo computacional e os recursos dependentes de língua utilizados.

Sugere-se que o motivo principal do SuPor-2 superar os demais pode ser explicado pelo fato de combinar características diversas de SA por meio de um modelo de aprendizado de máquina Bayesiano, que se evidenciou mais promissor para o ranking de sentenças. Embora tanto o SuPor original quanto o ClassSumm também sigam uma estratégia semelhante, as diferenças principais são:

a) O SuPor utiliza a representação binária das características, empobrecendo o poder de decisão do modelo de aprendizado;

b) O ClassSumm não emprega características mais elaboradas como o SuPor-2 e o SuPor, que mapeiam métodos completos de SA como o de Cadeias Lexicais e Mapa de Relacionamentos em características. Além disso, o ClassSumm adota um processo de discretização de características antes da utilização do classificador Naïve-Bayes. No SuPor-2, essa discretização não é necessária pois o classificador trata características numéricas diretamente.

Tabela 7-1 – Quadro comparativo de sistemas de SA para o Português do Brasil

Sistema	Modelo de Combinação de Características	Exigência de Treino	Número de Características	Tipo das Características	Custo [30] Computacional	Recursos Dependentes de Língua Natural
SuPor-2	Flexible-Bayes	Sim	11	Numéricas e Multinominais	Médio	*Stoplist* Léxico *Stemmer* Etiquetador morfossintático *Thesaurus*
CN-Voting (Antiqueira et al. 2009)	Votação	Não	14	Numéricas	Médio	*Stoplist* *Stemmer*
SuPor (Módolo 2003)	Naïve-Bayes	Sim	11	Binárias	Médio	*Stoplist* Léxico *Stemmer* Etiquetador morfossintático *Thesaurus*
ClassSumm (Larocca Neto et al. 2002)	Naïve-Bayes	Sim	13	Numéricas e Multinominais	Médio	*Stoplist* *Stemmer* Etiquetador morfossintático Lista de palavras indicativas
SABio (Orrú et al. 2006)	Rede Neural Multicamada	Sim	7	Multinominais	Alto	*Stoplist* *Stemmer* Lista de palavras indicativas
TF-ISF-Summ (Larocca Neto et al. 2000)	Não há	Não	1	Numérica	Baixo	*Stoplist* *Stemmer*
GistSumm (Pardo et al. 2003)	Não há	Não	1	Numérica	Baixo	*Stoplist* Léxico
NeuralSumm (Pardo 2003)	Rede Neural SOM	Sim	8	Multinominais	Alto	*Stoplist* *Stemmer* Lista de palavras indicativas

Cabe apontar que todos os sistemas comparados apresentam certo grau de dependência de língua natural. Os dependem menos são os que demandam apenas recursos como *stoplist* e *stemmer* ou Léxico, como o GistSumm e o TF-ISF. Já o SuPor, SuPor-2 e ClassSumm dependem de mais recursos, como por exemplo um etiquetador morfossintático. Entretanto, a dependência encontra-se apenas nos recursos, amplamente disponíveis para o Inglês e cada vez mais sendo disponibilizados para outras línguas.

2) Uma outra contribuição com relação à evolução de índices foi a construção dos modelos baseados no TextRank, descritos na Seção 5.2. Através da

[30] Custo estimado por nós com base nos métodos empregados em cada sumarizador

incorporação de recursos linguísticos relativamente simples, as versões originais propostas por Mihalcea (2005) foram superadas, conforme se mostrou na Seção 6.3. Além disso, Antiqueira et al. (2009) também reportam que essas duas versões superaram o melhor modelo de redes complexas numa avaliação conduzida sobre o corpus TeMário-2003 e com a ROUGE-1 como métrica em foco. Entretanto, não foram conduzidos testes de significância estatística.

Agrupam-se ainda contribuições adicionais deste trabalho de mestrado em três categorias: implementações, contribuições teóricas e publicações acadêmicas.

a) Implementações

- Desenvolvimento do SuPor-2 a partir de uma reengenharia sobre o SuPor original e acoplamento à ferramenta WEKA, em linguagem Object Pascal (Delphi). Esse sumarizador foi implementado como um software independente, podendo ser disponibilizado para utilização;
- Desenvolvimento de modelos aperfeiçoados do TextRank, em linguagem Object Pascal (Delphi). Ambos os modelos foram desenvolvidos como softwares independentes e podem ser disponibilizados para utilização;
- Desenvolvimento de modelos de características combinadas do SuPor-2 e da área de Redes Complexas, em linguagem Perl. Em função da dependência do cálculo das medidas de redes complexas, não disponíveis livremente, a distribuição e utilização desses sumarizadores é restrita;
- Desenvolvimento do SuPor-2 Fuzzy em linguagem Java e Object Pascal (Delphi). O software é independente e pode ser disponibilizado para utilização.

b) Contribuições Teóricas

- A proposição do modelo de mapeamento dos métodos e medidas originalmente utilizadas no SuPor em características numéricas e multinomiais que vieram a ser utilizadas no SuPor-2 e mais posteriormente no SuPor-2 Fuzzy de forma adaptada;
- A exploração e proposição para SA de métodos de seleção automática de características;

- A exploração e construção de modelos de SA utilizando grandes números de características. Pelo nosso conhecimento, não há trabalho que tenha explorado além de 37 características;
- A proposição de um modelo de ranking nebuloso de sentenças.

c) Publicações acadêmicas

- Desenvolvimento do SuPor-2 (Leite e Rino 2006);
- Desenvolvimento dos modelos baseados no TextRank (Leite et al. 2007);
- Exploração combinada de características do SuPor-2 e da área Redes Complexas (Leite e Rino 2008);
- Desenvolvimento do SuPor-2 Fuzzy (Leite e Rino 2009).

7.3 Limitações

a) Como já citado, as avaliações conduzidas foram objetivas e automáticas, não focando na qualidade e textualidade dos extratos produzidos. Não foi feita, nenhuma avaliação manual dos extratos produzidos. Há também algumas críticas em trabalhos mais recentes sobre a efetividade de medidas automáticas, particularmente a ROUGE que foi empregada (e.g., Liu e Liu 2008);

b) Não foram conduzidas avaliações extrínsecas dos modelos de SA aqui propostos, por exemplo, verificando-se a potencialidade de aplicação em tarefas correlatas como a categorização de textos e QA;

c) Em algumas comparações entre sumarizadores não se calculou a significância estatística em função da indisponibilidade de alguns dados. Em alguns casos onde se calculou, não se obteve significância estatística nas diferenças entre os resultados de alguns modelos. Somente se essa significância fosse verificada, poderia-se afirmar mais formalmente a superioridade de um modelo em relação a outro. Nesse sentido, conforme já citado, a utilização de um corpus maior e a disponibilidade de mais sumários de referência poderiam ser úteis na verificação da significância estatística dos resultados.

d) Embora do ponto de vista teórico os modelos computacionais aqui propostos sejam independentes de língua, algumas características de SA

demandam para seu cálculo recursos dependentes de língua. A adaptação para outras línguas, como o inglês, exigiria a substituição de recursos como *léxico*, *stoplist*, *stemmer, thesaurus* e etiquetador morfossintático no caso do SuPor-2 por exemplo. A Tabela 6-6 resume os recursos necessários para os modelos propostos neste trabalho.

e) A implementação do método de Cadeias Lexicais no SuPor-2 é limitada, assim como no SuPor, devido a ausência para o Português do Brasil de uma WordNet completa. Em seu lugar, é utilizado um thesaurus (Dias-da-Silva et al. 2000) que não contempla todas as relações semânticas de uma WordNet e pode limitar a identificação de cadeias lexicais fortes.

f) Embora possam ser adaptados para tal finalidade, os modelos de SA aqui propostos realizam a sumarização monodocumento apenas.

g) Todas as características e modelos explorados para SA só foram avaliados para SA de textos jornalísticos.

7.4 Possíveis Trabalhos Futuros

Sugerem-se aqui alguns estudos e extensões que podem ser feitos em busca do refinamento dos modelos de SA, divididos em trabalhos práticos e teóricos. Além disso, as próprias limitações nas avaliações conduzidas podem ser objeto de trabalho futuro. Por exemplo, conduzindo-se avaliações manuais de textualidade ou avaliações *glass-box* dos modelos, buscando-se evidenciar os critérios de decisão adotados em cada um, conforme levantado na Seção 6.1.

7.4.1 Possíveis Trabalhos Futuros Práticos

a) Melhoria de usabilidade dos sistemas desenvolvidos, com a finalidade de disponibilização pública;

b) Portabilidade dos sumarizadores desenvolvidos para outras línguas, como o Inglês, através da substituição dos recursos dependentes de língua natural;

c) Aplicações direcionados dos sumarizadores desenvolvidos nesse trabalho, como por exemplo a integração com mecanismos de recuperação da informação e Q&A na Internet. Alguns exemplos:

- Sumarizacão de notícias sobre uma empresa listada na Bolsa de Valores para auxiliar investidores
- Sumarização de opiniões sobre artigos vendidos online
- Sumarização de grandes bases textuais: textos jurídicos e médicos

7.4.2 Possíveis Trabalhos Futuros Teóricos

a) **Balanceamento de Classes**

Na geração do treinamento segundo a abordagem de Kupiec et al. (1995), o sistema obtém a classe a partir da comparação com o extrato de referência. Se os extratos de referência têm, em média, 30% do tamanho dos textos-fonte, espera-se que, em média, 30% das tuplas (sentenças rotuladas) tenham a classe *True*. Ou seja, há um desbalanceamento, em que a classe dominante é a *False* (as sentenças não pertencem ao extrato). Muitos classificadores têm seu desempenho prejudicado em função desse desbalanceamento (Liu 2004)

Larocca Neto et al. (2000) também identificaram o problema, com o classificador C4.5, e comprovaram que os resultados são bastante prejudicados quando não há nenhum mecanismo para balancear o treinamento. Sugerem, para isso, a abordagem mais comum na literatura que é eliminar parte das tuplas da classe dominante ou replicar aleatoriamente tuplas da classe de menor ocorrência.

No caso dos modelos de SA baseados em AM e propostos aqui, as técnicas de tratamento de classes desbalanceadas poderiam ser investigadas também.

b) **Redes Bayesianas**

Neste trabalho, sugeriu-se que os classificadores mais adequados para SA baseada em AM devem ser os que são capazes de gerar bons rankings. Nesse quesito, os classificadores probabilísticos parecem ter vantagem.

Uma possibilidade de trabalho futuro é explorar Redes Bayesianas (e.g., Cooper e Herskovits 1992) que, embora ainda sigam a mesma ideia do aprendizado Bayesiano, envolvem técnicas mais sofisticadas para superar algumas desvantagens do Naïve-Bayes, como a da independência condicional que impõe a necessidade de todas as características serem estatisticamente independentes.

c) **Ensembles de Rankings**

A questão principal na SA baseada em AM é como gerar adequadamente o ranking de sentenças. Prati (2006) discute modelos de geração de *ensembles* de rankings. Ou seja, agregar a contribuição de vários classificadores num modelo de votação para produzir um único ranking, que pode ser útil na seleção das sentenças mais relevantes.

d) **Método de Seleção Automática de Características**

Pelo nosso conhecimento, o CFS tem sido o método *filter* de seleção de subconjuntos de características mais utilizado recentemente. Dada a influência que o processo de seleção de características pode representar para a SA e o fato de a Hipótese 2 proposta neste trabalho não ter sido verificada em vários modelos, outros métodos de seleção de características poderiam ser avaliados. Como exemplo, Prassad et al. (2004) propõe o método FSS_ICA com resultados próximos e em alguns casos superiores ao CFS. Além disso, segundo os autores, o FSS_ICA busca selecionar características estatisticamente independentes entre si, o que gera um desempenho melhor quando usado com o classificador Naïve-Bayes, dada a premissa de independência desse classificador. Outro método mais recente é o

INTERACT[31] (Zhao e Liu 2007) que também apresentou na avaliação dos autores desempenho ligeiramente superior ao CFS.

e) **Regras de Classificação**

Em substituição aos classificadores tradicionais explorados, como o Bayesiano, poderiam-se explorar modelos de ranking de sentenças baseados em regras de classificação (e.g., Witten e Frank 2005). A vantagem da utilização de modelos desse tipo é que poderiam permitir mais facilmente a compreensão dos critérios de decisão delineados por meio do Aprendizado de Máquina.

f) **Refinamento do Cálculo da Similaridade entre Sentenças nos Métodos Baseados em Grafos**

As variações do TextRank produzidas neste trabalho refinaram o modo como as similaridades entre as sentenças são calculadas e obtiveram desempenho superior ao método original. Particularmente, a variação TextRank+Thesaurus foi a mais promissora.

Como continuidade dessa linha, poderia-se em vez de utilizar um thesaurus, dependente de língua, buscar de forma automática essas relações da Internet. Para isso, poderia-se partir do trabalho de Turney (2001), que propõe a utilização dos resultados de pesquisas feitas em algum buscado como fonte de informações. Outra possibilidade seria explorar conjuntos de palavras relacionadas providos pela ferramenta GoogleSets (labs.google.com/sets). Essa ferramenta gera conjuntos de palavras relacionadas a partir de conjuntos menores.

[31] Implementação disponível em http://www.public.asu.edu/~huanliu/INTERACT/INTERACTsoftware.html (Setembro/2010)

REFERÊNCIAS BIBLIOGRÁFICAS

ALBERT, R., BARABÁSI, A.L. 2002. Statistical mechanics of complex networks. Rev. Mod. Phys. 74, 47–97.

ANTIQUEIRA, L. 2007. Desenvolvimento de técnicas baseadas em redes complexas para sumarização extrativa de textos. Dissertação de Mestrado. Universidade de São Paulo - ICMC, São Carlos.

ANTIQUEIRA, L., OLIVEIRA JR., O.N., DA F. COSTA, L., NUNES, M.G.V. 2009. A Complex Network Approach to Text Summarization. Information Sciences 179, 584-599.

BALAGE FILHO, P.P., PARDO, T.A.S., NUNES, M.G.V. 2006. Estrutura Textual e Multiplicidade de Tópicos na Sumarização Automática: o Caso do Sistema GistSumm. Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação. Universidade de São Paulo, São Carlos, 18p.

BARZILAY, R., ELHADAD, M. 1999. Using Lexical Chains for Text Summarization. In: Advances in Automatic Text Summarization, I. MANI, M.T. MAYBURY Eds. MIT Press, Cambridge, 111-121.

BATAGELJ, V., ZAVERSNIK, M. 1999. Partitioning approach to visualization of large networks. In: Proc. of the Graph Drawing: 7th International Symposium (GD'99), Lecture Notes in Computer Science 1731, 90-98.

BIESIADA, J., DUCH, W., KACHEL, A., MACZKA, K., PALUCHA, S. 2005. Feature ranking methods based on information entropy. In: Proceedings of the International Conference on Research in Electrotechnology and Applied Informatics, Katowice, Poland.

BOHN, R.E., SHORT, J.E. 2009. How Much Information? 2009 Report on American Consumers. University of California, San Diego, 36p.

BRIN, S., PAGE, L. 1998. The anatomy of a large-scale hypertextual Web search engine. In: Computer Networks and ISDN Systems, 7.

CLAUSET, A., NEWMAN, M.E.J., MOORE, C. 2004. Finding community structure in very large networks. Phys. Rev. E 70, 66-111.

COLLOVINI, S., CARBONEL, T., FUCHS, J. T., COELHO, J.C., RINO, L., VIEIRA, R. 2007. Summit: Um corpus anotado com informações discursivas visando à sumarização automática. In: Proc. of the V Workshop on Information and Human Language Technology (TIL'2007) - XXVII Congresso da Sociedade Brasileira de Computação (SBC'2007), Rio de Janeiro-RJ.

COOPER, G., HERSKOVITS, E. 1992. A Bayesian Method for the Induction of Probabilistic Networks from Data. Machine Learning 9, 309-347.

COSTA, L.F., DA ROCHA, L.E.C. 2006. A generalized approach to complex networks. Eur. Phys. J. B 50, 237-242, cond-mat/0408076.

COSTA, L.F., KAISER, M., HILGETAG, C. 2006. Beyond the average: detecting global singular nodes from local features in complex networks. Physics 0607272

COSTA, L.F., RODRIGUES, F.A., TRAVIESO, G., VILLAS BOAS, P.R. 2006. Characterization of complex networks: A survey of measurements. In: cond-mat/0505185.

DEMSAR, J. 2006. Statistical Comparisons of Classifiers over Multiple Data Sets. Journal of Machine Learning Research 7, 1-30.

DIAS-DA-SILVA, B.C., MORAES, H.R.D., OLIVEIRA, M.F.D., HASEGAWA, R., AMORIM, D., PASSCHOALINO, C., NASCIMENTO, A.C. 2000. Construção de um thesaurus eletrônico para o português do Brasil. In: Proc. of the V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR 2000), Atibaia-SP, 1-10.

EBERHART, R., SHI, Y. 2007. Computation Intelligence: Concepts to Implementations. Morgan Kaufman, New York.

EDMUNDSON, H.P. 1969. New methods in automatic extracting. Journal for Computing Machinery 16, 264-285.

EFRON, B., HASTIE, T., JOHNSTONE, I.M., TIBSHIRANI., R. 2004. Least angle regression. Annals of Statistics 32, 407-499.

FAYYAD, U., IRANI, K. 1993. Multi-interval discretization of continuous-valued attributes for classification learning. In: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'93), Chambery/France, 1022-1027.

GALANIS, D., MALAKASIOTIS, P. 2008. AUEB at TAC 2008. In: Proc. of the TAC - Text Analysis Conference, Maryland, USA.

GELBUKH, A., SIDOROV, G. 2001. Zipf and Heaps Laws - Coefficients Depend on Language. In: Proc. Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2001), Mexico City, Springer-Verlag, 332–335.

HALL, A.M. 2000. Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning. In Proc. of 17th International Conference on Machine Learning, Morgan Kaufmann, San Francisco-CA, 359-366.

HALLIDAY, M.A.K., HASAN, R. 1976. Cohesion in English. Longman.

HAYKIN, S. 1999. Neural Networks: A Comprohensive Foundation. Prentice-Hall.

HEARST, M.A. 1993. TextTiling: A Quantitative Approach to Discourse Segmentation. Technical Report. University of California, Berkeley, 24p.

HERRERA, F., LOZANO, M., VERDEGAY, J.L. 1993. Genetic Algorithm Applications to Fuzzy Logic Based Systems. In: Proc. of the 9th Polish-Italian and 5th Polish-Finnish Symposium on Systems Analysis and Decision Support in Economics and Technology, Warsaw, 125-134.

JOHN, G., LANGLEY, P. 1995. Estimating continuous distributions in Bayesian classifiers. In Proc. of the 11th Conference on Uncertainty in Artificial Intelligence, P. BESNARD, S. HANKS Eds., Quebec, Canada, 338-345.

JUNIOR, J.C., IMAMURA, C.Y.M., REZENDE., S.O. 2001. Avaliação de um Algoritmo de Stemming para a Língua Portuguesa. In: Proc. of the 2nd Congress of Logic Applied to Technology (LABTEC'2001), São Paulo, 267-274.

KIANI-B, A., AKBARZADEH-T, M.R. 2006. Automatic Text Summarization Using: Hybrid Fuzzy GA-GP. In: Proc. of the IEEE Congress on Evolutionary Computation, Vancouver, Canada, IEEE Press, 5465-5471.

KLEINBERG, J.M. 1999. Authoritative sources in hyperlinked environment. Journal of the ACM 46, 604-632.

KLIR, G.J., YUAN, B. 1995. Fuzzy Sets and Fuzzy Logic – Theory and Applications. Pretince Hall, New Jersey.

KOHONEN, T. 1990. The self-organizing map. IEEE 78, 14.

KUPIEC, J., PETERSEN, J., CHEN, F. 1995. A trainable document summarizer. In: Proc. of the 18th ACM-SIGIR Conference on Research & Development in Information Retrieval, E.A. FOX, P. INGWERSEN, R. FIDEL Eds., Seatlle, WA, USA, 68-73.

LAROCCA NETO, J., FREITAS, A.A., KAESTNER, C.A.A. 2002. Automatic text summarization using a machine learning approach. In: Proc. of 16th Brazilian Symposium on Artificial Intelligence (SBIA'02), Lecture Notes in Artificial Intelligence 2057, 205-215.

LAROCCA NETO, J., SANTOS, A.D., KAESTNER, C.A.A., FREITAS, A.A. 2000. Document clustering and text summarization. In Proc. of the 4th Int. Conf. Practical Applications of Knowledge Discovery and Data Mining, Manchester, UK, 41–55.

LAROCCA NETO, J., SANTOS, A.D., KAESTNER, C.A.A., FREITAS, A.A. 2000. Generating Text Summaries through the Relative Importance of Topics. In: Proc. of the 15th Brazilian Symposium on Artificial Intelligence (SBIA'2000), Lecture Notes in Artificial Intelligence 1952, Springer-Verlag, 300-309.

LEITE, D.S., RINO, L.H.M. 2006. Selecting a Feature Set to Summarize Texts in Brazilian Portuguese. In: Proc. of the International Joint Conference IBERAMIA/SBIA

2006, Ribeirão Preto-SP, Lecture Notes in Computer Science 4140, Springer-Verlag, 462-471.

LEITE, D.S., RINO, L.H.M. 2008. Combining Multiple Features for Automatic Text Summarization through Machine Learning. In: Proc. of the International Conference on Computational Processing of Portuguese Language (PROPOR'2008), Aveiro, Portugal, Lecture Notes in Artificial Intelligence 5190, Springer-Verlag, 122-132.

LEITE, D.S., RINO, L.H.M. 2009. A Genetic Fuzzy Automatic Text Summarizer. In: Proc. of VII Encontro Nacional de Inteligência Artificial - XXIX Congresso da Sociedade Brasileira de Computação (CSBC 2009), Bento Gonçalves-RS, 1-10.

LEITE, D.S., RINO, L.H.M., PARDO, T.A.S., NUNES, M.G.V. 2007. Extractive Automatic Summarization: Does more linguistic knowledge make a difference? In: Proc. of the Workshop on TextGraphs-2: Graph-Based Algorithms for Natural Language Processing (NAACL 2007), Rochester-NY, C. BIEMANN, I. MATVEEVA, R. MIHALCEA, D. RADEV Eds. Association of Computational Linguistics, 17-24.

LI, S., OUYANG, Y., WANG, W., SUN, B. 2007. Multi-Document Summarization Using Support Vector Regression. In Proceedings of the Document Understanding Conference, Rochester-NY.

LIN, C., HOVY, E.H. 2003. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In Proc. of the Language Technology Conference (HLT-NAACL 2003), Edmonton, Canadá, 71-78.

LIU, A. 2004. The Effect of Oversampling and Undersampling on Classifying Imbalanced Text Datasets. Máster Thesis. University of Texas.

LIU, F., LIU, Y. 2008. Correlation between ROUGE and Human Evaluation of Extractive Meeting Summaries. In: Proc. of the 46th Annual Meeting of the Association for Computational Linguistics (ACL2008), Columbus, Ohio, USA, 201-204.

LUHN, H. 1958. The automatic creation of literature abstracts. IBM Journal of Research and Development 2, 159-165.

MANI, I. 2001. Automatic Summarization. John Benjamin's Publishing Company, Amsterdam.

MIHALCEA, R. 2005. Language Independent Extractive Summarization. In: Proc. of the 43th Annual Meeting of the Association for Computational Linguistics (ACL2005), Ann Arbor, MI, 49-52.

MILLER, G.A., BECKWITH, R., FELLBAUM, C., GROSS, D., MILLER, K. 1990. Introduction to WordNet: An On-line Lexical Database. International Journal of Lexicography 3, 235-244.

MITCHEL, T.M. 1997. Machine Learning. McGraw Hill.

MÓDOLO, M. 2003. SuPor: an Environment for Exploration of Extractive Methods for Automatic Text Summarization for Portuguese (in Portuguese). MSc. Dissertation. Departamento de Computação UFSCar, São Carlos, SP, Brasil.

NENKOVA, A., PASSONNEAU, R., MCKEOWN, K. 2007. The Pyramid Method: Incorporating human content selection variation in summarization evaluation. ACM Transactions on Speech and Language Processing 4(2), 1-23.

NETO, J.L., ALEXANDRE, N., SANTOS, D., KAESTNER, C.A.A., FREITAS, A.A., NIEVOLA, J.C. 2000. A Trainable Algorithm for Summarizing News Stories. In Proc. of the 4th European Conference on Principles and Practice of Knowledge Discovery in Databases, Lyon, France

ORRÚ, T., ROSA, J.L., ANDRADE NETTO, M.L. 2006. SABio: An Automatic Portuguese Text Summarizer Through Artificial Neural Networks in a More Biologically Plausible Model. In: Proc. of the International Conference on Computational Processing of Portuguese Language (PROPOR'2006), Itatiaia - RJ, Springer-Verlag, 11-20.

PARDO, T.A.S., RINO, L.H.M. 2003. TeMário: Um Corpus para Sumarização Automática de Textos. Série de Relatórios do NILC. NILC-TR-03-09. São Carlos-SP, Outubro, 13p.

PARDO, T.A.S., RINO, L.H.M., NUNES, M.G.V. 2003. GistSumm: A Summarization Tool Based on a New Extractive Method. In: Proc. of the International Conference on Computational Processing of Portuguese Language (PROPOR'2003), Algarve-FCHS, Faro, Portugal, Lecture Notes in Artificial Intelligence 2721, Springer-Verlag, 210-218.

PARDO, T.A.S., RINO, L.H.M., NUNES, M.G.V. 2003. NeuralSumm: Uma Abordagem Conexionista para a Sumarização Automática de Textos. In: Proc. of IV Encontro Nacional de Inteligência Artificial (ENIA'2003) - XXII Congresso Nacional da Sociedade Brasileira de Computação, Campinas-SP.

PARDO, T.A.S., RINO, L.H.M. 2004. Descrição do GEI - Gerador de Extratos Ideais para o Português do Brasil. Série de Relatórios do NILC. NILC-TR-04-07. São Carlos-SP, Agosto, 10p.

PINGALI, P., R.K., VARMA, V. 2007. IIIT Hyderabad at DUC 2007. In: Proc. of the Document Understanding Conference, Rochester, NY, USA.

PORTER, M.F. 1980. An Algorithm for Suffix Stripping. In Program, 130-137.

PRASAD, M., SOWMYA, A., KOCH, I. 2004. Efficient feature selection based on independent component analysis. In: Proceedings of the Intelligent Sensors, Sensor Networks and Information Processing Conference, 427-432.

PRATI, R.C. 2006. Novas abordagens em aprendizado de máquina para a geração de regras, classes desbalanceadas e ordenação de casos. Tese de Doutorado. Universidade de São Paulo - ICMC, São Carlos-SP.

QUINLAN, J.R. 1993. C4.5 Programs for machine learning. Morgan-Kaufman, San Mateo.

RINO, L.H.M., PARDO, T.A.S, SILLA JR., C.N., KAESTNER, C.A.A., POMBO, M 2004. A Comparison of Automatic Summarization Systems for Brazilian Portuguese Texts. In: Proc. f XVII Brazilian Symposium on Artificial Intelligence (SBIA'04), São Luís, Maranhão, Brazil, Lecture Notes in Computer Science 3171, Springer-Verlag, 235-244.

SALTON, G., BUCKLEY, C. 1988. Term-Weighting Approaches in Automatic Text Retrieval. Information Processing & Management 24, 513-523.

SALTON, G., SINGHAL, A., MITRA, M., BUCKLEY, C. 1997. Automatic Text Structuring and Summarization. Information Processing & Management 33, 193-207.

SCHILDER, F., KONDADADI, R., LEIDNER, J.L., CONRAD, J.G. 2008. Thomson Reuters at TAC 2008: Aggressive Filtering with FastSum for Update and Opinion Summarization. In: Proc. of the TAC - Text Analysis Conference, Maryland, USA.

SHANNON, C.E. 1948. A Mathematical Theory of Communication Bell System Tech. J., 379–423.

SPÄRCK JONES, K. 1997. Summarising: Where are we now? Where should we go? In: Proc. of the Intelligent Scalable Text Summarization Workshop, ACL/EACL'97 Joint Conference, Madrid, Spain.

SPÄRCK JONES, K. 1999. Automatic summarizing: Factors and directions. In Advances in Automatic Text Summarization, I.M. MANI, M. T Ed. MIT Press, 1-12.

SPÄRCK JONES, K. 2007. Spärck Jones, K. Information Processing and Management 43, 1449-1481.

SPÄRCK JONES, K., GALLIERS, J.R. 1996. Evaluating Natural Language Processing Systems. Lecture Notes in Artificial Intelligence 1083.

TRATZ, S., HOVY, E.H. 2008. Summarization Evaluation Using Transformed Basic Elements. In: Proc. of Text Analytics Conference (TAC-08), Gaithersburg, MD.

TURNEY, P. 2001. Mining the web for synonyms. In: Proc. of the 12th European Conference on Machine Learning (ECML'01), Lecture Notes in Computer Science 2167, Springer-Verlag, 491-502.

VAPNIK, V. 1995. The Nature of Statistical Learning Theory. Springer-Verlag.

VAPNIK, V.N. 1998. Statistical learning theory. John Wiley and Sons, New York.

WANG, L., MENDEL, J. 1992. Generating fuzzy rules by learning from examples. IEEE Trans. on SMC 22, 414-427.

WATTS, D.J., STROGATZ, S.H. 1998. Collective dynamics of 'small-world' networks. Nature 393, 440–442.

WITTEN, I.H., FRANK, E. 2005. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, San Francisco.

WONG, K.-F., WU, M., LI, W. 2008. Extractive Summarization Using Supervised and Semi-Supervised Learning. In: Proc. of the 22nd International Conference on Computational Linguistics (Coling 2008), Manchester, UK, 985-992.

ZADROZNY, B., ELKAN, C. 2001. Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. In: Proc. of the Eighteenth International Conference on Machine Learning, Morgan Kaufmann, 609-616.

ZHANG, H., SU, J. 2004. Naive Bayesian Classifiers for Ranking. . In Proc. of the 15th European Conference on Machine Learning (ECML'04), Pisa, Italy, Springer-Verlag, 501-512.

ZHAO, Z., LIU, H. 2007. Searching for Interacting Features. In Proc. of the International Joint Conference on Artificial Intelligence (IJCAI'07), 1156-1161.

ZIPF, G.K. 1949. Human Behaviour and the Principle of Least-Effort. Addison-

Apêndice A

Exemplo de Utilização do Modelo de Classificação Bayesiana com Características Binárias

Como exemplo de aplicação da fórmula de Bayes para SA, considere o pequeno conjunto de treinamento da tabela seguinte. Cada linha dessa tabela representa uma sentença processada do corpus de *treinamento*. Considere também uma taxa de compressão de 10% e um conjunto de três características em análise. Algumas dessas tuplas foram selecionadas de um conjunto real de treinamento obtido para o corpus TeMário-2003.

C_1	C_2	C_3	**Presença no extrato**
False	False	False	False
False	True	False	False
False	False	True	**True**
False	True	False	False
True	False	False	**True**
True	True	False	False
False	True	False	**True**
False	True	False	False

Suponha que, após a aplicação dos três métodos ao texto, uma certa sentença obteve os valores *True*, *False* e *True* para as características C_1, C_2, e C_3, respectivamente. O cálculo da probabilidade dessa sentença pertencer ao extrato é mostrado a seguir.

1) Cálculo de $P(s \in E)$

$$P(s \in E) = \frac{100 - 90}{100} = \frac{1}{10}$$

2) Cálculo de $P(C_j \mid s \in E)$

$$P(C_1 = True \mid s \in E) = 1/3$$

$$P(C_2 = False \mid s \in E) = 2/3$$

$$P(C_3 = True \mid s \in E) = 1/3$$

3) Cálculo de $P(C_j)$

$$P(O_2 = True) = 2/8 = 1/4$$

$$P(O_3 = False) = 3/8$$

$$P(O_5 = True) = 1/8$$

4) Cálculo de $P((s \in E) \mid (C_1 = True, C_2 = False, C_3 = True)$

$$P((s \in E) \mid (O_1 = True, O_2 = False, O_3 = True) = \frac{\left(\frac{1}{3} \times \frac{2}{3} \times \frac{1}{3}\right) \times \left(\frac{1}{10}\right)}{\frac{1}{4} \times \frac{3}{8} \times \frac{1}{8}} = \frac{512}{810} \approx 63{,}21\%$$

Assim, a probabilidade dessa sentença pertencer ao extrato é de 63,21%.

Apêndice B

Exemplos de Cálculo das Medidas Information Gain e Qui-Quadrado

Como exemplo de uso da medida *Information Gain* e Qui-quadrado, considere o seguinte conjunto de treino, considerando-se 11 características hipotéticas:

C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	**Presença no extrato**
False	False	False	True	False	False	False	False	False	False	False	False
False	False	True	True	False	False	False	False	False	False	False	False
False	False	False	True	True	True	True	False	False	False	False	**True**
False	False	True	True	False	False	True	False	True	False	False	False
False	True	False	True	False	True	False	False	True	False	True	**True**
False	True	True	True	False	False	False	False	False	False	False	False
False	False	True	False	False	True	False	False	False	False	False	**True**
False	False	True	True	False	False	False	False	False	False	False	False

-2003.

Note que a característica C_1 parece pouco informativa, pois o seu valor para todas as tuplas é *False*, e a classe (*presença no extrato*) ora é *False*, ora é *True*. O contrário parece ocorrer com a característica C_6: em todas as tuplas em que ela assume o valor *True*, a classe também é *True*. E em todas as tuplas em que ela assume o valor *False*, a classe também é *False*. Assim, intuitivamente, a característica C_6 parece ser muito mais relevante que a C_1. O cálculo do *IG* é mostrado a seguir:

- Para C_1

$$IG(\text{Pr}\,esencaExtrato, C_1) = H(\text{Pr}\,esencaExtrato) - H(\text{Pr}\,esencaExtrato, C_1)$$

1) Cálculo da entropia da classe:

$P(\text{Pr } esencaExtrato = true) = 3/8$

$P(\text{Pr } esencaExtrato = false) = 5/8$

$$H(\text{Pr } esencaExtrato) = -\frac{3}{8}\times\log_2\left(\frac{3}{8}\right) - \frac{5}{8}\times\log_2\left(\frac{5}{8}\right) = 0.95 \text{ bit}$$

2) Cálculo da entropia condicional

$P(C_1 = false) = 1$

$P(C_1 = true) = 0$

$$P(\text{Pr } esencaExtrato = false \mid C_1 = false) = \frac{5}{8}$$

$$P(\text{Pr } esencaExtrato = true \mid C_1 = false) = \frac{3}{8}$$

$P(\text{Pr } esencaExtr\,ato = false \mid C_1 = true) = 0$

$P(\text{Pr } esencaExtrato = true \mid C_1 = true) = 0$

$$H(\text{Pr } esencaExtrato, C_1) = -\left(1\times\left(\frac{3}{8}\times\log_2\left(\frac{3}{8}\right) + \frac{5}{8}\times\log_2\left(\frac{5}{8}\right)\right)\right) = 0.95 \text{ bit}$$

3) Cálculo do ganho de informação

$IG(\text{Pr } esencaExtrato, C_1) = 0.95 - 0.95 = 0 \text{ bit}$

Como era esperado, não há ganho de informação em se saber o valor do da característica C_1. A incerteza *(H)* sobre o valor que a classe assume é a mesma depois de se saber o valor dessa característica.

- Para C_6

$$IG(\text{Pr } esencaExtrato, C_6) = H(\text{Pr } esencaExtrato) - H(\text{Pr } esencaExtrato, C_6)$$

1) Cálculo da entropia da classe:

$P(\Pr\ esencaExtrato = true) = 3/8$

$P(\Pr\ esencaExtrato = false) = 5/8$

$$H(\Pr\ esencaExtrato) = -\frac{3}{8}\times\log_2\left(\frac{3}{8}\right) - \frac{5}{8}\times\log_2\left(\frac{5}{8}\right) = 0.95 \text{ bit}$$

2) Cálculo da entropia condicional

$$P(C_6 = false) = \frac{5}{8}$$

$$P(C_6 = true) = \frac{3}{8}$$

$P(\Pr\ esencaExtrato = false \mid C_6 = false) = 1$

$P(\Pr\ esencaExtrato = true \mid C_6 = false) = 0$

$P(\Pr\ esencaExtrato = false \mid C_6 = true) = 0$

$P(\Pr\ esencaExtrato = true \mid C_6 = true) = 1$

$$H(\Pr\ esencaExtrato, C_6) = -\left(\frac{5}{8}\times(1\times 0 + 0) + \frac{3}{8}\times(1\times 0 + 0)\right) = 0 \text{ bits}$$

3) Cálculo do ganho de informação

$IG(\Pr\ esencaExtrato, C_1) = 0.95 - 0 = 0.95$ bits

Neste caso, nota-se que a característica C_6 tem o máximo ganho de informação possível e a incerteza é drasticamente reduzida quando ela é considerada. Com certeza, essa característica deveria ser selecionada para o processo de aprendizado de máquina.

A seguir, exemplifica-se o cálculo do qui-quadrado. Inicialmente, constrói-se a seguinte tabela, que sintetiza os valores necessários.

Presença no extrato (classe)	**Característica C_6**		**Total**
	True	**False**	
True	3	0	3
False	0	5	5
Total	3	5	8

Aplica-se, então, a estatística:

1) Cálculo das frequências esperadas

$$E_{11} = \frac{3 \times 3}{8} = \frac{9}{8}$$

$$E_{21} = \frac{5 \times 3}{8} = \frac{15}{8}$$

$$E_{12} = \frac{5 \times 3}{8} = \frac{15}{8}$$

$$E_{22} = \frac{5 \times 5}{8} = \frac{25}{8}$$

2) Cálculo do χ^2

$$\chi^2 = \frac{\left(3 - 9/8\right)^2}{9/8} + \frac{\left(0 - 15/8\right)^2}{15/8} + \frac{\left(0 - 15/8\right)^2}{15/8} + \frac{\left(5 - 25/8\right)^2}{25/8} = 8$$

Apêndice C

Análise de Significância Estatística na Comparação Entre Sistemas de Sumarização Automática

Para a comparação e a análise dos resultados de sistemas de sumarização automática pode ser útil a verificação de que as diferenças encontradas nos resultados são estatisticamente significantes. Em outras palavras, busca-se provar que as diferenças nos resultados não se devem meramente ao acaso. Em estatística, isso equivale a aplicar um teste de hipóteses para verificar se as médias indicadas pelas amostras (resultados dos sistemas, no caso) são iguais ou não.

O teste para pares de sumarizadores permite analisar a significância estatística no desempenho de dois sumarizadores. Suponha que se tenha utilizado quatro sumarizadores num dado experimento (S1, S2, S3, S4). Os possíveis pares a serem analisados com o teste estatístico são: S1S2, S1S3, S1S4, S2S3, S2S4 e S3S4).

Na análise de um par de sumarizadores, têm-se duas hipóteses, então:

H0: é a hipótese nula, ou seja, a que se quer rejeitar: *As diferenças entre os dois sistemas não são significativas.*

H1: *As diferenças entre os sistemas são significativas.*

O método *t-student* emparelhado (*matched-pair t-student test*) é o mais usado para teste de significância estatística entre pares de modelos, que em nosso caso são sumarizadores. O emparelhamento refere-se ao fato de que ambos os

sumarizadores trabalham exatamente com os mesmos dados, ou seja, cada texto é sumarizado por ambos. O teste também é recomendado, de maneira geral, para a avaliação de algoritmos de aprendizagem no campo do Aprendizado de Máquina, conforme aponta Mitchel (1997).

No contexto da avaliação de sumários (ou extratos) automáticos, as premissas do teste *t-student* emparelhado são:

i) O teste só pode ser aplicado para uma única métrica de avaliação de sumários por vez. Por exemplo, se num experimento utilizou-se a medida *Recall* ROUGE-1 e ROUGE-2, serão necessários dois testes. Não é possível realizar o teste para várias métricas simultaneamente. Cada teste foca em apenas uma métrica de avaliação de sumários. Vale lembrar que o resultado do teste pode ser significativo considerando uma métrica de avaliação de sumários e não significativo se considerada uma outra métrica.

ii) Assume-se uma distribuição normal da população. Isso equivale ao seguinte: Seja X a variável aleatória que representa a medida de avaliação (por exemplo, ROUGE-1) de todos os sumários possíveis de serem produzidos por um sistema de sumarização, ou seja, de toda a população de sumários. Então, X segue uma distribuição normal. Isso pode ser verificado pelo teste de Shapiro-Wilk que verifica se os dados seguem uma distribuição normal.

Entretanto, quando o tamanho da amostra for considerado grande ($n > 30$), o Teorema Central do Limite garante que as médias amostrais serão aproximadamente normalmente distribuídas, e tenderão a uma distribuição normal à medida que o tamanho de amostra crescer. Então pode-se ter uma variável original com uma distribuição muito diferente da Normal, mas ao se tomar várias amostras grandes desta distribuição, e então se construir um histograma das médias amostrais, a forma se parecerá como uma curva Normal. Nessa situação, o t-teste ainda produz resultados confiáveis (Mitchel 1997) e pode-se ignorar o teste de normalidade de Shapiro-Wilk.

A fórmula do teste *t-student* é a seguinte:

$$t = (\overline{A} - \overline{B}) \sqrt{\frac{n \times (n-1)}{\sum_{i=1}^{n} \left((A_i - \overline{A}) - (B_i - \overline{B}) \right)^2}}$$

em que:

n é o tamanho da amostra, ou o número de pares de sumários automáticos gerados pelos sistemas A e B;

(*n – 1)* é o termo que recebe o nome especial de grau de liberdade do modelo;

$\overline{A}$ e $\overline{B}$ são as médias das medidas de avaliação dos sumários para os sistemas A e B, respectivamente;

A_i e B_i são as medidas de avaliação do sumário *i*, dos sistemas A e B, respectivamente.

Na aplicação do teste, pode-se proceder de duas formas: ou define-se previamente um nível de significância desejado para a tarefa sob teste (Caso I) ou busca-se um valor que indica a probabilidade de significância estatística, o chamado p-valor (Caso II).

Caso I. Define-se um nível de significância estatística, calcula-se o valor *t* usando a fórmula (I) e compara-se com o valor crítico de *t* apresentado na tabela seguinte. O uso de uma tabela de valores críticos para *t* é usual em estatística para evitar o cálculo integral envolvido nas fórmulas mais gerais. Assim, determina-se o valor crítico (células destacadas) buscando-se a linha cujo grau de liberdade é *n* – 1 e a coluna for o nível de significância estabelecido. Geralmente, essas tabelas, que são trazidas nos livros de Estatística, não são completas. Por isso, pode ocorrer de não se encontrar o valor de t crítico mais adequado. Por exemplo, para n = 9 não temos t crítico tabelado. Nessa situação, pode-se optar por fazer uma interpolação entre os t críticos apresentados para os graus de liberdade 5 e 10, presentes na tabela, ou adotar a solução analítica do problema, explorada no Caso II.

Se o valor de t calculado pela fórmula do teste for maior que o valor crítico encontrado, então as diferenças são estatisticamente significantes e a hipótese nula é rejeitada. Frequentemente, o valor α = 0,05 é utilizado para a análise de

significância, mas outros valores menos comuns para α podem ser usados, como mostra a tabela seguinte:

Grau de liberdade n-1	**Nível de Significância (α)**			
	0,1	**0,05**	**0,025**	**0,005**
1	3,07768	6,31375	1,27062	6,36567
2	1,88562	2,91999	4,30265	9,92484
3	1,63774	2,35336	3,18245	5,84091
4	1,53321	2,13185	2,77645	4,60409
5	1,47588	2,01505	2,57058	4,03214
10	1,37218	1,81246	2,22814	3,16927
30	1,31042	1,69726	2,04227	2,75000
100	1,29007	1,66023	1,98397	2,62589
∞	1,28156	1,64487	1,95999	2,57588

Caso II. Em vez de se utilizar valores tabelados para verificar se há significância estatística, calcula-se diretamente um número que indica a probabilidade de significância, o chamado p-valor, pela seguinte fórmula em que B indica a função Beta do cálculo integral. Quanto mais próximo de 0 for o p-valor, maior a probabilidade de significância.

$$p-valor = \frac{1}{\sqrt{n-1}\times \mathrm{B}\left(\frac{1}{2},\frac{n-1}{2}\right)}\int_{-t}^{t}\left(1+\frac{x^2}{n-1}\right)^{-\frac{n}{2}}dx$$

$$B(x,y)=\int_0^t\left(t^{x-1}(1-t)^{y-1}\right)dt$$

A fórmula do p-valor é incorporada a vários pacotes estatísticos e, assim, também no Microsoft Excel, sendo de uso bastante frequente para indicar o nível para o qual as diferenças entre os sistemas avaliados são significantes.

Exemplo numérico

Dados os resultados de avaliação de dois sistemas extrativos para um corpus de teste com 3 textos-fonte, mostrados na tabela a seguir, verifica-se se as diferenças são significativas num nível de significância de 0,05 (5%), segundo o Caso I. Por fins didáticos, o tamanho da amostra é pequeno (3 extratos) e não segue uma distribuição normal. Dessa forma, não se deveria confiar num teste t quando apenas 3 extratos forem produzidos e não seja dada nenhuma evidência que a distribuição segue uma distribuição normal.

Extrato	**Recall ROUGE-1**	
	Sistema A	**Sistema B**
1	0,59	0,39
2	0,58	0,44
3	0,57	0,45

Para:

$n = 3$ (número de extratos)

$n - 1 = 2$ (graus de liberdade)

$\overline{A} = 0{,}5800$

$\overline{B} = 0{,}4266$

o valor t calculado pela fórmula do teste é $\boldsymbol{t = 6{,}3790}$.

$$t = (0{,}5800 - 0{,}4266) \sqrt{\frac{3 \times (3-1)}{((0{,}59 - 0{,}5800) - (0{,}39 - 0{,}4266))^2 + ((0{,}58 - 0{,}5800) - (0{,}44 - 0{,}4266))^2 + ((0{,}57 - 0{,}5800) - (0{,}45 - 0{,}4266))^2}}$$

Buscando o valor crítico na tabela da distribuição t, vê-se que ele é menor que o valor encontrado (2,9200 < 6,3790), então se rejeita a hipótese nula e conclui-se que as diferenças são estatisticamente significantes.

Pode-se também calcular o p-valor (Caso II). Com o auxílio da função TESTET do Microsoft Excel, obtemos o *p-valor* **0,02370**.

Apêndice D

Comparação Entre Características do SuPor-2 e de Redes Complexas por Métricas de Feature Selection

A tabela a seguir mostra uma comparação por relevância das características do SuPor-2 e de características de redes complexas. As medidas consideradas foram Information Gain, Qui-quadro e Gain-Ration, todas normalizadas. A última coluna mostra a média entre as medidas:

#	Nome da Característica	Fonte	Information Gain	Qui-quadrado	Gain-Ratio	Média
7	Frequência das Palavras (pré-processamento usando quadrigramas)	SuPor-2	0.0335	133.0897	0.0316	0.9258
19	Locality Index	RC	0.0372	151.7118	0.0206	0.8841
28	Dilation (nível 3, ponderada)	RC	0.0316	127.7179	0.0233	0.8097
3	Tamanho da sentença	SuPor-2	0.0312	125.9432	0.0208	0.7757
32	Hubs (ordenado pela localidade e com corte de grau)	RC	0.0234	96.4134	0.0243	0.6777
13	Degree (variação com ponderação)	RC	0.0259	107.2528	0.0185	0.6626
37	Communities	RC	0.0256	99.4917	0.0196	0.6545
16	Minimal Paths	RC	0.0211	88.8322	0.0255	0.6527
14	Clustering Coefficient	RC	0.0260	103.4627	0.0176	0.6454
17	Minimal Paths (variação pesos complementares)	RC	0.0243	102.8557	0.0189	0.6424
30	Hubs (ordenado pelo grau)	RC	0.0243	99.0859	0.0189	0.6352
6	Frequência das Palavras (pré-processamento usando stemming)	SuPor-2	0.0252	97.8700	0.0178	0.6284
33	K-Cores (ordernado pela localidade)	RC	0.0255	102.1496	0.0161	0.6222
12	Degree	RC	0.0237	94.1983	0.0152	0.5793
1	Cadeias Lexicais (Divisão em dos tópicos pelos parágrafos)	SuPor-2	0.0207	88.7588	0.0176	0.5658
5	Posição	SuPor-2	0.0246	104.2210	0.0100	0.5549

18	Minimal Paths (variação pesos inversos)	RC	0.0215	89.7305	0.0150	0.5485
15	Clustering Coefficient (variação com ponderação)	RC	0.0214	88.2064	0.0143	0.5369
24	Dilation (nível 3)	RC	0.0156	59.6082	0.0199	0.4808
20	Locality Index (modificada)	RC	0.0149	56.4646	0.0209	0.4785
11	Importância dos Tópicos (pré-processamento usando quadrigramas)	SuPor-2	0.0142	59.7611	0.0178	0.4465
2	Cadeias Lexicais (Divisão em tópicos usando TextTiling)	SuPor-2	0.0142	61.5379	0.0161	0.4320
10	Importância dos Tópicos (pré-processamento usando stemming)	SuPor-2	0.0138	57.5941	0.0160	0.4191
31	Hubs (ordenado pela localidade)	RC	0.0108	39.8245	0.0207	0.4026
36	W-Cuts (ordenado pelo grau)	RC	0.0109	40.3451	0.0204	0.4017
35	W-Cuts (ordernado pela localidade)	RC	0.0109	40.4415	0.0202	0.3990
34	K-Cores (ordernado pelo grau)	RC	0.0107	39.7669	0.0196	0.3895
26	Dilation (nível 2, ponderada)	RC	0.0098	36.0488	0.0203	0.3809
22	Dilation (nível 2)	RC	0.0096	35.2634	0.0199	0.3733
25	Dilation (nível 3, acumulativa)	RC	0.0107	40.2246	0.0177	0.3711
21	Matching Index	RC	0.0103	38.6329	0.0172	0.3584
29	Dilation (nível 3, ponderada, acumulativa)	RC	0.0061	26.8536	0.0225	0.3508
4	Nomes próprios	SuPor-2	0.0088	36.9571	0.0109	0.2753
8	Mapa de Relacionamentos (pré-processamento usando stemming)	SuPor-2	0.0072	29.0575	0.0041	0.1712
9	Mapa de Relacionamento (pré-processamento usando quadrigramas)	SuPor-2	0.0020	8.3579	0.0013	0.0500
23	Dilation (nível 2, acumulativa)	RC	0	0	0	0
27	Dilation (nível 2, ponderada, acumulativa)	RC	0	0	0	0

Apêndice E

Resultados Por Texto da Avaliação Conjunta Sobre o TeMário-2006

ROUGE-1

Texto	SuPor-2	TextRank+StemmingStopRem	TextRank+Thesaurus	SuPor-2 LogReg	SuPor-2 Fuzzy
br94ab03-22.txt	0.6089	0.6114	0.6139	0.5965	0.5668
br94ab07-51.txt	0.6239	0.5373	0.6224	0.6254	0.6224
br94ab14-59.txt	0.5633	0.5241	0.6205	0.5934	0.6115
br94ag11-16.txt	0.5714	0.4914	0.5714	0.5286	0.5743
br94de04-14.txt	0.6472	0.5390	0.5108	0.5887	0.6104
br94de05-25.txt	0.7857	0.6893	0.6143	0.7679	0.6679
br94de18-13.txt	0.6583	0.5877	0.6355	0.6492	0.6014
br94de25-11.txt	0.6124	0.5383	0.5598	0.5742	0.5718
br94fe8-50.txt	0.6871	0.6085	0.6775	0.6161	0.6468
br94jl24-18.txt	0.6756	0.6059	0.6247	0.6971	0.6247
br94ma15-20.txt	0.7039	0.6201	0.5224	0.6927	0.6592
br94ma15-21.txt	0.6586	0.6085	0.5888	0.6525	0.6419
br94ma15-33.txt	0.7339	0.6630	0.7386	0.7575	0.6850
br94ma22-40.txt	0.6259	0.5000	0.5611	0.6130	0.6111
br94mr20-44.txt	0.7554	0.6104	0.6926	0.7056	0.7121
br94mr25-20.txt	0.7514	0.6714	0.6314	0.7429	0.7457
br94no06-12.txt	0.5877	0.5154	0.5351	0.5461	0.5943
br94no13-11.txt	0.5447	0.4638	0.5085	0.5000	0.5638
br94no18-15.txt	0.7969	0.7005	0.7318	0.7917	0.7422
br94no20-13.txt	0.6441	0.6250	0.6123	0.6356	0.5848
br94ou02-14.txt	0.6232	0.5665	0.5739	0.6010	0.6207
br94ou09-16.txt	0.7270	0.6558	0.6558	0.6795	0.6647
br94ou16-16.txt	0.6933	0.6677	0.6230	0.6805	0.6070
br94ou23-12.txt	0.6364	0.5281	0.5087	0.6402	0.5880
br94ou29-07.txt	0.7500	0.6321	0.6972	0.7825	0.6911
br94se11-20.txt	0.6716	0.6029	0.6593	0.6961	0.6373
br94se27-02.txt	0.7732	0.7500	0.7345	0.7655	0.7062
ce94ab10-26.txt	0.7288	0.5490	0.6699	0.6765	0.7124
ce94de09-23.txt	0.7266	0.6836	0.6602	0.7188	0.6504
ce94ju23-04.txt	0.6181	0.5844	0.4873	0.6097	0.6076
ce94mr1-09.txt	0.7784	0.7326	0.7519	0.8006	0.7090

ce94mr27-29.txt	0.7053	0.6798	0.6641	0.6798	0.6523
ce94ou04-84.txt	0.8126	0.7242	0.6779	0.8000	0.7368
ce94ou07-02.txt	0.7531	0.6855	0.6682	0.7516	0.6604
ce94se05-35.txt	0.7366	0.6687	0.7119	0.7387	0.6399
ce94se13-72.txt	0.7326	0.6457	0.7022	0.7239	0.6544
ce94se21-72.txt	0.6655	0.5773	0.5881	0.6583	0.6745
co94ab03-16.txt	0.6250	0.5799	0.6389	0.6389	0.6424
co94ab24-03.txt	0.5952	0.4360	0.5294	0.5675	0.5363
co94ag07-09.txt	0.6515	0.5492	0.6780	0.6023	0.6023
co94ag21-08.txt	0.6709	0.6262	0.5559	0.6741	0.6486
co94ag28-09.txt	0.5138	0.5092	0.4771	0.5000	0.5459
co94ag28-15.txt	0.7425	0.6301	0.5260	0.7014	0.6192
co94ag28-16.txt	0.6630	0.5452	0.4411	0.6384	0.5945
co94de04-23.txt	0.4960	0.5318	0.5318	0.4960	0.5278
co94de11-15.txt	0.5884	0.5244	0.5274	0.5976	0.6098
co94de16-28.txt	0.6727	0.6007	0.5827	0.6727	0.6835
co94de31-11.txt	0.7023	0.6488	0.4682	0.6756	0.6020
co94ja30-12.txt	0.6111	0.5648	0.6111	0.6019	0.5926
co94jl07-23.txt	0.6777	0.5515	0.5083	0.6445	0.5748
co94jl12-46.txt	0.7518	0.7007	0.6423	0.7299	0.6204
co94jl12-51.txt	0.5221	0.5141	0.5623	0.4659	0.5663
co94jl23-43.txt	0.6496	0.4957	0.6325	0.6239	0.5897
co94jl31-08.txt	0.6879	0.4894	0.6064	0.6028	0.6418
co94ma15-03.txt	0.5672	0.5112	0.6231	0.5448	0.5746
co94ma29-10.txt	0.6245	0.5852	0.6332	0.5546	0.5284
co94mr6-09.txt	0.5561	0.5051	0.5663	0.5816	0.5714
co94no01-08.txt	0.6550	0.5687	0.6200	0.6307	0.6226
co94no13-09.txt	0.7040	0.6570	0.6715	0.6787	0.6137
co94no27-17.txt	0.6587	0.5689	0.5659	0.5988	0.6557
co94ou02-30.txt	0.6375	0.5418	0.6653	0.6932	0.6534
co94ou22-04.txt	0.5261	0.4673	0.5458	0.5294	0.5686
co94se04-10.txt	0.5234	0.4723	0.5830	0.5489	0.5192
di94ab03-03.txt	0.6159	0.5224	0.6260	0.5894	0.6240
di94ab03-14.txt	0.6250	0.5933	0.6083	0.6100	0.6817
di94ag09-08.txt	0.7262	0.6928	0.6478	0.7195	0.7546
di94de04-13.txt	0.6620	0.6600	0.5666	0.6740	0.7097
di94de25-12.txt	0.6667	0.5983	0.6517	0.6560	0.6838
di94fe13-14.txt	0.7841	0.7170	0.6415	0.7757	0.7044
di94fe20-10.txt	0.7526	0.7339	0.6944	0.7817	0.6944
di94fe27-11.txt	0.6954	0.6433	0.6293	0.7114	0.6473
di94ja09-16.txt	0.6988	0.6437	0.5059	0.6772	0.6063
di94ja16-14.txt	0.5717	0.5323	0.5591	0.5932	0.6075
di94ja30-11.txt	0.7234	0.6824	0.6639	0.7172	0.6373
di94jl01-09.txt	0.7480	0.7183	0.7064	0.7401	0.7321
di94jl17-04.txt	0.7347	0.6527	0.6031	0.7481	0.7004
di94jl31-16.txt	0.6989	0.6431	0.6673	0.6673	0.6599
di94ju19-08.txt	0.6653	0.5690	0.6151	0.6548	0.6088
di94ju26-08.txt	0.6765	0.6144	0.6618	0.6569	0.6226
di94ju29-24.txt	0.7086	0.6468	0.6380	0.6623	0.6600
di94ma04-07.txt	0.6448	0.5282	0.6230	0.6485	0.6029
di94ma08-15.txt	0.6917	0.6073	0.6532	0.6477	0.6991
di94ma15-05.txt	0.6049	0.5843	0.6217	0.5861	0.6292

di94ma15-18.txt	0.7218	0.6579	0.6617	0.7087	0.6974
di94ma22-17.txt	0.6636	0.6238	0.6238	0.6519	0.7009
di94mr13-09.txt	0.6885	0.6211	0.5228	0.6594	0.6357
di94mr20-20.txt	0.6616	0.5755	0.5774	0.6348	0.6042
di94mr27-13.txt	0.7048	0.6667	0.6158	0.7150	0.7328
di94mr6-16.txt	0.6295	0.6205	0.5848	0.6205	0.5960
di94no06-16.txt	0.6836	0.5944	0.6224	0.6906	0.6678
di94ou16-17.txt	0.6887	0.6330	0.6206	0.6825	0.6742
di94se04-17.txt	0.6829	0.7000	0.6244	0.7098	0.6781
di94se19-16.txt	0.6613	0.5737	0.6521	0.6590	0.5807
mu94ab17-18.txt	0.7682	0.7107	0.7452	0.7548	0.7126
mu94ab17-25.txt	0.3715	0.3466	0.3847	0.3627	0.3524
mu94ag21-16.txt	0.7347	0.6118	0.7215	0.7434	0.7851
mu94ag28-27.txt	0.6957	0.6609	0.6667	0.6841	0.6240
mu94de25-07.txt	0.6861	0.5329	0.5912	0.6642	0.6263
mu94fe6-13.txt	0.7150	0.6359	0.6227	0.6903	0.6771
mu94ma08-12.txt	0.7233	0.6588	0.7280	0.7327	0.6887
mu94ma15-18.txt	0.6873	0.6100	0.6512	0.7062	0.7320
mu94ma22-25.txt	0.7136	0.6661	0.6579	0.7054	0.6759
mu94no27-18.txt	0.7115	0.5997	0.6608	0.7185	0.7098
mu94ou23-18.txt	0.7027	0.5081	0.6180	0.6775	0.6559
op94ab03-01.txt	0.5476	0.4966	0.5306	0.5918	0.5646
op94ab07-01.txt	0.7033	0.5714	0.6291	0.6676	0.7088
op94fe27-01.txt	0.5473	0.5444	0.5385	0.4911	0.5237
op94fe6-01.txt	0.5894	0.5810	0.5335	0.5447	0.5447
op94ju01-09.txt	0.6887	0.6459	0.6693	0.6770	0.6265
op94ju01-10.txt	0.6105	0.4947	0.5526	0.5395	0.6026
op94ma28-02.txt	0.6440	0.4799	0.5851	0.6347	0.5480
op94mr27-01.txt	0.5852	0.5341	0.5568	0.5796	0.5568
op94ou30-02.txt	0.5471	0.5181	0.6087	0.5326	0.5833
op94se25-01.txt	0.5866	0.4803	0.5433	0.5394	0.5394
td94ab03-01.txt	0.6120	0.4973	0.5792	0.5847	0.5574
td94ab03-08.txt	0.6220	0.5041	0.5772	0.6179	0.5447
td94ag28-01.txt	0.5756	0.5252	0.5462	0.5840	0.5714
td94fe13-02.txt	0.5845	0.4331	0.4718	0.5528	0.5352
td94fe20-03.txt	0.5947	0.5198	0.6256	0.5551	0.5815
td94fe20-09.txt	0.5856	0.3919	0.5856	0.5946	0.6126
td94fe27-02.txt	0.6426	0.5787	0.5575	0.6468	0.6298
td94fe27-09.txt	0.7108	0.5783	0.6466	0.6988	0.6586
td94fe6-10.txt	0.5800	0.5250	0.6500	0.6450	0.6050
td94ja08-04.txt	0.5938	0.5195	0.5195	0.5508	0.5664
td94ja15-04.txt	0.5714	0.4619	0.5143	0.5476	0.6333
td94ja16-13.txt	0.5980	0.5377	0.5829	0.6181	0.5980
td94ja22-01.txt	0.5909	0.5657	0.4141	0.6061	0.5657
td94ja23-02.txt	0.6474	0.4632	0.5263	0.6211	0.6368
td94ja23-10.txt	0.6742	0.4607	0.6067	0.6854	0.6067
td94ja23-11.txt	0.5707	0.3854	0.5902	0.5268	0.5073
td94ja30-10.txt	0.5388	0.4977	0.6073	0.5753	0.6804
td94ja9-02.txt	0.5800	0.5733	0.6533	0.6133	0.7133
td94jl10-08.txt	0.6158	0.5480	0.6441	0.5424	0.6441
td94jl31-02.txt	0.4878	0.4675	0.3699	0.4675	0.4959
td94jl31-05.txt	0.6636	0.6774	0.5899	0.6544	0.6774

td94jl31-10.txt	0.7183	0.5305	0.5869	0.6432	0.6197
td94ju12-03.txt	0.7023	0.5302	0.4744	0.6512	0.5721
td94ju12-17.txt	0.6863	0.5245	0.5686	0.7108	0.5441
td94ju26-05.txt	0.6094	0.4740	0.6094	0.5677	0.6354
td94ma01-02.txt	0.6603	0.6346	0.4936	0.6603	0.5256
td94ma08-02.txt	0.6084	0.4940	0.6265	0.6024	0.6325
td94ma08-13.txt	0.4849	0.5152	0.4546	0.5303	0.5455
td94mr13-03.txt	0.5931	0.5844	0.5931	0.6061	0.6061
td94mr13-09.txt	0.7195	0.6787	0.5385	0.6697	0.5294
td94mr13-13.txt	0.7267	0.5988	0.6686	0.7151	0.5698
td94mr27-09.txt	0.5433	0.5192	0.4231	0.4952	0.5673
td94no20-01.txt	0.6205	0.5026	0.5385	0.6205	0.5897
td94ou09-01.txt	0.6429	0.4706	0.6555	0.6387	0.5630
td94ou16-01.txt	0.8232	0.7374	0.7424	0.7626	0.6818
td94se11-03.txt	0.6301	0.5087	0.5491	0.5838	0.6532

ROUGE-2

Texto	SuPor-2	TextRank+StemmingStopRem	TextRank+Thesaurus	SuPor-2 LogReg	SuPor-2 Fuzzy
br94ab03-22.txt	0.1588	0.1712	0.1712	0.1489	0.1315
br94ab07-51.txt	0.2212	0.1226	0.2153	0.2018	0.2018
br94ab14-59.txt	0.1420	0.1178	0.1571	0.1631	0.1571
br94ag11-16.txt	0.1777	0.1519	0.1863	0.1891	0.1719
br94de04-14.txt	0.3167	0.2343	0.1323	0.2451	0.2560
br94de05-25.txt	0.3506	0.2522	0.2004	0.3417	0.2165
br94de18-13.txt	0.3082	0.2055	0.2511	0.3128	0.2169
br94de25-11.txt	0.2374	0.2182	0.1727	0.2182	0.2206
br94fe8-50.txt	0.3000	0.2212	0.2962	0.2539	0.2808
br94jl24-18.txt	0.2769	0.2285	0.1989	0.2742	0.1962
br94ma15-20.txt	0.2913	0.2073	0.0896	0.2857	0.2185
br94ma15-21.txt	0.3146	0.2781	0.2188	0.3237	0.2964
br94ma15-33.txt	0.3139	0.2366	0.3297	0.3470	0.2587
br94ma22-40.txt	0.2783	0.2041	0.2115	0.2597	0.2208
br94mr20-44.txt	0.3731	0.2842	0.3384	0.3536	0.3059
br94mr25-20.txt	0.4155	0.3381	0.2751	0.4212	0.4699
br94no06-12.txt	0.2242	0.1736	0.1495	0.1956	0.2132
br94no13-11.txt	0.2154	0.1365	0.2090	0.1898	0.2090
br94no18-15.txt	0.3812	0.2637	0.3029	0.3760	0.3394
br94no20-13.txt	0.2612	0.2739	0.2633	0.2824	0.2123
br94ou02-14.txt	0.2247	0.1901	0.1432	0.2148	0.2346
br94ou09-16.txt	0.2679	0.2113	0.2262	0.2560	0.2143
br94ou16-16.txt	0.3141	0.2628	0.2404	0.3173	0.2468
br94ou23-12.txt	0.2946	0.1919	0.1744	0.3004	0.2597
br94ou29-07.txt	0.3238	0.2464	0.2770	0.3890	0.2953
br94se11-20.txt	0.2752	0.2383	0.2604	0.3022	0.2678
br94se27-02.txt	0.3566	0.3127	0.3075	0.3902	0.2972
ce94ab10-26.txt	0.3607	0.2361	0.2984	0.3410	0.3541
ce94de09-23.txt	0.3190	0.2701	0.2485	0.3092	0.2505
ce94ju23-04.txt	0.2093	0.1713	0.1184	0.2135	0.1713

ce94mr1-09.txt	0.3550	0.2707	0.3003	0.3698	0.2840
ce94mr27-29.txt	0.3406	0.3228	0.2697	0.3327	0.3051
ce94ou04-84.txt	0.4916	0.3924	0.2911	0.4599	0.3882
ce94ou07-02.txt	0.3118	0.2394	0.2189	0.3370	0.2173
ce94se05-35.txt	0.3443	0.2639	0.2990	0.3505	0.2763
ce94se13-72.txt	0.3922	0.2941	0.3050	0.3900	0.2789
ce94se21-72.txt	0.2703	0.2180	0.2144	0.2721	0.2919
co94ab03-16.txt	0.3415	0.2892	0.3484	0.3868	0.3659
co94ab24-03.txt	0.2917	0.1667	0.2604	0.2778	0.2431
co94ag07-09.txt	0.4297	0.3156	0.3612	0.3840	0.3194
co94ag21-08.txt	0.3910	0.3173	0.2019	0.3782	0.3622
co94ag28-09.txt	0.1336	0.1982	0.1152	0.1336	0.2028
co94ag28-15.txt	0.4121	0.3050	0.1868	0.3736	0.3104
co94ag28-16.txt	0.2747	0.2115	0.1593	0.2637	0.2610
co94de04-23.txt	0.1315	0.1952	0.1952	0.1474	0.1634
co94de11-15.txt	0.2875	0.2355	0.2202	0.3211	0.3119
co94de16-28.txt	0.3610	0.2744	0.2708	0.3466	0.3249
co94de31-11.txt	0.2852	0.2785	0.1544	0.2718	0.2248
co94ja30-12.txt	0.2512	0.2093	0.2419	0.2326	0.2605
co94jl07-23.txt	0.3600	0.2800	0.2167	0.3400	0.2633
co94jl12-46.txt	0.4066	0.3626	0.2857	0.4103	0.3150
co94jl12-51.txt	0.2702	0.2500	0.2782	0.2218	0.2540
co94jl23-43.txt	0.3562	0.2103	0.3262	0.3562	0.2747
co94jl31-08.txt	0.3559	0.1815	0.2669	0.2918	0.3167
co94ma15-03.txt	0.2921	0.2060	0.2247	0.2921	0.2622
co94ma29-10.txt	0.2983	0.2719	0.3070	0.2588	0.1930
co94mr6-09.txt	0.1949	0.1487	0.1744	0.2154	0.2154
co94no01-08.txt	0.3757	0.3405	0.3351	0.3730	0.3730
co94no13-09.txt	0.3986	0.3225	0.3768	0.3913	0.3515
co94no27-17.txt	0.3093	0.2312	0.1982	0.2763	0.3393
co94ou02-30.txt	0.3000	0.2600	0.2880	0.3560	0.3200
co94ou22-04.txt	0.2689	0.2033	0.2066	0.2623	0.2721
co94se04-10.txt	0.2137	0.1838	0.2607	0.2222	0.2051
di94ab03-03.txt	0.2322	0.1976	0.2526	0.2383	0.2464
di94ab03-14.txt	0.2855	0.2421	0.2521	0.3072	0.3623
di94ag09-08.txt	0.4231	0.4331	0.3127	0.4197	0.4615
di94de04-13.txt	0.3068	0.3207	0.2590	0.3406	0.4004
di94de25-12.txt	0.3319	0.2463	0.2827	0.3062	0.3426
di94fe13-14.txt	0.4643	0.3782	0.3088	0.4475	0.3845
di94fe20-10.txt	0.3604	0.3396	0.3063	0.3896	0.3000
di94fe27-11.txt	0.3253	0.2771	0.2430	0.3514	0.2872
di94ja09-16.txt	0.3215	0.2544	0.1302	0.3195	0.2229
di94ja16-14.txt	0.2190	0.2208	0.2065	0.2585	0.2585
di94ja30-11.txt	0.3265	0.2772	0.2854	0.3347	0.2485
di94jl01-09.txt	0.3738	0.3380	0.3320	0.3917	0.3459
di94jl17-04.txt	0.3805	0.3384	0.3098	0.4207	0.3786
di94jl31-16.txt	0.3575	0.3240	0.3073	0.3371	0.3110
di94ju19-08.txt	0.2704	0.2159	0.2600	0.2663	0.2453
di94ju26-08.txt	0.3601	0.2619	0.2995	0.3568	0.3044
di94ju29-24.txt	0.3297	0.2478	0.3186	0.3319	0.2611
di94ma04-07.txt	0.3412	0.2719	0.2993	0.3759	0.3047
di94ma08-15.txt	0.3787	0.3217	0.3162	0.3346	0.3934

di94ma15-05.txt	0.2739	0.3171	0.2889	0.2683	0.3039
di94ma15-18.txt	0.4313	0.3729	0.3484	0.4011	0.3785
di94ma22-17.txt	0.3208	0.2459	0.2436	0.3162	0.3045
di94mr13-09.txt	0.3084	0.2336	0.2080	0.2865	0.2774
di94mr20-20.txt	0.3755	0.2720	0.2529	0.3506	0.2701
di94mr27-13.txt	0.3852	0.2832	0.2321	0.3571	0.3546
di94mr6-16.txt	0.2774	0.2103	0.2148	0.2796	0.2461
di94no06-16.txt	0.3433	0.2890	0.2592	0.3608	0.3643
di94ou16-17.txt	0.3388	0.2851	0.2727	0.3347	0.3099
di94se04-17.txt	0.2861	0.2983	0.2494	0.3203	0.3447
di94se19-16.txt	0.2356	0.1778	0.2587	0.2748	0.1963
mu94ab17-18.txt	0.4415	0.3589	0.3436	0.4204	0.3647
mu94ab17-25.txt	0.0471	0.0397	0.0485	0.0427	0.0412
mu94ag21-16.txt	0.3604	0.2615	0.3187	0.3868	0.4308
mu94ag28-27.txt	0.3010	0.2893	0.2466	0.2971	0.2602
mu94de25-07.txt	0.3363	0.2047	0.2661	0.3056	0.2822
mu94fe6-13.txt	0.4026	0.3201	0.2327	0.3960	0.3416
mu94ma08-12.txt	0.3228	0.2409	0.2976	0.3291	0.2803
mu94ma15-18.txt	0.3890	0.3012	0.2960	0.4234	0.4148
mu94ma22-25.txt	0.3098	0.2738	0.2246	0.3246	0.2787
mu94no27-18.txt	0.3695	0.2855	0.3047	0.3835	0.3800
mu94ou23-18.txt	0.3520	0.1534	0.2527	0.3069	0.3394
op94ab03-01.txt	0.1741	0.1399	0.1468	0.2218	0.2150
op94ab07-01.txt	0.3692	0.2342	0.2645	0.3333	0.4050
op94fe27-01.txt	0.2433	0.2255	0.2255	0.2463	0.2344
op94fe6-01.txt	0.1765	0.1877	0.1569	0.1625	0.1653
op94ju01-09.txt	0.3398	0.2969	0.3242	0.3164	0.2813
op94ju01-10.txt	0.2005	0.1187	0.1794	0.1478	0.2084
op94ma28-02.txt	0.3416	0.1770	0.3261	0.3509	0.2422
op94mr27-01.txt	0.1771	0.1600	0.1771	0.1657	0.1543
op94ou30-02.txt	0.2109	0.2218	0.2764	0.2327	0.2836
op94se25-01.txt	0.1937	0.1028	0.1383	0.1818	0.1779
td94ab03-01.txt	0.3077	0.1868	0.2582	0.2582	0.2198
td94ab03-08.txt	0.3633	0.2163	0.2980	0.3633	0.2939
td94ag28-01.txt	0.2869	0.2490	0.2785	0.3291	0.2996
td94fe13-02.txt	0.3251	0.2403	0.2156	0.3180	0.3074
td94fe20-03.txt	0.2699	0.1858	0.3053	0.2257	0.2478
td94fe20-09.txt	0.2534	0.1855	0.2670	0.3258	0.3032
td94fe27-02.txt	0.4188	0.3248	0.2992	0.4444	0.3761
td94fe27-09.txt	0.3468	0.2298	0.2782	0.3307	0.3145
td94fe6-10.txt	0.3116	0.3015	0.3618	0.4523	0.3568
td94ja08-04.txt	0.2980	0.2157	0.2235	0.2706	0.2667
td94ja15-04.txt	0.2297	0.1914	0.2010	0.2297	0.3397
td94ja16-13.txt	0.3333	0.2980	0.3030	0.3586	0.2727
td94ja22-01.txt	0.3503	0.3300	0.1574	0.3553	0.3147
td94ja23-02.txt	0.3439	0.2011	0.2064	0.3439	0.3333
td94ja23-10.txt	0.4237	0.2542	0.3842	0.4181	0.3955
td94ja23-11.txt	0.3284	0.1667	0.3235	0.3284	0.2794
td94ja30-10.txt	0.3073	0.2110	0.2982	0.3716	0.4220
td94ja9-02.txt	0.3289	0.3289	0.3222	0.3893	0.4631
td94jl10-08.txt	0.3921	0.3125	0.3580	0.3352	0.3864
td94jl31-02.txt	0.2163	0.2163	0.0980	0.2204	0.2286

td94jl31-05.txt	0.4074	0.4583	0.3056	0.3982	0.4120
td94jl31-10.txt	0.4198	0.2736	0.2972	0.3915	0.3585
td94ju12-03.txt	0.4252	0.2243	0.2523	0.4019	0.2664
td94ju12-17.txt	0.4335	0.2808	0.3301	0.4384	0.2611
td94ju26-05.txt	0.3194	0.2304	0.3403	0.2932	0.3613
td94ma01-02.txt	0.4258	0.4129	0.1871	0.4258	0.2129
td94ma08-02.txt	0.3212	0.1697	0.2667	0.3212	0.3212
td94ma08-13.txt	0.2081	0.2690	0.1726	0.2538	0.2741
td94mr13-03.txt	0.2696	0.2652	0.2609	0.2739	0.2609
td94mr13-09.txt	0.5227	0.4727	0.3591	0.4909	0.3091
td94mr13-13.txt	0.3801	0.2690	0.3392	0.3801	0.2105
td94mr27-09.txt	0.2464	0.2416	0.1159	0.2174	0.2560
td94no20-01.txt	0.3402	0.2268	0.2526	0.3969	0.3299
td94ou09-01.txt	0.4177	0.2785	0.4177	0.4641	0.3165
td94ou16-01.txt	0.6294	0.5279	0.4822	0.5685	0.4873
td94se11-03.txt	0.3256	0.1802	0.1977	0.2791	0.3081

Apêndice F

Resultado da seleção de características pelo CFS para os conjuntos SuPor-2, RC e SuPor-2 U RC

Os resultados a seguir foram obtidos por meio método CFS do WEKA:

1. Conjunto: SuPor-2 U RC (união das 37 características do SuPor-2 e de redes complexas)

CFS Selected attributes: 14

Lexical Chains (Paragraphs)
Sentence Length
Position
Frequency or Words (stemming)
Frequency or Words (4-grams)
Importance of Topics (stemming)
Importance of Topics (4-grams)
Clustering Coefficient
Hubs (Degree)
Hubs (Locality)
Dilations (level 2, weighted)
Locality Index (modified)
W-cuts (Degree)
W-cuts (Locality)

2. Conjunto: RC (Redes Complexas)

CFS Selected attributes: 13

K-cores (Locality)
Clustering Coefficient
Communities
Hubs (Degree)
Hubs (sorted by locality, with degree cut))
Hubs (Locality)
Dilations (level 2)
Dilations (level 2, weighted)
Locality Index (modified)
Locality Index
W-Cuts (Degree)
W-Cuts (Locality)
Clustering Coefficient (Weighted)

3. Conjunto: SuPor-2

CFS Selected attributes: 1,3,5,6,7,11 : 6

Lexical Chains (Paragraphs)
Sentence Length
Position
Frequency of Words (Stemming)
Frequency of Words (4-grams)
Importance of Topics (4-grams)

Apêndice G

Exemplos de Sumários Produzidos Pelos Sistemas

A seguir, mostram-se exemplos de texto-fontes e sumários gerados pelos modelos SuPor-2, TextRank+StopRem+Stemming, TextRank+Thesaurus e SuPor-2 Fuzzy. Os textos-fontes e sumários manuais são parte integrante do corpus TeMário-2006.

EXEMPLO 1 (br94de25-11.txt) – texto de opinião

- **Texto-fonte br94de25-11.txt**

O efeito Gutenberg

MARCELO LEITE

Um leitor qualificado e perspicaz sugeriu-me outro dia uma pergunta difícil: como se comportará a Folha em relação ao governo Fernando Henrique Cardoso? No maniqueísmo inerente ao jornalismo, só haveria uma alternativa: ou amor ou ódio.

A questão é pertinente, dada a notória proximidade do jornal com o presidente eleito. Até setembro de 1992, FHC mantinha uma coluna semanal na pág. 1-2, publicada às quintas-feiras. Tal relação de colaboração só foi interrompida porque o senador peessedebista se tornou chanceler de Itamar Franco.

(Segundo praxe da Folha, um colunista não pode simultaneamente ocupar ou candidatar-se a cargo no Executivo. Nesta condição, sua coluna correria o risco de

transformar-se em tribuna para defesa de um interesse privado –a reputação como governante.)

Fernando Henrique não foi o único tucano a ocupar esse espaço, conhecido na Redação como coluna vertical. Depois de amanhã deverá ser publicado o último texto do futuro ministro do Planejamento José Serra.

Pertinente, a questão não é porém nova. O próprio retrospecto das colunas de ombudsman aponta para uma simpatia espúria:

Durante a campanha eleitoral, minha antecessora apontou fernandohenriquismo do jornal;

Ao estrear, emiti a opinião de que este e outros diários tinham mesmo henricado;

A 30 de outubro, na coluna Lua-de-mel na Europa, critiquei a condescendência da Folha com o presidente eleito.

Quando FHC enfim se lançou ao primeiro ato de governo, montar seu ministério, temi pelo pior.

No episódio da escolha de Pedro Malan para ministro da Fazenda, intencionalmente vazada para repórteres, os jornais evidenciaram sua tibieza. Com arrogância, FHC desqualificou as manchetes de 1º de dezembro, dizendo que era um ministério Gutenberg (referência a Johannes Gutenberg, que inventou a imprensa de tipos móveis no século 15).

Ficou por isso mesmo. Em outras épocas, a Folha teria posto a boca no trombone, denunciando a tentativa de manipulação.

FHC seguiu a seu modo a receita do seu sucessor na Fazenda, aquele premiado com a embaixada em Roma pela ajuda ao candidato: esconder o que é ruim (as pressões para indicar Serra no lugar de Malan) e faturar o que é bom (a imagem favorável de Malan).

No último dia 14, perguntei em minha crítica interna da edição –documento distribuído diariamente na Redação– se o termo loteamento também não se aplicaria às negociações em curso, em especial as tratativas com o fisiológico PMDB. Afinal, eram um tanto semelhantes às entabuladas por Itamar dois anos antes e desancadas pelo jornal.

Dois dias depois, uma chamada na capa do jornal anunciava: FHC cede a pressão e loteia ministério. Ao elogiar a iniciativa crítica, no entanto, fiz uma ressalva:

"Faltou mencionar um ponto importante, na análise das 'pressões': FHC teria condições, sem contemplar PMDB, de fazer a reforma constitucional (ou pelo menos fiscal) exigida por todos, inclusive esta Folha?

É uma espécie de outro lado –neste caso, da questão. A necessidade de criticar o emprego de métodos políticos atrasados, como a distribuição de cargos, não desobriga de outra, a de eventualmente reconhecer que pode não haver outra moeda no mercado para negociar a estabilidade.

A palavra-chave do comportamento que a Folha deve observar frente ao governo –qualquer governo– é equilíbrio. Sem simpatia nem rancor.

O perigo das relações estremecidas, como no caso FHC-Folha, são as hiper-reações resultantes de encontrões fortuitos.

Foi o que sucedeu com o sociólogo Luciano Martins, amigo de FHC e organizador de um convescote acadêmico em Brasília. Na véspera do seminário, ele tinha dado entrevista à Folha e falado da crise do Estado-Nação, publicada sob título Acabou o Estado nacional, diz tucano.

Era um exagero, mas confesso que nem me chamou a atenção. Por vaidade, ou cioso das diferenciações que matizam o pensamento, seu ofício, Martins chiou.

Em carta ao Painel do Leitor, expôs suas divergências e levou troco imediato, na forma de uma atordoante Nota da Redação:

"Por serem resumos extremamente condensados, os títulos jornalísticos quase nunca comportam filigranas como esta que tanto preocupa o missivista. Para Luciano Martins, o conceito de Estado nacional não acabou, mas está em crise. E daí? A imprensa deve melhorar seus títulos, não há dúvida. Mas os intelectuais agora transformados em aprendizes de políticos ajudariam muito se começassem a falar de maneira categórica ou, pelo menos, clara."

O reflexo desse destempero pôde ser visto pelo público no próprio Painel do Leitor, 11 dias depois: quatro cartas de protesto, nenhuma de apoio ao jornal, nenhuma nova nota justificando ou se desculpando pela anterior.

Os leitores estão certos. Se o jornal acha que intelectuais não têm nada de importante ou compreensível para dizer, não deveria insistir em entrevistá-los. Se entrevista, tem de cobrar clareza durante a conversa; depois, só lhe resta ser fiel ao que dizem.

Atritos como esse são exceção. No geral, a relação entre tucanos e repórteres é afável. Sua melhor expressão é o "off", um acordo entre fonte e jornalista para manter a primeira no anonimato.

Na última terça-feira, o colunista Luís Nassif levantou questões pertinentes sobre o abuso dessa modalidade de investigação. Seu alvo eram as muitas reportagens abusivamente atribuídas à famosa equipe econômica.

Aproveitei a deixa para anotar que a distorção afetava grande parte, talvez a maior, do noticiário sobre o governo Itamar Cardoso. No caso deste jornal, sem que as reportagens respeitassem norma do "Novo Manual da Redação", que manda identificar o "off" com a expressão "a Folha apurou".

Foi o caso, entre outros, da notícia sobre a escolha de Malan para a Fazenda (ironizada e depois confirmada). E também da indicação de Bresser para o Itamaraty (manchetada e depois revista).

Trata-se de uma distorção, sim. Embora a prática jornalística brasileira sugira o "off" como ferramenta básica de repórteres, ele contraria o direito à informação. Deve ser encarado como exceção, e nunca oferecido pelo próprio repórter, muito menos aceito, se o confidente não tiver motivos sólidos para manter-se em sigilo.

Não me parece que o anúncio a conta-gotas do "ministério possível" de FHC, todo ele em "off", se enquadre nessa exigência.

A identificação da fonte é crucial para a credibilidade de uma informação. O jornalista que admite a exceção não pode esconder do leitor que se trata de um "off", pelo simples fato de que o interesse no anonimato pode comprometer aquilo que se revela.

Afinal, não foi para esconder informações que Gutenberg inventou a imprensa.

O ombudsman estará de folga até o dia 2. Se você tiver alguma reclamação, deixe recado na secretária eletrônica ou mande fax. Na volta, respondo.

- **Sumário manual br94de25-11.txt**

O efeito Gutenberg -- MARCELO LEITE. Um leitor me sugeriu uma pergunta difícil: como se comportará a Folha em relação ao governo Fernando Henrique Cardoso. Se o jornalismo é maniqueísta, só podia ser de amor ou ódio. Faz sentido a pergunta, já que a coluna semanal de FHC na Folha condicionava uma proximidade com o jornal. Com a sua nomeação a chanceler, interrompeu-a, seguindo praxe da Folha. Não foi o único tucano a ocupar esse espaço. E o retrospecto que fiz em gestão anterior do meu cargo mostrou compadrio durante a campanha eleitoral, relação que eu critiquei na Folha e em outros diários. Na escolha de Pedro Malan para o ministério da Fazenda, houve vazamento intencional à imprensa. Dada a indiferença dos jornais, Fernando Henrique reagiu com arrogância. Se fosse em outras épocas, a Folha teria retrucado duramente. Na minha crítica interna --- documento distribuído diariamente na Redação ---, ao tratar da questão de indicação de cargo, perguntei se o loteamento não estava sendo incluído nas negociações com o PMDB, como tinha ocorrido antes no governo Itamar – procedimento tão criticado pela Folha. O resultado veio dois dias depois: "FHC cede a pressão e loteia ministério". Elogiei a iniciativa, mas com uma ressalva : a de que a crítica aos métodos atrasados de distribuição de cargos deveria ser acompanhada do reconhecimento de, às vezes,não há outro caminho para negociar a estabilidade. O estremecimento das relações entre FHC e Folha resultou num desdobramento indigesto: o sociólogo Luciano Martins, amigo do presidente, falou numa entrevista ao jornal sobre crise Estado-Nação. O título que saiu foi "Acabou o Estado nacional. " A divergência do sociólogo quanto à interpretação mereceu uma resposta despropositada. O repórter disse que era impossível resumir um texto e contemplar, ao mesmo tempo, ideias menores como aquela que o missivista reclamava. E acrescentou que, se a imprensa deve melhorar seus títulos, os intelectuais aprendizes de político deviam expressar-se com mais clareza. Os próprios leitores não apoiaram a Folha, com o que concordo. Se o jornal acha que eles não têm algo importante a dizer, não deve entrevistá-los. No entanto, esses atritos não acontecem sempre entre o jornal e os tucanos. Até, com mais frequência do que seria esperado, os repórteres preservam o anonimato dos entrevistados.

- **Sumário SuPor-2 br94de25-11.txt**

Um leitor qualificado e perspicaz sugeriu-me outro dia uma pergunta difícil: como se comportará a Folha em relação ao governo Fernando Henrique Cardoso?

A questão é pertinente, dada a notória proximidade do jornal com o presidente eleito.

Até setembro de 1992, FHC mantinha uma coluna semanal na pág 1-2, publicada às quintas-feiras.

Tal relação de colaboração só foi interrompida porque o senador peessedebista se tornou chanceler de Itamar Franco.

Fernando Henrique não foi o único tucano a ocupar esse espaço, conhecido na Redação como coluna vertical.

A 30 de outubro, na coluna Lua-de-mel na Europa, critiquei a condescendência da Folha com o presidente eleito.

Quando FHC enfim se lançou ao primeiro ato de governo, montar seu ministério, temi pelo pior.

No episódio da escolha de Pedro Malan para ministro da Fazenda, intencionalmente vazada para repórteres, os jornais evidenciaram sua tibieza.

Com arrogância, FHC desqualificou as manchetes de 1º de dezembro, dizendo que era um ministério Gutenberg (referência a Johannes Gutenberg, que inventou a imprensa de tipos móveis no século 15).

No último dia 14, perguntei em minha crítica interna da edição – documento distribuído diariamente na Redação – se o termo loteamento também não se aplicaria às negociações em curso, em especial as tratativas com o fisiológico PMDB.

"Faltou mencionar um ponto importante, na análise das 'pressões': FHC teria condições, sem contemplar PMDB, de fazer a reforma constitucional (ou pelo menos fiscal) exigida por todos, inclusive esta Folha?"

A necessidade de criticar o emprego de métodos políticos atrasados, como a distribuição de cargos, não desobriga de outra, a de eventualmente reconhecer que pode não haver outra moeda no mercado para negociar a estabilidade.

A palavra-chave do comportamento que a Folha deve observar frente ao governo – qualquer governo – é equilíbrio.

O perigo das relações estremecidas, como no caso FHC-Folha, são as hiper-reações resultantes de encontrões fortuitos.

Se o jornal acha que intelectuais não têm nada de importante ou compreensível para dizer, não deveria insistir em entrevistá-los.

No caso deste jornal, sem que as reportagens respeitassem norma do "Novo Manual da Redação", que manda identificar o "off" com a expressão "a Folha apurou".

Embora a prática jornalística brasileira sugira o "off" como ferramenta básica de repórteres, ele contraria o direito à informação.

- **Sumário SuPor-2 Fuzzy br94de25-11.txt**

MARCELO LEITE

Um leitor qualificado e perspicaz sugeriu-me outro dia uma pergunta difícil: como se comportará a Folha em relação ao governo Fernando Henrique Cardoso?

A questão é pertinente, dada a notória proximidade do jornal com o presidente eleito.

(Segundo praxe da Folha, um colunista não pode simultaneamente ocupar ou candidatar-se a cargo no Executivo. Nesta condição, sua coluna correria o risco de transformar-se em tribuna para defesa de um interesse privado – a reputação como governante)

Ao estrear, emiti a opinião de que este e outros diários tinham mesmo henricado;

Quando FHC enfim se lançou ao primeiro ato de governo, montar seu ministério, temi pelo pior.

No episódio da escolha de Pedro Malan para ministro da Fazenda, intencionalmente vazada para repórteres, os jornais evidenciaram sua tibieza.

Ficou por isso mesmo.

Em outras épocas, a Folha teria posto a boca no trombone, denunciando a tentativa de manipulação.

FHC seguiu a seu modo a receita do seu sucessor na Fazenda, aquele premiado com a embaixada em Roma pela ajuda ao candidato: esconder o que é ruim (as pressões para indicar Serra no lugar de Malan) e faturar o que é bom (a imagem favorável de Malan).

No último dia 14, perguntei em minha crítica interna da edição – documento distribuído diariamente na Redação – se o termo loteamento também não se aplicaria às negociações em curso, em especial as tratativas com o fisiológico PMDB.

Afinal, eram um tanto semelhantes às entabuladas por Itamar dois anos antes e desancadas pelo jornal.

Ao elogiar a iniciativa crítica, no entanto, fiz uma ressalva:

A necessidade de criticar o emprego de métodos políticos atrasados, como a distribuição de cargos, não desobriga de outra, a de eventualmente reconhecer que pode não haver outra moeda no mercado para negociar a estabilidade.

Sem simpatia nem rancor.

Foi o que sucedeu com o sociólogo Luciano Martins, amigo de FHC e organizador de um convescote acadêmico em Brasília.

Em carta ao Painel do Leitor, expôs suas divergências e levou troco imediato, na forma de uma atordoante Nota da Redação:

Para Luciano Martins, o conceito de Estado nacional não acabou, mas está em crise.

O reflexo desse destempero pôde ser visto pelo público no próprio Painel do Leitor, 11 dias depois: quatro cartas de protesto, nenhuma de apoio ao jornal, nenhuma nova nota justificando ou se desculpando pela anterior.

Embora a prática jornalística brasileira sugira o "off" como ferramenta básica de repórteres, ele contraria o direito à informação.

Se você tiver alguma reclamação, deixe recado na secretária eletrônica ou mande fax.

- **Sumário TextRank+StopRem+Stemming br94de25-11.txt**

Um leitor qualificado e perspicaz sugeriu-me outro dia uma pergunta difícil: como se comportará a Folha em relação ao governo Fernando Henrique Cardoso?

(Segundo praxe da Folha, um colunista não pode simultaneamente ocupar ou candidatar-se a cargo no Executivo.

Fernando Henrique não foi o único tucano a ocupar esse espaço, conhecido na Redação como coluna vertical.

Depois de amanhã deverá ser publicado o último texto do futuro ministro do Planejamento José Serra.

Com arrogância, FHC desqualificou as manchetes de 1º de dezembro, dizendo que era um ministério Gutenberg (referência a Johannes Gutenberg, que inventou a imprensa de tipos móveis no século 15).

Afinal, eram um tanto semelhantes às entabuladas por Itamar dois anos antes e desancadas pelo jornal.

Dois dias depois, uma chamada na capa do jornal anunciava: FHC cede a pressão e loteia ministério.

A necessidade de criticar o emprego de métodos políticos atrasados, como a distribuição de cargos, não desobriga de outra, a de eventualmente reconhecer que pode não haver outra moeda no mercado para negociar a estabilidade.

Na véspera do seminário, ele tinha dado entrevista à Folha e falado da crise do Estado-Nação, publicada sob título Acabou o Estado nacional, diz tucano.

O reflexo desse destempero pôde ser visto pelo público no próprio Painel do Leitor, 11 dias depois: quatro cartas de protesto, nenhuma de apoio ao jornal, nenhuma nova nota justificando ou se desculpando pela anterior.

Se o jornal acha que intelectuais não têm nada de importante ou compreensível para dizer, não deveria insistir em entrevistá-los.

Se entrevista, tem de cobrar clareza durante a conversa; depois, só lhe resta ser fiel ao que dizem.

Atritos como esse são exceção.

No caso deste jornal, sem que as reportagens respeitassem norma do "Novo Manual da Redação", que manda identificar o "off" com a expressão "a Folha apurou".

Deve ser encarado como exceção, e nunca oferecido pelo próprio repórter, muito menos aceito, se o confidente não tiver motivos sólidos para manter-se em sigilo.

O jornalista que admite a exceção não pode esconder do leitor que se trata de um "off", pelo simples fato de que o interesse no anonimato pode comprometer aquilo que se revela.

Afinal, não foi para esconder informações que Gutenberg inventou a imprensa.

- **Sumário TextRank+Thesaurus br94de25-11.txt**

A questão é pertinente, dada a notória proximidade do jornal com o presidente eleito.

Nesta condição, sua coluna correria o risco de transformar-se em tribuna para defesa de um interesse privado – a reputação como governante)

No episódio da escolha de Pedro Malan para ministro da Fazenda, intencionalmente vazada para repórteres, os jornais evidenciaram sua tibieza.

Com arrogância, FHC desqualificou as manchetes de 1º de dezembro, dizendo que era um ministério Gutenberg (referência a Johannes Gutenberg, que inventou a imprensa de tipos móveis no século 15).

Em outras épocas, a Folha teria posto a boca no trombone, denunciando a tentativa de manipulação.

FHC seguiu a seu modo a receita do seu sucessor na Fazenda, aquele premiado com a embaixada em Roma pela ajuda ao candidato: esconder o que é ruim (as pressões para indicar Serra no lugar de Malan) e faturar o que é bom (a imagem favorável de Malan).

No último dia 14, perguntei em minha crítica interna da edição – documento distribuído diariamente na Redação – se o termo loteamento também não se aplicaria às negociações em curso, em especial as tratativas com o fisiológico PMDB.

"Faltou mencionar um ponto importante, na análise das 'pressões': FHC teria condições, sem contemplar PMDB, de fazer a reforma constitucional (ou pelo menos fiscal) exigida por todos, inclusive esta Folha?

A necessidade de criticar o emprego de métodos políticos atrasados, como a distribuição de cargos, não desobriga de outra, a de eventualmente reconhecer que pode não haver outra moeda no mercado para negociar a estabilidade.

O perigo das relações estremecidas, como no caso FHC-Folha, são as hiper-reações resultantes de encontrões fortuitos.

Na véspera do seminário, ele tinha dado entrevista à Folha e falado da crise do Estado-Nação, publicada sob título Acabou o Estado nacional, diz tucano.

Em carta ao Painel do Leitor, expôs suas divergências e levou troco imediato, na forma de uma atordoante Nota da Redação:

O reflexo desse destempero pôde ser visto pelo público no próprio Painel do Leitor, 11 dias depois: quatro cartas de protesto, nenhuma de apoio ao jornal, nenhuma nova nota justificando ou se desculpando pela anterior.

Sua melhor expressão é o "off", um acordo entre fonte e jornalista para manter a primeira no anonimato.

Na última terça-feira, o colunista Luís Nassif levantou questões pertinentes sobre o abuso dessa modalidade de investigação.

No caso deste jornal, sem que as reportagens respeitassem norma do "Novo Manual da Redação", que manda identificar o "off" com a expressão "a Folha apurou".

EXEMPLO 2 (td94ab03-08.txt) – texto de divulgação

- **Texto-fonte td94ab03-08.txt**

Maquetes crescem com mercado imobiliário

Trabalhos são vendidos por até US$ 15 mil e podem ser também usados em projetos educacionais e artísticos

CLÁUDIA RIBEIRO MESQUITA

Free-lance para a Folha

Maquetes são como "bolas de cristal" que antecipam, de forma tridimensional, edifícios, parques, usinas, cenários, projetos educacionais e culturais e os mais variados tipos de produtos. Seu grande filão é o mercado imobiliário, que, quando aquecido –como está ocorrendo este ano–, agita freneticamente os artesãos das oficinas.

Maquetes de prédios e conjuntos residenciais respondem por mais de 80% dos pedidos –e são as mais bem pagas. Uma maquete simples, de um prédio de 20 andares, por exemplo, pode custar entre US$ 4.000 e US$ 7.000.

Outros modelos mais complexos chegam a valer o dobro, como uma maquete do projeto de um conjunto residencial em Campinas (interior de São Paulo), o Bougamville, encomendada à Kenji Maquetes por US$ 15 mil.

O objetivo desse tipo de modelo é promocional, para auxiliar na venda dos imóveis. "A função da maquete, nesse caso, é elucidar ao leigo o que foi projetado

em duas dimensões e instigá-lo", diz Kenji Furuyama, 61. Há 32 anos no mercado, Kenji conta com 15 empregados e fatura por mês cerca de US$ 30 mil, 20% dos quais computados como lucro.

Segundo ele, as despesas com mão-de-obra ficam em quase 70%. "Meus funcionários recebem um salário e uma comissão de 30% em cada trabalho que executam", diz.

Kenji começou a trabalhar com maquetes aos 19 anos na Kevel, uma das poucas maquetarias de São Paulo no ano de 1954. Por ali passaram também dois outros maquetistas da cidade, Adhemir Fogassa e Achiles Maimoni. Os três aprenderam o ofício na prática.

"A formação de um maquetista é aleatória", afirma o professor Júlio Katinsky, do departamento de história da Faculdade de Arquitetura e Urbanismo da USP. Segundo ele, muitos começaram com aeromodelismo e, hoje, são profissionais bem remunerados.

São poucos os que se dedicam a essas miniaturas. Em São Paulo, de acordo com os maquetistas, deve haver cerca de 60 profissionais. Quem está no ramo não reclama. Um autônomo, em um bom mês, pode faturar até US$ 6.000.

Mário Segall, 39, abriu seu escritório em 93. Segundo ele, o investimento para montar a oficina ficou em torno de US$ 10 mil. Alguns equipamentos foram trazidos de Londres. Sua capacidade de produção é de quatro a cinco maquetes por mês. Em meses de pico, Segall afirma que fatura cerca de US$ 6.000, e seu lucro beira os 25%. "As maquetes, em Londres, são respeitadas como parte do projeto", conta. "Aqui, nem tanto."

Roberto Cardoso, arquiteto recém-formado, começou a fazer maquetes para os projetos de faculdade e, hoje, a maior parte de sua renda vem delas. Segundo ele, dá para lucrar, em média, US$ 1.000 por mês. Mas em 92, por exemplo, ele e mais um grupo de maquetistas receberam, cada um, US$ 5.000 por 45 dias de trabalho para a produção de uma maquete do projeto de despoluição do rio Tietê, apresentada na Eco 92.

- **Sumário manual td94ab03-08.txt**

Maquetes crescem com mercado imobiliário

Trabalhos são vendidos por até US$15 mil e podem ser também usados em projetos educacionais e artísticos

CLÁUDIA RIBEIRO MESQUITA - Free-lance para a Folha

Maquetes são antecipações tridimensionais dos mais variados tipos de construções e produtos. Seu carro-chefe é o mercado imobiliário, que responde por mais de 80% dos pedidos. Uma maquete comum de um prédio de 20 andares oscila de US$4.000 a US$7.000.

A maquete auxilia a venda , esclarecendo ao leigo o que foi projetado. Kenji Furuyama, há 32 anos no mercado, com 15 empregados, fatura mensalmente cerca de US$30.000 , com um lucro de 20%. Segundo ele, as despesas com mão-de-obra orçam em quase 70%.

O professor Júlio Katinsky , do departamento de história da Faculdade de Arquitetura e Urbanismo da USP , afirma que o maquetista tem formação aleatória, muitos começando pelo aeromodelismo. São profissionais bem remunerados. O mercado não é concorrido em São Paulo. Um autônomo pode ganhar US$6.000 num mês .

Mário Segall abriu seu escritório em 93, investindo na oficina cerca de RS$10 mil. Num bom mês , fatura US$6.000 , com aproximadamente 25% de lucro.

Roberto Cardoso, arquiteto recém-formado, começou a fazer maquetes para os projetos da faculdade e, hoje, tem nelas uma razoável fonte de renda.

- **Sumário SuPor-2 td94ab03-08.txt**

Maquetes crescem com mercado imobiliário

Trabalhos são vendidos por até US$ 15 mil e podem ser também usados em projetos educacionais e artísticos

Maquetes são como "bolas de cristal" que antecipam, de forma tridimensional, edifícios, parques, usinas, cenários, projetos educacionais e culturais e os mais variados tipos de produtos.

Outros modelos mais complexos chegam a valer o dobro, como uma maquete do projeto de um conjunto residencial em Campinas (interior de São Paulo), o Bougamville, encomendada à Kenji Maquetes por US$ 15 mil.

"A função da maquete, nesse caso, é elucidar ao leigo o que foi projetado em duas dimensões e instigá-lo ", diz Kenji Furuyama, 61.

Kenji começou a trabalhar com maquetes aos 19 anos na Kevel, uma das poucas maquetarias de São Paulo no ano de 1954.

"A formação de um maquetista é aleatória", afirma o professor Júlio Katinsky, do departamento de história da Faculdade de Arquitetura e Urbanismo da USP.

Roberto Cardoso, arquiteto recém-formado, começou a fazer maquetes para os projetos de faculdade e, hoje, a maior parte de sua renda vem delas.

Mas em 92, por exemplo, ele e mais um grupo de maquetistas receberam, cada um, US$ 5.000 por 45 dias de trabalho para a produção de uma maquete do projeto de despoluição do rio Tietê, apresentada na Eco 92.

- **Sumário SuPor-2 Fuzzy td94ab03-08.txt**

Trabalhos são vendidos por até US$ 15 mil e podem ser também usados em projetos educacionais e artísticos

CLÁUDIA RIBEIRO MESQUITA

Maquetes são como "bolas de cristal" que antecipam, de forma tridimensional, edifícios, parques, usinas, cenários, projetos educacionais e culturais e os mais variados tipos de produtos.

Maquetes de prédios e conjuntos residenciais respondem por mais de 80 % dos pedidos – e são as mais bem pagas.

Uma maquete simples, de um prédio de 20 andares, por exemplo, pode custar entre US$ 4.000 e US$ 7.000.

Outros modelos mais complexos chegam a valer o dobro, como uma maquete do projeto de um conjunto residencial em Campinas (interior de São Paulo), o Bougamville, encomendada à Kenji Maquetes por US$ 15 mil.

Há 32 anos no mercado, Kenji conta com 15 empregados e fatura por mês cerca de US$ 30 mil, 20 % dos quais computados como lucro.

Mário Segall, 39, abriu seu escritório em 93.

Alguns equipamentos foram trazidos de Londres.

Sua capacidade de produção é de quatro a cinco maquetes por mês.

Em meses de pico, Segall afirma que fatura cerca de US$ 6.000, e seu lucro beira os 25 %.

"As maquetes, em Londres, são respeitadas como parte do projeto ", conta.

"Aqui, nem tanto "

- **Sumário TextRank+StopRem+Stemming td94ab03-08.txt**

Trabalhos são vendidos por até US$ 15 mil e podem ser também usados em projetos educacionais e artísticos

Maquetes são como "bolas de cristal" que antecipam, de forma tridimensional, edifícios, parques, usinas, cenários, projetos educacionais e culturais e os mais variados tipos de produtos.

Outros modelos mais complexos chegam a valer o dobro, como uma maquete do projeto de um conjunto residencial em Campinas (interior de São Paulo), o Bougamville, encomendada à Kenji Maquetes por US$ 15 mil.

"A função da maquete, nesse caso, é elucidar ao leigo o que foi projetado em duas dimensões e instigá-lo ", diz Kenji Furuyama, 61.

Há 32 anos no mercado, Kenji conta com 15 empregados e fatura por mês cerca de US$ 30 mil, 20 % dos quais computados como lucro.

Kenji começou a trabalhar com maquetes aos 19 anos na Kevel, uma das poucas maquetarias de São Paulo no ano de 1954.

Em São Paulo, de acordo com os maquetistas, deve haver cerca de 60 profissionais.

"As maquetes, em Londres, são respeitadas como parte do projeto ", conta.

Mas em 92, por exemplo, ele e mais um grupo de maquetistas receberam, cada um, US$ 5.000 por 45 dias de trabalho para a produção de uma maquete do projeto de despoluição do rio Tietê, apresentada na Eco 92.

- **Sumário TextRank+Thesaurus td94ab03-08.txt**

Maquetes são como "bolas de cristal" que antecipam, de forma tridimensional, edifícios, parques, usinas, cenários, projetos educacionais e culturais e os mais variados tipos de produtos.

Seu grande filão é o mercado imobiliário, que, quando aquecido – como está ocorrendo este ano –, agita freneticamente os artesãos das oficinas.

Outros modelos mais complexos chegam a valer o dobro, como uma maquete do projeto de um conjunto residencial em Campinas (interior de São Paulo), o Bougamville, encomendada à Kenji Maquetes por US$ 15 mil.

O objetivo desse tipo de modelo é promocional, para auxiliar na venda dos imóveis.

Há 32 anos no mercado, Kenji conta com 15 empregados e fatura por mês cerca de US$ 30 mil, 20 % dos quais computados como lucro.

Segundo ele, as despesas com mão-de-obra ficam em quase 70 %.

"A formação de um maquetista é aleatória", afirma o professor Júlio Katinsky, do departamento de história da Faculdade de Arquitetura e Urbanismo da USP.

Roberto Cardoso, arquiteto recém-formado, começou a fazer maquetes para os projetos de faculdade e, hoje, a maior parte de sua renda vem delas.

Mas em 92, por exemplo, ele e mais um grupo de maquetistas receberam, cada um, US$ 5.000 por 45 dias de trabalho para a produção de uma maquete do projeto de despoluição do rio Tietê, apresentada na Eco 92.

EXEMPLO 3 (di94mr20-20.txt) – texto sobre política

- **Texto-fonte di94mr20-20.txt**

Os custos sociais do liberalismo suicida

Países centrais tomam consciência da gravidade dos problemas gerados por uma política liberal irresponsável

MARIA DA CONCEIÇÃO TAVARES

Especial para a Folha

Finalmente, políticos e intelectuais dos países centrais começam a se dar conta da gravidade dos problemas sociais e econômicos gerados por mais de uma década de um liberalismo irresponsável, dogmático e anárquico.

Esta tardia tomada de consciência se manifesta no encontro de cúpula dos ministros do Trabalho dos países centrais em Detroit (o Job Summit) e em recentes declarações de renomados e respeitáveis economistas conservadores. Pela primeira vez, o G-7 se reúne para discutir o problema do desemprego em massa nos países

desenvolvidos, que não pára de crescer, lançando uma parcela cada vez maior da população na marginalidade.

Intimamente ligada a este processo está a questão da deslocalização, onde setores e até comunidades inteiras são destruídas, pois suas indústrias deixaram de ser competitivas num ambiente de globalização financeira e abertura comercial indiscriminada.

A combinação de taxas de desemprego crescente com a decadência econômica de regiões onde ocorre a deslocalização gera um quadro social terrível, cujas consequências são bem conhecidas.

Não falam, é claro, os liberais como os da revista "The Economist", que ainda no número da semana passada repetem a ladainha de que o problema do desemprego é resultado da rigidez do mercado de trabalho dos países desenvolvidos, em particular os europeus. A solução, como sempre, seria aumentar a "flexibilidade" do mercado de trabalho, com a retirada do seguro-desemprego e demais empecilhos ao livre jogo das forças de mercado.

Em outras palavras, o problema do desemprego viria do fato de que as economias centrais, no que diz respeito ao mercado de trabalho, são liberais de menos e a solução seria mais liberalismo.

Depois de anos de crescente "flexibilização" do mercado de trabalho, acompanhado de grande aumento e não de diminuição do desemprego, é natural que os governos e até alguns liberais de renome comecem a desconfiar que a solução para os males sociais causados pelo liberalismo irresponsável não seja mais liberalismo.

Em um artigo recente, o professor Maurice Allais, que recebeu o Prêmio Nobel de Economia em 1988 por suas contribuições à teoria neoclássica (teoria de onde a fé liberal busca obter credibilidade "científica"), faz um ataque frontal à aplicação, nas condições contemporâneas, da doutrina das vantagens comparativas.

Segundo ele, esta "só é aplicável sob condicões altamente restritivas, particularmente se as taxas de câmbio correspondem ao equilíbrio das balanças comerciais e se as vantagens comparativas são permanentes, o que em geral não é o caso". Allais, talvez por vício profissional, ou sentimento de impotência ante a realidade, se esqueceu de mencionar a necessidade da hipótese de pleno emprego.

Na maioria dos casos, o resultado da política liberal foi uma enorme destruição de empregos locais, em troca de uma pequena redução no preço do

produto para o consumidor e um grande custo fiscal para a sociedade toda, sobretudo para os próprios consumidores que mativeram-se empregados.

Os custos sociais estão hoje em evidência em toda parte. Um relatório recente da OIT prevê para o final da década taxas de desemprego em torno de 30% para os países desenvolvidos. Esta situação e a falta de perspectiva para os mais jovens cria um caldo de cultura propício à marginalidade e aos movimentos de extrema direita, visíveis em toda a Europa.

Frente a esta situação de catástrofe social, o ex-liberal Maurice Allais recomenda o fechamento comercial do mercado comum europeu, através do controle quantitativo de importações dos países extra-comunitários. No caso de a CEE não adotar francamente uma política de bloco, frontalmente contrária às regras do Gatt, recomenda que a França o faça sozinha. Na verdade, apesar da retórica liberal, é esta a prática corrente nos Estados Unidos e no Japão em matéria de comércio de mercadorias que ameaçam suas indústrias.

De outro lado, renascem também as propostas utópicas onde há os que, como Ricardo Petrella –em recente artigo no "Le Monde Diplomatique"–, esperam que a ONU no seu próximo encontro de cúpula sobre a questão social, a ser realizado em Copenhague em 1995, estabeleça as bases para uma nova ordem econômica e financeira mundial!

Independentemente do caráter conservador ou utópico e da viabilidade técnica ou política de quaisquer destas propostas, é um consolo saber que as pessoas estão reaprendendo que a solução para o problema do desemprego, resultante da modernização conservadora e dos excessos do liberalismo, não pode ser simplesmente mais liberalismo.

Enquanto isso, chega ao Brasil lady Margaret Thatcher, símbolo do que há de pior no liberalismo socialmente irresponsável e é aplaudida de pé pela nata do empresariado brasileiro.

As classes produtoras brasileiras não tomam juízo. Pagam US$ 100 mil para ouvir um show requentado da pseudo-rainha de um ex-império, cuja indústria entrou em decadência há 100 anos. Enquanto isto, sabotam, em nome do "livre mercado", mais um plano de estabilização, apesar de supostamente apoiarem o ministro como candidato.

Melhor fossem em caravana a Washington (e não a Nova York) verificar "in locu" as duas caras do consenso na capital do império. Na verdade, o que deviam

escutar e estudar são os planos de reestruturação da indústria e a reforma do sistema de saúde, privado e público, que o governo dos Estados Unidos está aplicando para melhorar a situação interna do seu país.

Não deveriam impressionar-se tanto com as receitas e pressões do FMI e do secretário do Tesouro norte-americano sobre o Brasil e muito menos deslumbrar-se com a performance de uma atriz coadjuvante. Se prestassem atenção ao que está ocorrendo com as mudanças na economia norte-americana, ficariam surpresos, por exemplo, com o grau de estatização do novo programa de telecomunicações.

Talvez aprendessem também que o aumento de produtividade sistêmica é incompatível com o sucateamento do Estado e não implica, do lado empresarial, simplesmente aumentar o desemprego e subir os preços.

Finalmente, concluiriam que o governo americano não está baixando os impostos nem desregulando sua economia, mas regulando-a mais intensamente do que nunca, para enfrentar a concorrência dos países asiáticos e do Japão.

Ao mesmo tempo, o "Consenso de Washington" pretende obter da América Latina um déficit comercial, através de uma sobrevalorização da nossa moeda, o que permitiria aos Estados Unidos reequilibrar a curto prazo suas contas externas.

Isto significa que o Brasil, o último país a resistir ao novo ajuste, que é o oposto do de 1982/83, deve submeter-se à dolarização e promover a toque de caixa e no segredo dos gabinetes a reforma constitucional, no capítulo da ordem econômica, numa direção supostamente liberal, o que sustentaria novo ciclo de endividamento.

Mas seria pedir demais às classes produtoras brasileiras, interessadas apenas no botim imediato, que tomassem consciência do seu destino e do destino da nação. Provincianos e deslumbrados pela mídia, parecem não saber o que acontece no mundo e são incapazes de pensamento estratégico.

Continuam viciados numa ideologia liberal suicida, preocupados apenas com os seus desejos incontidos de ganância especulativa e patrimonial, que vão custar ao governo, este ano, mais de US$ 10 bilhões em juros internos. Somando os juros da dívida externa (cuja negociação ainda não terminou), o próprio FMI estima em 5,7% do PIB (mais de US$ 22 bilhões) a conta global de juros, uma cifra inacreditável, cuidadosamente oculta pela equipe econômica, e superior ao impacto fiscal ocorrido no auge da crise da dívida externa!

É por isso que o "ajuste fiscal" nunca termina e que o processo de privatização é uma farsa sinistra.

Na verdade, como disse recentemente Clovis Rossi nesta Folha, estamos precisando mesmo é de uma "ruptura democrática" que exponha o nosso empresariado aos ventos da negociação e da verdadeira produtividade e que termine de vez com o seu caráter de parasitas financeiros.

O saneamento do Estado e o cuidado com o povo, seguramente não cabem a eles e sim ao avanço da consciência e do desejo de cidadania do próprio povo, particularmente na escolha de seus representantes no Congresso e dos futuros governos da nação.

MARIA DA CONCEIÇÃO TAVARES, 63, é economista, professora emérita da Universidade Federal do Rio de Janeiro (UFRJ) e professora associada da Universidade de Campinas (Unicamp).

- **Sumário manual di94mr20-20.txt**

Os custos sociais do liberalismo suicida

Países centrais tomam consciência da gravidade dos problemas gerados por uma política liberal irresponsável

MARIA DA CONCEIÇÃO TAVARES- Especial para a Folha

'Até que enfim , políticos e intelectuais dos países centrais começam a se conscientizar da gravidade dos problemas sociais e econômicos gerados por mais de uma década de um liberalismo irresponsável e dogmático.

Foi o que se observou no encontro de cúpula dos ministros do Trabalho dos países centrais em Detroit e em recentes declarações de respeitáveis economistas conservadores. Pela primeira vez, o G-7 se reúne para discutir a questão do desemprego em massa, marginalizando grande parte da população.

Isso é fruto da perda de competitividade de muitas empresas, vítimas da globalização financeira e da abertura comercial indiscriminada.

Os liberais não falam sobre isso e atribuem o desemprego à rigidez do mercado de trabalho dos países desenvolvidos. Propõem a flexibilidade do mercado do trabalho, retirando vantagens já conquistadas pelos trabalhadores.

Mas , depois de anos de flexibilização com aumento do desemprego, governos e até alguns renomados liberais começam a desconfiar que mais liberalismo não cura liberalismo.

O professor Maurice Allais, Prêmio Nobel de Economia, ataca a aplicação, nas condições contemporâneas, da doutrina das vantagens comparativas.Segundo ele, ela só é aplicável sob condições altamente restritivas.

Na maioria dos casos, a política liberal só destruiu empregos. Um relatório recente da OIT prevê taxas de desemprego por volta de 30% para os países desenvolvidos . Com a falta de perspectiva para os mais jovens desenvolve-se um clima de marginalidade propício aos movimentos de extrema direita.

Diante desse quadro catastrófico, o ex-liberal Maurice Allais recomenda o fechamento comercial do mercado comum europeu, através do controle quantitativo das importações dos países fora da CEE.

Enquanto isso, Margaret Thatcher, o que há de pior no liberalismo socialmente irresponsável, é vivamente aplaudida pelo nosso empresariado.

Se nossos empresários prestassem atenção no que ocorre nos EUA , se surpreenderiam , por exemplo, com a estatização do novo programa de telecomunicações. Compreenderiam que o governo americano está , mais do que nunca, regulando sua economia para enfrentar a concorrência dos países asiáticos e dos japoneses.

Também o "Consenso de Washington" quer ampliar o déficit comercial da América Latina com a sobrevalorização da nossa moeda, para reequilibrar suas contas externas. Para tanto, o Brasil deve submeter-se à dolarização e promover a reforma constitucional no capítulo sobre a economia , numa direção supostamente liberal.

Esses empresários , anestesiados pela ideologia liberal, continuam preocupados somente com a ambição especulativa e patrimonial, que vai custar ao governo, só neste ano, mais de US$ 10 bilhões de juros internos.

Como disse o jornalista Clovis Rossi na Folha , precisamos é de uma "ruptura democrática" que dê um banho de negociação e de verdadeira produtividade nos nossos empresários e elimine de vez seu caráter de parasitas financeiros.

- **Sumário SuPor-2 di94mr20-20.txt**

Finalmente, políticos e intelectuais dos países centrais começam a se dar conta da gravidade dos problemas sociais e econômicos gerados por mais de uma década de um liberalismo irresponsável, dogmático e anárquico.

Esta tardia tomada de consciência se manifesta no encontro de cúpula dos ministros do Trabalho dos países centrais em Detroit (o Job Summit) e em recentes declarações de renomados e respeitáveis economistas conservadores.

Pela primeira vez, o G-7 se reúne para discutir o problema do desemprego em massa nos países desenvolvidos, que não pára de crescer, lançando uma parcela cada vez maior da população na marginalidade.

Intimamente ligada a este processo está a questão da deslocalização, onde setores e até comunidades inteiras são destruídas, pois suas indústrias deixaram de ser competitivas num ambiente de globalização financeira e abertura comercial indiscriminada.

Não falam, é claro, os liberais como os da revista "The Economist", que ainda no número da semana passada repetem a ladainha de que o problema do desemprego é resultado da rigidez do mercado de trabalho dos países desenvolvidos, em particular os europeus.

Em outras palavras, o problema do desemprego viria do fato de que as economias centrais, no que diz respeito ao mercado de trabalho, são liberais de menos e a solução seria mais liberalismo.

Depois de anos de crescente "flexibilização" do mercado de trabalho, acompanhado de grande aumento e não de diminuição do desemprego, é natural que os governos e até alguns liberais de renome comecem a desconfiar que a solução para os males sociais causados pelo liberalismo irresponsável não seja mais liberalismo.

Em um artigo recente, o professor Maurice Allais, que recebeu o Prêmio Nobel de Economia em 1988 por suas contribuições à teoria neoclássica (teoria de onde a fé liberal busca obter credibilidade" científica "), faz um ataque frontal à aplicação, nas condições contemporâneas, da doutrina das vantagens comparativas.

Na maioria dos casos, o resultado da política liberal foi uma enorme destruição de empregos locais, em troca de uma pequena redução no preço do

produto para o consumidor e um grande custo fiscal para a sociedade toda, sobretudo para os próprios consumidores que mativeram-se empregados.

Frente a esta situação de catástrofe social, o ex-liberal Maurice Allais recomenda o fechamento comercial do mercado comum europeu, através do controle quantitativo de importações dos países extra-comunitários.

De outro lado, renascem também as propostas utópicas onde há os que, como Ricardo Petrella – em recente artigo no "Le Monde Diplomatique" –, esperam que a ONU no seu próximo encontro de cúpula sobre a questão social, a ser realizado em Copenhague em 1995, estabeleça as bases para uma nova ordem econômica e financeira mundial!

Independentemente do caráter conservador ou utópico e da viabilidade técnica ou política de quaisquer destas propostas, é um consolo saber que as pessoas estão reaprendendo que a solução para o problema do desemprego, resultante da modernização conservadora e dos excessos do liberalismo, não pode ser simplesmente mais liberalismo.

Finalmente, concluiriam que o governo americano não está baixando os impostos nem desregulando sua economia, mas regulando-a mais intensamente do que nunca, para enfrentar a concorrência dos países asiáticos e do Japão.

- **Sumário SuPor-2 Fuzzy di94mr20-20.txt**

Pela primeira vez, o G-7 se reúne para discutir o problema do desemprego em massa nos países desenvolvidos, que não pára de crescer, lançando uma parcela cada vez maior da população na marginalidade.

Intimamente ligada a este processo está a questão da deslocalização, onde setores e até comunidades inteiras são destruídas, pois suas indústrias deixaram de ser competitivas num ambiente de globalização financeira e abertura comercial indiscriminada.

A combinação de taxas de desemprego crescente com a decadência econômica de regiões onde ocorre a deslocalização gera um quadro social terrível, cujas consequências são bem conhecidas.

Não falam, é claro, os liberais como os da revista "The Economist", que ainda no número da semana passada repetem a ladainha de que o problema do desemprego é resultado da rigidez do mercado de trabalho dos países

desenvolvidos, em particular os europeus.

A solução, como sempre, seria aumentar a "flexibilidade" do mercado de trabalho, com a retirada do seguro-desemprego e demais empecilhos ao livre jogo das forças de mercado.

Depois de anos de crescente "flexibilização" do mercado de trabalho, acompanhado de grande aumento e não de diminuição do desemprego, é natural que os governos e até alguns liberais de renome comecem a desconfiar que a solução para os males sociais causados pelo liberalismo irresponsável não seja mais liberalismo.

Na verdade, apesar da retórica liberal, é esta a prática corrente nos Estados Unidos e no Japão em matéria de comércio de mercadorias que ameaçam suas indústrias.

Na verdade, o que deviam escutar e estudar são os planos de reestruturação da indústria e a reforma do sistema de saúde, privado e público, que o governo dos Estados Unidos está aplicando para melhorar a situação interna do seu país.

Se prestassem atenção ao que está ocorrendo com as mudanças na economia norte-americana, ficariam surpresos, por exemplo, com o grau de estatização do novo programa de telecomunicações.

Ao mesmo tempo, o "Consenso de Washington" pretende obter da América Latina um déficit comercial, através de uma sobrevalorização da nossa moeda, o que permitiria aos Estados Unidos reequilibrar a curto prazo suas contas externas.

Continuam viciados numa ideologia liberal suicida, preocupados apenas com os seus desejos incontidos de ganância especulativa e patrimonial, que vão custar ao governo, este ano, mais de US$ 10 bilhões em juros internos.

É por isso que o "ajuste fiscal" nunca termina e que o processo de privatização é uma farsa sinistra.

O saneamento do Estado e o cuidado com o povo, seguramente não cabem a eles e sim ao avanço da consciência e do desejo de cidadania do próprio povo, particularmente na escolha de seus representantes no Congresso e dos futuros governos da nação.

- **Sumário TextRankStemming+StopRem di94mr20-20.txt**

Países centrais tomam consciência da gravidade dos problemas gerados por uma política liberal irresponsável

Finalmente, políticos e intelectuais dos países centrais começam a se dar conta da gravidade dos problemas sociais e econômicos gerados por mais de uma década de um liberalismo irresponsável, dogmático e anárquico.

Esta tardia tomada de consciência se manifesta no encontro de cúpula dos ministros do Trabalho dos países centrais em Detroit (o Job Summit) e em recentes declarações de renomados e respeitáveis economistas conservadores.

Não falam, é claro, os liberais como os da revista "The Economist", que ainda no número da semana passada repetem a ladainha de que o problema do desemprego é resultado da rigidez do mercado de trabalho dos países desenvolvidos, em particular os europeus.

Em outras palavras, o problema do desemprego viria do fato de que as economias centrais, no que diz respeito ao mercado de trabalho, são liberais de menos e a solução seria mais liberalismo.

Depois de anos de crescente "flexibilização" do mercado de trabalho, acompanhado de grande aumento e não de diminuição do desemprego, é natural que os governos e até alguns liberais de renome comecem a desconfiar que a solução para os males sociais causados pelo liberalismo irresponsável não seja mais liberalismo.

Na maioria dos casos, o resultado da política liberal foi uma enorme destruição de empregos locais, em troca de uma pequena redução no preço do produto para o consumidor e um grande custo fiscal para a sociedade toda, sobretudo para os próprios consumidores que mativeram-se empregados.

De outro lado, renascem também as propostas utópicas onde há os que, como Ricardo Petrella – em recente artigo no "Le Monde Diplomatique" –, esperam que a ONU no seu próximo encontro de cúpula sobre a questão social, a ser realizado em Copenhague em 1995, estabeleça as bases para uma nova ordem econômica e financeira mundial!

Independentemente do caráter conservador ou utópico e da viabilidade técnica ou política de quaisquer destas propostas, é um consolo saber que as pessoas estão reaprendendo que a solução para o problema do desemprego,

resultante da modernização conservadora e dos excessos do liberalismo, não pode ser simplesmente mais liberalismo.

Na verdade, o que deviam escutar e estudar são os planos de reestruturação da indústria e a reforma do sistema de saúde, privado e público, que o governo dos Estados Unidos está aplicando para melhorar a situação interna do seu país.

Finalmente, concluiriam que o governo americano não está baixando os impostos nem desregulando sua economia, mas regulando-a mais intensamente do que nunca, para enfrentar a concorrência dos países asiáticos e do Japão.

Somando os juros da dívida externa (cuja negociação ainda não terminou), o próprio FMI estima em 5,7 % do PIB (mais de US$ 22 bilhões) a conta global de juros, uma cifra inacreditável, cuidadosamente oculta pela equipe econômica, e superior ao impacto fiscal ocorrido no auge da crise da dívida externa!

- **Sumário TextRank+Thesaurus di94mr20-20.txt**

Esta tardia tomada de consciência se manifesta no encontro de cúpula dos ministros do Trabalho dos países centrais em Detroit (o Job Summit) e em recentes declarações de renomados e respeitáveis economistas conservadores.

Intimamente ligada a este processo está a questão da deslocalização, onde setores e até comunidades inteiras são destruídas, pois suas indústrias deixaram de ser competitivas num ambiente de globalização financeira e abertura comercial indiscriminada.

Não falam, é claro, os liberais como os da revista "The Economist", que ainda no número da semana passada repetem a ladainha de que o problema do desemprego é resultado da rigidez do mercado de trabalho dos países desenvolvidos, em particular os europeus.

Depois de anos de crescente "flexibilização" do mercado de trabalho, acompanhado de grande aumento e não de diminuição do desemprego, é natural que os governos e até alguns liberais de renome comecem a desconfiar que a solução para os males sociais causados pelo liberalismo irresponsável não seja mais liberalismo.

Em um artigo recente, o professor Maurice Allais, que recebeu o Prêmio Nobel de Economia em 1988 por suas contribuições à teoria neoclássica (teoria de

onde a fé liberal busca obter credibilidade" científica "), faz um ataque frontal à aplicação, nas condições contemporâneas, da doutrina das vantagens comparativas.

Na maioria dos casos, o resultado da política liberal foi uma enorme destruição de empregos locais, em troca de uma pequena redução no preço do produto para o consumidor e um grande custo fiscal para a sociedade toda, sobretudo para os próprios consumidores que mativeram-se empregados.

Esta situação e a falta de perspectiva para os mais jovens cria um caldo de cultura propício à marginalidade e aos movimentos de extrema direita, visíveis em toda a Europa.

Independentemente do caráter conservador ou utópico e da viabilidade técnica ou política de quaisquer destas propostas, é um consolo saber que as pessoas estão reaprendendo que a solução para o problema do desemprego, resultante da modernização conservadora e dos excessos do liberalismo, não pode ser simplesmente mais liberalismo.

Na verdade, o que deviam escutar e estudar são os planos de reestruturação da indústria e a reforma do sistema de saúde, privado e público, que o governo dos Estados Unidos está aplicando para melhorar a situação interna do seu país.

Isto significa que o Brasil, o último país a resistir ao novo ajuste, que é o oposto do de 1982/83, deve submeter-se à dolarização e promover a toque de caixa e no segredo dos gabinetes a reforma constitucional, no capítulo da ordem econômica, numa direção supostamente liberal, o que sustentaria novo ciclo de endividamento.

Somando os juros da dívida externa (cuja negociação ainda não terminou), o próprio FMI estima em 5,7 % do PIB (mais de US$ 22 bilhões) a conta global de juros, uma cifra inacreditável, cuidadosamente oculta pela equipe econômica, e superior ao impacto fiscal ocorrido no auge da crise da dívida externa!

O saneamento do Estado e o cuidado com o povo, seguramente não cabem a eles e sim ao avanço da consciência e do desejo de cidadania do próprio povo, particularmente na escolha de seus representantes no Congresso e dos futuros governos da nação.