Universidade Estadual de Campinas
Faculdade de Engenharia Elétrica e Computação
Departamento de Engenharia de Computação e Automação Industrial

SABIO: Abordagem Conexionista Supervisionada para Sumarização Automática de Textos

TÉLVIO ORRÚ

Orientador: Prof. Dr. Márcio Luiz de Andrade Netto
(DCA-FEEC-Unicamp)

Co-Orientador: Prof. Dr. João Luís Garcia Rosa
(Ceatec - PUC-Campinas)

Dissertação de Mestrado apresentada à Faculdade de Engenharia Elétrica e de Computação da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Mestre em Engenharia Elétrica. Área de concentração: **Engenharia de Computação**

Banca Examinadora: Prof. Dr. Márcio Luiz de Andrade Netto (DCA/FEEC-UNICAMP)
Prof. Dr. Ricardo Ribeiro Gudwin (DCA/FEEC-UNICAMP)
Prof. Dr. Fernando José Von Zuben (DCA/FEEC-UNICAMP)
Prof. Dr. Edson Françozo (IEL-UNICAMP)

Campinas
Agosto de 2005

FICHA CATALOGRÁFICA ELABORADA PELA
BIBLIOTECA DA ÁREA DE ENGENHARIA - BAE - UNICAMP

Orrú, Télvio
Or7s SABIO: abordagem conexionista supervisionada para sumarização automática de textos / Télvio Orrú. --Campinas, SP: [s.n.], 2005.

Orientadores: Márcio Luiz de Andrade Netto, João Luís Garcia Rosa
Dissertação (Mestrado) - Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação.

1. Redes neurais (Computação). 2. Processamento da linguagem natural (Computação). I. Andrade Netto, Márcio Luiz de. II. Rosa, João Luis Garcia. III. Universidade Estadual de Campinas. Faculdade de Engenharia Elétrica e de Computação. IV. Título.

Titulo em Inglês: SABIO: Supervised connectionist approach to automatic text summarization
Palavras-chave em Inglês: Artificial neural networks, Natural language processing
Área de concentração: Engenharia de Computação
Titulação: Mestre em Engenharia Elétrica
Banca examinadora: Ricardo Ribeiro Gudwin, Fernando José Von Zuben e Edson Françozo
Data da defesa: 26/08/2005

Resumo

Propõe-se, neste projeto, a criação de uma ferramenta computacional para geração de novos sumários a partir de novos textos-fonte, por meio do uso de abordagem conexionista (Redes Neurais Artificiais). Dentre as contribuições que este trabalho pretende trazer à área de Processamento de Línguas Naturais, destaca-se a abordagem biologicamente mais plausível da arquitetura e do treinamento conexionistas para a sumarização automática. Utilizou-se esta abordagem para o treinamento da rede pois acredita-se que este tratamento poderá trazer ganhos em relação à eficiência computacional quando comparado aos modelos conexionistas considerados biologicamente implausíveis

Palavras-chave: redes neurais artificiais, processamento de línguas naturais, abordagem conexionista biologicamente mais plausível, sumarizadores automáticos de textos.

Abstract

It is proposed here an implementation of a computational tool to generate new summaries from new source texts, by means of a connectionist approach (artificial neural networks). Among other contributions that this work intends to bring to natural language processing, it is highlighted the use of biologically more plausible connectionist architecture and training for automatic summarization. The choice relies on the expectation that it may bring an increase in computational efficiency when compared to the so-called biologically implausible algorithms.

Keywords: artificial neural networks, natural language processing, biologically more plausible connectionist approach, automatic text summarizers.

Agradecimentos

Agradeço ao professor Márcio Luiz de Andrade Netto pela orientação e exposição de seus comentários de forma objetiva; ao professor João Luís Garcia Rosa, da PUC-Campinas, pela co-orientação e pela grandiosa disposição, prontidão e interesse demonstrados nesta caminhada.

Agradeço à Universidade Estadual de Campinas e particularmente à Faculdade de Engenharia Elétrica e de Computação pela oportunidade de realização deste trabalho.

Destaco a gratidão aos meus pais e ao meu irmão pelo apoio e pela tolerância neste período no qual estive “distante” de alguns compromissos pessoais para a realização deste trabalho.

Agradeço a Deus por permitir a conclusão desta etapa.

Dedico esta dissertação aos meus Pais
e ao meu Irmão.

Índice

CAPÍTULO 1 - INTRODUÇÃO 1

CAPÍTULO 2 - CONCEITOS BÁSICOS 5

2.1. Sumarização de Textos 5
2.2. Sumarização Automática de Textos 7
2.3. Redes Neurais Artificiais 13
2.3.1. Um algoritmo de treinamento biologicamente implausível (Backpropagation) 16
2.3.2. Um algoritmo de treinamento biologicamente mais plausível (GeneRec) 22

CAPÍTULO 3 - IMPLEMENTAÇÕES 31

3.1. Implementações Existentes 31
3.1.1. Variação do $A_{r\text{-}p}$ 31
3.1.2. "Temperando" o Backpropagation 34
3.1.3. Bio-Pred 35
3.2. O sistema proposto SABIO 38
3.2.1. Características do SABIO 38

CAPÍTULO 4 - RESULTADOS 47

4.1. Comparações entre a Medida-F na ausência de traços utilizado no SABIO 49
4.2. Comparações entre o algoritmo de treinamento *Backpropagation* e o algoritmo de treinamento *GeneRec* para o sistema proposto (SABIO) 51
4.3. Comparações entre sumarizadores automáticos de textos 53
4.3.1. SABIO & Sumarizadores Automáticos de Textos 54
4.3.2. SABIO e NeuralSumm 58
4.3.3. SABIO (GR) e SABIO (BP) / Dependência do Corpus 59

CAPÍTULO 5 - CONCLUSÃO 63

APÊNDICES 67

Apêndice A - *Stoplist* utilizada no SABIO 69

Apêndice B - Algoritmos 71

Apêndice C - Telas 73

Apêndice D - Características técnicas e funcionais do SABIO 89

Apêndice E - Gráficos 91

Apêndice F - Exemplos de Extratos Gerados 95

BIBLIOGRAFIA 103

Lista de Tabelas

Tabela 3.1. – Comparação entre o algoritmo conexionista Backpropagation, considerado biologicamente implausível e um algoritmo considerado biologicamente mais plausível (Backpropagation com "tempero") (SCHRAUDOLPH & SEJNOWSKI, 1996) 34

Tabela 3.2.-Representação da matriz dos traços para treinamento da rede utilizada pelo SABIO.. 44

Tabela 4.1. – Procedimento para validação: validação cruzada em pacotes de 10 grupos, não enviesada ("*10-fold cross validation, non-biasing*")..... 55

Tabela 4.2.–Comparação entre Sumarizadores Automáticos de Textos (adaptada de RINO *et al.*, 2004)..... 56

Tabela A.1. – Stoplist utilizada pelo SABIO..... 69

Lista de Figuras

Figura 2.1 – Passo 1 da ativação do Backpropagation 17
Figura 2.2 – Propagação do padrão entre as camadas na fase de ativação do Backpropagation .. 18
Figura 2.3 – Momento em que a RNA gera uma saída 18
Figura 2.4 – Diferença entre saída desejada e a saída real sendo retropropagada da camada de saída para a camada escondida 19
Figura 2.5 – Retropropagação do erro no Backpropagation 20
Figura 2.6 – Representação parcial das duas fases de GeneRec (O´REILLY, 1996; ROSA, 2002a) 23
Figura 2.7 – Passo 1 da fase "menos" do GeneRec 24
Figura 2.8 – Passo 2 da fase "menos" do GeneRec 24
Figura 2.9 – Passo 3 da fase "menos" do GeneRec 25
Figura 2.10 – Passo 1 da fase "mais" do GeneRec 26
Figura 2.11 – Passo 2 da fase "mais" do GeneRec 26
Figura 3.1 – Curvas de aprendizado dos algoritmos Backpropagation e $A_{r\text{-}p}$ (MAZZONI *et al*., 1991) 32
Figura 3.2. – Comparação entre o erro apresentado com o treinamento através do algoritmo Backpropagation e com o algoritmo $A_{r\text{-}p}$ (MAZZONI *et al*., 1991) 33
Figura 3.3. - Comparação entre os sistemas Bio-Pred1, Bio-Pred2 e Pred-DR para a sentença de entrada "*The wolf frightened the girl*" (ROSA, 2002a) 36
Figura 3.4. - Comparação entre os sistemas Bio-Pred1, Bio-Pred2 e Pred-DR, para a sentença de entrada "*The stone broke the vase*" (ROSA, 2002a) 37
Figura 3.5. – Escala utilizada (traços 1, 5 e 7) para classificação de sentenças no SABIO. 41
Figura 4.1. – Redução da Medida-F quando desconsiderado um traço da rede do SABIO. 49
Figura 4.2. – Melhor performance encontrada nos testes efetuados para o SABIO. 52
Figura 4.3 – Comportamento do erro mínimo para o melhor desempenho encontrado (quanto ao erro mínimo) 53
Figura 4.4. – Variação da Medida-F quanto aos valores iniciais dos pesos das conexões da RNA 56
Figura 4.5. – Comparação das taxas de cobertura, precisão e Medida-F entre as duas versões do SABio (GR e BP) e o NeuralSumm. 58
Figura 4.6 – Comparação Medida-F entre as duas versões do SABIO (GR e BP) e o NeuralSumm nos corpora CorpusDT e no TeMário. 59
Figura 4.7 – Comparação da Medida-F para verificação da dependência do corpus utilizado para o treinamento da RNA. Taxa de compressão de 70%. 60
Figura D.1 - Estrutura das pastas para utilização do SABIO 90
Figura E.1 - SABio com 10 neurônios na camada escondida e taxa de aprendizagem de 0.10. 91
Figura E.2 - SABio com 10 neurônios na camada escondida e taxa de aprendizagem de 0.25. 92
Figura E.3 – SABio com 10 neurônios na camada escondida e taxa de aprendizagem de 0.35. 92
Figura E.4 – SABio com 10 neurônios na camada escondida e taxa de aprendizagem de 0.45. 93
Figura E.5 – Relação entre a taxa de compressão e a taxa de cobertura. 94
Figura E.6 – Relação entre a taxa de compressão e a taxa de precisão. 94

CAPÍTULO 1

INTRODUÇÃO

Nota-se que o crescente volume de informações disponíveis e a escassez de tempo para a leitura de textos longos faz com que a sociedade moderna busque por resumos e manchetes de notícias ao invés de textos completos.

No interesse em realizar a sumarização automática de textos, propõe-se o sistema SABIO (Sumarizador Automático com Arquitetura e Aprendizado Conexionistas Biologicamente mais Plausíveis) por meio do uso de redes neurais artificiais com treinamento considerado biologicamente mais[1] plausível. Apresenta-se uma arquitetura em princípio inédita[2] para esta aplicação tendo em vista características interessantes como: a) o treinamento da rede neural com tratamento considerado biologicamente mais plausível; b) o conjunto de treinamento supervisionado (pares de entrada e saída) formado por traços[3] das sentenças dos textos-fontes e das sentenças dos extratos ideais, que embora sejam valores binários não estão limitados a "pertence" ou "não pertence", podendo representar valores intermediários de uma lógica multivalorada.

Para apresentar o SABIO, esta dissertação está organizada em cinco capítulos, sendo:

1. Introdução;

2. Conceitos Básicos: Dividido em três itens:

[1] Chama-se atenção para o termo "mais" de "biologicamente mais plausível", visto que este termo expressa (neste trabalho) apenas a comparação entre um algoritmo considerado "biologicamente mais plausível" com um algoritmo considerado "biologicamente implausível". Em nenhum momento afirma-se que a arquitetura utilizada é "biologicamente plausível".

[2] Fez-se busca por Sumarizadores Automáticos e não foi encontrado nenhum que utilize este tipo de arquitetura.

[3] Define-se traços como as características (*"features")* - que são analisadas na fase de treinamento – e que representam as sentenças dos textos que são usados no SABIO. Os traços utilizados no SABIO estão descritos no capítulo 3.

2.1. Sumarização de Textos: traz conceitos fundamentais sobre a sumarização de textos, apresentando a justificativa e a motivação na escolha desta aplicação (sumarização automática de textos) para este projeto,

2.2. Sumarização Automática de Textos: apresenta definições sobre a Sumarização Automática de Textos explicando tipos e classificações de sumários bem como diferenciando áreas correlatas da sumarização de textos. Comenta-se sobre algumas tentativas existentes para o processo automático de sumarização de textos que utilizam desde métodos estatísticos até redes neurais artificiais,

2.3. Redes Neurais Artificiais: traz uma descrição superficial de conceitos de redes neurais artificiais citando alguns algoritmos existentes para treinamento da rede, procurando exibir as diferenças entre as características do algoritmo de treinamento tradicional (Backpropagation) e o algoritmo de treinamento de RNA utilizado neste trabalho (GeneRec). Ainda neste item justifica-se, baseado em textos encontrados na literatura, o porquê do Backpropagation (tradicional algoritmo supervisionado de treinamento de redes conexionistas) ser considerado biologicamente implausível;

3 . Implementações: Embora escassas as aplicações que utilizam redes neurais artificiais para a sumarização automática de textos, pode-se encontrar trabalhos que propõem (em outras aplicações) comparações entre o treinamento de redes neurais com tratamento considerado biologicamente mais plausível e o treinamento de redes neurais com tratamento considerado biologicamente implausível. Neste capítulo, além de se fazer referências a aplicações com esta proposta, apresenta-se o sistema SABIO que faz uso de redes neurais artificiais para realizar a sumarização automática de textos;

4. Resultados: Apresentam-se comparações (para a aplicação proposta) entre o algoritmo de treinamento de redes neurais artificiais Backpropagation e o algoritmo de treinamento de redes neurais artificiais GeneRec;

5. Conclusões: Faz-se algumas considerações sobre o estado da arte da "Sumarização Automática de Textos" e de "Redes Neurais Artificiais com treinamento considerado biologicamente mais plausível". Ainda neste capítulo expõe-se o ponto de vista do autor sobre o modelo de treinamento de redes neurais artificiais utilizado para a aplicação de sumarização automática de textos, onde são relatadas algumas situações e experiências encontradas no processo de desenvolvimento do sistema SABIO.

CAPÍTULO 2

CONCEITOS BÁSICOS

2.1. Sumarização de Textos

A sumarização de textos é o processo de produzir uma versão mais curta de um texto-fonte (MANI & MAYBURY, 1999). Pode-se, através deste processo, obter um extrato (do termo *extract*, em inglês) ou um sumário (do termo *abstract*, em inglês). Extratos são criados pela justaposição de sentenças do texto-fonte consideradas importantes; sumários, diferentemente de extratos, alteram a estrutura e/ou o conteúdo das sentenças originais, fundindo-as e/ou reescrevendo-as, para generalizar ou especificar as informações (MANI, 2001).

A pesquisa deste tema apóia-se no interesse crescente da sociedade moderna na busca de manchetes de jornais e/ou revistas que propõem trazer temas atuais resumidos ao invés de textos completos, principalmente pela falta de tempo das pessoas nos dias atuais.

Assumindo que o foco deste trabalho é a sumarização automática de textos, comenta-se na seqüência características de um sumário. Tem-se duas visões em um sumário: do ponto de vista do leitor (usuário do sumário) e do ponto de vista do escritor (o criador do sumário). Este último tem a tarefa de condensar um texto-fonte para que a idéia principal do texto seja possível ser transmitida com o sumário criado. Sendo assim, as principais premissas da sumarização podem ser citadas (PARDO *et al.*, 2003b):

- Está disponível um texto, aqui denominado *texto-fonte*, que deve ser condensado,
- A afirmação de que o objeto a ser sumarizado constitui um texto implica, adicionalmente, na existência de (PARDO *et al.*, 2003b):

a) uma idéia central – o tópico principal do texto – sobre a qual constrói-se a trama textual;

b) um conjunto de unidades de informação que têm relação com a idéia central em desenvolvimento;

c) um objetivo comunicativo central que, implícita ou explicitamente, direciona tanto a seleção das unidades de informação quanto a seleção da forma como a informação será estruturada, para estabelecer a idéia pretendida;

d) um enredo, tecido em função das escolhas antes citadas, visando transmitir a idéia central de forma coerente, com intenção de atingir o objetivo comunicativo pretendido.

Considerando estes conceitos pode-se atribuir a principal função da sumarização de textos como a tarefa de identificar o que é relevante no texto e, então, traçar o novo enredo, a partir do conteúdo disponível, preservando sua idéia central sem transgredir o significado original pretendido (PARDO *et al.*, 2003b).

SPARCK JONES (1993) classifica os sumários em duas categorias: indicativos e informativos. Sumários indicativos, como já é sugerido em seu nome, apenas permitem transmitir uma idéia do que os textos-fonte se referem, não podendo, portanto substituí-los, pois necessariamente não preservam o que têm de mais importante (em conteúdo e estrutura). Sumários informativos (ou autocontidos), por sua vez, devem conter os aspectos principais dos textos-fonte, dispensando a leitura dos mesmos.

Devido a esta classificação, ficam bastantes distintas as aplicações para cada tipo, bem como a avaliação da sua qualidade. Como exemplo, pode-se utilizar sumários indicativos para classificação de documentos bibliográficos de forma que indiquem o conteúdo do texto-fonte – neste caso servem de indexadores.

Em sumários informativos a aplicação é como meio de informação ao usuário. A avaliação da eficiência destes sumários não é uma tarefa trivial visto que – na maioria dos casos – depende do auxílio do ser humano.

De modo geral, é mais fácil produzir automaticamente sumários indicativos do que informativos. A condensação do texto-fonte com a preservação de seu conteúdo mais relevante, ou seja, a geração de um sumário informativo, é denominada sumarização de textos.

2.2. Sumarização Automática de Textos

Do ponto de vista computacional três operações básicas podem descrever o processo de sumarização: análise, seleção de conteúdo e generalização (ARETOULAKI,1996).

Existem várias tentativas para automatizar o processo de sumarização de textos, utilizando desde métodos estatísticos até Redes Neurais Artificiais. Dentre estes trabalhos, pode-se encontrar:

- LUHN (1958) sugere o uso de informações estatísticas derivadas do cálculo da freqüência das palavras e sua distribuição no texto para calcular uma "medida relativa de significância";
- EDMUNDSON (1969) propõe o método "seleção computacional de sentenças com maior potencial de transmitir ao leitor a substância do documento". A principal evolução em relação ao trabalho de Lunh é que Edmundson considera as palavras sinalizadoras[4];
- POLLOCK & ZAMORA (1975) sugerem a necessidade de se restringir domínios e o cruzamento de sentenças com o título da obra;
- KUPIEC *et al*. (1995) propõem um sumarizador extrativo que agrupa as potenciais características a partir da comparação do conteúdo do texto-

[4] "*Cue words*" em inglês. Exemplos de palavras sinalizadoras: importante, significativo e bastante.

fonte com o conteúdo de seus respectivos extratos construídos manualmente. Este trabalho é muito relevante também por motivar a busca por técnicas mais robustas, como a Sumarização Automática ("SA") baseada em *corpora*[5];

- TEUFEL & MOENS (2002) realizam uma análise de traços superficiais em sentenças de textos científicos que podem indicar sua relevância para compor um sumário/extrato;
- LAROCCA NETO *et al*. (2002), de maneira semelhante a Teufel & Moens, consideram um conjunto mais amplo de traços que abrange (a) a distribuição de palavras e medidas estatísticas, (b) as medidas de coesão das sentenças e (c) a estrutura argumentativa do texto a sumarizar. Os textos utilizados são artigos de revistas sobre computação em geral.

Enquanto os três primeiros trabalhos citados tratam de obras clássicas em SA, os três últimos utilizam classificação bayesiana, ou seja, assume-se que para classificar um determinado objeto existe um conjunto de atributos que permitem calcular a probabilidade do objeto pertencer a uma determinada classe.

Destaque especial faz-se aos trabalhos que utilizam Redes Neurais Artificiais (RNA), pois aqui também esta ferramenta será utilizada:

- PARDO *et al.* (2003b), em "NeuralSumm", utiliza uma RNA tipo SOM (*self-organizing map*), que foi treinada para classificar sentenças de um texto-fonte de acordo com seu grau de importância e assim produzir o extrato correspondente. Esta rede organiza as informações em grupos de similaridade em função dos traços apresentados;
- ARETOULAKI (1996) utiliza RNA com treinamento considerado biologicamente implausível (Backpropagation) - diferentemente da

[5] *Corpora* é o plural de *corpus*. *Corpus* (do latim "corpo") é definido como qualquer coleção que contenha mais de um texto (FELTRIM *et al*., 2001).

abordagem aqui proposta. Aretoulaki faz uma análise detalhada de textos jornalísticos e científicos de vários domínios para reconhecer traços genéricos que representam bem o conteúdo das sentenças de qualquer gênero textual[6] e que podem ser aprendidas por uma rede neural. Como traços da RNA, Aretoulaki considera características de diversas naturezas, desde superficiais até o conjunto de regras do idioma, que possam ser identificados em uma análise textual.

É importante lembrar que por um longo período de tempo a exploração de métodos para SA ficou estagnada, devido à impossibilidade técnica para implementá-los (limitações de hardware, software e a falta de repositórios lingüísticos) (PARDO *et al.*, 2003b).

MANI (2001) faz uma diferenciação entre Sumarização Automática e áreas correlatas:

- Recuperação de Documentos: para uma certa "chave de busca", visa produzir uma coleção de documentos relevantes, sem necessariamente condensá-los;
- Indexação: visa identificar termos convenientes para a recuperação de informação;
- Extração de informação: não necessariamente tem a condensação de informação como restrição fundamental;
- Mineração de textos: sua principal função é identificar, nos textos, informações singulares, e não necessariamente informações principais.

Não é raro encontrar trabalhos que fazem citações ao emprego de anotação e etiquetagem ("*annotation*" e "*tagging*") quando pretendem analisar os diferentes sentidos das palavras.

[6] Gêneros textuais referem-se aos temas dos textos utilizados (JURAFSKY *&* MARTIN, 2000)

MANI & MAYBURY (1999) afirmam que “o problema é determinar a contribuição relativa de diferentes características, condição altamente dependente do gênero textual”.

Extrair automaticamente de um corpus todas as ocorrências de uma determinada palavra seria bastante fácil, porém tal processo não faria a distinção entre os diferentes sentidos das palavras encontradas. Um “anotador” humano precisaria inspecionar todas as ocorrências e distinguir os diferentes sentidos com o auxílio de um dicionário. Sendo assim, o “anotador” registra a relação entre a ocorrência da palavra e o correspondente sentido encontrado em um dicionário. Esse processo de anotação semântica é conhecido também como “etiquetagem”, do inglês “*tagging*” (FELLBAUM *et al.*, 2001).

FELLBAUM *et al* (2001) afirmam que um dicionário descreveria a “etiquetagem” como a forma como “os anotadores inspecionam a ocorrência de polissemia em uma seqüência de textos, interpretam, determinam os significados e igualam esses a um dicionário”.

Na abordagem deste trabalho não será usada a técnica de etiquetagem, visto que o custo computacional é relativamente alto para realizar o processamento semântico (FELLBAUM *et al.*, 2001).

Há sumarizadores automáticos de textos que utilizam técnicas para descoberta das sentenças mais importantes no texto-fonte (PARDO, 2002; PARDO, 2003b). Tais abordagens são, em geral, estatísticas, pois procuram posicionar as sentenças de acordo com o grau de freqüência que as palavras contidas nas mesmas representam no texto a que pertence. Por ponderar as sentenças conforme o grau de freqüência de repetição que as palavras da mesma se repetem no texto e por acreditar que as sentenças que possuem pontuação mais

alta expressam a idéia principal do texto, estas sentenças são também referenciadas como "*gist sentences*"[7].

Para a avaliação de sumarizadores automáticos de textos comumente são empregadas algumas taxas, a saber:

- Taxa da cobertura → calculada por meio da quantidade das sentenças selecionadas corretamente dividida pela quantidade total de sentenças corretas;

- Taxa de precisão → calculada por meio da quantidade das sentenças selecionadas corretamente dividida pela quantidade total de sentenças selecionadas;

- Medida-F → produto da taxa de precisão e taxa de cobertura dividido pelo produto da taxa de precisão e taxa de cobertura. O resultado desta operação é multiplicado por 2.

Matematicamente, pode-se representar como:

$$TaxaCobertura = \frac{Qtde.deSentençasSelecionadasCorretamente}{Qtde.TotaldeSentençasCorretas}$$

$$Taxa\ Pr ecisão = \frac{Qtde.deSentençasSelecionadasCorretamente}{Qtde.TotaldeSentençasSelecionadas}$$

$$Medida - F = 2 * \left(\frac{Taxa\ Pr ecisão * TaxaCobertura}{Taxa\ Pr ecisão + TaxaCobertura} \right)$$

[7] "*Gist*" em inglês. Em português: significado mais importante de alguma coisa, essência (Michaelis, versão eletrônica, 1999).

Percebe-se que quanto maior a Medida-F obtida melhor pode ser considerado o extrato gerado pois para o cálculo desta taxa são consideradas as taxas de cobertura e precisão.

Resumidamente: os sistemas automáticos capazes de efetuar a condensação[8], com a preservação do conteúdo mais relevante do texto-fonte, denominados sistemas para Sumarização Automática de textos, costumam ser avaliados, e comparados entre si, utilizando as taxas de cobertura, precisão e medida-F, comentadas neste capítulo.

[8] Entende-se como condensação o "resumo ou a síntese" (Michaelis, versão eletrônica, 1999).

2.3. Redes Neurais Artificiais

As RNAs tiveram sua origem a partir do esforço de pesquisadores em entender o funcionamento do cérebro humano e então desenvolver modelos matemáticos para representá-lo. A intenção inicial seria que estas redes operassem de modo similar ao cérebro humano podendo tomar decisões, aprender e lembrar de forma semelhante (ou melhor) que o cérebro humano, porém isto ainda não ocorre devido à complexidade e ao não conhecimento do funcionamento do cérebro humano de forma detalhada.

As principais áreas envolvidas para implementação de RNA's são: estatística, teoria da informação, teoria de sistemas lineares e não-lineares, teoria da computação, álgebra linear, aproximação de funções, processamento de sinais, controle de processos e otimização de sistemas (HAYKIN, 1999).

As RNAs artificiais são baseadas na interconexão de unidades de processamento simples e similares, denominadas neurônios artificiais.

As RNAs podem apresentar soluções para problemas que possuam multidimensionalidade e variáveis sujeitas a interações não-lineares, desconhecidas ou matematicamente intratáveis (DE CASTRO & VON ZUBEN, 1998). Além disto, uma interessante característica das RNAs é sua capacidade de "descobrir soluções gerais". Considerando que tenha aprendido através de um conjunto de exemplos, as RNAs podem produzir saídas corretas através de entradas que não tinham sido apresentadas às mesmas.

O tratamento fornecido por uma RNA através de entradas que geram saídas é feito por meio de treino iterativo, comumente com exemplos de pares de entrada/saída. Um aprendizado é reconhecido como "ótimo" quando o erro médio da rede é minimizado. A minimização do erro é obtida por meio da modificação

dos pesos das conexões, assim pode-se obter a reprodução das saídas desejadas quando uma nova entrada é recebida pela RNA.

Existem basicamente três paradigmas de aprendizado para RNA, a saber: aprendizado supervisionado, aprendizado por reforço e aprendizado não-supervisionado. Neste trabalho optou-se por utilizar o aprendizado supervisionado com a arquitetura do perceptron de múltiplas camadas (*multi-layer perceptron* - "MLP") – pois a sumarização automática de textos refere-se a um problema de classificação e reconhecimento de padrões[9], para o qual sabe-se que o aprendizado supervisionado pode contribuir para encontrar a solução (DE CASTRO & VON ZUBEN, 1998).

Considerando a arquitetura de rede que será utilizada ("multicamada"), deve-se definir a maneira como a rede será treinada e para isto existem vários algoritmos de treinamento supervisionados, entre eles (DE CASTRO, 2003):

- algoritmo padrão (BP) $\rightarrow$ Neste algoritmo (Backpropagation) o erro (diferença entre a saída desejada e saída obtida) é calculado. Este erro é retropropagado da camada de saída até a camada de entrada alterando os pesos das conexões das unidades da camada de saída e das camadas intermediárias. Detalhes deste algoritmo são comentados no capítulo 4;
- método do gradiente (GRAD) $\rightarrow$ o erro deve apresentar um mínimo em função do parâmetro que o causa. Dentre os métodos que utilizam diferenciação e busca, o método do gradiente é o mais simples para obtenção da direção d_i, pois utiliza apenas informações de primeira ordem. Na *i*-ésima iteração, a direção d_i é definida como a direção de módulo unitário de maior decrescimento da função *J.*

$$d = -\frac{\nabla J(\theta)}{\left\|\nabla J(\theta)\right\|}$$

A lei de ajuste do método do gradiente é dada por:

[9] No sub-item 3.2.1. exemplifica-se que a sumarização automática de texto refere-se a um problema de classificação e reconhecimento de padrões.

$$\theta_{i+1} = \theta i - \alpha_i \frac{\nabla J(\theta_i)}{\left\|\nabla J(\theta_i)\right\|}$$

onde:

$J \rightarrow$ é a superfície de erro do problema de aproximação;

$\theta \rightarrow$ é um vetor de parâmetros;

$\alpha \rightarrow$ é um escalar que define o passo de ajuste.

- gradiente conjugado (GC) $\rightarrow$ efetua a busca local diferenciando do Backpropagation pelo cálculo dos gradientes e conseqüentemente pelas correções de pesos. Os métodos do gradiente conjugado possuem sua estratégia baseada no modelo geral de otimização apresentado no algoritmo padrão e do gradiente, mas escolhem a direção de busca d_i, o passo α_i e o coeficiente de momento b_i mais eficientemente utilizando informações de segunda ordem;
- Método das Secantes de um Passo - *One-step Secant* (OSS) $\rightarrow$ algoritmo de segunda ordem. Esse método mostrou ser eficiente para dados gerados, sendo usado para cálculo da média e da diferença da estimativa (BATTITI, 1992). O termo método de secante provém do fato de que as derivadas são aproximadas por secantes avaliadas em dois pontos da função (neste caso a função é o gradiente). Uma vantagem deste método apresentado é que sua complexidade é de ordem $O(P)$, ou seja, é linear em relação ao número P de parâmetros (DE CASTRO, 2003).

O treinamento de redes neurais com várias camadas pode ser entendido como um caso especial de aproximação de funções, para o qual não é levado em consideração nenhum modelo explícito dos dados (SHEPHERD, 1997). Não se deve esquecer que alguns dos métodos de treinamento de RNAs citados acima são conhecidos como “Métodos de Segunda Ordem” e são considerados a maneira

mais eficiente de se fazer o treinamento de redes neurais do tipo MLP (SHEPHERD, 1997). Estes métodos recorrem a um rigor matemático baseado em modelos de otimização não-linear irrestrita bem definidos, não apresentando assim um vínculo natural com a inspiração biológica inicialmente proposta para as RNAs (DE CASTRO, 2003).

A escolha do paradigma de aprendizado supervisionado para o desenvolvimento deste projeto foi influenciada pelo fato de se dispor de pares de entrada e saída para o treinamento da Rede Neural Artificial.

Neste trabalho dar-se-á foco ao algoritmo de treinamento padrão (Backpropagation) e ao algoritmo de treinamento considerado biologicamente mais plausível (GeneRec).

Sabe-se da existência de outros algoritmos considerados biologicamente mais plausíveis para o treinamento de RNAs, porém optou-se pelo GeneRec visto que foram encontradas em maior quantidade - embora em outras aplicações - comparações entre este algoritmo e o Backpropagation, que exibem um desempenho superior do GeneRec sobre o Backpropagation (O´REILLY, 1996; ROSA, 2002a; ROSA, 2002b)[10].

2.3.1. Um algoritmo de treinamento biologicamente implausível (Backpropagation)

Conceitos e Características

O algoritmo Backpropagation é muito utilizado nos dias atuais como um dos algoritmos conexionistas supervisionados mais eficientes computacionalmente.

[10] Detalhes sobre algumas implementações existentes que propõem comparações entre métodos de treinamento de RNAs podem ser encontrados no sub-item 3.1.

O Backpropagation reapareceu em 1986 (RUMELHART *et al.*, 1986) e apresenta um modelo matemático que "supera" a limitação imposta anteriormente por MINSKY & PAPERT (1969), para quem as RNAs não eram capazes de solucionar problemas que não fossem linearmente separáveis. Um exemplo clássico é a função ou-exclusivo (XOR).

Basicamente o procedimento do Backpropagation está descrito nas seguintes etapas:

Etapa 1: Um padrão é apresentado à camada de entrada da rede. Este padrão é propagado para a camada escondida. A figura 2.1 ilustra esta etapa.

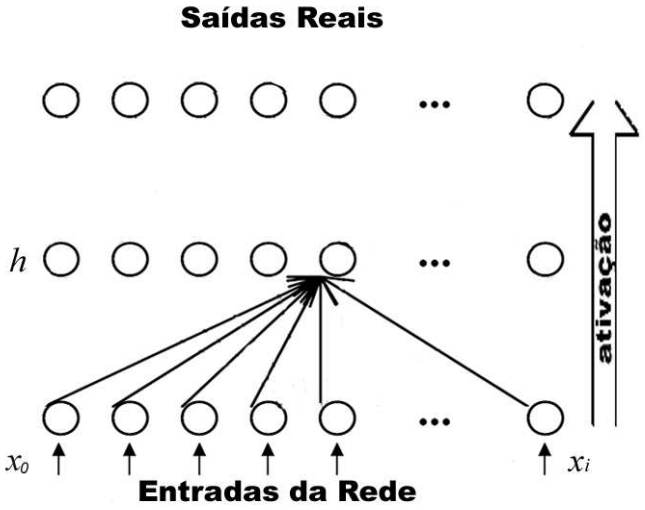


Figura 2.1 – Passo 1 da ativação do Backpropagation

A função de ativação σ é sigmoidal. A equação 1 mostra o cálculo da ativação escondida.

$$h_i = \sigma\left(\sum_{i=0}^{A} w_{ij}.x_i\right) \quad (1)$$

Este padrão é propagado camada por camada conforme ilustrado na figura 2.2.

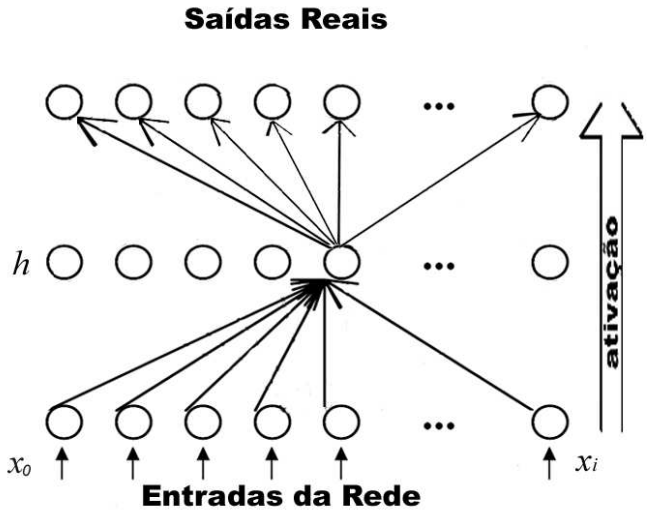


Figura 2.2 – Propagação do padrão entre as camadas na fase de ativação do Backpropagation

A figura 2.3 mostra a rede quando a camada de saída produz um resultado.

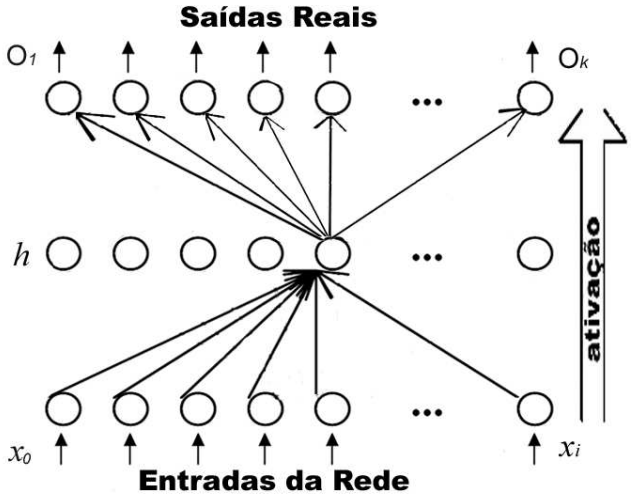


Figura 2.3 – Momento em que a RNA gera uma saída

A equação 2 mostra o cálculo da saída real o_k .

$$o_k = \sigma(\sum_{i=0}^{B} w_{jk}.h_i) \quad (2)$$

Etapa 2: A saída obtida é comparada à saída desejada para esse padrão particular. Se estas não forem iguais, o erro (diferença entre ambas) é calculado. O erro é retropropagado a partir da camada de saída, conforme ilustra a figura 2.4.

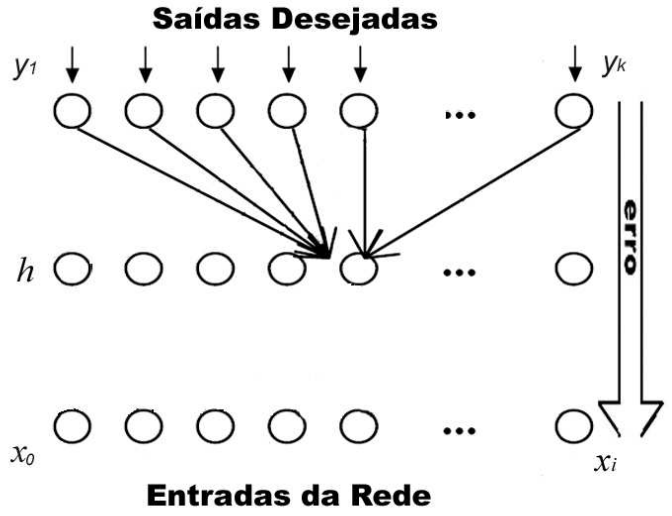


Figura 2.4 – Diferença entre saída desejada e a saída real sendo retropropagada da camada de saída para a camada escondida

A equação 3 mostra o cálculo do erro das unidades na camada de saída (representado por $\delta 2_j$) e a equação 4 mostra o cálculo do erro das unidades na camada escondida (representado por $\delta 1_j$).

$$\delta 2_j = o_j(1 - o_j)(y_j - o_j) \quad (3)$$

$$\delta 1_j = h_j(1 - h_j).soma3 \qquad (4)$$

onde

$$soma3 = \sum_{i=1}^{K} \delta 2_i . w_{jk}$$

Os pesos das conexões das unidades da camada de saída e das camadas escondidas vão sendo modificados conforme o erro é retropropagado (figura 2.5).

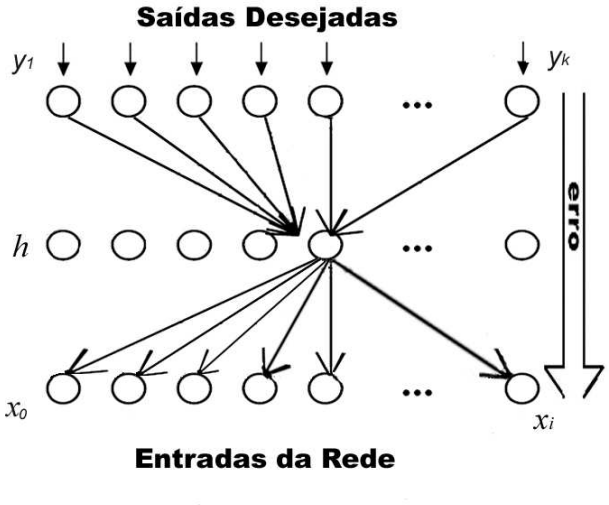


Figura 2.5 – Retropropagação do erro no Backpropagation

Para tornar o aprendizado possível, ocorre a atualização dos pesos sinápticos representados nas equações 5 e 6.

$$\Delta w_{jk} = \eta . \delta 2_j . h_j \quad (5)$$

$$\Delta w_{ij} = \eta.\delta 1_j.x_j \quad (6)$$

onde:

η => Taxa de Aprendizagem
x_i => entradas
o_k => saída atual
w_{ij} => pesos sinápticos (camada entrada e camada escondida)
w_{jk} => pesos sinápticos (camada escondida e camada saída)
y_k => saída desejada
h_j => ativação escondida

Por que o Backpropagation é considerado biologicamente implausível?

O algoritmo Backpropagation é considerado biologicamente implausível (CRICK, 1989; HAYKIN, 1999). As justificativas apresentadas para tal afirmação são:

- ⇒ Supondo que o Backpropagation ocorresse no cérebro, o erro seria propagado através do dendrito do neurônio pós-sináptico para o axônio e então para o dendrito do neurônio pré-sináptico. Isto é improvável (KANDEL *et al.*, 1995);
- ⇒ Pesquisas mostram que os pesos sinápticos são modificados para permitir o aprendizado, porém não desta maneira. Acredita-se que os pesos mudam usando apenas informação local da sinapse. Assim o Backpropagation mostra-se fisiologicamente implausível (O´REILLY, 1996).

2.3.2. Um algoritmo de treinamento biologicamente mais plausível (GeneRec)

Conceitos e Características

O algoritmo GeneRec foi desenvolvido em 1996 por O´REILLY respeitando propriedades de um algoritmo de treinamento de rede neural artificial biologicamente mais plausível.

É fundamental nos modelos biologicamente mais plausíveis que a representação adotada seja distribuída (ou seja, o conceito é representado através de várias unidades da arquitetura conexionista), pois acredita-se que o córtex cerebral utilize representação distribuída. Cada unidade da representação distribuída pode ser a representação de um "traço", com informação criada por combinações de algumas características. Inclui-se como benefício da representação distribuída a maior eficiência, robustez, precisão e a capacidade para representar similaridade entre relacionamentos. A maior eficiência proveniente da representação distribuída pode-se explicar quando são comparadas as unidades das redes com as letras de um alfabeto. Algumas combinações de poucas letras podem representar um grande número de palavras como pode ocorrer com diferentes combinações dos conjuntos de unidades de uma rede representando um grande número de informações. A robustez da representação distribuída tem origem da redundância de se ter cada item representado por muitas unidades. A precisão pode ser representada através de valores graduados, onde um valor é codificado pelas relativas magnitudes de várias unidades amplamente afinadas. Finalmente, a capacidade para representar similaridade entre relacionamentos é representada pelas unidades compartilhadas envolvidas nas representações distribuídas de itens diferentes (O'REILLY, 1998).

Embora o GeneRec ("*Generalized Recirculation*") tenha forte influência do algoritmo *Recirculation* (HINTON & MCCLELLAND, 1988), sabe-se que a derivação do GeneRec - exibido em O'REILLY, 1996 - depende de vários conceitos de algoritmos de aprendizagem, incluindo o Backpropagation (RUMELHART *et al.*, 1986), o algoritmo Almeida-Pineda para redes recorrentes e o algoritmo de aprendizado de *Hebbian* usado na máquina de *Boltzmann*.

O GeneRec faz uso de duas fases: "menos" e "mais" (figura 2.6.).

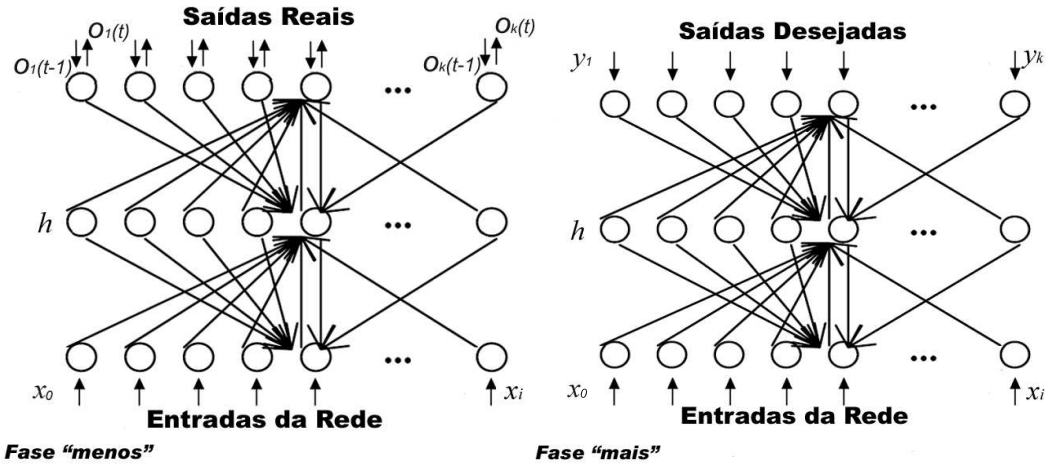


Figura 2.6 – Representação parcial das duas fases de GeneRec (O´REILLY, 1996; ROSA, 2002a)

Fase "menos"

Quando as entradas x_i são apresentadas à camada de entrada e ocorre a propagação desse estímulo para a camada escondida (propagação *bottom-up*) (figura 2.7).

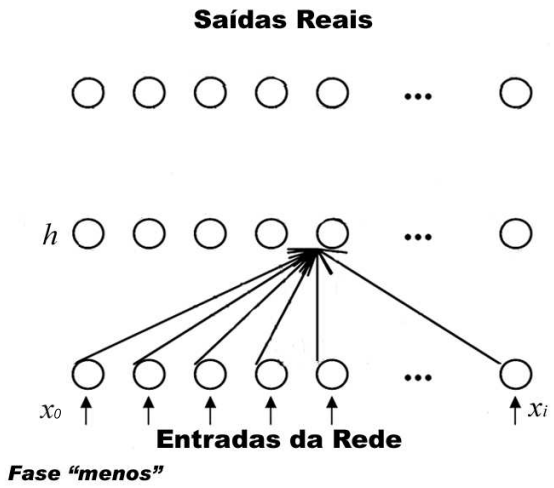


Figura 2.7 – Passo 1 da fase "menos" do GeneRec

Ocorre também a propagação da saída anterior o_k para a camada escondida (propagação *top-down*) (figura 2.8).

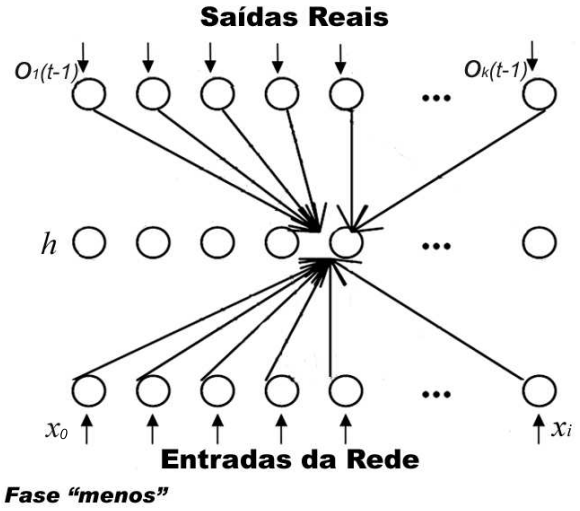


Figura 2.8 – Passo 2 da fase "menos" do GeneRec

Então a ativação escondida "menos" – fase menos – (h_j^-) é gerada (soma das propagações *bottom-up* e *top-down*). A função de ativação σ é sigmoidal. A equação 7 mostra o cálculo da ativação escondida.

$$h_j^-(t) = \sigma\left(\sum_{i=0}^{A} w_{ij}(t).x_i(t) + \sum_{k=1}^{C} w_{jk}(t).o_k(t-1)\right) \quad (7)$$

Finalmente, a saída real o_k no instante atual é gerada através da propagação da ativação da fase "menos" para a camada de saída (figura 2.9).

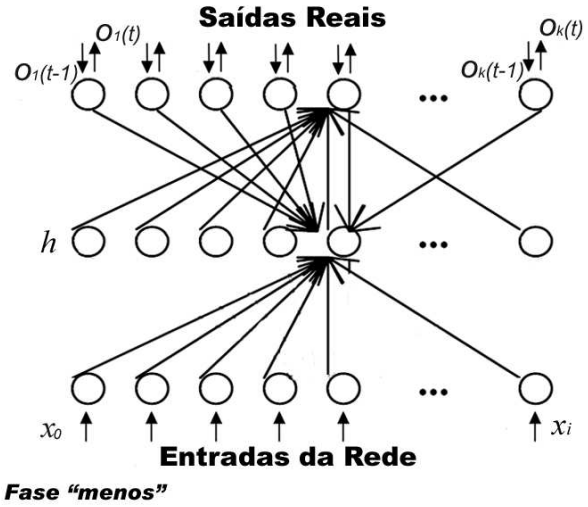


Figura 2.9 – Passo 3 da fase "menos" do GeneRec

A equação 8 mostra a saída real o_k sendo calculada (O´REILLY, 1996).

$$o_k(t) = \sigma\left(\sum_{j=1}^{B} w_{jk}(t).h_j^-(t)\right) \quad (8)$$

Nota-se que a arquitetura é bidirecional.

x_i => entradas

o_k => saída atual

w_{ij} => pesos sinápticos (camada entrada e camada escondida)

w_{jk} => pesos sinápticos (camada escondida e camada saída)

y_k => saída desejada

h_j^- => ativação escondida da fase menos

Fase "mais"

As entradas x_i são apresentadas novamente para a camada de entrada e então ocorre a propagação para a camada escondida (*bottom-up*) (figura 2.10.).

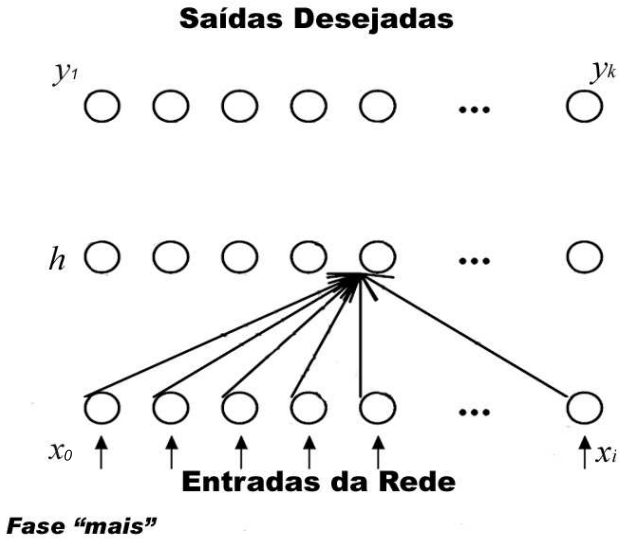


Figura 2.10 – Passo 1 da fase "mais" do GeneRec

Ocorre também a propagação da saída desejada y_k para a camada escondida (*top-down*) (figura 2.11).

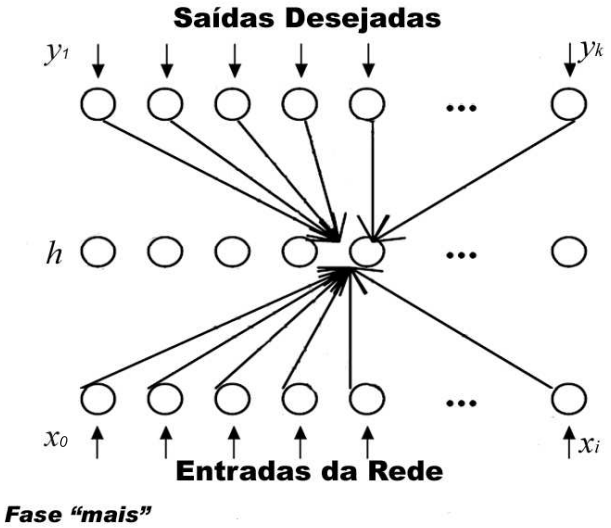


Figura 2.11 – Passo 2 da fase "mais" do GeneRec

Em seguida a ativação escondida “mais” é gerada, somando essas duas propagações (equação 9) (O´REILLY, 1996)

$$h_j^+(t) = \sigma(\sum_{i=0}^{A} w_{ij}(t).x_i(t) + \sum_{k=1}^{C} w_{jk}(t).y_k(t)) \quad (9)$$

x_i => entradas

w_{ij} => pesos sinápticos (camada entrada e camada escondida)

w_{jk} => pesos sinápticos (camada escondida e camada saída)

y_k => saída desejada

Para tornar o aprendizado possível, ocorre a atualização dos pesos sinápticos *w,* baseados em x_i, h^-_j, h^+_j, o_k e y_k (equações 10 e 11).

$$\Delta w_{jk}(t) = \eta.(y_k(t) - o_k(t)).h_j^-(t) \quad (10)$$

$$\Delta w_{ij}(t) = \eta.(h_j^+(t) - h_j^-(t)).x_i(t) \quad (11)$$

η => *Taxa de Aprendizagem*

Nota-se que somente após a atualização dos pesos (equações 10 e 11) obtém-se a arquitetura mostrada nas figura 2.6.

É importante lembrar que a arquitetura bidirecional em alguns casos – principalmente em problemas simples - pode fazer com que o GeneRec demore mais para convergir que o Backpropagation (“sobrecarga”) (O´REILLY, 1996). Porém em problemas cognitivos mais complexos, por exemplo, em Processamento de Línguas Naturais, isto não é verificado conforme constatado em ROSA (2002a).

Por que o GeneRec é considerado biologicamente mais plausível?

Conforme O´REILLY & MUNAKATA (2000), existem evidências de que o córtex cerebral seja conectado de forma bidirecional onde a representação distribuída predomina. Sendo assim, modelos conexionistas biologicamente mais plausíveis devem possuir algumas das seguintes características:

Representação distribuída: pode-se obter generalização e redução do tamanho da rede se a representação distribuída for adotada, visto que as conexões entre um conjunto de unidades são capazes de suportar um amplo número de diferentes padrões e criar novos conceitos sem alocação de novo hardware;

Competição inibitória: uma espécie de o vencedor ganha todos ("*winner-takes-all")*; aqueles neurônios que estão próximos do "vencedor" recebem um estímulo negativo destacando desta maneira o neurônio vencedor. Durante uma inibição lateral, um neurônio excita um interneurônio inibitório que faz uma conexão de realimentação com o primeiro neurônio (O´REILLY, 1998);

Propagação de ativação bidirecional: As camadas escondidas recebem estímulos das camadas de entrada e das camadas de saídas. A bi-direcionalidade da arquitetura é necessária para simular sinapses elétricas que ocorrem no cérebro e que podem ser bidirecionais (KANDEL *et al*., 1995);

Aprendizado de tarefa dirigido a erros: No GeneRec o erro é calculado por meio da diferença local nas sinapses, baseado em propriedades neurofisiológicas (O´REILLY, 1998), ao contrário do Backpropagation, que requer a propagação de sinais de erros.

Nota-se que o GeneRec possui características que o torna biologicamente mais plausível (sub-item 2.3.2). No próximo capítulo exibe-se os resultados de alguns trabalhos que propõem comparações entre algoritmos considerados

biologicamente implausíveis e algoritmos considerados biologicamente mais plausíveis.

CAPÍTULO 3

IMPLEMENTAÇÕES

Pode-se encontrar na literatura algumas citações e implementações que sugerem a comparação entre a eficiência de algoritmos conexionistas convencionais e algoritmos considerados biologicamente mais plausíveis para treinamento de RNAs (MAZZONI *et al.*,1991; SCHRAUDOLPH & SEJNOWSKI, 1996; ROSA, 2002a)

Porém, nos dias atuais, é escassa a aplicação de Sumarização Automática de Textos utilizando RNAs (ARETOULAKI, 1996; PARDO *et al.*, 2003b) e não se encontrou sumarizadores automáticos de textos que utilizam RNAs com treinamento considerado biologicamente mais plausível.

Considerando este fato, comenta-se na seqüência três comparações existentes entre métodos de treinamento em aplicações não relacionadas a sumarização automática de textos sendo que as duas primeiras comparações, que embora não utilizem o GeneRec, também utilizam algoritmos considerados biologicamente mais plausíveis e a terceira aplicação citada utiliza o algoritmo GeneRec para o treinamento da RNA.

3.1. Implementações Existentes

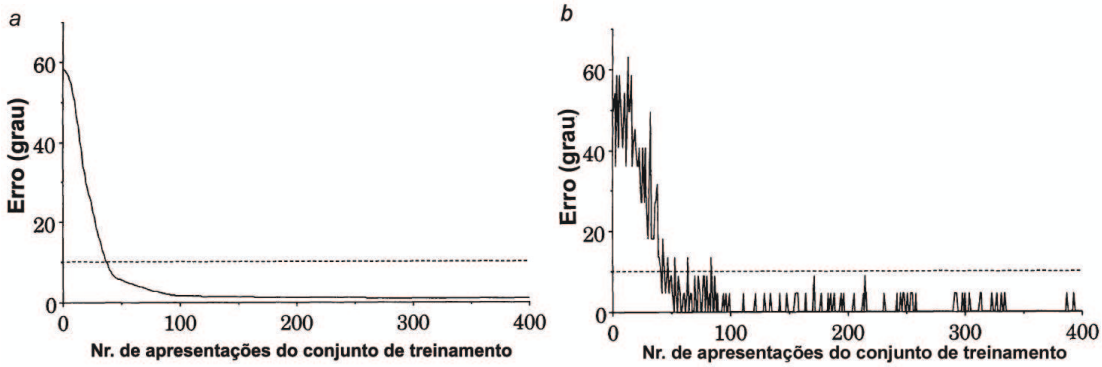
3.1.1. Variação do A_{r-p}

MAZZONI *et al.* (1991) propõem uma comparação (com aparente interesse em provar que há algoritmos com eficiência semelhante e, talvez, melhor que o tradicional algoritmo considerado biologicamente implausível) entre o Backpropagation e um algoritmo considerado biologicamente mais plausível, descrito como uma variação do A_{r-p} "*associative reward-penalty*". As principais alterações feitas são: a) todas as unidades da rede recebem um sinal de reforço, cujo valor depende da comparação entre a saída corrente da rede e a saída

desejada; b) a unidade usa apenas a informação local para realizar a sinapse ou o ajuste do peso da conexão.

O treinamento desta rede é similar ao da arquitetura do modelo de ZIPSER & ANDERSEN (1988), porém usando um algoritmo de treinamento considerado biologicamente mais plausível. Possui três camadas completamente conectadas com arquitetura *feed-forward*. A camada de entrada (sensorial) foi modelada de acordo com as características dos neurônios da área 7a[11] utilizada em estudos de ZIPSER & ANDERSEN (1988). As camadas escondidas e a camada de saída consistem de elementos estocásticos binários, que produzem a saída "1" com uma probabilidade dada pela função logística da soma dos pesos de entrada e uma saída "0" em outros casos.

Em um dos experimentos apresentados neste artigo, o autor exibe graficamente o comportamento da rede neural quando utilizado o $A_{r\text{-}p}$ e quando utilizado o Backpropagation para o treinamento. Na figura 3.1.a mostra-se a curva do aprendizado com o uso do Backpropagation, enquanto em 3.1.b mostra-se a curva do aprendizado com o uso do $A_{r\text{-}p}$.



(Backpropagation) ($A_{r\text{-}p}$)

Figura 3.1 – Curvas de aprendizado dos algoritmos Backpropagation e $A_{r\text{-}p}$ (MAZZONI *et al.*, 1991)

O método utilizado para o treinamento da RNA com o $A_{r\text{-}p}$ atualiza os pesos das conexões entre os neurônios usando apenas as informações locais às

[5] A área "7a" possui poderosas conexões da região cortical com áreas visuais.

sinapses tornando a busca pela solução mais randômica do que com o Backpropagation (MAZZONI *et al.*, 1991).

Conforme comentado, a intenção do autor parece ser provar a existência de algoritmos biologicamente mais plausíveis com grau de satisfação semelhante ao do Backpropagation. A conclusão do autor foi que, no mínimo, há uma igualdade na eficiência, com um pequeno ganho quando usado um algoritmo considerado biologicamente mais plausível para treinamento de redes neurais artificiais, como pode ser mostrado na figura 3.2.

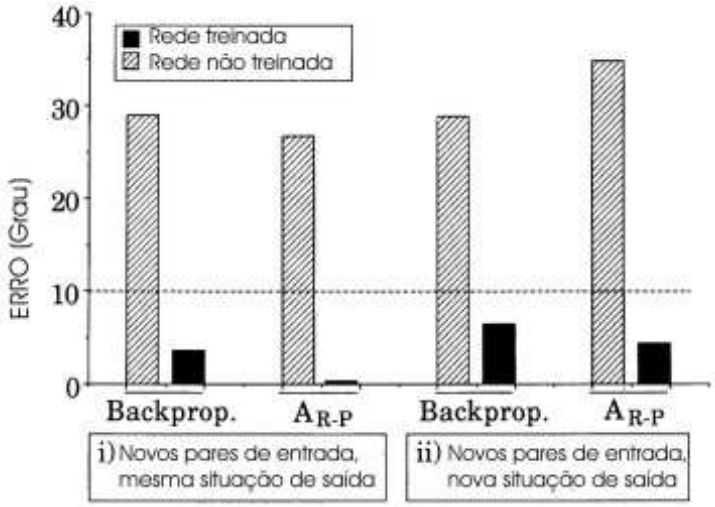


Figura 3.2. – Comparação entre o erro apresentado com o treinamento através do algoritmo Backpropagation e com o algoritmo $A_{r\text{-}p}$ (MAZZONI *et al.*, 1991)

A figura 3.2. mostra o erro obtido com o uso do Backpropagation e com o uso do algoritmo $A_{r\text{-}p}$ antes e depois de treinar com 40 novos pares de entrada: quando apresentadas (i) 40 novas entradas randômicas com a mesma situação de saída do conjunto de treinamento, e (ii) 40 novas entradas randômicas com nova situação de saída.

É possível notar nas duas situações representadas na figura 3.2., que quando há treinamento da rede, o A_{r-p} apresenta menor erro além de possuir tendência a convergência mais rápida (quando comparado ao Backpropagation).

O autor conclui que o algoritmo A_{r-p} conseguiu desempenhar o ajuste da rede para a precisão desejada (erro mínimo de 10^{-2}).

3.1.2. "Temperando" o Backpropagation

SCHRAUDOLPH & SEJNOWSKI (1996) elaboraram algumas variações no algoritmo Backpropagation que tendem a uma aproximação a um modelo biologicamente mais plausível. Os autores concluem que os resultados obtidos nas suas implementações, que propõem comparar a eficiência computacional entre um algoritmo de treinamento considerado biologicamente mais plausível com o Backpropagation, são satisfatórios.

O termo "temperando" o Backpropagation utilizado pelos autores é baseado em modelar as atividades e sinais de erros do Backpropagation como variáveis randômicas independentes.

Os experimentos dos autores são aplicados ao "*batch learning & momentum*" e à regra delta-bar-delta (JACOBS, 1988). Os resultados são apresentados na tabela 3.1.

Algoritmo	**batch & momentum**			**delta-bar-delta**		
	η	*Média (épocas)*	*Desvio Padrão*	η	*Média (épocas)*	*Desvio Padrão*
convencional	$3 * 10^{-3}$	2438	1153	$3 * 10^{-4}$	696	218
com "tempero"	$1 * 10^{-2}$	339	95	$3 * 10^{-2}$	89,6	11,8

η -> taxa de aprendizagem

Tabela 3.1. – Comparação entre o algoritmo conexionista Backpropagation, considerado biologicamente implausível e um algoritmo considerado biologicamente mais plausível (Backpropagation com "tempero") (SCHRAUDOLPH & SEJNOWSKI, 1996)

É possível notar uma eficiência maior no algoritmo modificado ("com tempero"), considerando que a tabela 3.1 exibe o número de épocas necessárias para a convergência. Fica constatado que além do número de épocas e do desvio padrão serem reduzidos - quando comparados ao Backpropagation - também a melhor taxa de aprendizagem é maior no algoritmo modificado

3.1.3. Bio-Pred

ROSA (2002a) elaborou um sistema denominado Bio-Pred, que utiliza o algoritmo de treinamento GeneRec, para previsão da próxima palavra em uma sentença. Abaixo algumas características e comparações extraídas deste artigo:

-Bio-Pred1 -> GeneRec com 24.000 ciclos de treinamento
-Bio-Pred2 -> GeneRec com 4.057 ciclos de treinamento
-Pred-DR -> Backpropagation com 24.000 ciclos de treinamento

Para o treinamento foram apresentadas cem (100) sentenças inteiras diferentes entre si.

Parâmetros utilizados:

Entradas: Palavras de uma sentença (uma de cada vez)
Saídas...: A próxima palavra
Taxa Aprendizagem (η) -> 0,25
Erro máximo aceitável (e) -> 0,02

Caso 1: *Saída em relação à sentença "The wolf frightened the girl"*: Quando informada a palavra *wolf* os índices de acerto para a previsão de *frightened* como próxima palavra são:

Bio Pred1 => 81,3%
Bio-Pred2 => 62,8%

Pred-DR => 78,0%

Quando inserida a segunda palavra (*frightened*), os resultados obtidos são (previsão de que a próxima palavra será *girl*):

Bio Pred1 => 82,6%

Bio-Pred2 => 81,2%

Pred-DR => 76,7%

Quando inserida a palavra *girl* todas as versões retornaram com 100% de exatidão a informação de "fim da sentença". Veja a figura 3.3.

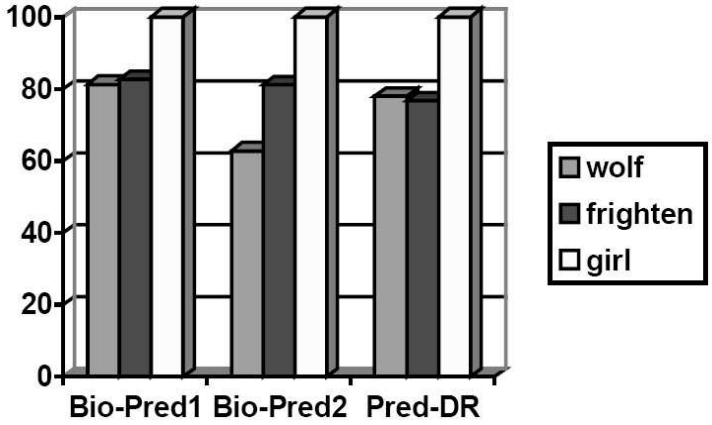


Figura 3.3. - Comparação entre os sistemas Bio-Pred1, Bio-Pred2 e Pred-DR para a sentença de entrada "*The wolf frightened the girl*" (ROSA, 2002a)

Caso 2.: *Saída em relação a sentença "The stone broke the vase"*: Quando informada a palavra *stone* os resultados obtidos são (previsão de que a próxima palavra será *break*):

Bio Pred1 => 66,1%

Bio-Pred2 => 74,5%

Pred-DR => 84,9%

Neste caso, a eficiência menor no algoritmo usando GeneRec deve-se ao fato de que a palavra *break* não é facilmente processada pois admite um, dois ou três operandos (ROSA, 2002a).

Quando informada a segunda palavra (*break*) os resultados obtidos são (previsão de que a próxima palavra será *vase*):

Bio Pred1 => 91,7%

Bio-Pred2 => 44,8%

Pred-DR => 67,9%

A predição do final da sentença não foi dada com 100% em nenhum dos casos, visto que "*break*" permite diferentes números de operandos. Por exemplo, o verbo "quebrar" permite: "Quem quebrou o que" ou "Quem quebrou o que com o que".

Porém, quando informada a última palavra (*vase*) os resultados obtidos quanto a previsão do final da sentença são:

Bio Pred1 => 66,2%

Bio-Pred2 => 71,5%

Pred-DR => 83,8%

Veja a figura 3.4.:

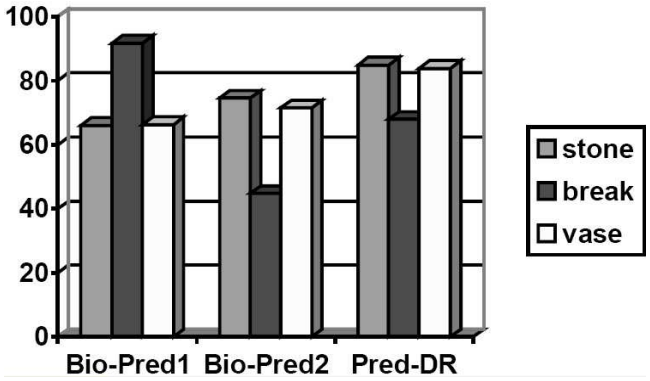


Figura 3.4. - Comparação entre os sistemas Bio-Pred1, Bio-Pred2 e Pred-DR, para a sentença de entrada "*The stone broke the vase*" (ROSA, 2002a)

3.2. O sistema proposto SABIO

Existem vários trabalhos realizados que propõem sumarização automática de texto utilizando *corpora* já disponíveis. Empregam desde abordagens simbólicas até abordagens utilizando Redes Neurais Artificiais. Encontraram-se implementações que utilizam Redes Neurais Artificiais com treinamento biologicamente implausível, diferentemente do que se propõe neste trabalho.

Neste trabalho propõe-se a construção do SABIO, onde "SA" se refere à *S*umarização *A*utomática e "Bio" faz referência ao tratamento conexionista considerado *bio*logicamente mais plausível. Este sistema produz extratos, utilizando Redes Neurais Artificiais com treinamento considerado biologicamente mais plausível.

Na seqüência comenta-se as características do SABIO.

3.2.1. Características do SABIO

A rede neural do SABIO foi treinada com sentenças de um corpus denominado "TeMário" que contém 100 textos jornalísticos, totalizando 61.412 palavras. O número médio de palavras por texto é de 613. 60 textos constam do jornal *on-line* Folha de São Paulo e estão distribuídos igualmente nas seções Especial, Mundo e Opinião; os 40 textos restantes foram publicados no Jornal do Brasil também *on-line*, e estão também uniformemente distribuídos nas seções Internacional e Política (PARDO & RINO, 2003a).

O corpus TeMário é composto por textos-fonte, sumários autênticos e extratos ideais. Sumários autênticos são produzidos pelos próprios autores dos textos-fonte, sendo resultantes de um processo de reescrita do conteúdo que o escritor "julga" ser mais relevante. Raramente há correspondência explícita entre textos-fonte e os sumários autênticos.

Pela razão da não existência de correspondência explícita entre textos-fonte e os sumários autênticos e o fato da necessidade de humanos para tal tarefa torná-la custosa e demorada, utiliza-se no sistema proposto os chamados extratos[12] ideais. Extratos ideais costumam usar a medida do co-seno (SALTON, 1989): para cada sentença do sumário autêntico, procura-se a sentença correspondente no texto-fonte mais semelhante. Isto é feito pela co-ocorrência de palavras: quanto mais palavras do sumário autêntico uma sentença do texto-fonte tem, maior sua chance de apresentar o mesmo conteúdo da sentença do sumário, podendo, portanto, ser utilizada para compor o sumário ideal (PARDO & RINO, 2003a).

Apesar de utilizar apenas o processamento lexical e não outras etapas do processamento lingüístico tais como análises sintática e semântica para desenvolver um sumário, o SABIO produziu indícios de bons resultados.

Na escolha entre diversos corpora, optou-se pelo TeMário considerando a facilidade em realizar uma futura comparação com outros sumarizadores que também utilizam este corpus e também considerou-se que as conferências internacionais de avaliação de sumarizadores automáticos, como a SUMMAC (*text **SUMMA**rization evaluation **C**onference*) e a DUC (***D**ocument **U**nderstanding **C**onference*) têm utilizado textos jornalísticos.

Entretanto, algumas adaptações (citadas na seqüência) podem fazer o SABIO utilizar outros corpora para o treinamento da rede.

A restrição do domínio em corpora é importante visto que quanto mais abrangente for o aspecto da língua que se deseja investigar, maiores serão as dificuldades em se obter uma amostra que seja representativa. Ao contrário, se se

[12] Aqui “sumários ideais” e “extratos ideais” serão tratados sem distinção, embora haja diferenças entre os mesmos (ver capítulo 2). Os extratos ideais possuem, em média, 30% do tamanho do texto-fonte.

restringir a investigação, maiores serão as chances de se coletar uma amostra realmente representativa (FELTRIN *et al.,* 2001).

Na fase do treinamento, para cada sentença[13] dos textos-fonte, são analisados sete traços que representam as sentenças de entrada. Estas sentenças codificadas são associadas a uma saída desejada que é classificada conforme o grau de importância[14] da sentença no extrato-ideal, sendo os valores possíveis: "Nenhuma", "Pequena", "Razoavelmente Pequena", "Média", "Razoavelmente Grande" ou "Grande" freqüência. A classificação da freqüência das sentenças que representam as saídas desejadas é obtida através do método para descoberta da "*gist sentence*"[15]. O SABIO está programado para analisar os sete traços e associar as sentenças codificadas a uma saída desejada (fase de classificação de padrões na RNA) sem a necessidade de intervenção humana.

Após a fase do treinamento novos extratos poderão ser gerados através da indicação de textos-fonte a sumarizar. Tal processo denomina-se como "fase de realização", que visa analisar cada sentença do texto-fonte a sumarizar e associá-las à saída gerada pela RNA (fase de reconhecimento de padrões).

Para todas as sentenças é aplicado o "*case folding",* ou seja, as letras das palavras são transformadas em letras minúsculas, com o objetivo de uniformização (WITTEN *et al.,* 1994).

Os traços utilizados, suas descrições e valores possíveis são (PARDO *et al.*, 2003b):

[13] No SABIO, o final da sentença pode ser indicado através dos sinais de pontuação tradicionais: ponto final, ponto de exclamação e ponto de interrogação.
[14] Acredita-se que as sentenças que possuem termos mais freqüentes nos textos-fonte possam apresentar uma importância maior no texto. Outros sumarizadores (citados neste trabalho) também utilizam métodos estatísticos e cálculo de freqüência de termos na tentativa de encontrar grau de importância das sentenças.
[15] Para saber qual é a "*gist sentence*" utiliza-se abordagem mencionada em GistSumm (PARDO *et al.*, 2002).

1. *Tamanho da sentença*: sentenças longas normalmente apresentam maior conteúdo informativo, sendo, portanto, relevantes para o texto (KUPIEC *et al.*, 1995); Para calcular o tamanho da sentença é considerada a quantidade das palavras que pertencem a mesma, exceto as palavras contidas no "*StopList*"[16]. As sentenças do texto estão classificadas em uma escala de quatro patamares, conforme exibe a figura 3.5: a variável "Pequena" representa o tamanho da menor sentença do texto; a variável "Média" representa a média do tamanho das sentenças do texto; a variável "Grande" representa o tamanho da maior sentença do texto;

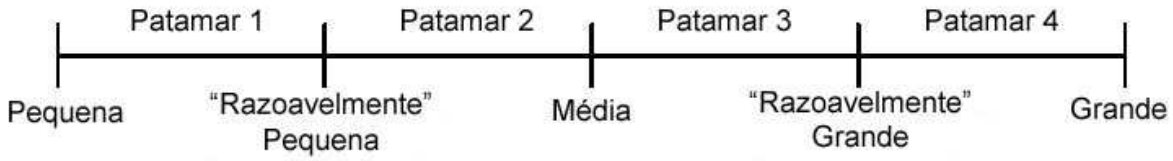


Figura 3.5. – Escala utilizada (traços 1, 5 e 7) para classificação de sentenças no SABIO.

2. *Posição da sentença no texto*: a posição da sentença pode indicar sua relevância (ARETOULAKI, 1996); No SABIO, semelhante ao NeuralSumm (PARDO *et al.*, 2003b), considera-se que uma sentença pode estar no "início" (primeiro parágrafo[17]), no "fim" (último parágrafo) ou no "meio" (parágrafos restantes) do texto. Caso haja apenas um parágrafo no texto será considerado que a(s) sentença(s) estão no "início";

3. *Posição da sentença no parágrafo a que pertence*: a posição da sentença no parágrafo também pode indicar sua relevância (BAXENDALE, 1958); No SABIO, considera-se que uma sentença pode estar no "início" (primeira sentença), no "fim" (última sentença) ou no "meio" (posições restantes) do parágrafo; Quando o parágrafo possui apenas uma sentença considera-se que esta está no "início";

[16] *StopList* é uma lista de palavras muito comuns ou que são consideradas irrelevantes para um texto por possuírem pouco valor semântico (JURAFSKY & MARTIN, 2000). O *stoplist* utilizado neste trabalho encontra-se no Apêndice A.
[17] No SABIO, o final do parágrafo é indicado pela tecla *"enter".*

4. *Presença de palavras da "gist sentence" na sentença*: sentenças que possuem palavras da *gist sentence*, isto é, a sentença que melhor expressa a idéia principal do texto, tendem a ser relevantes, pois contribuem para a idéia principal do texto (PARDO *et al.*, 2002);

5. *Pontuação da sentença com base na distribuição das palavras do texto*: sentenças com alta pontuação normalmente são relevantes para o texto (BLACK & JOHNSON, 1988). A pontuação de cada sentença é calculada através da soma do número de ocorrências de cada uma de suas palavras no texto todo dividido pelo número de palavras da sentença, sendo que o resultado obtido nesta operação será relacionado ao patamar mencionado no primeiro traço;

6. *TF-ISF da sentença*: sentenças com alto valor de TF-ISF (*Term Frequency – Inverse Sentence Frequency*) são sentenças representativas do texto (LAROCCA NETO *et al.*, 2000). Para cada palavra de uma sentença, a medida TF-ISF é calculada pela fórmula:

$$TF - ISF(w) = F(w) \times \frac{\log n}{S(w)}$$

Onde: F(w) é a freqüência da palavra w na sentença
n é o número de palavras da sentença a que w pertence
S(w) é o número de sentenças em que w aparece

O valor TF-ISF de uma sentença é a média dos valores TF-ISF de cada uma de suas palavras, sendo que o resultado obtido nesta operação será relacionado ao patamar mencionado no primeiro traço;

7. *Presença de palavras indicativas na sentença*: palavras indicativas (*cue words*) normalmente indicam a importância do conteúdo das sentenças (PAICE, 1981). Esse traço é o único dependente de língua e do gênero e domínio do texto. Devido ao SABIO estar adaptado para a língua portuguesa utiliza-se as mesmas

palavras indicativas do NeuralSumm (PARDO *et al.*, 2003b), também adaptado à língua portuguesa, que são: avaliação, conclusão, método, objetivo, problema, propósito, resultado, situação e solução.

Pode-se representar a matriz para treinamento da rede como mostrado na tabela 3.2.

Tabela 3.2-Representação da matriz dos traços para treinamento da rede utilizada pelo SABIO.

	Traço 1	Traço 2	Traço 3	Traço 4	Traço 5	Traço 6	Traço 7	Saída Desejada
VALORES POSSÍVEIS	Representados por cinco bits, sendo: Pequena (1 0 0 0 0) Razoavelmente pequena (0 1 0 0 0) Média (0 0 1 0 0) Razoavelmente grande (0 0 0 1 0) Grande (0 0 0 0 1)	Representados por três bits, sendo: Início (1 0 0) Meio (0 1 0) Fim (0 0 1)	Representados por três bits, sendo: Início (1 0 0) Meio (0 1 0) Fim (0 0 1)	Representados por dois bits, sendo: Contém (1 0) Não contém (0 1)	Representados por cinco bits, sendo: Pequena (1 0 0 0 0) Razoavelmente pequena (0 1 0 0 0) Média (0 0 1 0 0) Razoavelmente grande (0 0 0 1 0) Grande (0 0 0 0 1)	Representados por cinco bits, sendo: Pequena (1 0 0 0 0) Razoavelmente pequena (0 1 0 0 0) Média (0 0 1 0 0) Razoavelmente grande (0 0 0 1 0) Grande (0 0 0 0 1)	Representados por dois bits, sendo: Contém (1 0) Não contém (0 1)	***Freqüência da Sentença no Sumário:*** Nenhuma (1 0 0 0 0 0) Pequena (0 1 0 0 0 0) Razoavelmente pequena (0 0 1 0 0 0) Média (0 0 0 1 0 0) Razoavelmente grande (0 0 0 0 1 0) Grande (0 0 0 0 0 1)

A representação binária dos traços do conjunto de treinamento do SABIO é inspirado no modelo de MCCLELLAND & KAWAMOTO (1986). Unidades binárias mutuamente exclusivas na representação conexionista são uma característica importante que facilita a generalização na RNA.

Embora os traços do SABIO sejam semelhantes aos utilizados no NeuralSumm (PARDO *et al.*, 2003b), o SABIO difere em vários pontos, tais como:

SABIO	**NeuralSumm**
Utiliza Redes Neurais Artificiais com a arquitetura do perceptron de múltiplas camadas ("MLP");	Utiliza Redes Neurais Artificiais do tipo SOM (*self-organizing map*) (KOHONEN, 1982);
Utiliza um algoritmo de treinamento da rede supervisionado considerado biologicamente mais plausível;	Utiliza algoritmo de treinamento da rede não supervisionado;
Não é necessária a utilização de juízes lingüistas para que haja a indicação de quais são as sentenças "mais" importantes ou "menos" importantes do texto-fonte, eliminando esta tarefa exaustiva e demorada. Tal indicação é feita pelo SABIO utilizando os extratos ideais[18] por meio do método da "*gist, sentence*"	Para o treinamento utilizou-se 10 juízes lingüistas computacionais e falantes nativos do Português do Brasil. Estes classificaram manualmente as sentenças do texto-fonte conforme seu grau de importância, podendo ser: essencial, complementar ou supérflua;
Utiliza o corpus "TeMário" (PARDO & RINO, 2003a). Este contém 100 textos jornalísticos, totalizando 61.412 palavras.	Utiliza o corpus "CorpusDT" (FELTRIM *et al.*, 2001). Este contém 10 textos científicos, totalizando 530 palavras.

[18] Lembre-se que para gerar os extratos ideais não é necessária a utilização de juízes lingüistas. Ao invés disso, utiliza-se os sumários autênticos que foram gerados pelos próprios autores dos textos fontes e que já estão disponíveis no *corpus*.

Embora em menor quantidade, os traços utilizados no SABIO são parecidos com os traços utilizados em vários outros sumarizadores automáticos de textos (KUPIEC *et al*., 1995; LAROCCA NETO *et al*., 2002; TEUFEL & MOENS, 2002; PARDO *et al*., 2003b; MÓDOLO, 2003). Infelizmente há poucos indicativos sobre quais são os traços mais relevantes para a tarefa da sumarização (LAROCCA NETO *et al*., 2002).

Até o momento não se conhece nenhum sumarizador automático de textos que utilize RNAs com tratamento considerado biologicamente mais plausível. Por este motivo, na primeira parte deste capítulo, fez-se citações a aplicações em outras áreas que utilizam o treinamento biologicamente mais plausível para RNAs. Na segunda parte deste capítulo especificou-se as características do sistema elaborado: o SABIO. No próximo capítulo são exibidos resultados das várias comparações elaboradas pelo SABIO.

CAPÍTULO 4

RESULTADOS

Sugere-se neste capítulo a verificação da influência dos traços utilizados no SABIO em relação à Medida-F e algumas comparações entre o tradicional algoritmo de treinamento Backpropagation e o GeneRec, bem como comparações entre o SABIO e outros sumarizadores automáticos de textos.

Para conseguir comparações eficazes com trabalhos relacionados à SA utilizou-se taxas de compressão[19] de 70% e 80%, sendo que estas são as taxas mais freqüentes de uso, quando os sumários são gerados. Embora não haja um consenso sobre o tamanho ideal de um sumário, considera-se que os sumários devem manter entre 5% a 30% do conteúdo do texto-fonte (Mani, 2001). Sendo assim, a maioria das sentenças é desprezada nos sumários gerados. Sabendo de tal fato, fez-se um balanceamento[20] no conjunto de treinamento da RNA visto que haveria um número desproporcional de sentenças no patamar de freqüência/"importância" nenhuma[21], podendo causar dificuldades no aprendizado da RNA.

Faz-se necessário lembrar que foram feitos testes na ordem em que as sentenças foram apresentadas para o treinamento da RNA. Lembra-se também que foram feitos testes preliminares quanto aos valores iniciais dos pesos das conexões e, em geral, obteve-se o melhor desempenho quando estes foram iniciados aleatoriamente com valores entre -0.9 a +0.9. Os resultados que são

[19] Taxa de compressão corresponde à quantidade do texto-fonte que será reduzida. Por exemplo, com a taxa de compressão de 70% o sumário gerado terá 30% das sentenças do texto-fonte.
[20] Para efetuar o "balanceamento" no conjunto de treinamento todos os patamares da saída desejada foram nivelados para que tivessem a mesma quantidade de sentenças. Para obter o nivelamento excluiu-se aleatoriamente pares de entrada/saída desejada (quando necessário).
[21] Sentenças que estão no patamar "nenhuma importância" são as sentenças do texto-fonte que não aparecem no extrato ideal. Considerando a taxa de compressão utilizada, a maioria das sentenças está nesta classificação.

apresentados são da rede treinada com um conjunto de treinamento sem ordem pré-definida, ou seja, as sentenças estão “misturadas”.

4.1. Comparações entre a Medida-F na ausência de traços utilizado no SABIO

Neste primeiro item dos resultados procurou-se verificar a sensibilidade da RNA do SABIO em relação aos traços utilizados.

Para verificar a influência de cada um dos sete traços do SABIO elaborou-se um teste comparativo entre a rede com os sete traços (SABIO "original") e a rede com a ausência de um dos traços (exclui-se um traço de cada vez). Sendo assim, comparou-se a rede treinada originalmente com 7 traços com a rede treinada com 6 traços. A figura 4.1 exibe o resultado deste teste.

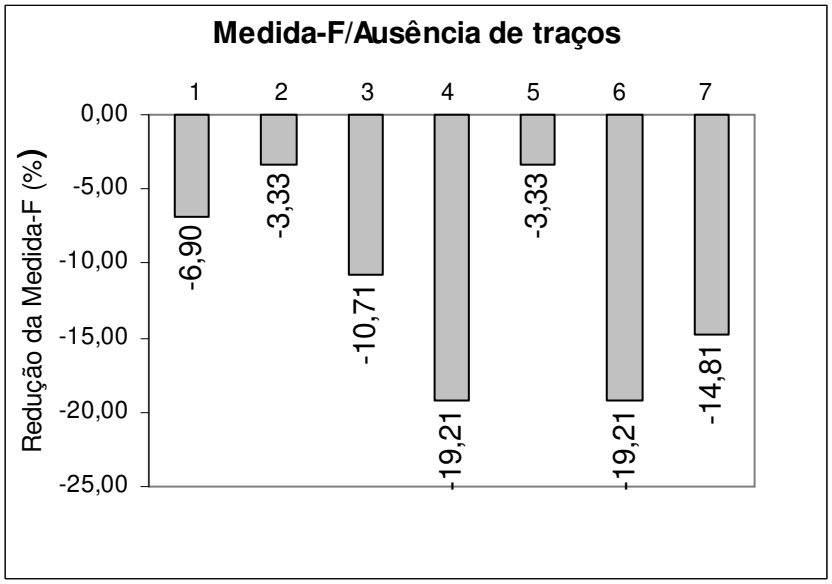


Figura 4.1. – Redução da Medida-F quando desconsiderado um traço da rede do SABIO.

Convém lembrar neste momento os traços utilizados pelo SABio:

1. Tamanho da sentença;
2. Posição da sentença no texto;
3. Posição da sentença no parágrafo a que pertence;
4. Presença de palavras da "gist sentence" na sentença;

5. Pontuação da sentença com base na distribuição das palavras do texto;

6. TF-ISF da sentença;

7. Presença de palavras indicativas na sentença.

Percebe-se que os traços relacionados à freqüência das palavras nas sentenças e nos textos (traços 4 e 6) são os traços que, quando ausentes, mais reduzem a Medida-F. Como em outros trabalhos relacionados à sumarização automática (LUHN, 1958; LAROCCA NETO *et al.*, 2000; PARDO *et al.*, 2002) acredita-se também que os termos mais freqüentes possuem realmente uma maior importância no texto.

Destaca-se também a influência do traço 7, o único dependente de língua, gênero e domínio do texto. Este traço motiva, para trabalhos futuros, o emprego de análises sintática e semântica dos textos a sumarizar.

4.2. Comparações entre o algoritmo de treinamento *Backpropagation* e o algoritmo de treinamento *GeneRec* para o sistema proposto (SABIO)

Neste item dos resultados procurou-se comparar o GeneRec com o Backpropagation na RNA do SABIO. Sendo assim, nesta fase não se compara os resultados do SABIO com outros trabalhos relacionados a SA.

Observou-se que quando a RNA do SABIO foi treinada através do GeneRec atingiu-se o erro mínimo[22] com uma quantidade menor de épocas quando comparado ao Backpropagation. Foram feitos testes com várias configurações na RNA, a saber:

Taxas de aprendizagem: 0.10, 0.25, 0.35 e 0.45;
Número de neurônios na camada escondida: 10, 11, 12, 20 e 25.

A figura 4.2. exibe a RNA do melhor resultado encontrado durante os testes[23] no Backpropagation e no Generec, nos quais a intenção foi comparar as quantidades de épocas e o tempo de processamento para convergência.

[22] A fórmula do erro é relacionada à derivada da função de ativação (sigmóide). Trata-se do erro quadrático médio. Para as comparações considerou-se que a rede convergiu quando o erro mínimo atingiu o valor de 10^{-3}.
[23] No apêndice deste trabalho encontram-se outras figuras com este experimento.

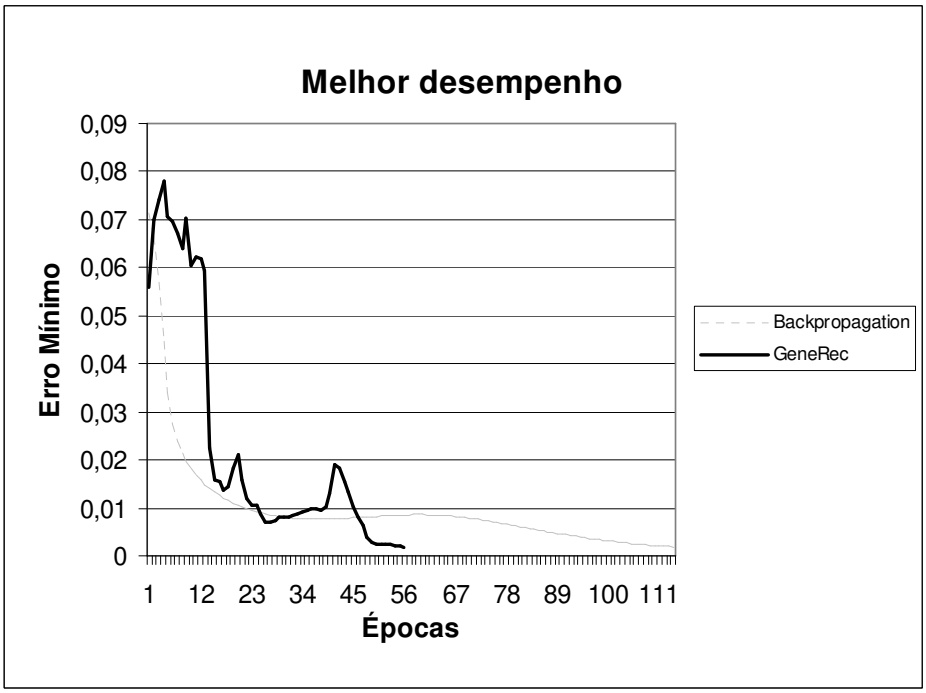


Figura 4.2. – Melhor performance encontrada nos testes efetuados para o SABIO.

As configurações utilizadas para obter o resultado exibido na figura 4.2 foram Backpropagation com 25 neurônios na camada intermediária e taxa de aprendizagem de 0.35 (convergiu em 115 épocas e 8 segundos); GeneRec com 11 neurônios na camada intermediária e taxa de aprendizagem de 0.45 (convergiu em 57 épocas e 3 segundos).

Embora os resultados exibidos na figura 4.2 mostram a comparação entre erros mínimos em relação a quantidade de épocas e tempo de processamento dos algoritmos GeneRec e Backpropagation, considera-se relevante comentar que a taxa de cobertura obtida através do treinamento realizado com o GeneRec foi superior à taxa de cobertura obtida através do treinamento realizado com o Backpropagation (a Medida-F foi respectivamente: 33,27 e 30,97).

Com a intenção de verificar o comportamento do erro para as próximas épocas estendeu-se a quantidade de épocas até a de número 250, considerando que a partir da época 227, para os dois algoritmos de treinamento, o erro

apresentava crescimento. A figura 4.3. mostra que o erro foi decrescente após atingir o erro mínimo estabelecido (10^{-3}) apenas para o treinamento efetuado por meio do algoritmo Backpropagation.

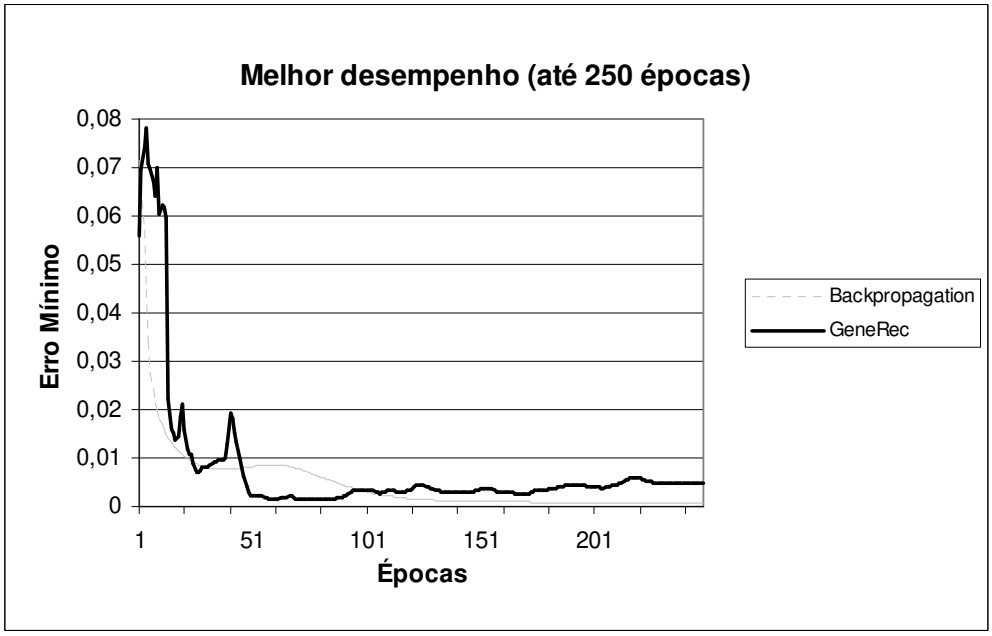


Figura 4.3 – Comportamento do erro mínimo para o melhor desempenho encontrado (quanto ao erro mínimo)

Procurou-se verificar a Medida-F na época 250 para os dois algoritmos e, embora o erro obtido apresentava-se menor com o treinamento por meio do algoritmo Backpropagation quando comparado ao GeneRec, a Medida-F foi maior quando a RNA foi treinada com o GeneRec (respectivamente, 32,89 e 32,12).

4.3. Comparações entre sumarizadores automáticos de textos

Neste item procura-se comparar o sistema SABIO com outros sumarizadores automáticos de textos. Avaliou-se também a taxa de cobertura e precisão dos sumários gerados através do SABIO. No sub-item 4.3.1 apresenta-se uma comparação entre o SABIO e mais cinco outros sumarizadores de textos utilizando o corpus TeMário. No sub-item 4.3.2 apresenta-se uma comparação

direta com o NeuralSumm (PARDO *et al.*, 2003b), uma vez que o SABIO possui traços semelhantes com o mesmo. No sub-item 4.3.3 apresenta-se a relação existente entre o conjunto de treinamento da RNA e o conjunto de teste no processo de sumarização automática de textos.

4.3.1. SABIO & Sumarizadores Automáticos de Textos

Para saber qual a melhor arquitetura que o SABIO poderia utilizar para atingir uma maior abrangência na cobertura e precisão foram feitos testes preliminares com várias[24] arquiteturas alterando quantidade de épocas, taxas de aprendizagem e quantidade de neurônios na camada intermediária[25]. A melhor arquitetura encontrada nos testes foi a utilizada para comparar o SABIO com outros sumarizadores automáticos de textos.

Para efetuar comparações entre o SABIO e outros sumarizadores automáticos foi usada a "Medida-F" (VAN RIJSBERGEN, 1979), visto que para obtê-la são consideradas as taxas de cobertura e precisão. Tal medida é freqüentemente usada quando se deseja comparar a eficiência de sumarizadores automáticos em relação ao sumário gerado (RINO *et al.,* 2004). Porém, convém lembrar que a Medida-F, além de não considerar a avaliação de aspectos semânticos, não considera o custo computacional / tempo para geração dos sumários.

Neste sub-item, para as comparações, procurou-se utilizar as mesmas condições de Rino *et al.*, 2004, ou seja:

a) taxa de compressão de 70%;

b) utilização do procedimento para validação "10-fold cross validation", ou seja: dividiu-se o TeMário em 10 grupos distintos de textos sendo que cada grupo

[24] Os testes preliminares foram feitos com: a) taxas de aprendizagem: 0.05, 0.15, 0.25, 0.35 e 0.45; b) Épocas: 2000 até 10000 (múltiplas de 2000); c) Neurônios na camada intermediária: 08, 16 e 22

[25] Para efetuar as comparações entre o tratamento considerado biologicamente mais plausível e o tratamento considerado biologicamente implausível utilizando a arquitetura do SABIO, fez-se adaptações no SABIO para que este fosse treinado também com o Backpropagation

possui dez textos diferentes entre si. Para o conjunto de treinamento utilizou-se 9 destes grupos formados e para o teste, um grupo. Desta maneira os textos utilizados para o teste não fazem parte do conjunto de treinamento. Sendo assim, efetuou-se 10 testes (um para cada grupo formado), medindo as taxas de cobertura, precisão e calculando a Medida-F. Após completar esta tarefa extraiu-se a média das taxas de cobertura, precisão e da Medida-F dos 10 experimentos efetuados. A tabela 4.1 ilustra o procedimento para validação.

Conjunto para treinamento	**Grupo para teste**
G01,G02,G03,G04,G05,G06,G07,G08 e G09	G10
G02,G03,G04,G05,G06,G07,G08,G09 e G10	G01
G03,G04,G05,G06,G07,G08,G09,G10 e G01	G02
G04,G05,G06,G07,G08,G09,G10,G01 e G02	G03
G05,G06,G07,G08,G09,G10,G01,G02 e G03	G04
G06,G07,G08,G09,G10,G01,G02,G03 e G04	G05
G07,G08,G09,G10,G01,G02,G03,G04 e G05	G06
G08,G09,G10,G01,G02,G03,G04,G05 e G06	G07
G09,G10,G01,G02,G03,G04,G05,G06 e G07	G08
G10,G01,G02,G03,G04,G05,G06,G07 e G08	G09

Tabela 4.1. – Procedimento para validação: validação cruzada em pacotes de 10 grupos, não enviesada ("*10-fold cross validation, non-biasing*").

Com os valores obtidos neste experimento foi possível comparar o SABIO com outros sumarizadores automáticos de textos que utilizaram exatamente o mesmo método e o mesmo corpus.

Conforme comentado anteriormente testou-se a variação dos resultados produzidos pelo SABIO quanto aos valores iniciais dos pesos das conexões (entre -0.1 a +0.1, -0.5 a +0.5 e -0.9 a +0.9). A figura 4.4. exibe a comparação entre a Medida-F.

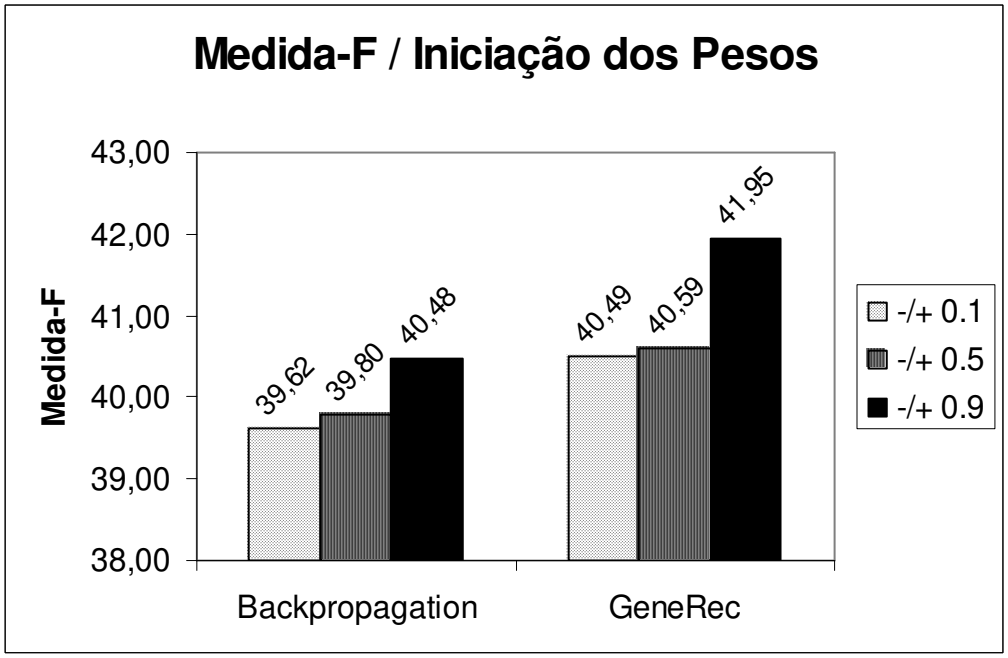


Figura 4.4. – Variação da Medida-F quanto aos valores iniciais dos pesos das conexões da RNA

A tabela 4.2 exibe o quadro comparativo entre os sumarizadores:

Performance dos sistemas (em %)				
Sumarizador	**Precisão**	**Cobertura**	**Medida-F**	**Diferença entre a Medida-F do SABIO (GR) (em %)**
Supor	44,9	40,8	42,8	1,90
ClassSumm	45,6	39,7	42,4	0,95
SABIO (GR)[26]	**43,8**	**40,3**	**42,0**	**-**
SABIO (BP)[27]	**42,4**	**38,7**	**40,5**	**-3,70**
From-Top	42,9	32,6	37,0	-13,51
TF-ISF-Summ	39,6	34,3	36,8	-14,13
GistSumm	49,9	25,6	33,8	-24,26
NeuralSumm	36,0	29,5	32,4	-29,63
Random order	34,0	28,5	31,0	-35,48

Tabela 4.2.–Comparação entre Sumarizadores Automáticos de Textos (adaptada de RINO *et al.*, 2004).

[26] SABIO treinado com o algoritmo GeneRec com 22 neurônios na camada escondida, taxa de aprendizagem de 0.25 e 4000 épocas (501 segundos).
[27] SABio treinado com o algoritmo Backpropagation com 16 neurônios na camada escondida, taxa de aprendizagem de 0.45 e 8000 épocas (896 segundos).

Considera-se que o resultado obtido pelo sistema SABIO foi satisfatório, visto que:

- O primeiro colocado, Supor – *Text **Su**mmarization in **Por**tuguese* (MÓDOLO, 2003) – obteve 1,90% de desempenho[28] superior ao SABIO, porém o Supor utiliza técnicas que tornam o custo computacional mais alto do que para o SABIO, tais como a utilização de cadeias lexicais e tesauro;
- O segundo colocado, ClassSumm – *Classification System* (LAROCCA *et al.*, 2002) – obteve 0,95% de desempenho superior ao SABIO e também utiliza técnicas que tornam o custo computacional mais alto do que para o SABIO, tais como análise semântica, similaridade da sentença com o título (os extratos ideais para o ClassSumm devem possuir títulos), análise de ocorrência de anáforas e etiquetagem.

Sendo assim, o SABIO aparece em terceiro colocado, tratando-se de uma boa classificação. Em quarto colocado aparece novamente o SABIO porém com adaptações para que a RNA fosse treinada com o algoritmo padrão Backpropagation. Acredita-se que desta maneira consegue-se efetuar um segundo teste comparativo entre os algoritmos de treinamento de RNA (considerando que o sub-item 4.1 descreve a primeira comparação neste trabalho: em relação ao erro mínimo e tempo de processamento para convergência).

Lembre-se novamente que o SABIO possui algumas semelhanças com o NeuralSumm (PARDO *et al.*, 2003b) quanto aos traços utilizados para formar o conjunto de treinamento (7 de 8 dos traços do NeuralSumm estão presentes no SABIO). Esta afirmação condiz com LAROCCA *et al.*, 2002 e PARDO *et al.*, 2003b, que afirmam que a busca por classificadores mais adequados para a Sumarização Automática de Textos é tão importante quanto a seleção dos traços que representam o problema em questão.

[28] Para medida de desempenho utilizou-se a “Medida-F”.

4.3.2. SABIO e NeuralSumm

Conforme descrito no sub-item 3.2.1, os traços utilizados pelo SABIO são semelhantes aos traços utilizados no NeuralSumm. Devido a esta característica, propõe-se neste sub-item uma comparação direta quanto às taxas de cobertura, precisão e Medida-F do NeuralSumm.

Para esta comparação ser realizada, o SABIO foi adaptado para utilizar o corpus "CorpusDT"[29] (FELTRIM *et al.*, 2001), visto que este é o corpus utilizado originalmente pelo NeuralSumm. A taxa de compressão foi de 80% (idêntica a usada no NeuralSumm). Testou-se várias[30] arquiteturas do SABIO. A figura 4.5. exibe o melhor resultado obtido do SABIO comparado ao NeuralSumm:

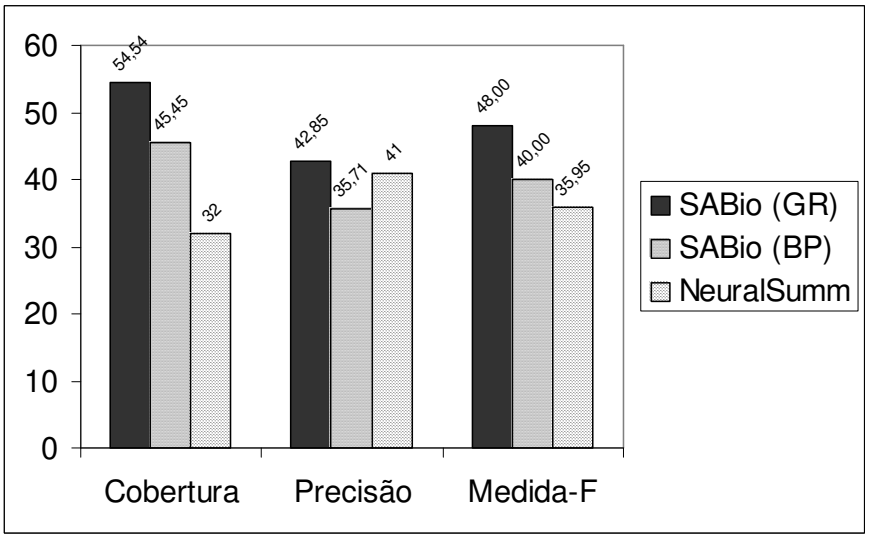


Figura 4.5. – Comparação das taxas de cobertura, precisão e Medida-F entre as duas versões do SABio (GR e BP) e o NeuralSumm. Considerando a Medida-F, o SABIO (GR)[31] treinado com o algoritmo GeneRec obteve a primeira colocação, seguido pelo SABIO (BP)[32] treinado com o algoritmo Backpropagation e, em terceira colocação, o NeuralSumm (48,00, 40,00 e 35.95 respectivamente).

[29] O CorpusDT contém 10 textos científicos, totalizando 530 palavras.
[30] Configurações: a) taxas de aprendizagem: 0.05, 0.15, 0.25, 0.35 e 0.45; b) Épocas: 2000 até 20000 (múltiplas de 2000); c) Neurônios na camada intermediária: 08, 16 e 22; d) Valores iniciais dos pesos das conexões da RNA entre: -0.1 a +0.1, -0.5 a +0.5 e -0.9 a +0.9.
[31] SABIO treinado com o algoritmo GeneRec com 08 neurônios na camada escondida, taxa de aprendizagem de 0.15 e 12000 épocas (79 segundos).
[32] SABIO treinado com o algoritmo Backpropagation com 16 neurônios na camada escondida, taxa de aprendizagem de 0.05 e 2000 épocas (12 segundos).

Diante dos resultados obtidos nesses experimentos, conclui-se que o SABIO obteve resultados satisfatórios quando utilizou outro corpus (Corpus-DT).

Embora o SABIO (GR) na melhor configuração encontrada (exibida na figura 4.5) utiliza uma quantidade maior de épocas quando comparado ao SABio (BP), constatou-se nos testes que o SABio (GR) obteve Medida-F equivalente ao SABio (BP) quando treinado com a mesma quantidade de épocas, ou seja, 2000 épocas.

A figura 4.6 sintetiza os resultados obtidos pelos sumarizadores automáticos de textos SABIO e o NeuralSumm:

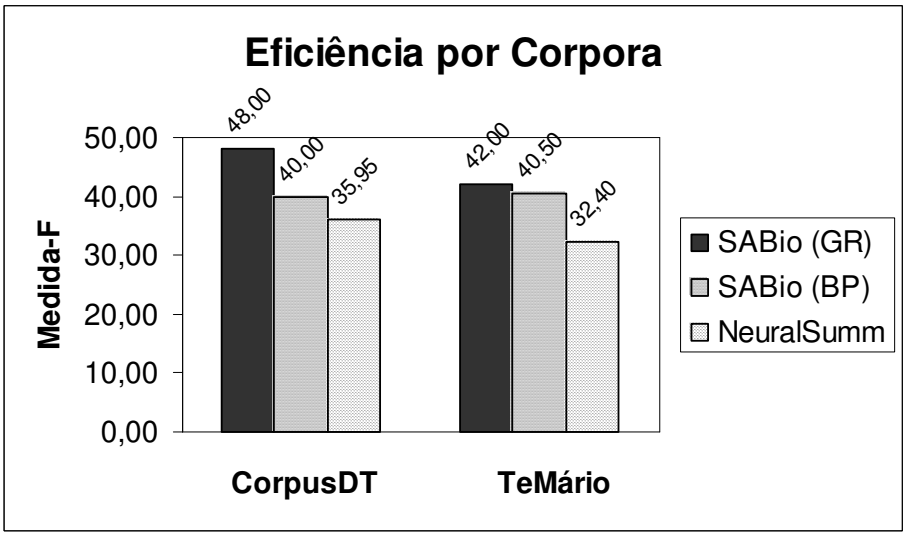


Figura 4.6 – Comparação Medida-F entre as duas versões do SABIO (GR e BP) e o NeuralSumm nos corpora CorpusDT e no TeMário.

4.3.3. SABIO (GR) e SABIO (BP) / Dependência do Corpus

Procura-se neste item avaliar a dependência do SABIO ao corpus utilizado para o treinamento da RNA. O teste foi feito na versão do SABIO com treinamento utilizando o algoritmo GeneRec.

Para efetuar este teste utilizou-se 2/3 dos textos do corpus TeMário para efetuar o treinamento da RNA e, na primeira fase, 10 textos do mesmo corpus

(diferentes dos textos utilizados no conjunto de treinamento) para teste. Na segunda fase deste teste utilizou-se 10 textos do CorpusDT para teste, ou seja, o treinamento da RNA foi efetuado com o corpus TeMário e o teste efetuado com o CorpusDT. Semelhantemente aos itens anteriores para saber qual a melhor arquitetura que o SABIO poderia utilizar para atingir uma maior abrangência na cobertura e precisão foram feitos testes preliminares com várias[33] arquiteturas alterando quantidade de épocas, taxas de aprendizagem e quantidade de neurônios na camada intermediária. A melhor arquitetura encontrada nos testes foi a utilizada para efetuar os testes. A figura 4.7 ilustra a Medida-F obtida neste experimento.

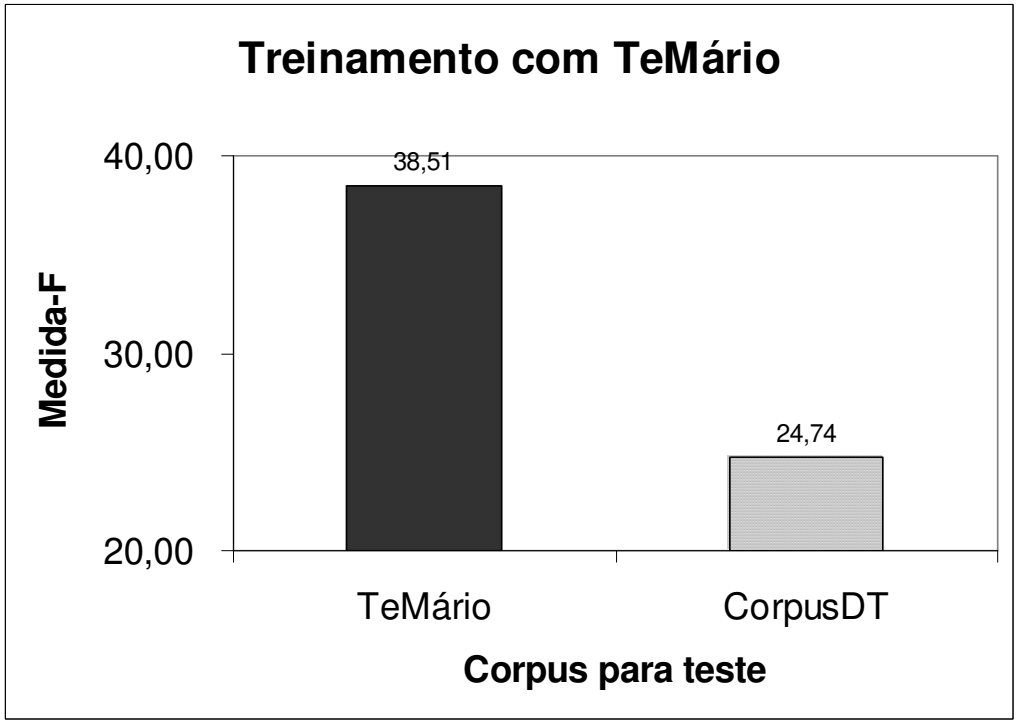


Figura 4.7 – Comparação da Medida-F para verificação da dependência do corpus utilizado para o treinamento da RNA. Taxa de compressão de 70%.

A figura 4.7. ilustra a dependência do SABIO quanto ao corpus (contexto) utilizado para o treinamento da RNA. Quando efetuados o treinamento e o teste com textos do corpus TeMário, a Medida-F apresenta valor de 38,51 e quando

[33] Os testes preliminares foram feitos com: a) taxas de aprendizagem: 0.05, 0.15, 0.25, 0.35 e 0.45; b) Épocas: 2000 até 10000 (múltiplas de 2000); c) Neurônios na camada intermediária: 08, 16 e 22

treinou-se o SABIO com textos do corpus TeMário e fez-se o teste com textos do corpus CorpusDT, a Medida-F reduziu para 24,74.

É possível observar por meio dos resultados exibidos neste capítulo que quando a RNA foi treinada com o algoritmo GeneRec para se obter a sumarização automática de textos ocorreram, em alguns casos, benefícios na velocidade de processamento (eficiência computacional) e na obtenção de maiores taxas de cobertura e precisão dos sumários gerados quando comparados ao Backpropagation.

CAPÍTULO 5

CONCLUSÃO

Após avaliar e comparar resultados obtidos através de sumarizadores automáticos de textos (citados no capítulo 4), notou-se a rápida e crescente evolução nas propostas de aplicações para construção destes sumarizadores. A crescente pesquisa neste tema apóia-se no interesse também crescente da sociedade moderna na busca de manchetes de jornais e/ou revistas que propõem trazer temas atuais resumidos ao invés de textos completos, principalmente pela falta de tempo das pessoas nos dias atuais.

Foi elaborado e implementado nesta dissertação um Sumarizador Automático de Textos: o SABIO. Este sistema utiliza Redes Neurais Artificiais em um modelo considerado biologicamente mais plausível. Considerando esta característica propõe-se uma comparação entre o algoritmo padrão de treinamento de RNA (Backpropagation) e um algoritmo considerado biologicamente mais plausível (GeneRec) na aplicação de sumarização automática de textos. Conforme exibido no capítulo 4, sabe-se de outros trabalhos que propõe a construção de Sumarizadores Automáticos de Textos e, alguns deles possuem valores de cobertura e precisão maiores que o SABIO, porém utilizam outras abordagens para o processo de sumarização, tais como:

1) Anotação semântica / Etiquetagem (FELLBAUM *et al.*, 2001);
2) Importância retórica (TEUFEL *&* MOENS., 2002);
3) Análises morfológica, semântica e pragmática (ARETOULAKI,1996);
4) Cadeias lexicais (SILBER *&* MCCOY., 2002);
5) Semântica ontológica (NIRENBURG & RASKIN, 2004).

Conforme constatado através das referências mencionadas, o uso destas abordagens podem colaborar para o aumento da cobertura e precisão dos sumários gerados, porém tais abordagens possuem custo computacional mais alto quando comparadas à abordagem utilizada no SABIO. Acredita-se que, em

trabalhos futuros, o uso de abordagens híbridas possam trazer ganhos substanciais no processo de sumarização.

É objetivo deste trabalho mostrar que se pode ter ganhos no processo de sumarização automática de textos quando utilizado o GeneRec, um modelo considerado biologicamente mais plausível, ao invés do Backpropagation, considerado um modelo biologicamente implausível, para o treinamento da RNA. Este trabalho não defende, em nenhum momento, o uso "isolado" da abordagem considerada biologicamente mais plausível como solução única para a sumarização automática de textos. Por este motivo acredita-se que o SABIO obteve ótimos resultados quando comparado com outros trabalhos que utilizam RNAs e resultados satisfatórios quando comparados com trabalhos que utilizam outras abordagens para a sumarização automática de textos.

Também não é intenção deste trabalho apontar fraquezas em sumarizadores automáticos existentes e nem efetuar comparações que possam afirmar que um sumarizador é "superior" ao outro. Para efetuar afirmações sobre qual é o "melhor" sumarizador seria necessário haver um consenso sobre quais características são as mais importantes para análise de um sumarizador automático. Entretanto, conforme mostrado no capítulo 4, encontra-se sumarizadores com taxas de cobertura e precisão maiores que o SABIO porém com eficiência computacional menor. Também é possível encontrar sumarizadores mais eficientes computacionalmente que o SABIO, como por exemplo o GistSumm (PARDO *et al.*, 2002), porém sua cobertura é inferior a do SABIO.

Acredita-se que para trabalhos futuros para a sumarização automática de textos pode-se considerar que:

- O modelo utilizado para o treinamento da RNA do SABIO pode trazer ganhos na eficiência computacional, taxas de cobertura, precisão e Medida-F;

- A investigação de outros traços podem ser incorporados ao treinamento da RNA trazendo ganhos nos resultados da aplicação;
- A quantificação da influência do classificador e da influência dos traços para o resultado produzido (extrato);
- A busca por outros métodos possibilitem (ou ao menos aproximem) de um treinamento universal da RNA, independentemente do tipo do corpus (jornalístico, científico, etc.);
- A lógica *fuzzy* pode ser considerada na codificação dos traços;
- A utilização, para a mesma aplicação, do uso do Backpropagation e do GeneRec simultaneamente;
- Avaliar as relações lexicais[34] e a qualidade semântica dos extratos gerados.

Acredita-se também que o tipo de abordagem utilizada neste trabalho poderá trazer ganhos em várias outras aplicações não relacionadas à sumarização automática. Tal esperança deve-se ao fato de se encontrar na literatura trabalhos que exibem comparações entre RNAs com algoritmos de treinamento biologicamente implausível e RNAs com algoritmos de treinamento biologicamente mais plausível (vide capítulo 3).

[34] Relações lexicais são as propriedades das palavras e dos conceitos que as ligam semanticamente (JURAFSKY *&* MARTIN, 2000)

APÊNDICES

Apêndice A - *Stoplist* utilizada no SABIO

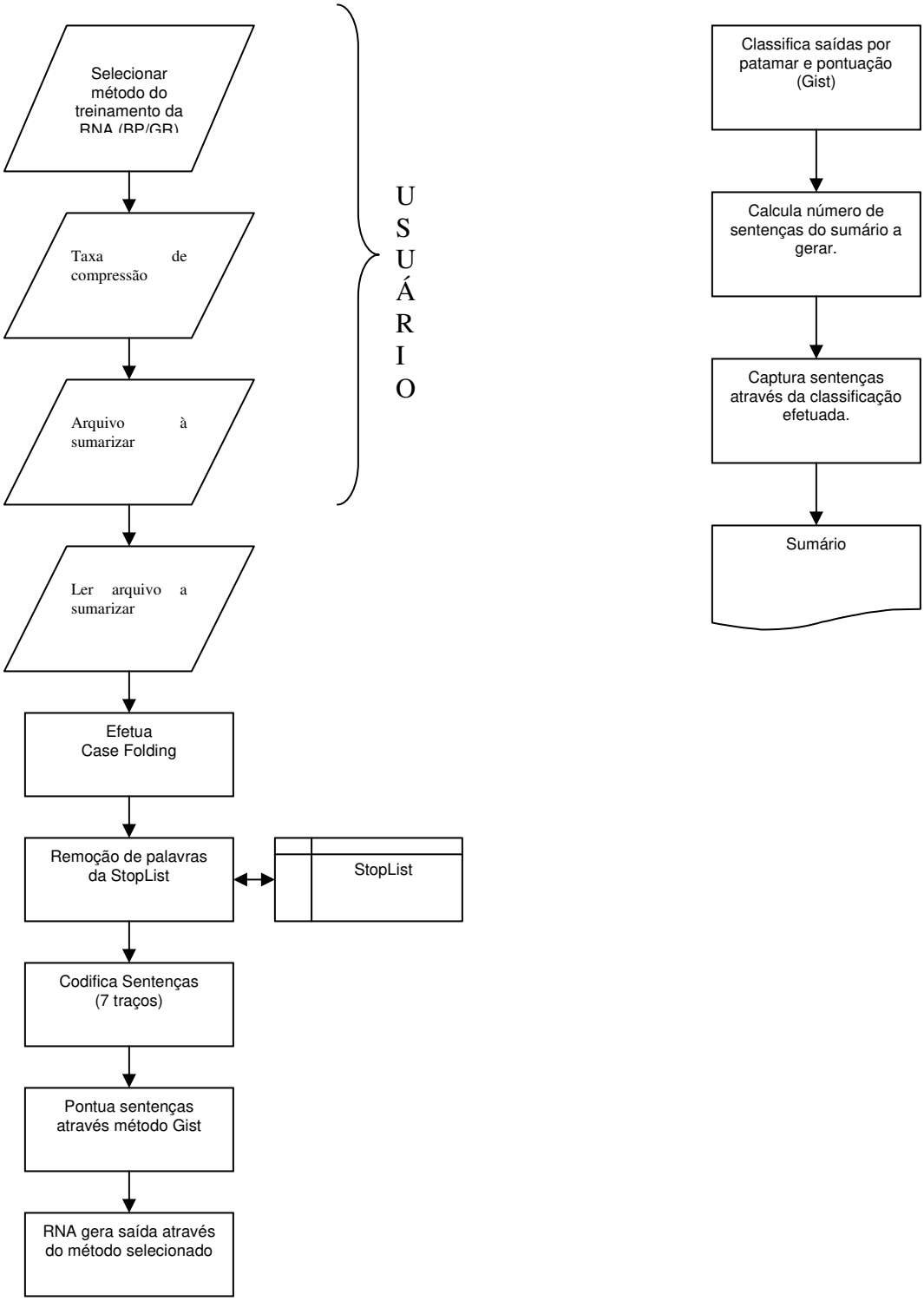
A tabela A.1. contém as palavras que compõe a *Stoplist* (CUNHA & CINTRA, 2001; PARDO *et al.*, 2002) que será usada no SABIO. Embora se tenha pesquisado outras *Stoplists*, optou-se por esta devido a mesma estar adaptada para textos jornalístico e científicos além de se considerar que possui uma quantidade razoável de palavras. Todas as palavras que constam desta *StopList* serão desconsideradas pelo SABIO. Todas as outras palavras terão um tratamento mais refinado, sendo consideradas palavras de maior importância dos textos.

A	À	Ah	Ai	Algo	Alguém	Algum	Alguma
Algumas	Alguns	Alô	Ambos	Ante	Ao	Após	Aquela
Aquelas	Aquele	Aqueles	Aquilo	As	Até	Bis	Cada
Certa	Certas	Certo	Certos	Chi	Com	Comigo	Conforme
Conosco	Consigo	Contigo	Contra	Convosco	Cuja	Cujas	Cujo
Cujos	Da	Das	De	Dela	Delas	Dele	Deles
Desde	Do	Dos	E	Eia	Ela	Elas	Ele
Eles	Em	Embora	Enquanto	entre	Essa	Essas	Esse
Esses	Esta	Estas	Este	Estes	Eu	Hem	Hum
Ih	Isso	Isto	Lhe	Lhes	Logo	Mas	Me
Mesmos	Meu	Meus	Mim	Minha	Minhas	Muita	Muitas
Muito	Muitos	Na	Nada	Nas	Nela	Nelas	Nele
neles	Nem	Nenhum	Nenhuma	Nenhumas	Nenhus	Ninguém	No
Nos	Nós	Nossa	Nossas	Nosso	Nossos	o	ó
ô	Oba	Oh	Olá	Onde	Opa	Ora	Os
ou	Outra	Outras	Outrem	Outro	Outros	Para	Per
perante	Pois	Por	Porém	Porque	Portanto	Pouca	Poucas
Pouco	Poucos	Próprios	Psit	Psiu	Quais	Quais	Quaisquer
Qual	Qualquer	Quando	Quanta	Quantas	Quanto	Quantos	Que
Quem	Se	Sem	Seu	Seus	Si	Sob	Sobre
Sua	Suas	Tanta	Tantas	Tanto	Tantos	Te	Teu
Teus	Ti	Toda	Todas	Todo	Todos	Trás	Tu
Tua	Tuas	Tudo	Ué	Uh	Ui	Um	Uma
Umas	Uns	Vária	Várias	Vário	Vários	Você	Vós
Vossa	Vossas	Vosso	Vossos				

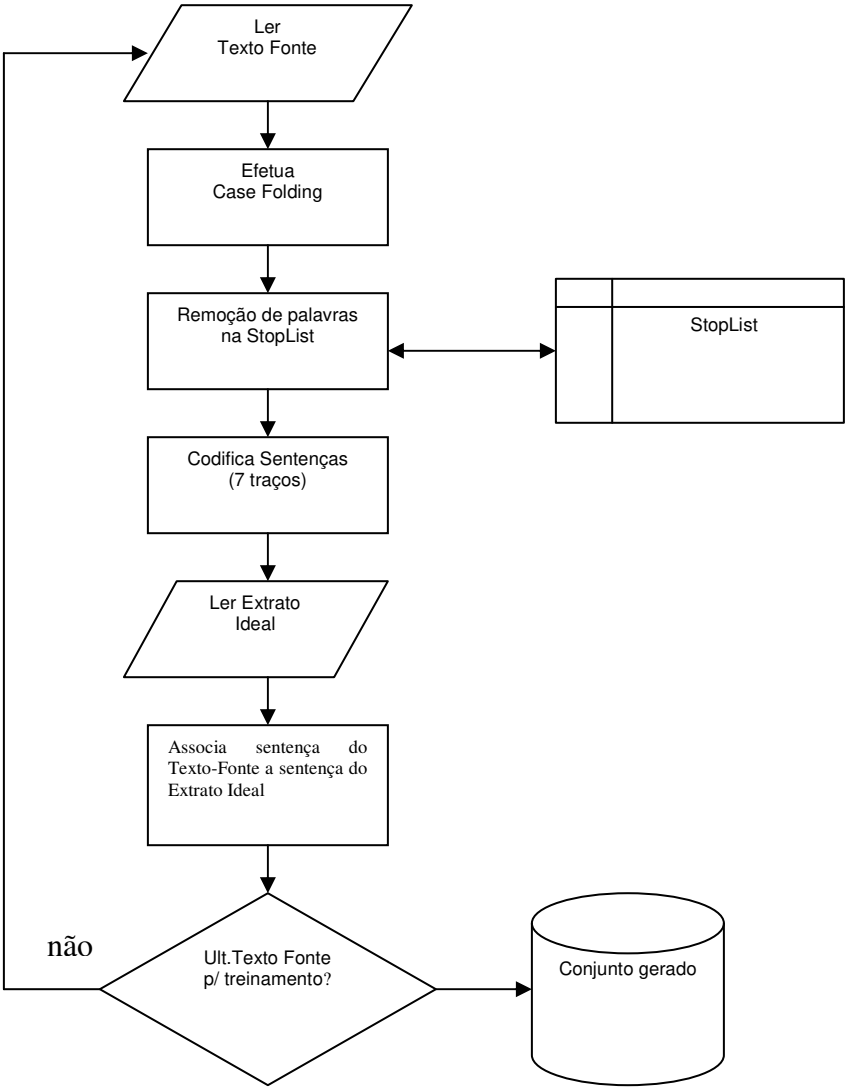
Tabela A.1. – Stoplist utilizada pelo SABIO

Apêndice B - Algoritmos

B.1. - Algoritmo sobre o processo de sumarização no SABIO

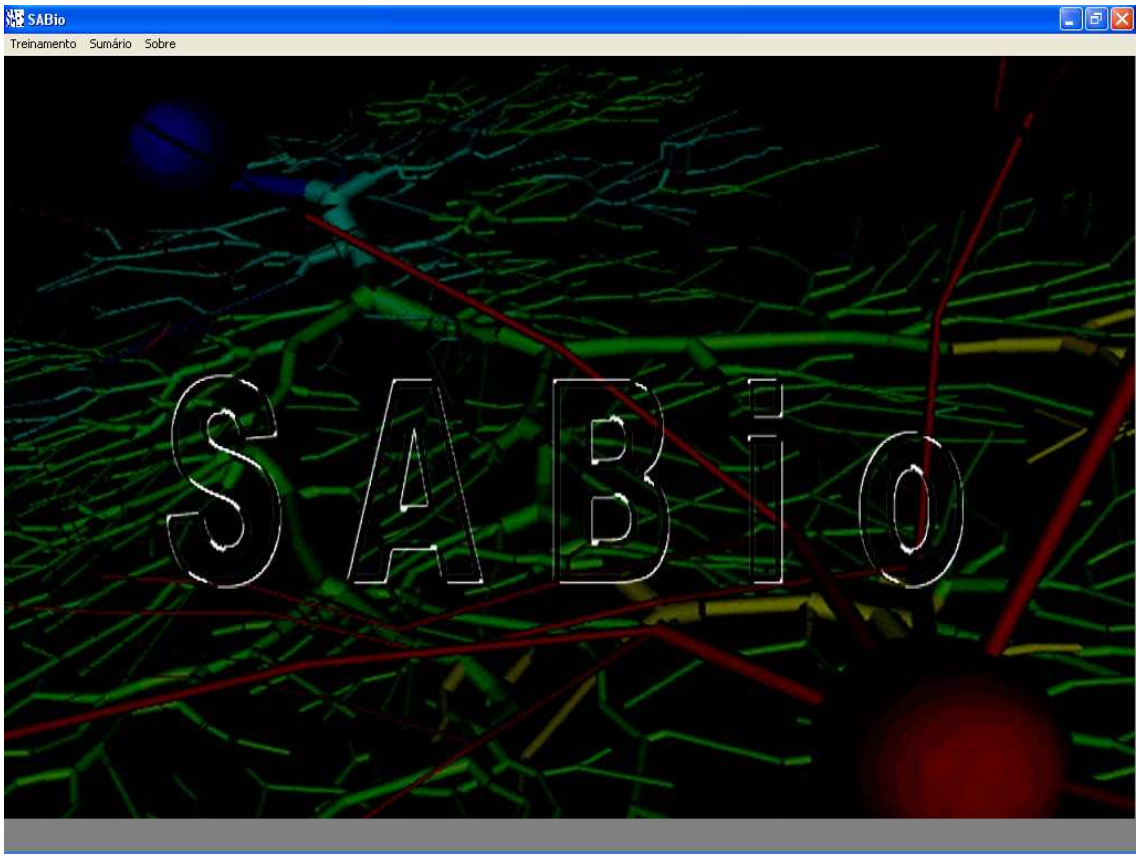


B.2. - Algoritmo para o conjunto treinamento

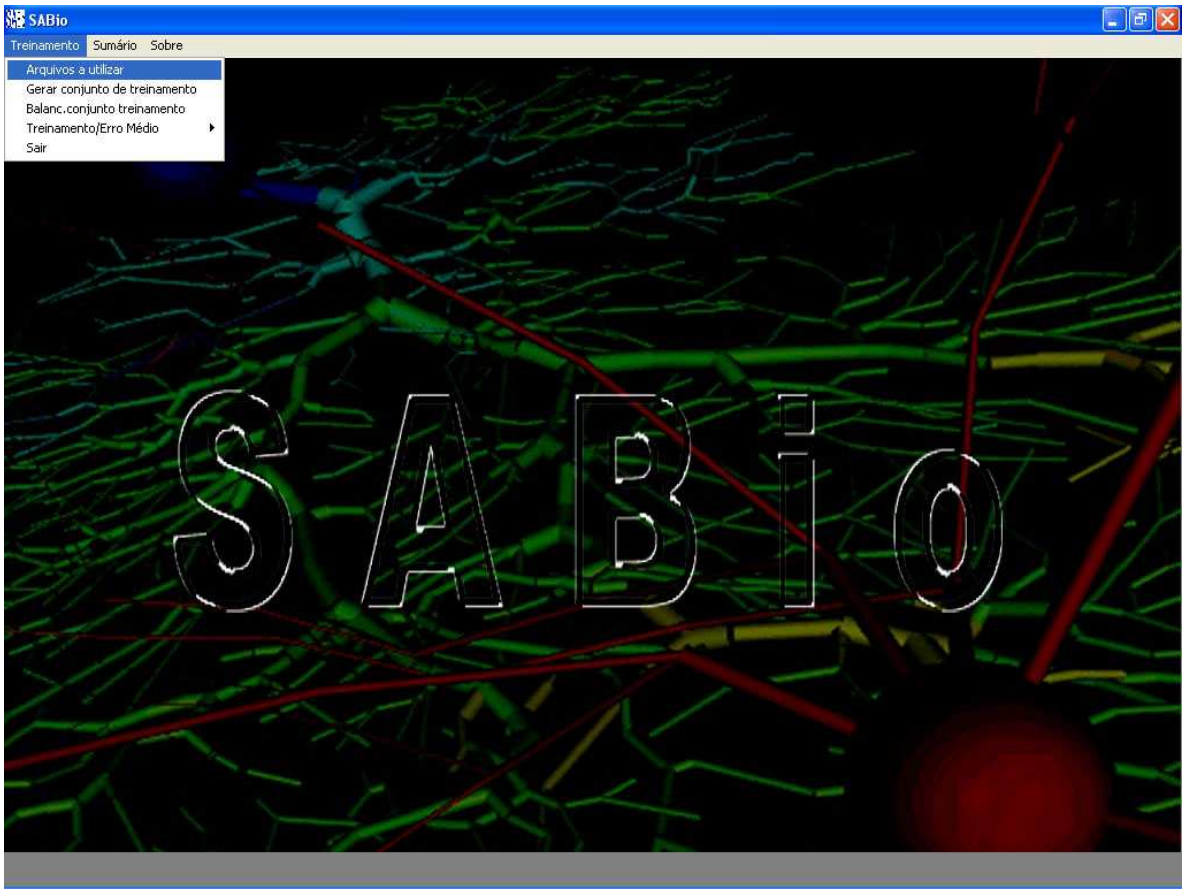


Apêndice C - Telas

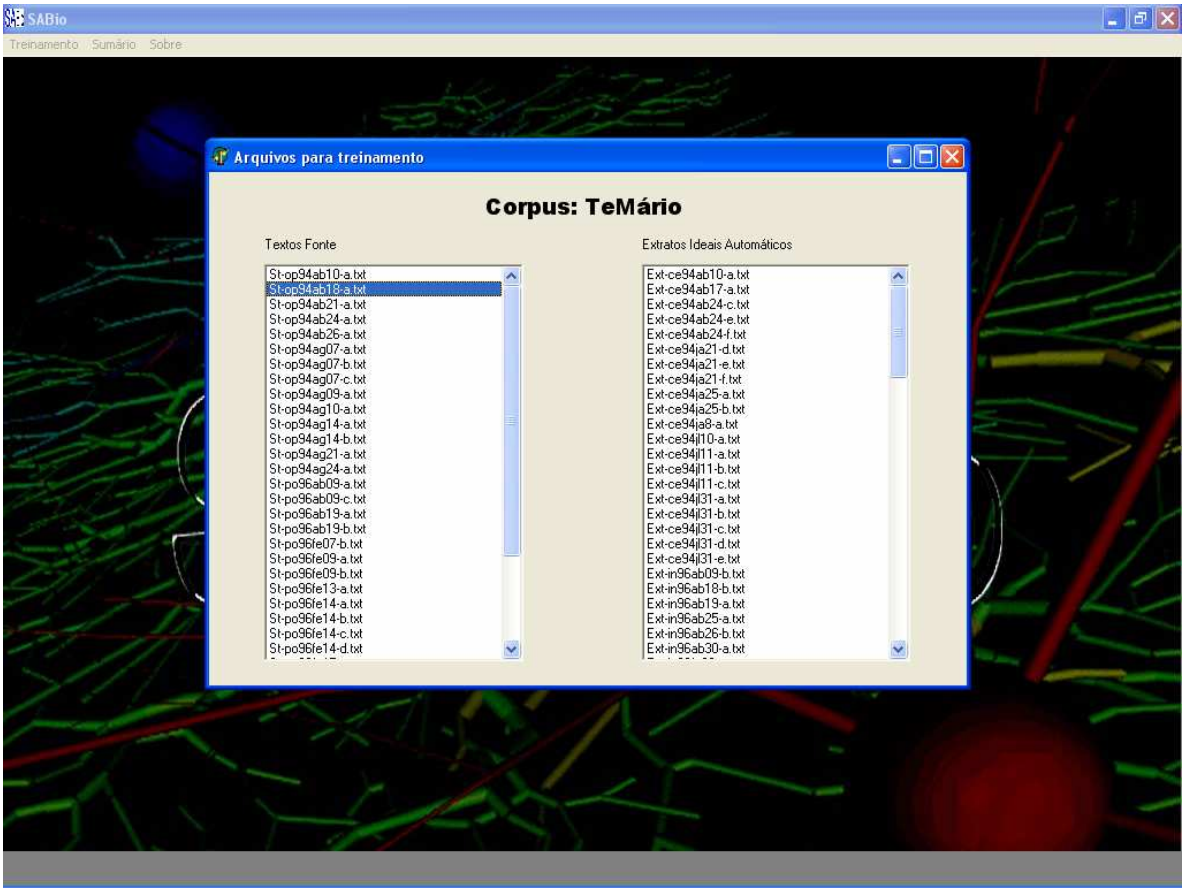
Tela principal do SABIO:



Menu de treinamento:

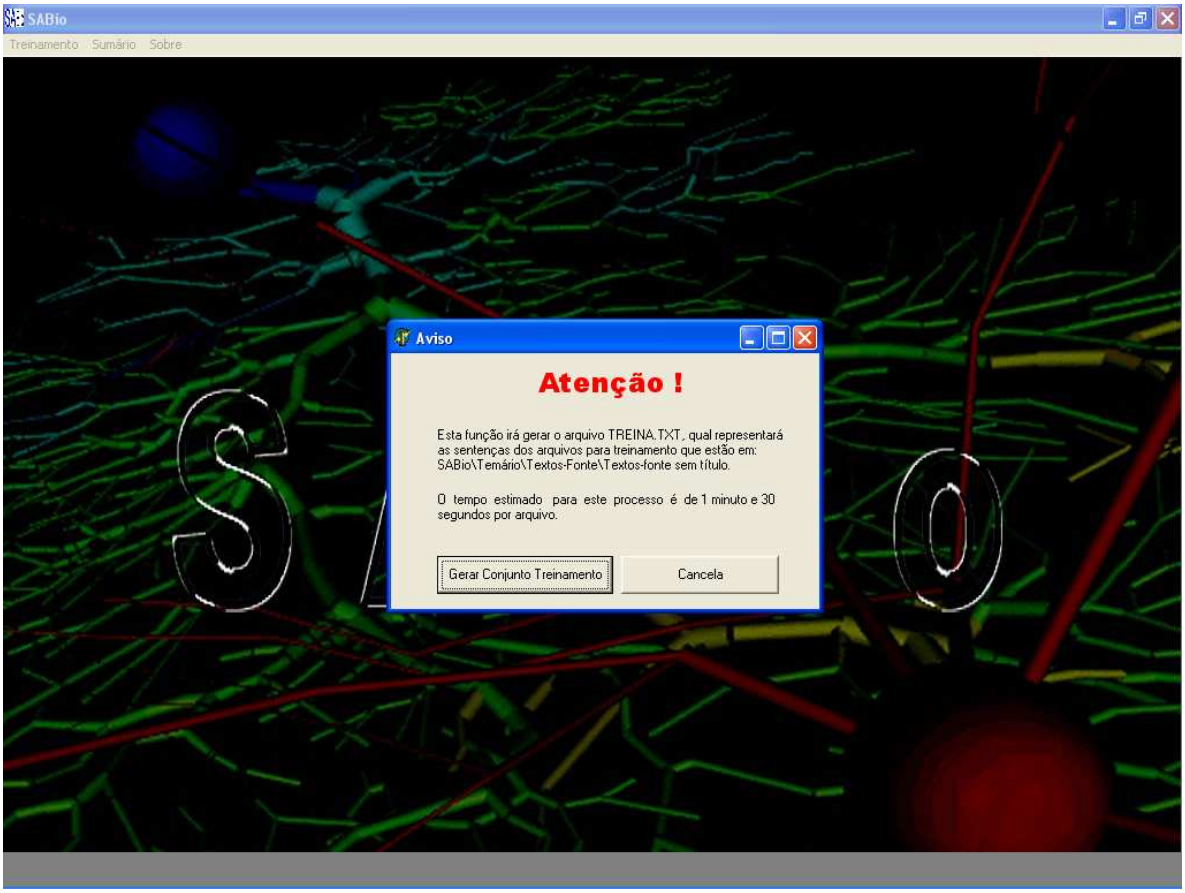


Opção.: "*Arquivos a utilizar*" (menu treinamento): O usuário escolhe o texto-fonte e será exibido o seu conteúdo deste arquivo juntamente com o respectivo extrato-ideal.

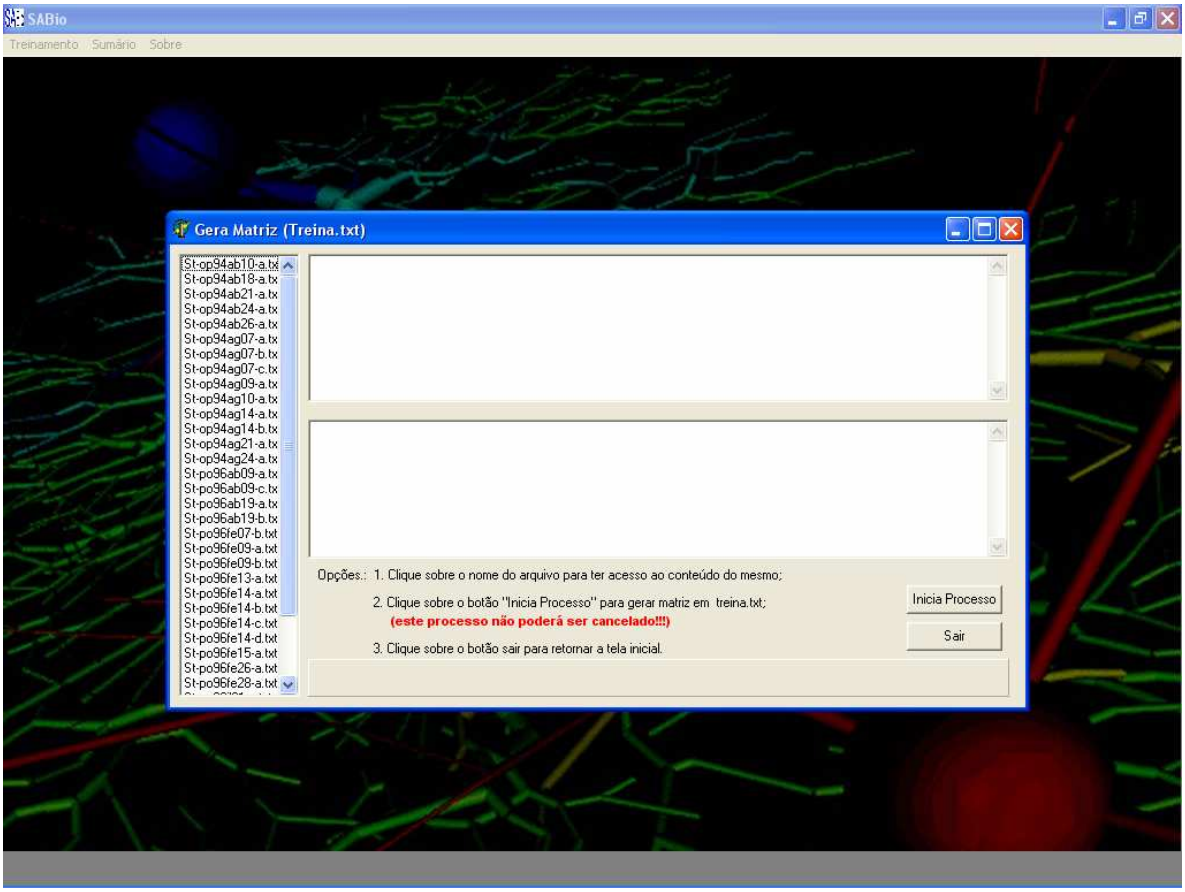


Opção.: *"gerar conjunto de treinamento"* (menu treinamento):

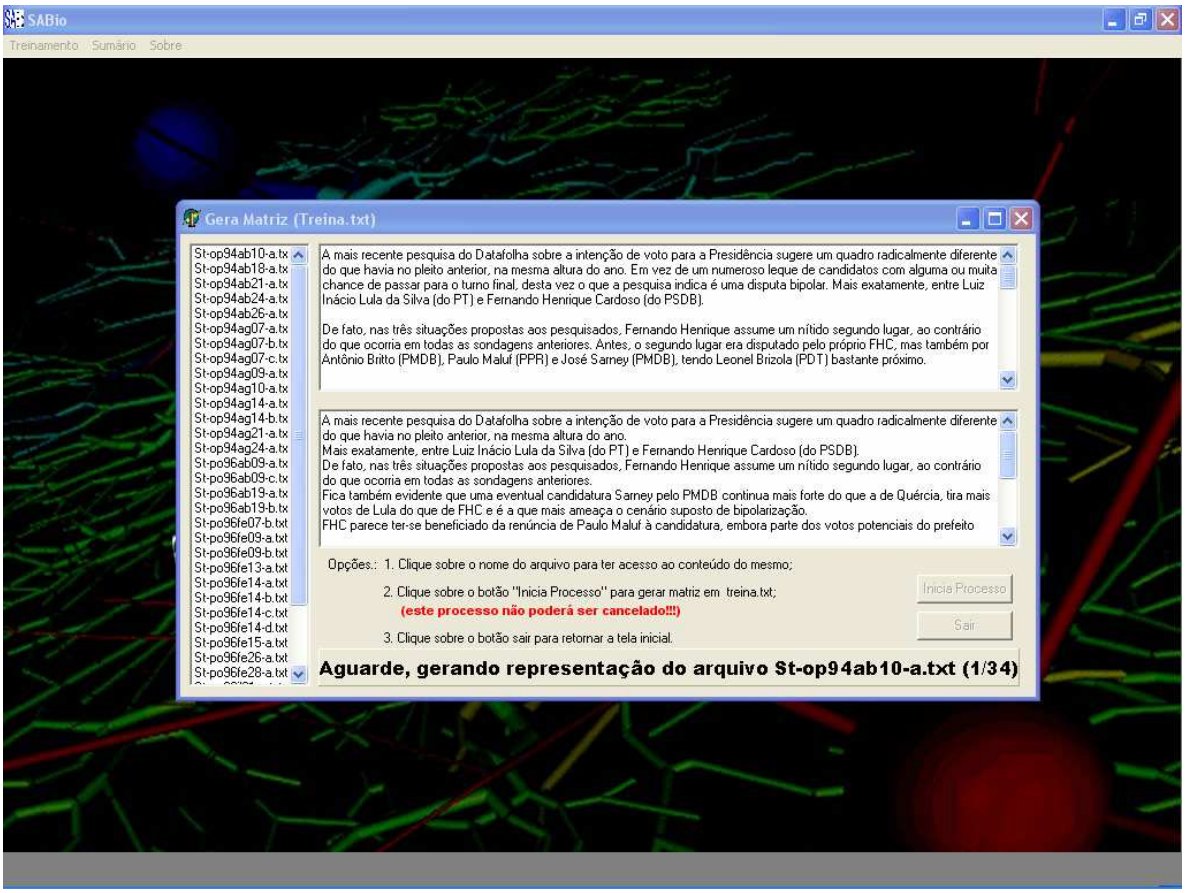
Passo 1: Alerta que será gerado um arquivo "Treina.txt" que representará as sentenças dos arquivos do conjunto para treinamento. Os arquivos utilizados para gerar esta representação serão os arquivos que estiverem localizados em \SABio\TeMário\Textos-Fonte\Texto-Fonte sem titulo:



Passo 2: Exibe os arquivos que formarão o conjunto de treinamento e faz algumas orientações sobre o processo de treinamento:



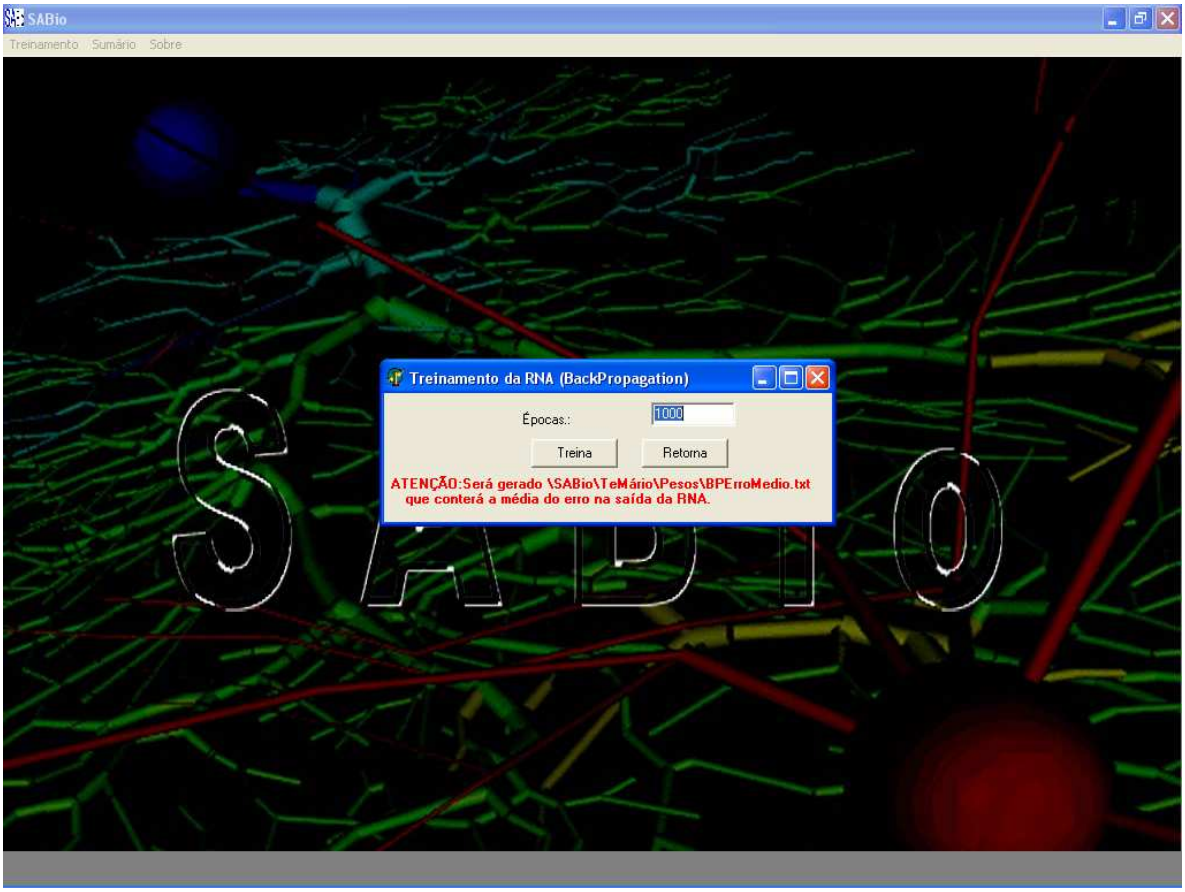
Passo 3: Quando o processo é iniciado exibe o conteúdo do texto-fonte e seu respectivo extrato-ideal. Este processo é finalizado quando o último arquivo pertencente ao conjunto de treinamento for codificado.



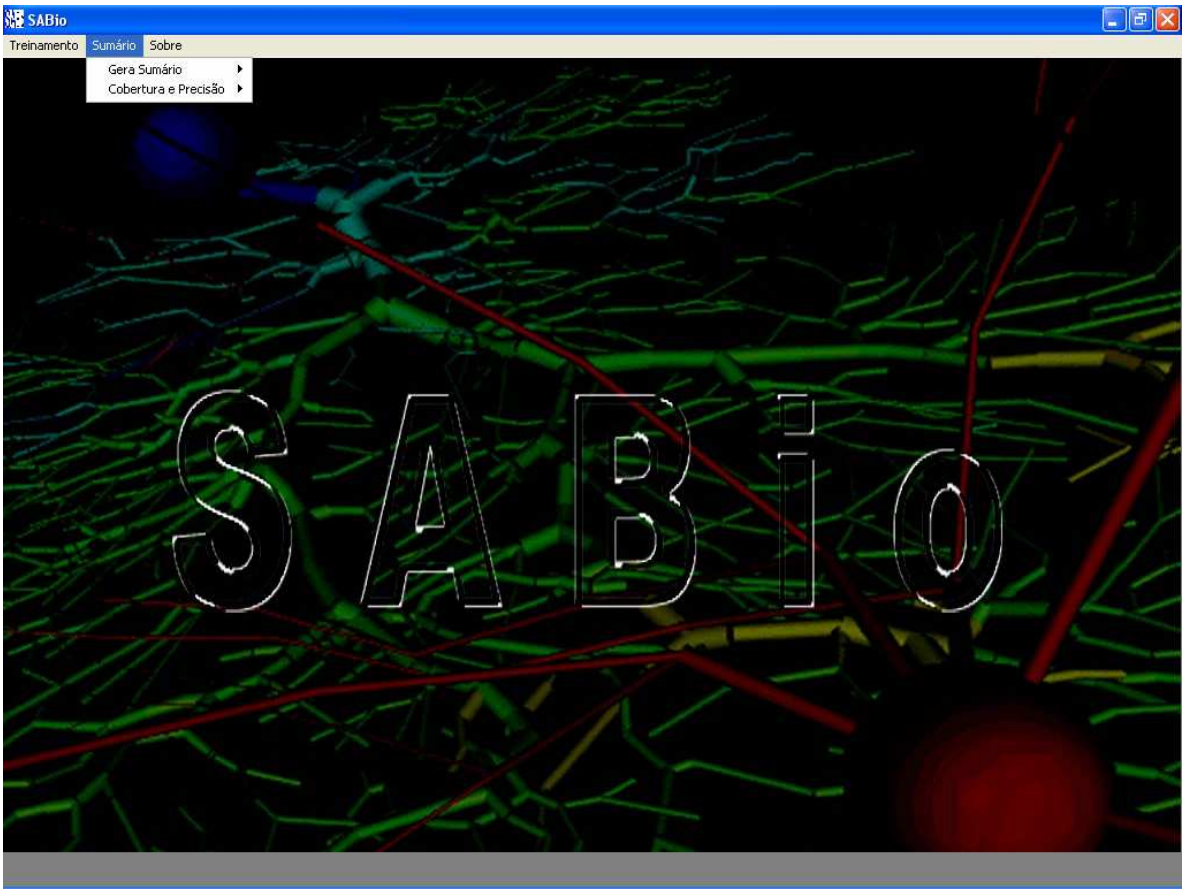
Opção.: *"Treinamento/Erro Médio"* (menu treinamento):

Passo 1: O usuário seleciona qual o método de treinamento da RNA (Backpropagation ou GeneRec) para o qual deseja obter informações sobre o erro médio:

Passo 2: O usuário informa para quantas épocas a RNA será treinada. O SABIO irá gerar em \SABio\TeMário\Pesos o arquivo BPErroMedio.txt (BackPropagation) ou GRErroMedio.txt (GeneRec) que conterá informações sobre o erro médio a cada 1000 épocas até o limite fornecido neste passo. Neste passo o SABIO grava em arquivos de controle do sistema os pesos das conexões (matrizes) a cada 1000 épocas até atingir o limite fornecido. Tal processo visa agilizar o processo de Sumarização.

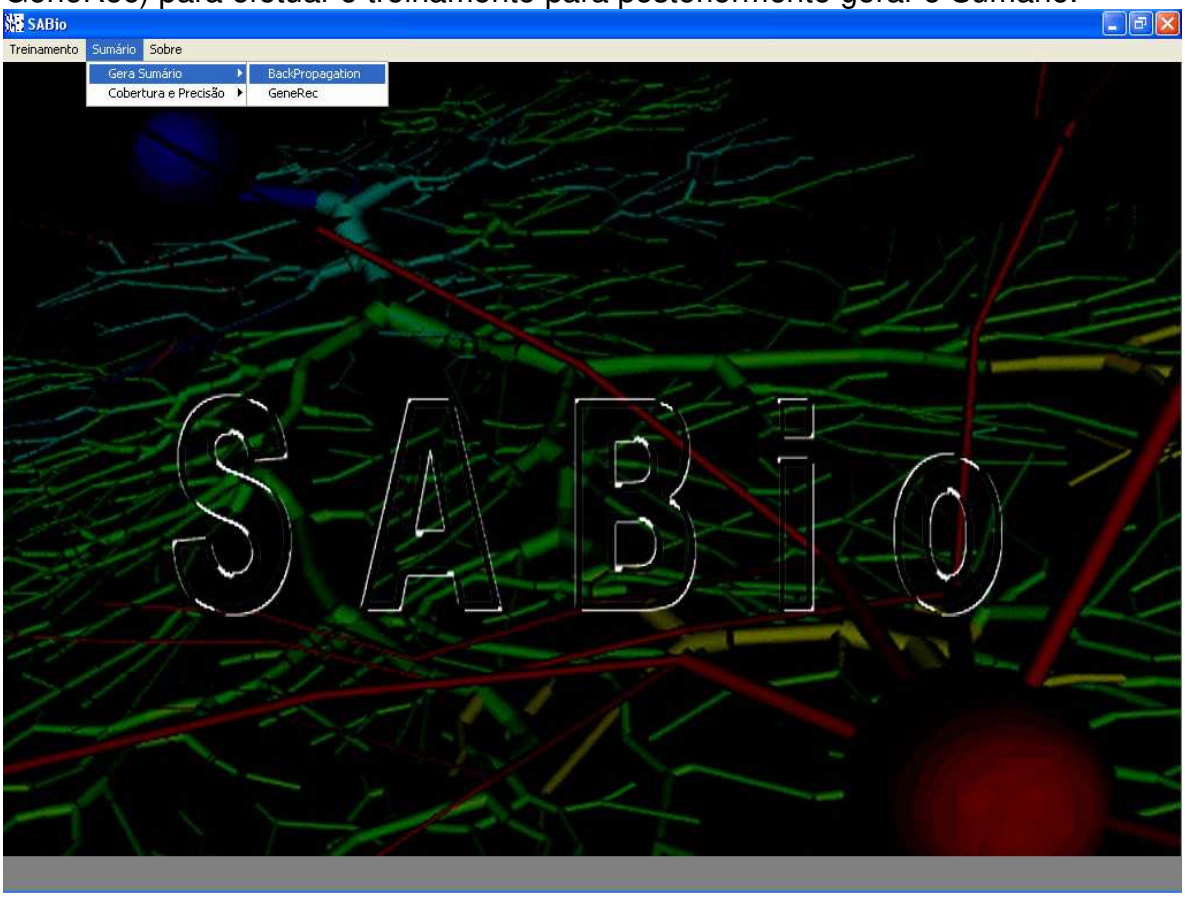


Menu “Sumário”

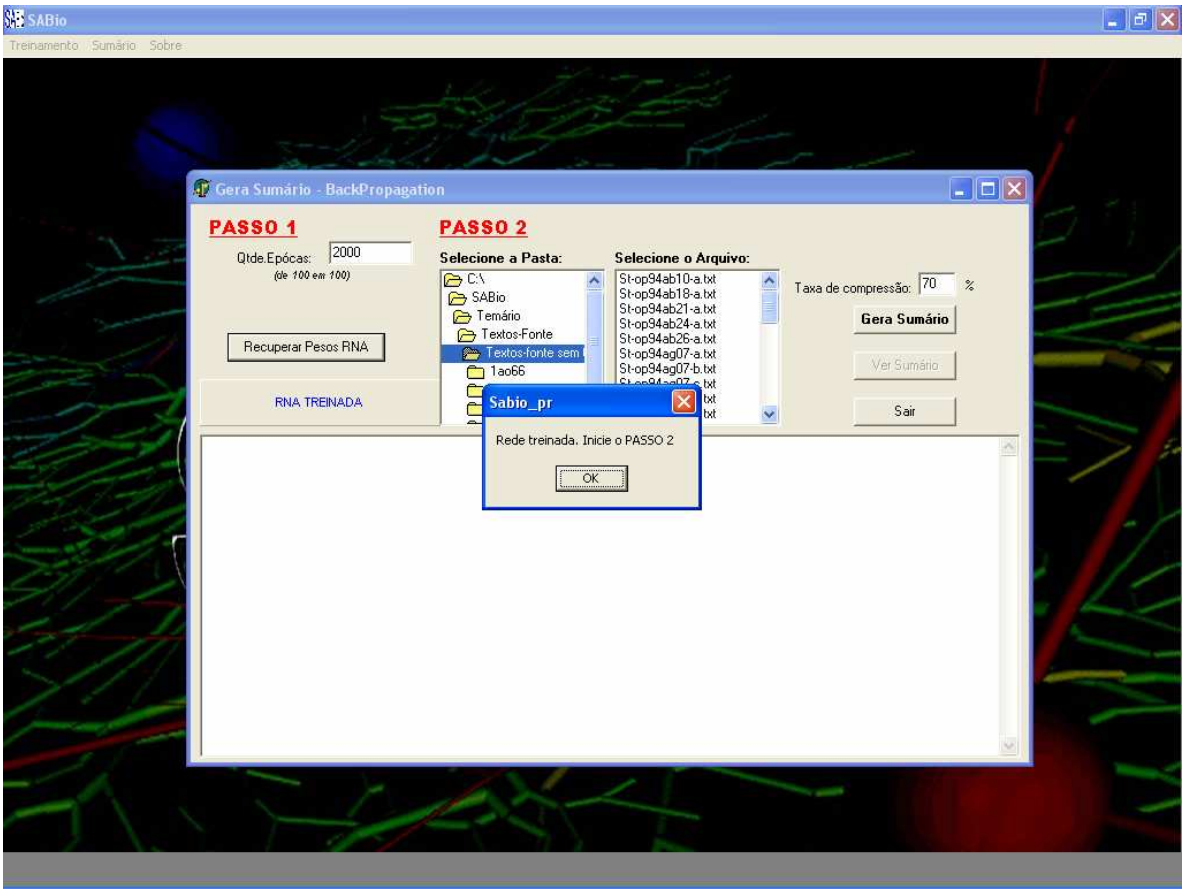


Opção: *"Gera Sumário"* (no menu sumário):

O usuário seleciona o método de treinamento da RNA (BackPropagation ou GeneRec) para efetuar o treinamento para posteriormente gerar o Sumário.

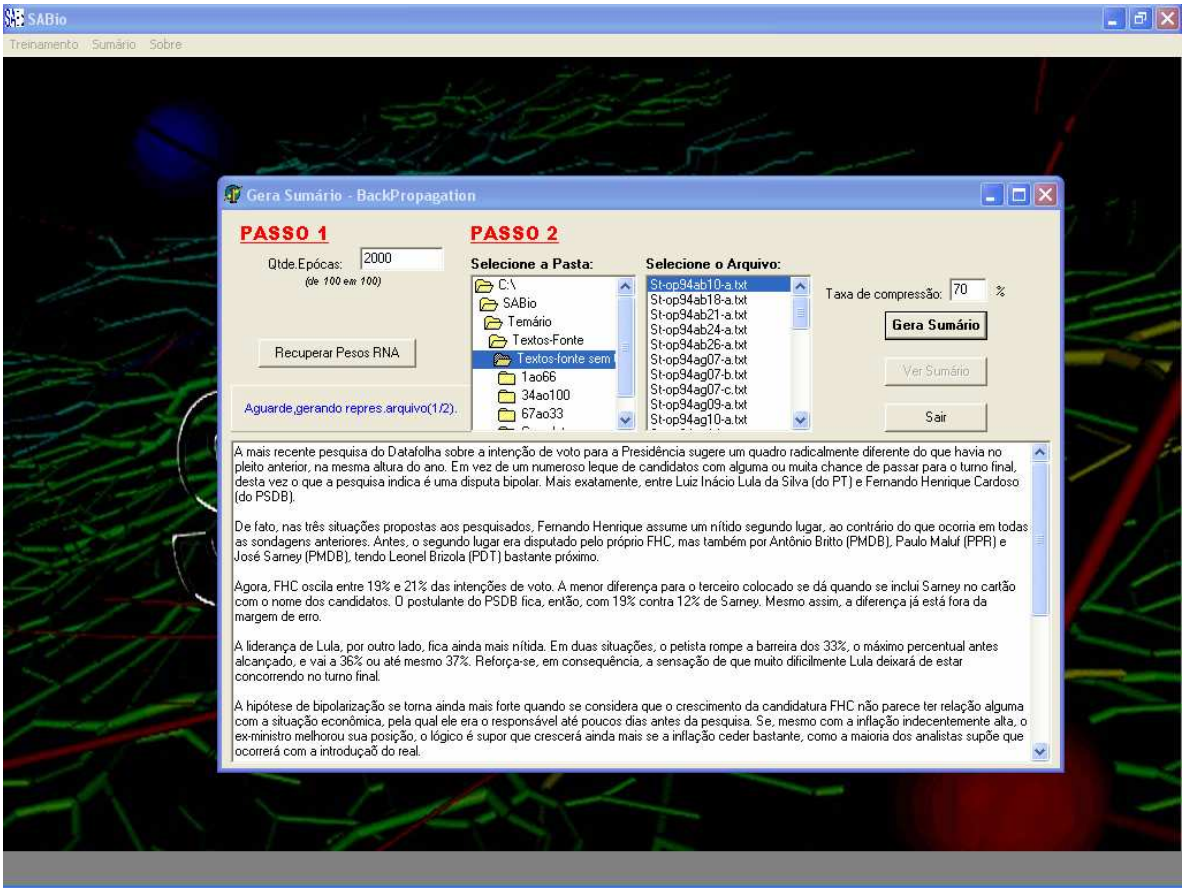


Passo 1: O usuário seleciona a quantidade de épocas para treinar a rede e clica em “Recuperar Pesos RNA”. “A mensagem Rede Treinada. Inicie o PASSO 2” é exibida

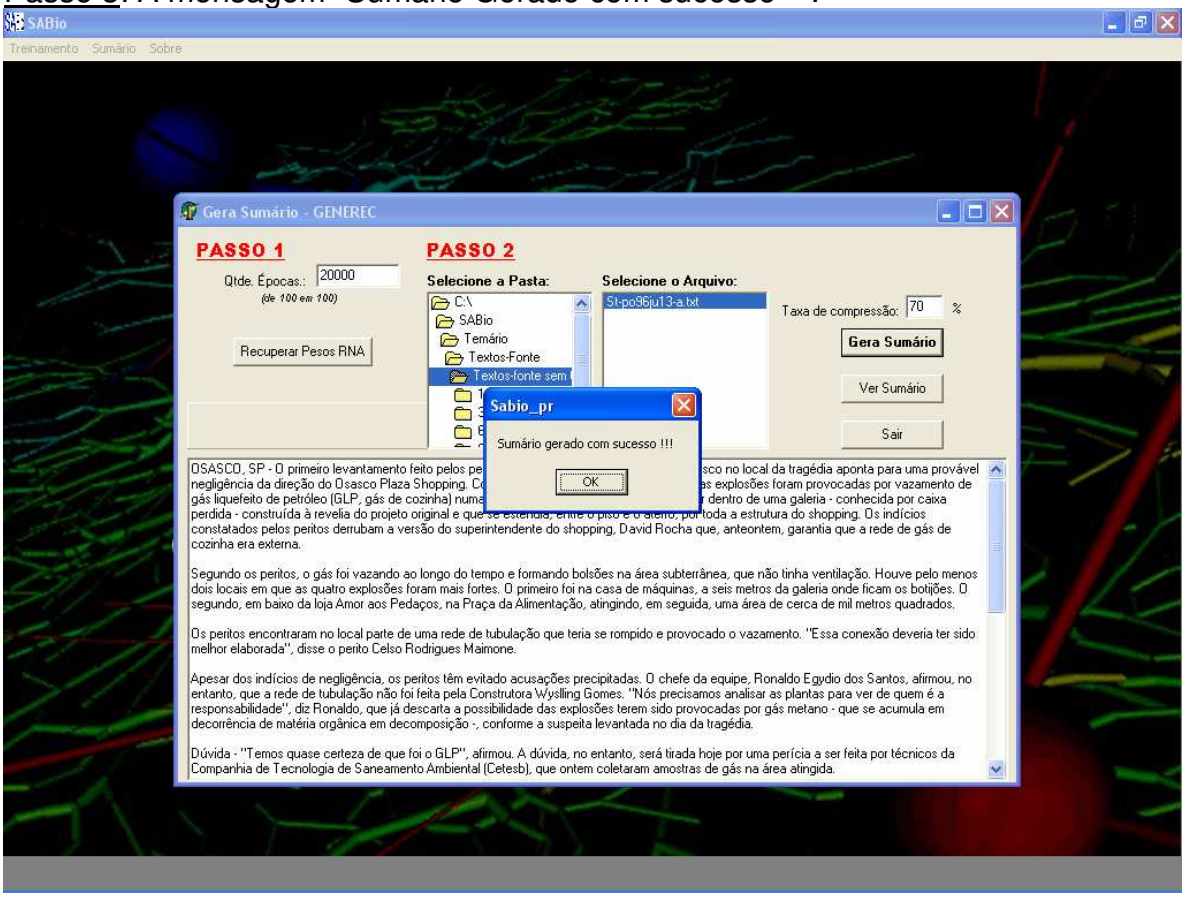


O SABIO foi treinado previamente até a época 10000 e as matrizes de pesos da RNA foram gravadas para todas épocas múltiplas de 1000. Tal processo visa agilizar o processo de treinamento na fase de Sumarização. Caso haja interesse em efetuar a sumarização com quantidade de épocas diferentes das já existentes, deve-se treinar a rede (opção treinamento).

Passo 2: O usuário informa o caminho e o arquivo do “novo” texto-fonte à sumarizar, escolhe a taxa de compressão e clica em “Gera Sumário”



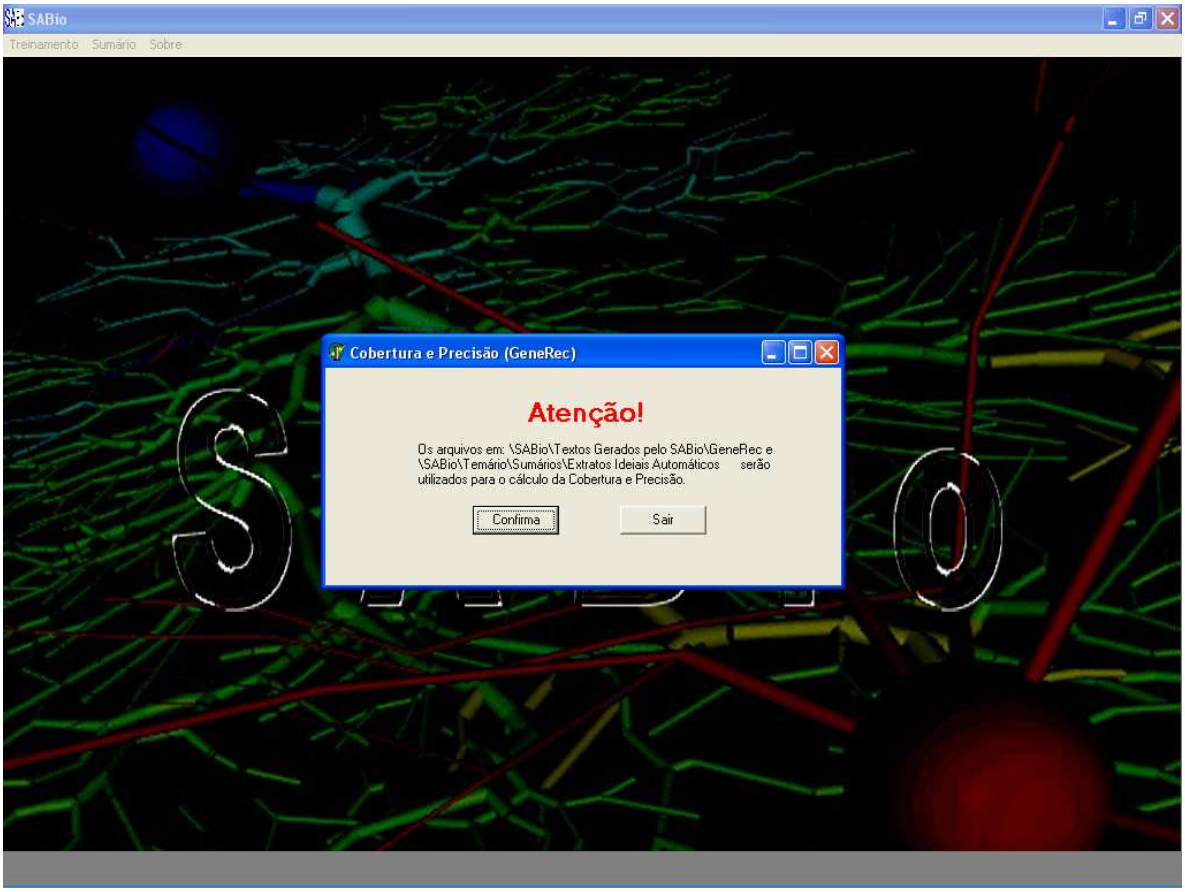
Passo 3: A mensagem “Sumário Gerado com sucesso”[35]:



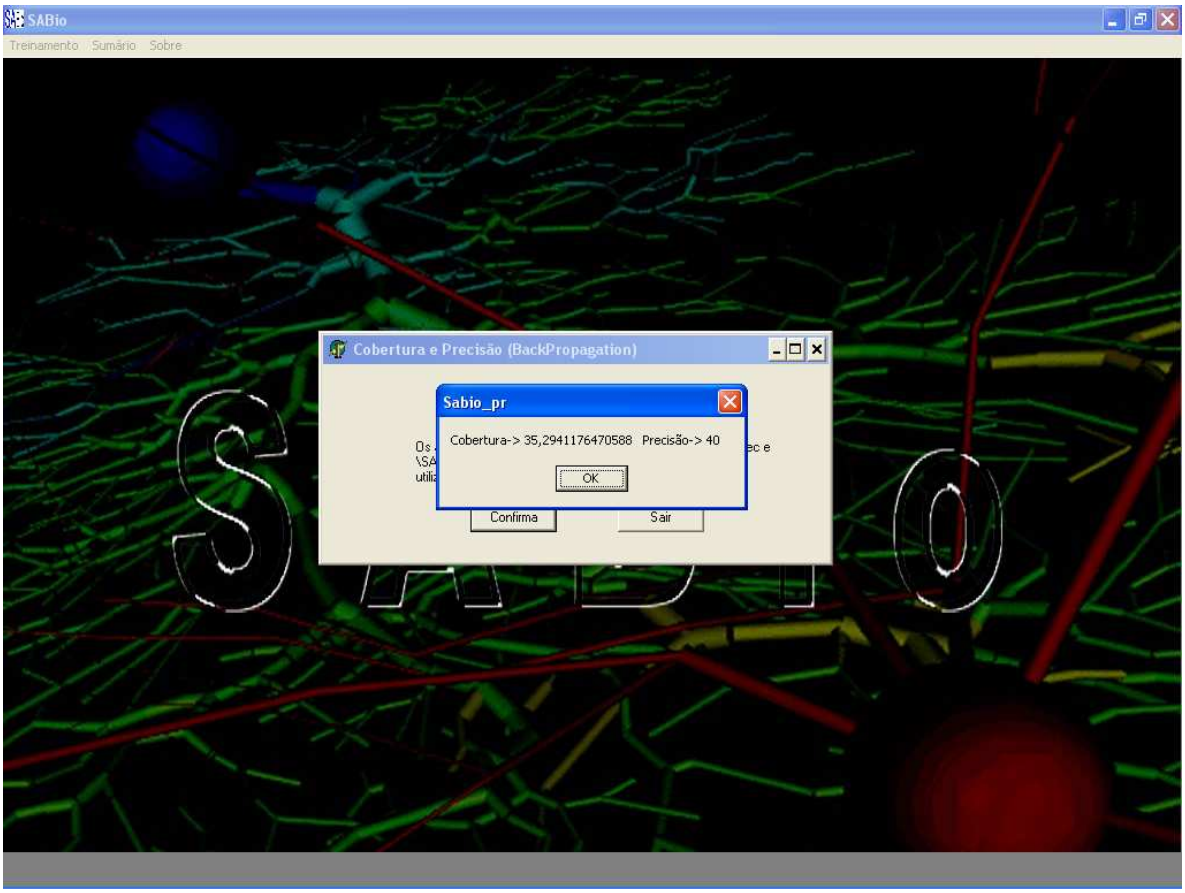
[35] O sumário gerado é um arquivo tipo texto que será gravado em \SABio\Textos Gerados pelo SABio\Backpropagation\ (quando a RNA for treinada através do Backpropagation) ou \SABio\Textos Gerado pelo SABio\GeneRec\ (quando a RNA for treinada através GeneRec); O nome do arquivo que representa o sumário gerado será semelhante ao do texto-fonte, acrescido de “BP-“ no inicio do nome (quando a RNA for treinada através do Backpropagation), ou acrescido de “Bio-“ no inicio do nome (quando a RNA for treinada através do GeneRec)

Opção.: *"Cobertura e Precisão"* (menu treinamento):

Passo 1: É exibida a mensagem que o valor da Cobertura e Precisão serão calculados através dos sumários já gerados no processo anterior. Os arquivos utilizados para o cálculo médio da Cobertura e Precisão são aqueles encontrados em \SABio\Textos Gerados pelo SABio\BackPropagation (quando calculado para o algoritmo do Backpropagation) ou \SABio\Textos Gerados pelo SABio\GeneRec (quando calculado para o algoritmo do GeneRec) com seus respectivos extratos ideais em \SABio\Temário\Sumários\Extratos Ideais Automáticos\.

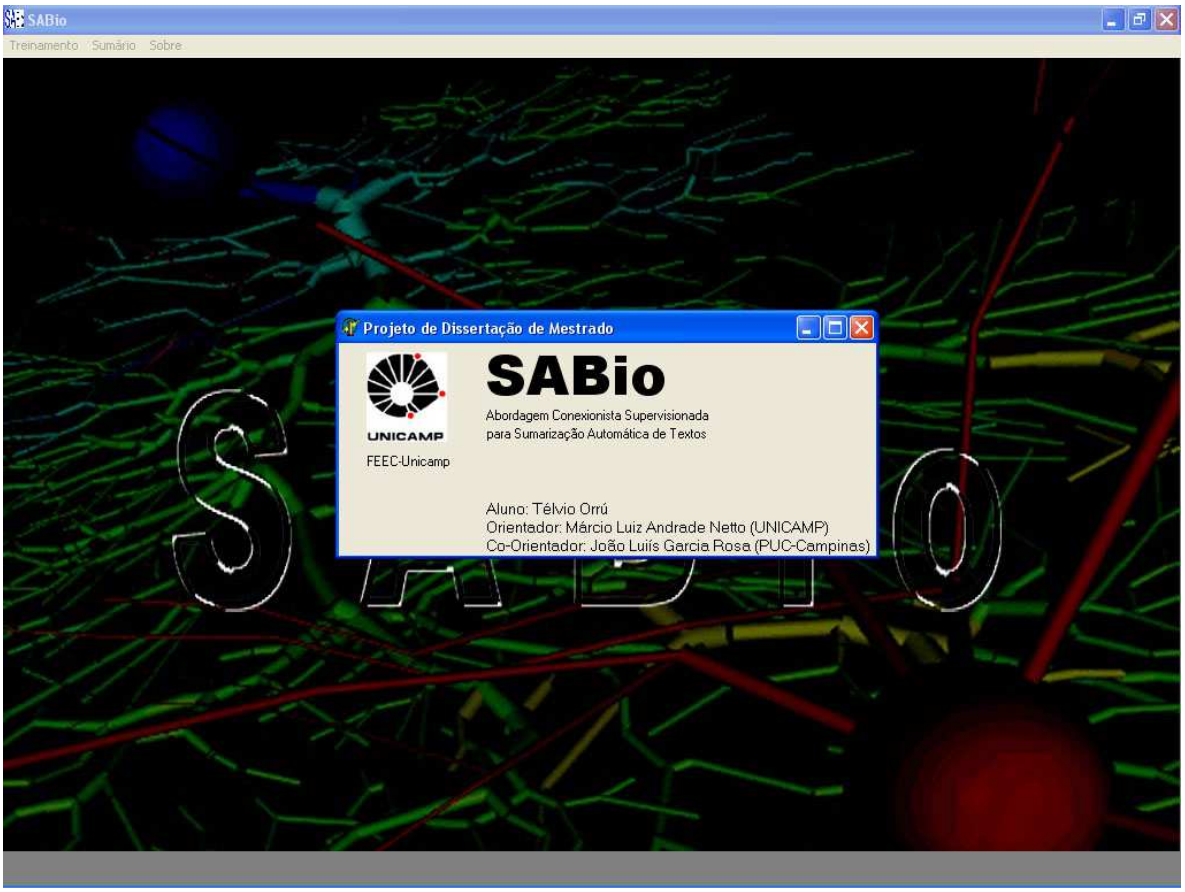


Passo 2: A Cobertura e a Precisão são exibidas:



Menu “Sobre”:

Opção.: SABIO:



Apêndice D - Características técnicas e funcionais do SABIO

Para a implementação do SABIO utilizou-se a linguagem de programação Borland Delphi versão 5. A escolha por esta linguagem deve-se à familiaridade do autor com a mesma, aliada à facilidade para utilização de matrizes neste dialeto.

O SABIO é compatível com qualquer computador que utilize o Microsoft Windows 98 (ou superior) como sistema operacional. Pode ser instalado em qualquer drive (local ou rede) desde que seja criada uma pasta exclusiva para ele (SABIO) com a seguinte estrutura:

Subpastas:

- TeMário → o corpus que foi utilizado (na estrutura que foi concebido);
- TeMário\Pesos → Durante a fase do treinamento, o SABIO irá gerar nesta pasta arquivos do tipo texto para armazenar valores numéricos que correspondem aos pesos que as matrizes assumem a cada 1000 épocas;
- Representa Arqs → O SABIO gera nesta pasta a representação dos arquivos que serão sumarizados (texto-fonte novo) para os 7 traços comentados. Tal processo visa agilizar a sumarização caso o mesmo arquivo venha a ser sumarizado mais que uma vez, porém com outra taxa de compressão;
- Textos Gerados pelo SABIO (BackPropagation/GeneRec)→ Todos os sumários gerados através do SABIO são gravados nesta pasta.

Na seqüência exibe-se a estrutura das pastas para o funcionamento do SABIO:

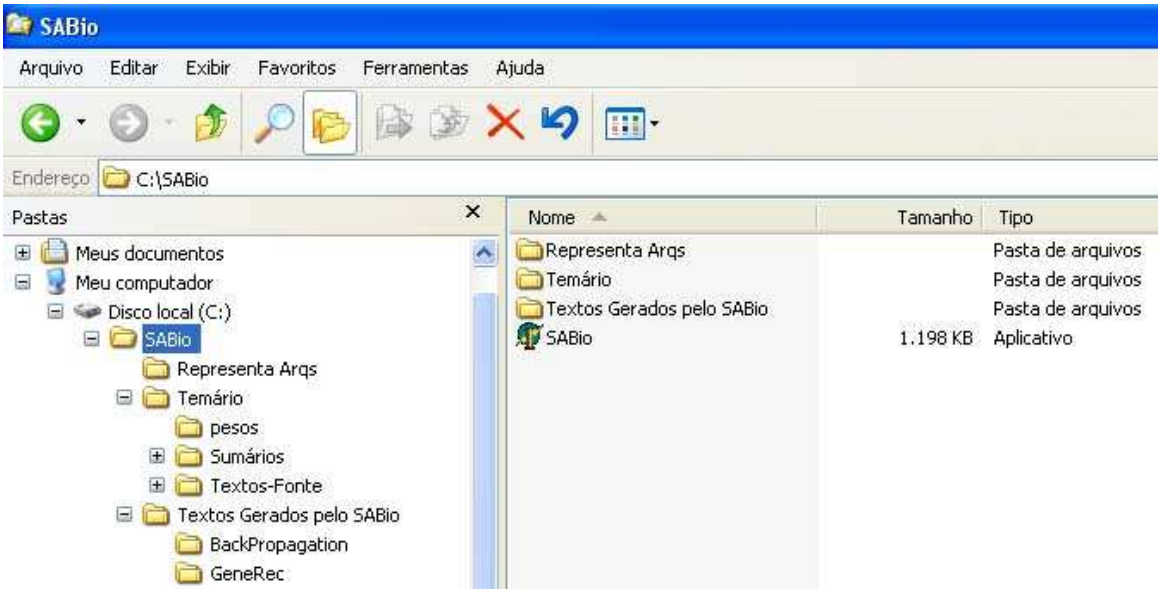


Figura D.1 - Estrutura das pastas para utilização do SABIO

O arquivo a ser sumarizado deve estar em formato “texto” para que o SABIO faça a sumarização. Através deste arquivo o respectivo sumário será gerado em um outro arquivo no formato “texto”, sendo que o nome deste arquivo dependerá do algoritmo de treinamento selecionado (Backpropagation ou GeneRec).

O critério adotado na fase de seleção das sentenças que irão compor o sumário foi a importância da sentença no texto-fonte a sumarizar, considerando o patamar mencionado no item 3.2.1 da dissertação. Caso haja empate na seleção das sentenças, prevalecerá aquela que obteve a maior pontuação (calculada através do método *Gist Sentence,* item 3.2.1).

Apêndice E - Gráficos

E.1 – Comparações de erros entre o algoritmo GeneRec e o algoritmo Backpropagation

Na seqüência são mostrados gráficos que exibem comparações entre o erro quadrático médio dos algoritmos GeneRec e do Backpropagation na fase de aprendizado[36] do SABIO com taxas de aprendizagem variadas:

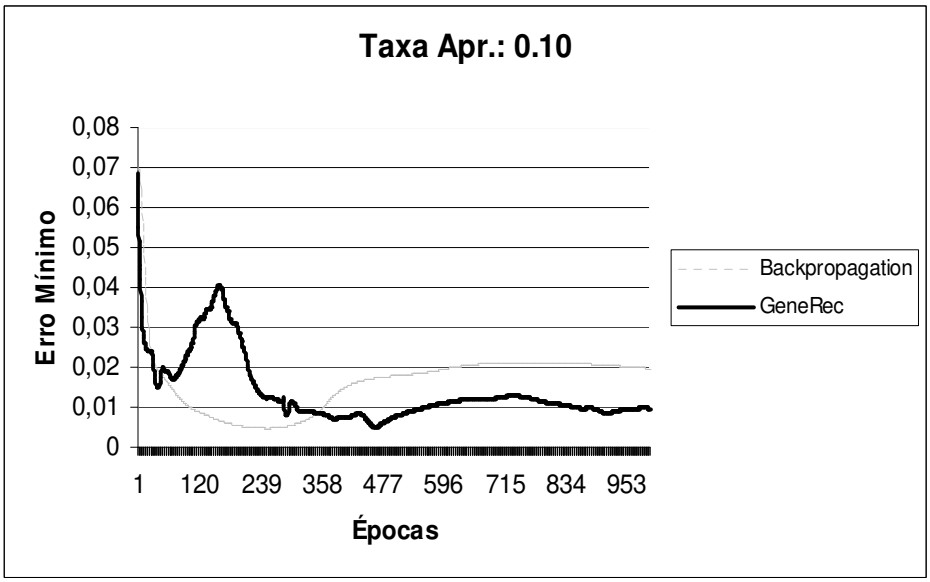


Figura E.1 - SABio com 10 neurônios na camada escondida e taxa de aprendizagem de 0.10. O erro mínimo estipulado não foi atingido, porém na época 1000 o GeneRec apresenta valor menor de erro quando comparado ao Backpropagation (0.009612 (69s[37]) e 0,019656 (68s), respectivamente).

[36] Utilizou-se 2/3 do corpus TeMário para a fase de treinamento, considerando que o 1/3 restante ficou para formulação de testes de Cobertura e Precisão (sub-capítulo 4.2).

[37] Tempo de processamento em segundos na época mencionada. Utilizou-se processador AMD Sempron 2.4Ghz.

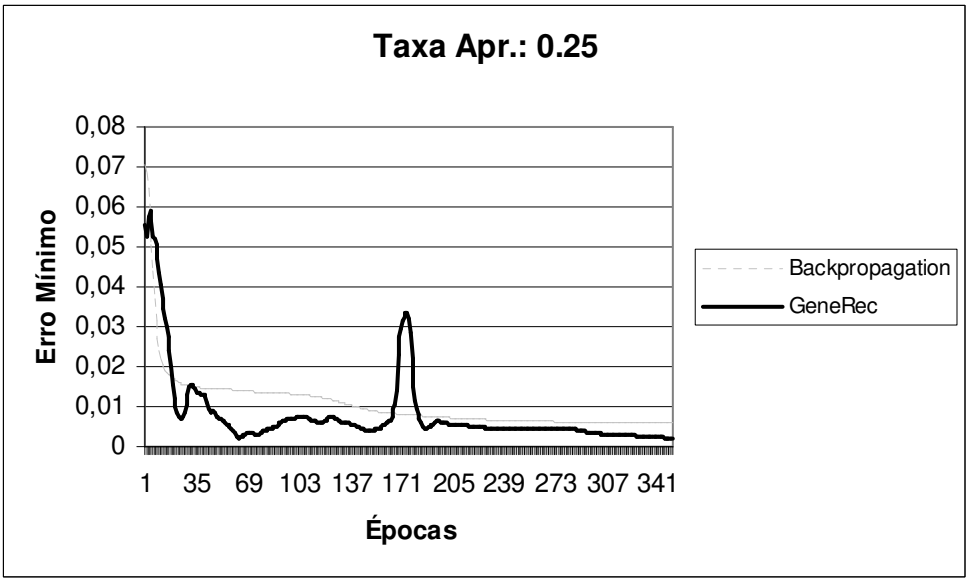


Figura E.2 - SABio com 10 neurônios na camada escondida e taxa de aprendizagem de 0.25. O erro mínimo estipulado foi atingido através do GeneRec (Época 352 (25s)) enquanto o BackPropagation não atingiu o erro mínimo (Época 1000 = 0.00339266 (68s))

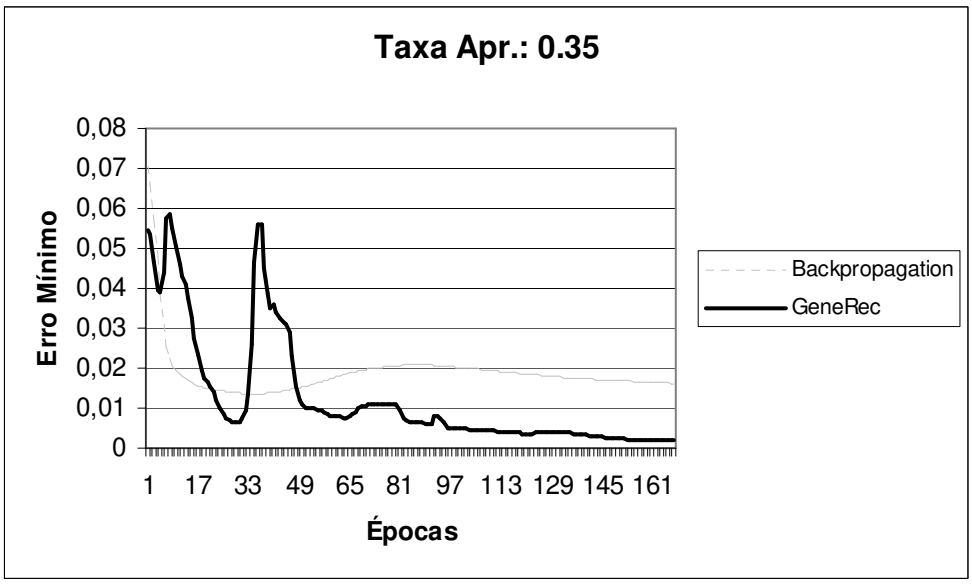


Figura E.3 – SABio com 10 neurônios na camada escondida e taxa de aprendizagem de 0.35. O erro mínimo estipulado foi atingido através do GeneRec (Época 169 (12s)) enquanto o BackPropagation não atingiu o erro mínimo (Época 1000 = 0.009189 (69s))

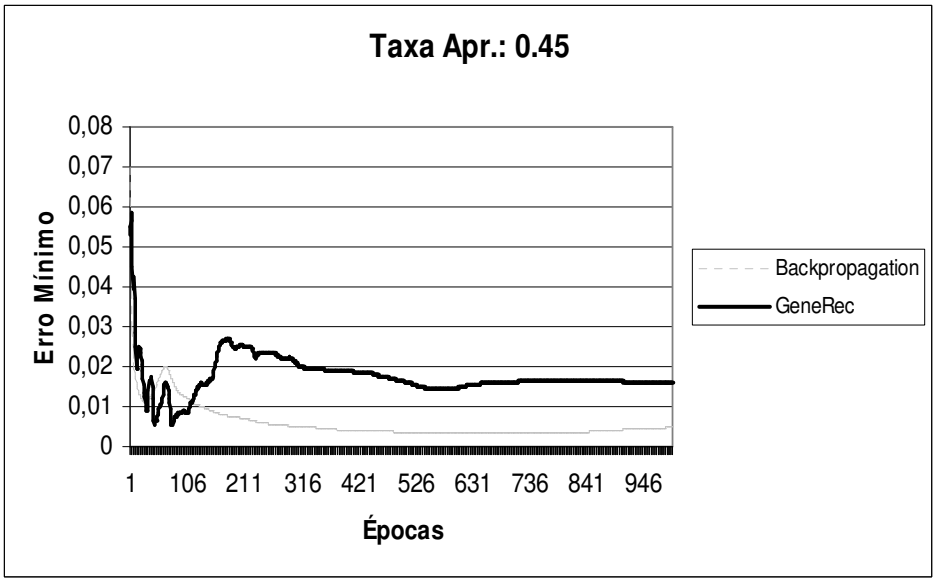


Figura E.4 – SABio com 10 neurônios na camada escondida e taxa de aprendizagem de 0.45. O erro mínimo estipulado não foi atingido, porém na época 1000 o Backpropagation apresenta valor menor de erro quando comparado ao GeneRec (0.004815 (69s) e 0,015873 (71s), respectivamente)

Considerando as arquiteturas exibidas nos gráficos 1 ao 4 (10 neurônios na camada intermediária), somente com a taxa de aprendizagem de 0.45 o Backpropagation obteve melhor desempenho que o GeneRec. Conforme foi comentado, várias outras arquiteturas de RNA foram testadas e, foi constatado que o GeneRec obtém melhor desempenho para algumas arquiteturas da RNA.

E.2. – Relação entre taxa de compressão do texto-fonte e taxas de cobertura e precisão.

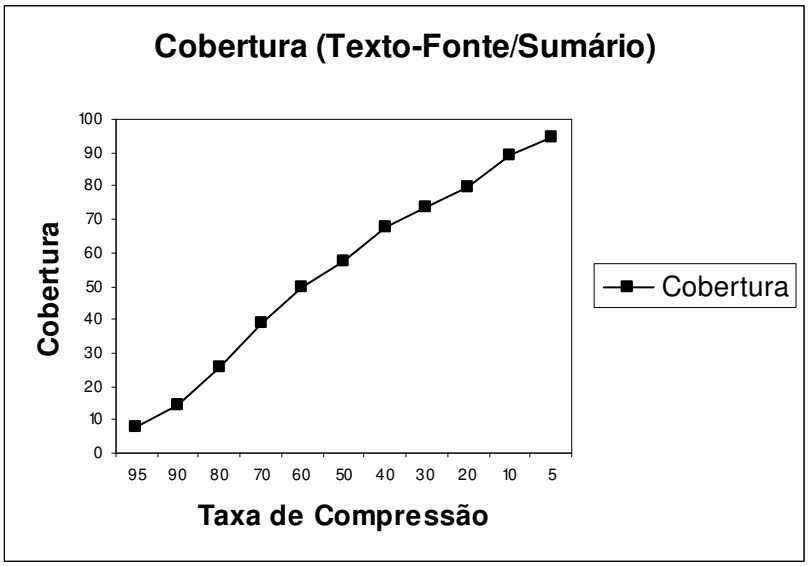


Figura E.5 – Relação entre a taxa de compressão e a taxa de cobertura. Observou-se nas simulações feitas através do SABIO que quando menor a taxa de compressão maior a taxa de cobertura.

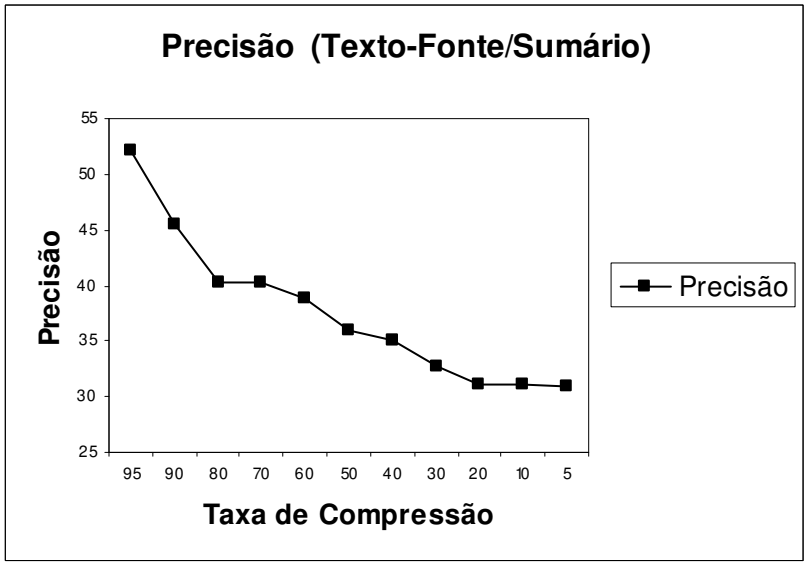


Figura E.6 – Relação entre a taxa de compressão e a taxa de precisão. Observou-se nas simulações feitas através do SABIO que quando maior a taxa de compressão maior a taxa de precisão. Isto pode demonstrar que o SABIO preserva, na maioria das vezes, as sentenças que possuem maior importância no sumário.

Apêndice F - Exemplos de Extratos Gerados

Texto-fonte (St-op94ab18-a)	Extrato (Bio-op94ab18-a)
Existe dentro do PT uma propensão eufórica negativa: "Lula já venceu!" Em política, a vitória parece segura em uma semana; pode converter-se em derrota na semana seguinte. Essa euforia, que tenho combatido, também pode impulsionar nossa perda. Acredito que Luiz Inácio Lula da Silva sabe disso melhor que eu. Se assumo a responsabilidade de um debate público, faço-o pensando no que significa o PT no cenário trágico e opressivo em que ele se insere. Apesar dos movimentos das elites das classes dominantes e dos partidos da ordem, parece evidente que Lula e a coligação dos partidos contestadores que o apóiam têm possibilidades de vencer. Não só a equação pessoal comprovou a ressonância de suas diretrizes políticas na massa dos excluídos, dos assalariados, dos estratos médios em proletarização, nas entidades radicais empenhadas na renovação da economia, da sociedade e da cultura, como reforçou a consciência crucial: ou Lula ou a repetição do passado com alterações cosméticas. Nesse sentido, independentemente dos resultados das urnas, Lula e suas forças sociais e políticas triunfaram. Poderão ocupar ou não o poder. O Brasil, todavia, não será o mesmo depois do vendaval que se aproxima. O teste procede da afoita fabricação de um candidato à Presidência da confiança do grande capital interno e estrangeiro, com seus mentores e suportes humanos, institucionais e financeiros. Como diria um provecto ex-reitor da USP, "querem salvar os bolsinhos". Lançaram-se ao embate eleitoral com o propósito político de manter seus privilégios, com o monopólio do poder estatal que eles pressupõem. Depois de Collor, essa resistência autoritária serve mais como advertência que como arma eleitoral. Ela desperta uma indagação inevitável nos jovens e adultos: por que essa exacerbação que explicita a predisposição de "esmagar Lula"?	Apesar dos movimentos das elites das classes dominantes e dos partidos da ordem, parece evidente que Lula e a coligação dos partidos contestadores que o apóiam têm possibilidades de vencer. Não só a equação pessoal comprovou a ressonância de suas diretrizes políticas na massa dos excluídos, dos assalariados, dos estratos médios em proletarização, nas entidades radicais empenhadas na renovação da economia, da sociedade e da cultura, como reforçou a consciência crucial: ou Lula ou a repetição do passado com alterações cosméticas. Depois de Collor, essa resistência autoritária serve mais como advertência que como arma eleitoral. Para o presente e o futuro, contudo, é essencial romper com o passado, com o fisiologismo, com o clientelismo e o privatismo de conteúdo patrimonialista, com o oportunismo político (tão destrutivo na direita, quanto no centro-esquerda), com as conciliações pelo alto contra o povo. A sorte de Lula está lançada e, com ela, os papéis construtivos do PT e seus aliados, neste momento decisivo.

Esta eleição ergue uma pergunta: convém ao PT ganhá-la ou encaixá-la na acumulação crescente de dinamismos de desgaste da ordem, que o favorece? Nenhum partido da esquerda persegue o segundo objetivo. Mas ganhar ou perder denotam algo relativo, na situação histórica vigente. Para o presente e o futuro, contudo, é essencial romper com o passado, com o fisiologismo, com o clientelismo e o privatismo de conteúdo patrimonialista, com o oportunismo político (tão destrutivo na direita, quanto no centro-esquerda), com as conciliações pelo alto contra o povo. Esses aspectos sinalizam o caminho de Lula, do PT e dos partidos fiéis a uma identidade política sólida. A esquerda autêntica compreende a natureza de seus compromissos com a transformação da ordem e a criação de uma sociedade nova. Acabamos de constatar quais são os vínculos de uma social-democracia improvisada com o mudancismo conservador. Se os eleitores se enganarem, consagrarão a República democrática de fachada. Para não correrem tais riscos, deverão bater-se por alvos claros e certeiros, que se definem no campo da esquerda. A sorte de Lula está lançada e, com ela, os papéis construtivos do PT e seus aliados, neste momento decisivo.	

Cobertura → 50,00%
Precisão → 60,00%
Medida-F → 54,54%

Texto-fonte (St-op94ab26-a)	Extrato (Bio-op94ab26-a)
Estudei engenharia, fui apaixonado por matemática e acabei doutor em economia. Não tenho, portanto, formação jurídica. Nessa área toco de ouvido e com parcimônia. Ainda assim me arriscarei a questionar algumas teses jurídicas sobre a revisão constitucional, pela relevância do tema. Para começar, reafirmo minha convicção de que o Brasil não conseguirá retomar seu desenvolvimento a médio e longo prazos, nem consolidar a democracia, sem reformar amplamente a Constituição de 1988. Aliás, o próprio texto constitucional previu a reforma. Uma das teses mais estapafúrdias levantadas por alguns juristas é que a revisão não poderia ser feita, pois estaria vinculada ao plebiscito sobre sistema de governo. Mantido o presidencialismo, nada haveria a revisar na Carta. A tese é estapafúrdia porque: 1) os dispositivos sobre o plebiscito e a revisão são independentes; 2) nada no texto constitucional restringe a revisão ao sistema de governo. Outra tese jurídica implausível do ponto de vista lógico é que apenas o Congresso atual, eleito em 1990, poderia fazer a revisão. Mas onde está dito ou subentendido que a revisão só poderia ser feita por um mesmo Congresso? O texto constitucional diz apenas que a revisão deveria ser feita "após" 5 de outubro de 1993, ou seja, poderia começar nessa data ou no ano 2000. Apesar da advertência de que, começando em 5 de outubro, a revisão certamente iria fracassar –por causa do ano supereleitoral de 1994, que alimenta o "ausentismo" dos parlamentares, a organização dos "contras" e o desinteresse do governo–, a maioria do Congresso decidiu iniciá-la logo, em vez de transferi-la para 1995. As previsões sombrias foram confirmadas e não se aprovou praticamente nada até agora. O prazo termina em 31 de maio e ainda se tenta votar algumas emendas, mas o fundamental ficará de fora. Que fazer? O lógico não seria encerrar a revisão, mas apenas desativá-la até serem feitas as eleições ou até o começo de 1995. Mas o relator Nélson Jobim acha que isso não é possível, porque	Para começar, reafirmo minha convicção de que o Brasil não conseguirá retomar seu desenvolvimento a médio e longo prazos, nem consolidar a democracia, sem reformar amplamente a Constituição de 1988. Uma das teses mais estapafúrdias levantadas por alguns juristas é que a revisão não poderia ser feita, pois estaria vinculada ao plebiscito sobre sistema de governo. A tese é estapafúrdia porque: 1) os dispositivos sobre o plebiscito e a revisão são independentes; 2) nada no texto constitucional restringe a revisão ao sistema de governo. Outra tese jurídica implausível do ponto de vista lógico é que apenas o Congresso atual, eleito em 1990, poderia fazer a revisão. O texto constitucional diz apenas que a revisão deveria ser feita "após" 5 de outubro de 1993, ou seja, poderia começar nessa data ou no ano 2000. Apesar da advertência de que, começando em 5 de outubro, a revisão certamente iria fracassar –por causa do ano supereleitoral de 1994, que alimenta o "ausentismo" dos parlamentares, a organização dos "contras" e o desinteresse do governo–, a maioria do Congresso decidiu iniciá-la logo, em vez de transferi-la para 1995.

uma emenda já foi promulgada, a do Fundo Social de Emergência. O Supremo Tribunal Federal não aceitaria a prorrogação, argumentando que o Congresso estaria adotando um novo método permanente de mudar a Lei Magna e, assim, a revisão não acabaria nunca. De fato, o Supremo nunca deliberou sobre o assunto. E a argumentação não convence, pois o Congresso Revisor poderia até alterar as normas de mudar a Constituição, que não configuram nenhuma cláusula pétrea. Poderia, portanto, introduzir na Constituição uma data encerrando a revisão. Li ou ouvi a opinião de numerosos juristas sobre a possibilidade de adiamento para 1995. A mais recente é o brilhante artigo do professor Miguel Reale no jornal "O Estado de S. Paulo" de sábado último. Também opinam na mesma direção Miguel Reale Jr., Fábio Comparato (embora prefira o Congresso revisor exclusivo), Saulo Ramos, Tércio Sampaio Ferraz, Celso Bastos e Manoel Alceu Afonso Ferreira. Parece que as únicas coisas que esses juristas têm em comum é serem paulistas e julgarem que a revisão pode ser adiada. Será que todos estão errados? Parece-me improvável, pois nenhum deles é formado em economia...	

Cobertura → 36,36%
Precisão → 66,67%
Medida-F → 47,06%

Texto-Fonte(St-po96fe15-a)	Extrato(Bio-po96fe15-a)
BRASÍLIA - Preocupado com a repercussão negativa da manutenção da aposentadoria privilegiada dos parlamentares, o presidente da Câmara, Luís Eduardo Magalhães (PFL-BA), decidiu ontem que o Instituto de Previdência dos Congressistas (IPC) será extinto. A idéia é acabar com o instituto através de lei ordinária a ser votada nos próximos 60 dias. Apesar de ter o apoio de todos os líderes, Luís Eduardo encontra fortes resistências dos deputados. "Acabar com o IPC é inaceitável. Os líderes não estão suficientemente respaldados para decidir a extinção imediata do instituto", afirmou o vice-líder do PPB, deputado Gérson Peres (PA). "Causa espanto tratar como um privilégio o instituto de previdência parlamentar. Isso existe em todos os países onde há democracia", disse o deputado Prisco Viana (PPB-BA). "É uma loucura extinguir o IPC. É uma posição radical e precipitada", argumentou o presidente do instituto, deputado Heráclito Fortes (PFL-PI). Depois do carnaval, Heráclito vai mandar um questionário para cada um dos 513 deputados com o objetivo de saber sua posição sobre o fim do IPC. "Os deputados e os líderes estão com medo da imprensa. É frescura achar que o IPC é um privilégio", argumentou o deputado Agnaldo Timóteo (PPB-RJ). O coro dos descontentes foi engrossado pelo deputado Nílson Gibson (PPB-PE), que começou a recolher assinaturas para a apresentação de um destaque mantendo o instituto. "Vamos ver no painel de votação quem é quem. Aqui há estrelas que são vestais e que fazem discursos contra o IPC, mas por baixo do pano trabalham pela manutenção do instituto", disse o relator da reforma da Previdência, deputado Euler Ribeiro (PMDB-AM). Pito - A operação para acabar com o instituto começou ontem cedo pela manhã, logo depois que Luís Eduardo leu os jornais. O presidente da Câmara ficou particularmente irritado com as declarações de Euler Ribeiro. Para manter o IPC em seu substitutivo, Euler alegou que o líder do PMDB, Michel Temer (SP), tinha sofrido pressões de vários parlamentares. "Como você dá um entrevista dessas?", cobrou	BRASÍLIA - Preocupado com a repercussão negativa da manutenção da aposentadoria privilegiada dos parlamentares, o presidente da Câmara, Luís Eduardo Magalhães (PFL-BA), decidiu ontem que o Instituto de Previdência dos Congressistas (IPC) será extinto. Os líderes não estão suficientemente respaldados para decidir a extinção imediata do instituto", afirmou o vice-líder do PPB, deputado Gérson Peres (PA). É frescura achar que o IPC é um privilégio", argumentou o deputado Agnaldo Timóteo (PPB-RJ). O coro dos descontentes foi engrossado pelo deputado Nílson Gibson (PPB-PE), que começou a recolher assinaturas para a apresentação de um destaque mantendo o instituto. Aqui há estrelas que são vestais e que fazem discursos contra o IPC, mas por baixo do pano trabalham pela manutenção do instituto", disse o relator da reforma da Previdência, deputado Euler Ribeiro (PMDB-AM). Isso é um assunto menor que não pode empatar a reforma", afirmou Luís Eduardo. Até ontem à noite, três partidos - PT, PDT e PSDB - fecharam questão a favor da extinção do IPC. Atualmente, o instituto gasta mensalmente R$ 3,8 milhões com o pagamento de 2757 pensões a 793 ex-parlamentares, 506 parentes de ex-parlamentares, 995 ex-funcionários e 463 parentes de ex-funcionários. Existem apenas 17 deputados e senadores que recebem o benefício máximo - R$ 8 mil - pago pelo instituto.

Luís Eduardo do relator. Imediatamente, todos os líderes foram convocados para uma reunião no gabinete do presidente da Câmara. Na avaliação dos líderes e de Luís Eduardo, a polêmica em torno do IPC poderia prejudicar a tramitação da reforma da Previdência. Além disso, o presidente da Central Única dos Trabalhadores (CUT), Vicente Paulo da Silva, o Vicentinho, foi enfático ao defender o fim das aposentadorias privilegiadas dos parlamentares. E a manutenção do instituto poderia ser um pretexto para CUT sair das negociações da reforma da Previdência. "Sou a favor do fim do IPC. É bom que a Câmara dê o exemplo. Isso é um assunto menor que não pode empatar a reforma", afirmou Luís Eduardo. Até ontem à noite, três partidos - PT, PDT e PSDB - fecharam questão a favor da extinção do IPC. O PFL e o PMDB, os dois maiores partidos da Câmara, vão consultar suas bancadas para tomar uma posição. O líder do PFL, deputado Inocêncio Oliveira (PE), garantiu, no entanto, que o partido também será a favor do fim do IPC. "Não tem força humana que salve o instituto. Já consultei 30 deputados pefelistas e apenas um protestou", afirmou. Direitos - Cauteloso, Michel Temer alegou que a extinção do IPC não depende apenas dos líderes. "A aposentadoria é um direito individual e, portanto, as bancadas têm que decidir sobre o fim do instituto", disse Temer. A lei propondo o fim do IPC preservará direitos adquiridos. Todos os parlamentares que tiverem oito anos de mandato e 50 anos de idade terão direito à aposentadoria ou, se preferirem, receberão devolução das contribuições pagas. Os outros receberão de volta as contribuições. Estima-se que os deputados que optarem pela devolução receberão cerca de R$ 40 mil por mandato. Pelos cálculos do governo, se todos os parlamentares optarem pela devolução serão gastos R$ 97 milhões. O IPC tem hoje patrimônio de R$ 150 milhões. Atualmente, o instituto gasta mensalmente R$ 3,8 milhões com o pagamento de 2757 pensões a 793 ex-parlamentares, 506 parentes de ex-parlamentares, 995 ex-funcionários e 463 parentes de ex-funcionários. Existem apenas \17 deputados e senadores que recebem o	

benefício máximo - R$ 8 mil - pago pelo instituto.	

Cobertura → 10,00%
Precisão → 11,11%
Medida-F → 10,53%

BIBLIOGRAFIA

1. Aretoulaki, M. (1996). **"COSY-MATS: A Hybrid Connectionist-Symbolic Approach To The Pragmatic Analysis of Texts For Their Automatic Summarisation"**. *PhD. Thesis*. University of Manchester.

2. Battiti, R. (1992). **"First and second order methods for learning: Between steepest descent and Newton's method"**. *Neural Computation 4*, pp. 141–166.

3. Baxendale, P.B. (1958). Machine-made index for technical literature – an experiment. *IBM Journal of Research and Development*, volume 2, pp. 354-365.

4. Black, W.J. & Johnson, F.C. (1988**). "A Practical Evaluation of Two Rule-Based Automatic Abstraction Techniques"**. *Expert Systems for Information Management*, volume 1, n. 3. Department of Computation. University of Manchester Institute of Science and Technology, pp. 159-177.

5. Chalmers, D. (1990). "**Why Fodor and Pylyshyn were wrong: the simplest refutation"**. *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society,* Cambridge, MA, pp. 340-347.

6. Cunha, C. & Cintra, L.F.L. (2001). **"Nova Gramática do Português Contemporâneo"**. 3ª. edição. Editora Nova Fronteira.

7. Crick, F.H.C. (1989). **"The Recent excitement about neural networks"**. *Nature 337*, pp. 129-132.

8. de Castro, L. N. (1998), **"Análise e Síntese de Estratégias de Aprendizado para Redes Neurais Artificiais"**, *Dissertação de Mestrado*, DCA – FEEC/UNICAMP, Campinas/SP, Brasil.

9. de Castro, L. N., & Von Zuben, F. J. (1998), **"Optimized Training Techniques for Feedforward Neural Networks**", Technical *Report – RT DCA 03/98*, FEEC/UNICAMP, Brazil.

10. de Castro, L.N. (2003). "**Fundamentals of Neurocomputing"**, *Technical Report - RT DCA 01/03*. Disponível para download em: http://www.dca.fee.unicamp.br/~lnunes.

11. Edmundson, H. (1969). **"New Methods in Automatic Abstracting"**. *Jornal of ACM*, volume 16, n. 2.

12. Fellbaum, C.; Palmer, M.; Dang, H.T.; Delfs L.; Wolf, S. (2001). "**Manual and Automatic Semantic Annotation with WordNet"**. In *Proceedings of the NAACL Workshop on WordNet and Other Lexical Resources: Applications, Customizations.* Pittsburgh, PA, pp. 3-10.

13. Feltrim, V.D.; Nunes, M.G.V.; Aluísio, S.M. (2001). "**Um corpus de textos científicos em Português para a análise da Estrutura Esquemática**". *Série de Relatórios do NILC. NILCTR-01-4.* Disponível para download em www.nilc.icmc.usp.br/nilc/pessoas/valeria.htm

14. Fodor, J.A & Pylyshyn, W (1988). **"Connectionism and Cognitive Architecture: A Critical Analysis"**. *Cognition*, volume 28, pp. 3-71.

15. Haykin, S (1999) . **"Neural Networks - A Comprehensive Foundation"**, 2nd edition. Prentice Hall, Upper Saddle River, New Jersey.

16. Hinton, G.E. & McClelland, J.L. 1988. **"Learning representations by recirculation"**, in D.Z. Anderson (Ed.). *Neural Information processing Systems*. New York: American Institute of Physics, pp. 358-366.

17. Jacobs R.A. (1988)**, "Increased Rates of Convergence Through Learning Rate Adaptation"**, *Neural Networks*, volume 1, pp. 295-307,

18. Jurafsky, D. & J. H. Martin (2000). **"Speech and language processing: An Introduction to Natural Language Processing"**, *Computational Linguistics and Speech Recognition*, Prentice-Hall.

19. Kandel, E. R., Schwartz, J. H., & Jessell, T. M. (1995). **"Essentials of Neural Science and Behavior"**. *Appleton & Lange*. Stamford, Connecticut.

20. Kohonen, T. (1982). **"Self-organized formation of topologically correct feature maps"**. *Biological Cybernetics*, volume 43, pp. 59-69.

21. Kupiec, J.; Pedersen, J.; Chen, F. (1995). **"A trainable document summarizer."** *ACM SIGIR,* volume 1, pp. 68-73.

22. Larocca Neto, J.; Santos, A.D.; Kaestner, C.A.A.; Freitas, A.A. (2000). **"Generating Text Summaries through the Relative Importance of Topics"**. In the *Proceedings of the International Joint Conference IBERAMIA/SBIA*, pp. 301-309. Atibaia, SP.

23. Larocca Neto, J.; Freitas, A.A; Kaestner, C.A.A. (2002). **"Automatic Text Summarization Using a Machine Learning Approach"**. In the *Proceedings of the 16th Brazilian Symposium on Artificial Intelligence*, pp. 205-215. Porto de Galinhas/Recife.

24. Luhn, H. P. (1958). **"The automatic creation of literature abstracts"**. *IBM Journal of Research and Development*, volume 2, pp. 159-165."

25. Mani, I. (2001). "**Automatic Summarization"**. *John Benjamins Publishing Co.*, Amsterdam.

26. Mani, I.; Maybury, M.T. (1999).. "**Advances in automatic text summarization."** *MIT Press*, Cambridge, MA.

27. Mazzoni, P; Andersen, R.A.; Jordan, M.I (1991). **"A more biologically plausible learning rule for neural networks"**. *Proceedings of the National Academy of Sciences*, USA, pp. 4433-4437.

28. McClelland, J. L. & Kawamoto, A. H. (1986). **"Mechanisms of Sentence Processing: Assigning Roles to Constituents of Sentences"**. In J. L. McClelland and D. E. Rumelhart (Eds.), *Parallel Distributed Processing – Explorations in the Microstructure of Cognition, Volume 2 – Psychological and Biological Models*. A Bradford Book, MIT Press, pp. 272-325

29. McClelland, J.L., & Rumelhart D.E. (1986), **"Parallel Distributed Processing: Explorations in the Microstructure of Cognition",** *Psychological and Bilogical Models - M.I.T. Press, Cambridge, Massachusetts,* volume 2.

30. Medler, D.A. (1998) **"A Brief History of Connectionism"** *Neural Computing Survey*, volume 1. pp. 61-101

31. Minsky, M.L., & Papert, S. A. (1969). **"Perceptrons"**. Cambridge, MA. MIT Press

32. Módolo, M. (2003). **"Supor: um Ambiente para a Exploração de Métodos Extrativos para a Sumarização Automática de textos em Português"**.*Tese de Mestrado.* Departamento de Computação, UFSCar.

33. Nirenburg, S. & Raskin, V. (2004). **"Ontological Semantics"**. Cambridge, MA: MIT Press (2004).

34. O´Reilly, R.C. (1996). **"Biologically Plausible Error-driven Learning using Local Activation Differences: The Generalized Recirculation Algorithm"**. *Neural Computation*, volume 8, n.5, pp.895-938.

35. O´Reilly, R.C. (1998). **"Six Principles for Biologically-Based Computational Models of Cortical Cognition, Trends in Cognitive Science"**, *Trends in Cognitive Science*, volume 2, pp. 455-462.

36. O'Reilly, R. C., & Munakata, Y. (2000). **"Computational Explorations in Cognitive Neuroscience – Understanding the Mind by Simulating the Brain"**. *A Bradford Book*, The MIT Press, Cambridge, Massachusetts, USA.

37. Paice, C. D. (1981). **"The automatic generation of literature abstracts: an approach based on the identification of self-indicating phrases**. *Proceedings of the 3rd annual ACM conference on Research and development in information retrieval*, Cambridge, England , pp. 172-191.

38. Pardo, T.A.S.; Rino, L.H.M.; Nunes, M.G.V. (2002). **"GistSumm: A Summarization Tool Base don a New Extractive Method"**. *6th Workshop on Computational Processing of the Portuguese Language* – Springer, volume 6, pp. 210-218.

39. Pardo, T.A.S., & Rino, L.H.M. (2003a). **"TeMário: Um Corpus para Sumarização Automática de Textos"**. *Relatório Técnico* – NILCTR-03-09 – ICMC-USP.

40. Pardo, T.A.S; Rino, L.H.M.; Nunes, M.G.V. (2003b). **"NeuralSumm: Uma Abordagem Conexionista para a Sumarização Automática de Textos"**. *Anais do IV Encontro Nacional de Inteligência Artificial*, Campinas-SP, volume 1, pp. 1-10.

41. Pollock, J. J. & Zamora, A. (1975). **"Automatic Abstracting Research at Chemical Abstracts Service".** *Journal of Chemical Information and Computer Sciences,* volume 15. n. 4, pp. 226-232.

42. Rino, L.H.M. & Pardo, T.A.S. (2003). **"A Sumarização Automática de Textos: Principais Características e Metodologias"**. *Anais do XXIII Congresso da Sociedade Brasileira de Computação,* Campinas-SP, volume 8, n. 3., pp. 203-245.

43. Rino, L.H.M.; Pardo, T.A.S.; Silla Jr., C.N.; Kaestner, C.A.; Pombo, M. (2004). **"A Comparison of Automatic Summarization Systems for Brazilian Portuguese Texts"**. In *the Proceedings of the XVII Brazilian Symposium on Artificial Intelligence - SBIA2004*, São Luís, Maranhão, Brazil, pp. 235-244..

44. Rohde, D.L.T., & Plaut, D.C. (1999). **"Language Acquisition in the Absence of Explicit Negative Evidence: How Important is Starting Small?".** *Cognition*, volume 72, pp. 67-109.

45. Rosa, J.L.G. (1993). **"Redes Neurais e Lógica Formal em Processamento de Linguagem Natural"**, *Dissertação de Mestrado,* DCA – FEEC/UNICAMP, Campinas/SP, Brasil

46. Rosa, J.L.G. (2001). **"An Artificial Neural Network Model Based on Neuroscience: Looking Closely at the Brain."** In V. Kurková, N.C. Steele. R. Neruda, and M.Karny (Eds). *Artificial Neural Nets and Genetic Algorihms - Proceedings of the International Conference in Prague*, Czech Republic - ICANNGA-2001. Springer-Verlag, pp. 138-141.

47. Rosa, J.L.G. (2002a). **"A Biologically Motivated Connectionist System for Predicting the Next Word in Natural Language Sentences"**, *Proceedings of the 2002 IEEE International Conference on Systems, Man and Cybernetics* – SMC'2002, Hammamet, Tunisia

48. Rosa, J.L.G. (2002b). **"A Biologically Inspired Connectionist System for Natural Language Processing"**. *Proceedings of the VII Brazilian Symposium on Neural Networks (SBRN´02)*. IEEE Computer Society Press. Brazil, pp. 243-248

49. Rosa, J.L.G. (2002c). **"Next Word Prediction in a Connectionist Distributed Representation System"**. *IEEE International Conference on Systems, Man and Cybernetics. Hammamet*, Tunisia, October 6-9.

50. Rumelhart D.E., McClelland J.L., & the PDP Research Group. (1986). **"Parallel Distributed Processing: Exploration in the Microstructure of Cognition"**, volume 1. *MIT - Press*, Cambridge, Massachussetts.

51. Salton, G. (1989) **"Automatic Text Processing. The Transformation, Analysis and Retrieval of Information by Computer"**. *Addison-Wesley*.

52. Schraudolph, N.N. & Sejnowski, T.J. (1996). **"Tempering Backpropagation Networks: Not All Weights are Created Equal"**. *Advances in Neural Information Processing Systems, MIT Press*, Cambridge, volume 8, pp. 563-569.

53. Shepherd A.J. (1997). **"Second-Order Methods for Neural Networks – Fast and Reliable Methods for Multi-Layer Perceptrons"**, *Springer*.

54. Silber, H.G. & McCoy, K.F. (2002). **"Efficiently Computed Lexical Chains as an Intermediate Representation for Automatic Text Summarization"**. *Computational Linguistics – MIT Press*, volume 28, n. 4, pp. 487-496.

55. Sparck Jones, K. (1999). **"Automatic Summarizing: factors and directions"**. In I. Mani and M. Maybury (eds.), *Advances in automatic text summarization, The MIT Press,* pp. 1-12

56. Teufel, S. & Moens, M. (2002). **"Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status"**. *Computational Linguistics,* volume 28, n. 4, pp. 409-445.

57. Zipser, D. & Andersen, R.A. (1988). **"A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons".** *Nature (London), 331,* pp. 679-684.

58. van Rijsbergen, C.J. (1979). **"Information Retrieval"**. *Butterworth*, London, Segunda Edição.

59. Witten, I.H.; Moffat, A.; Bell, T.C. (1994). **"Managing Gigabytes"**. Van Nostrand Reinhold. New York.