# Uso de Informações Lingüísticas na etapa de pré-processamento em Mineração de Textos

**Cassiana Fagundes da Silva, Fernando Santos Osório, Renata Vieira [1]**

[1]PIPCA – Universidade do Vale do Rio dos Sinos (UNISINOS)
Av. Unisinos, 950 – 93.022-000 – São Leopoldo – RS – Brazil

{cassiana,osorio,renata}@exatas.unisinos.br

***Abstract.*** *This dissertation proposes and evaluates the use of linguistic information in the pre-processing phase for text mining tasks applied to Portuguese texts. We present several experiments comparing our proposal to the usual techniques applied in the field. The results show that the use of linguistic information in the pre-processing phase brings some improvement for both text categorization and clustering.*

***Resumo.*** *Esta dissertação propõe e avalia o uso de informações lingüísticas na etapa de pré-processamento para as tarefas de Mineração de Textos aplicadas a textos em língua Portuguesa do Brasil. Vários experimentos foram realizados comparando nossa abordagem com as técnicas usualmente aplicadas neste campo. Os resultados mostram que o uso de informações lingüísticas na etapa de pré-processamento apresenta melhorias para ambas tarefas: categorização e agrupamento de textos.*

## 1. Introdução

Textos em língua natural podem ser modelados como base de dados uniformes de tal maneira que métodos similares aos usados na extração de conhecimento em base de dados podem ser aplicados aos textos. A adaptação destes métodos aos textos é conhecida como Mineração de Textos (Tan, 1999).

O objetivo deste trabalho é propor uma nova abordagem para a etapa de pré-processamento dos textos para tarefas de mineração. Usualmente, o pré-processamento é composto de etapas baseadas na abordagem *bag-of-words*, que utiliza técnicas simples de remoção de termos irrelevantes e redução de termos ao seu radical. Como alternativa, propomos o uso de informações lingüísticas na etapa de pré-processamento, para a seleção por categorias gramaticais (substantivos, adjetivos, nomes próprios) para duas tarefas de MT: categorização e agrupamento de textos.

A tarefa de categorização visa identificar os tópicos principais em um documento e associar este documento a uma ou mais categorias pré-definidas (Yang and Pedersen, 1997), enquanto que o agrupamento busca agrupar, um conjunto de exemplos de acordo com a similaridade ou dissimilaridade de seu conteúdo (Theodoridas and Koutroumbas, 1998).

Este artigo esta organizado como segue. A seção 2 apresenta os métodos usados para extração de conhecimento lingüístico no processo de mineração de textos. A metodologia utilizada para a realização dos experimentos é apresentada na seção 3. A

seção 4 apresenta a análise dos resultados e por fim, as conclusões são discutidas na seção 5.

## 2. Informações Lingüísticas em Mineração de Textos

A etapa de pré-processamento é normalmente baseada na abordagem *bag-of*-words, mas com as técnicas disponíveis de Processamento de Linguagem Natural (PLN) torna-se possível fazer um pré-processamento baseado em análise sintática ou analise semântica. Neste contexto, este trabalho abordou um pré-processamento baseado na análise sintática dos textos.

Para extração de informações lingüísticas (termos sintáticos), foi utilizado o analisador sintático PALAVRAS [Bick 2000]. A Figura 1 mostra a saída do analisador sintático para a sentença "*Janeiro começa com grandes liqüidações"*.

Outra ferramenta denominada PALAVRAS Xtractor [Gasperin et. al. 2003] foi utilizada para extrair as características dos textos analisados. Esta ferramenta converte a saída do analisador sintático em três arquivos XML contendo: uma lista de palavras do texto e seu identificador (Figura 2); informações morfo-sintática (POS) para cada palavra listada (Figura 3); e a estrutura das sentenças.

```
STA:fcl
=SUBJ:n('janeiro' M S)Janeiro
=P:v-fin('começar'PR 3S IND)começa
=ADVL:pp
==H:prp('com')com
==P<:np
===>N:adj('grande' F P)grandes
===H:n('liquidação'F P)liquidações
=.
```

**Figura 1. Saída PALAVRAS**

```
<words>
<word
id="word_1">Janeiro</word>
<word
id="word_2">começa</word>
<word id="word_3">com</word>
<word
id="word_4">grandes</word>
<word id="word_5">liquidações
</word>
<word id="word_6">.</word>
</words>
```

**Figura 2. Arquivo de *Words***

```
<words>
<word id="word_1">
<n canon="janeiro" gender="M" number="S"/>
</word>
<word id="word_2">
<v canon="começar">
<fin tense="PR" person="3S" mode="IND"/>
 ...
```

**Figura 3. Arquivo de POS**

A extração de combinações gramaticais é realizada utilizando folhas de estilos XSL[1] (*eXtensible Stylesheet Language*). As combinações gramaticais extraídas neste trabalho são as seguintes: substantivos; substantivos e adjetivos; substantivos e nomes próprios; substantivos, nomes próprios e adjetivos; e nomes próprios e adjetivos. As

[1] Linguagem desenvolvida pelo W3C disponível em: http://www.w3.org/Style/XSL/

listas de termos resultantes de acordo com cada combinação gramatical são passadas para a fase de preparação e seleção dos dados.

Na próxima seção são descritos os métodos aplicados aos experimentos para as tarefas de categorização e agrupamento de textos no processo de MT.

## 3. Experimentos

Os experimentos realizados neste trabalho avaliam e comparam tarefas de MT com base em dois diferentes métodos de pré-processamento: usual e baseado na extração de informações lingüísticas.

### 3.1 Corpus

O *corpus* utilizado para realização dos experimentos é um subconjunto do corpus fornecido pelo Núcleo Interinstitutional de Lingüística Computacional - NILC[2]. Foram escolhidos 855 textos extraídos da Folha Jornal de São Paulo do ano de 1994, e classificados em cinco seções (171 documentos por seção): Informática, Imóveis, Esporte, Política e Turismo.

### 3.2 Pré-processamento para Categorização de Textos

Para que o resultado do processo de categorização não seja tendencioso, o corpus foi organizado em 3-*folds*, visando verificar a variação dos resultados em diferentes distribuições de treino e teste. Para cada distribuição o corpus foi dividido em dois conjuntos: treino e teste (2/3 e 1/3 de documentos, respectivamente). Apesar de ser usual o uso de *10-fold cross validation*, a divisão em treino e teste em 2/3 e 1/3 adotada permite uma avaliação com um conjunto de teste mais expressivo.

Após a definição e divisão do corpus, os exemplos selecionados foram pré-processados contemplando as sub-etapas correspondentes aos métodos usuais de pré-processamento: análise léxica e remoção dos termos irrelevantes. Uma lista de termos irrelevantes (*stopwords*), contendo 476 termos (tais como: artigos, preposições, verbos ser e estar, pronomes, entre outros) foi utilizada. Dos termos resultantes, foram extraídos os afixos. Para esta extração foi adotado o algoritmo proposto por Martin Porter[3], que remove as letras finais dos termos da língua portuguesa que possuem a mesma variação morfológica e de flexão.

Para o pré-processamento usando informações lingüísticas foi adotada a metodologia proposta na seção 2, baseada na análise sintática dos textos e seleção de estruturas através de folhas de estilos XSL.

Para ambos pré-processamentos (métodos usuais e informações lingüísticas), os termos mais relevantes foram identificados nos conjuntos de treino com o auxílio de scripts implementando o cálculo de freqüência relativa e Tf-Idf. Para seleção dos termos mais relevantes adotou-se a técnica de truncagem. A representação dos documentos foi elaborada com base no espaço vetorial. Vetores locais foram construídos com base nos termos restantes do pré-processamento (métodos usuais ou informações lingüísticas). Cabe destacar que os vetores locais foram elaborados a partir dos *n* termos mais

[2] Disponível em http://www.nilc.icmp.usp.br/nilc
[3] Disponível para várias línguas em http://snowball.sourceforge.net

freqüentes do conjunto de treinamento correspondentes aos documentos de cada categoria. Dessa forma, os vetores globais foram construídos através da união dos vetores locais e serviram como índices para os vetores de cada exemplo e as posições correspondentes representaram a importância da mesma no documento.

Após a codificação dos exemplos, foram gerados os scripts para a ferramenta de suporte *Weka*[4]. As técnicas de aprendizado adotadas neste trabalho para o processo de categorização de textos são: Árvores de Decisão e Redes Neurais Artificiais.

### 3.3 Pré-processamento para Agrupamento de Textos

O pré-processamento no agrupamento de textos é constituído basicamente das mesmas etapas adotadas no processo de categorização. Porém, a distribuição do número de documentos do conjunto de treino e teste para as 3 versões são 570 e 285, respectivamente. Novamente os termos mais relevantes foram escolhidos utilizando o calculo freqüência relativa e selecionados através da técnica de truncagem.

Geralmente, todo processo de agrupamento está baseado em algum tipo de similaridade entre os documentos, pois os agrupa (ou separa-os) em grupos de documentos que possuam alguma semelhança entre si. A medida de similaridade adotada neste trabalho, é baseada na Distancia Euclidiana utilizando a função co-seno (Cole, 1998). Depois de realizado o cálculo de similaridade entre os vetores de documentos e gerados os vetores de entrada para a ferramenta *Weka*, estes são submetidos ao algoritmo *K-means* adotado com o intuito de agrupar os documentos similares. Os parâmetros utilizados no algoritmo, são os sugeridos pela ferramenta, com semente aleatória igual a 10 e o número de grupos igual a 5, correspondente as seções dos textos.

### 3.4 Avaliação e Interpretação dos Resultados

Para avaliação dos resultados no processo de categorização foram observados o Menor Erro de Classificação para os classificadores treinados por Árvores de Decisão e Redes Neurais, gerados a partir da matriz de confusão resultante como saída da ferramenta *Weka*. Na avaliação do processo de agrupamento também é utilizada a matriz de confusão com os exemplos pertencentes a cada grupo.

A comparação de desempenho entre as técnicas de aprendizado supervisionado e não-supervisonado é realizada por meio da comparação do desempenho do melhor classificador obtido por cada técnica para o corpus em termos do número de atributos contidos no vetor global. Por fim, para se determinar qual a técnica de aprendizado que melhor se adequa aos termos resultantes do pré-processamento baseado em métodos usuais e em informações lingüísticas, serão comparados os menores erros de classificação para a seleção dos termos em comum.

---

[4] Ferramenta formada por uma coleção de algoritmos de Aprendizado de Máquina para resolução de problemas reais de Mineração de Dados (MD), disponível em http://www.cs.waikato.ac.nz/ml/weka/

## 4. Resultados

### 4.1 Categorização de Textos

Para o processo de categorização foi adotada a categorização múltipla dos exemplos, onde uma classe de um conjunto de classes é definida para cada documento. Os algoritmos de aprendizado utilizados no treinamento do categorizador foram: Árvores de Decisão e Redes Neurais Artificiais implementados na ferramenta *Weka*. Para os experimentos foram variados os números de termos selecionados por categoria correspondentes aos 30, 60, 90, 120 e 150 termos que constituem os vetores globais.

#### 4.1.1 Pré-processamento baseado em Métodos Usuais

Os primeiros experimentos foram realizados com as Árvores de Decisão (ADs). A média dos resultados percentuais obtidos para as variações do corpus (V1, V2 e V3) são mostrados na Tabela 1.

**Tabela 1. Média do Erro de Classificação para as Variações do Corpus**

| | Número de Termos | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 30 | | 60 | | 90 | | 120 | | 150 | |
| Codificação | Erro | Nodos | Erro | Nodos | Erro | Nodos | Erro | Nodos | Erro | Nodos |
| F.Relativa | 21,64 | 70 | 21,99 | 73 | 20,47 | 71 | 20,35 | 74 | **19,77** | 73 |
| TF-IDF | 21,75 | 67 | 21,17 | 74 | 19,88 | 71 | **19,18** | 77 | 20,12 | 76 |

Analisando a Tabela anterior pode-se verificar que a variação do número de termos relevantes implicou em alguns casos no aumento e em outros na diminuição do erro de classificação. No entanto, os erros de classificação obtidos através das codificações: freqüência relativa e Tf-Idf, não apresentaram muita diferença, isso pode ser devido ao fato do corpus não apresentar uma grande variação no tamanho dos documentos. Assim, optou-se pelo uso da codificação por freqüência relativa nos experimentos restantes.

Os experimentos seguintes foram conduzidos com as Redes Neurais Artificiais (RNAs), variando o número de termos relevantes selecionados de acordo com o experimento anterior. A RNA utilizada foi a MLP, sendo aplicado o algoritmo de aprendizado *Backpropagation* (BP). Nos experimentos com o algoritmo BP, foi utilizado um valor de 0.9 para o *momentum,* 0.1 para a taxa de aprendizado e variado o número de neurônios na camada intermediária (8 e 16). Para cada topologia, foram realizadas 10 simulações, variando-se a semente aleatória e mantendo-se os demais parâmetros de configuração da rede.

A Tabela 2 apresenta os valores médios (para as 10 simulações) dos menores erros de generalização (%) obtidos no processo de aprendizagem para as três variações do corpus, Nº.Neurônios corresponde ao número de neurônios na camada intermediária.

Observando os resultados da Tabela 2, verifica-se que o menor erro (14,64%) foi obtido utilizando-se 8 neurônios na camada intermediária, e um número de 90 termos relevantes. Nota-se também uma grande variação no erro de generalização à medida que o número de exemplos de entrada da rede aumenta. Essa variação no erro é decorrente do aumento no número de exemplos de entrada, pois muitas informações semelhantes

são apresentadas a rede e esta não consegue distinguir o que pertence a uma classe distinta, classificando muitos padrões de forma errada.

**Tabela 2. Média do menor erro de generalização para as variações do corpus**

| | Número de Termos | | | | |
|---|---|---|---|---|---|
| | 30 | 60 | 90 | 120 | 150 |
| Nº.Neurônios | Erro | Erro | Erro | Erro | Erro |
| 8 | 20,26 | 16,56 | **14,64** | 32,32 | 55,37 |
| 16 | 20,68 | 16,74 | 16,19 | 24,28 | 54,23 |

Uma comparação entre as técnicas simbólicas e conexionistas foi realizada para comparar o resultado do pré-processamento baseado em métodos usuais em ambas técnicas de aprendizado (Tabela 3).

**Tabela 3. Comparação entre Árvores de Decisão e Redes Neurais %**

| Técnica | Menor erro | Média global |
|---|---|---|
| Árvores de Decisão | 19,18 | 20,84 |
| RNA MLP-BP | 14,64 | 27,12 |

Conforme comparação entre as técnicas de aprendizado, a melhor taxa de classificação para o pré-processamento baseado em métodos usuais, foi obtida com o algoritmo RNA MLP-BP. As taxas de erro das RNAs foram obtidas a partir do ajuste fino de seus parâmetros (com base em experimentos preliminares). Porém, observou-se uma maior dependência das RNAs aos seus parâmetros do que a apresentada pelas Árvores de Decisão, que apresentaram de uma forma mais geral um resultado mais estável. Comparando-se a média dos resultados dos experimentos com ADs e RNAs temos 20,84% e 27,12%, respectivamente.

### 4.1.2 Pré-processamento baseado em Informações Lingüísticas

Os experimentos realizados com a extração de informações lingüísticas na etapa de pré-processamento seguem as mesmas variações do experimento anterior, primeiramente são apresentados os resultados para a geração de Árvores de Decisão e logo após para as RNAs para cada uma das categorias gramaticais adotadas. A Tabela 4 mostra a média dos resultados, conforme a categoria gramatical testada.

A combinação gramatical Substantivos-Adjetivos foi a que apresentou o menor erro de generalização para os 90, 120 e 150 termos mais relevantes. Enquanto que as combinações gramaticais: Substantivos e Substantivos–Nomes próprios apresentam os resultados semelhantes ou piores aos métodos usuais de pré-processamento (Tabela 4).

Para o treinamento da RNA com o uso de informações lingüísticas adotamos os mesmos parâmetros dos experimentos realizados com o pré-processamento baseado em métodos usuais. A Tabela 5 apresenta os valores médios (para as 10 simulações) dos menores erros de generalização para as três variações do corpus de acordo com as categorias gramaticais. A seleção gramatical que apresenta o menor erro de generalização é a combinação de substantivos com os complementos adjetivos e nomes próprios.

**Tabela 4. Média da Taxa de Erro para as categorias gramaticais**

| | Número de Termos | | | | |
|---|---|---|---|---|---|
| Categorias Gramaticais | 30 | 60 | 90 | 120 | 150 |
| Substantivos | 24,91 | **21,75** | 23,98 | 23,51 | 22,69 |
| Substantivos- Nomes próprios | 24,09 | 24,56 | 22,80 | **22,45** | 22,80 |
| Substantivos-Adjetivos | 23,15 | 20,35 | **18,01** | 19,18 | 18,71 |
| Substantivos-Adjetivos-Nomes próprios | **20,82** | 22,92 | 20,94 | 21,05 | 21,17 |
| Adjetivos-Nomes próprios | **47,01** | 46,34 | 32,51 | 33,21 | 32,86 |

**Tabela 5. Média do Erro de Generalização**

| | Número de Termos | | | | |
|---|---|---|---|---|---|
| Categoria Gramatical/Números Neurônios | 30 | 60 | 90 | 120 | 150 |
| Substantivos /8 | 22,85 | 19,49 | 21,54 | 48,01 | 71,13 |
| Substantivos /16 | 23,09 | **19,22** | 20,27 | 54,38 | 67,54 |
| Substantivos-Adjetivos/8 | 21,86 | 19,66 | 21,59 | 34,82 | 56,38 |
| Substantivos-Adjetivos/16 | 22,56 | 19,18 | **19,05** | 32,35 | 56,52 |
| Substantivos-Nomes Próprios/8 | 22,27 | 19,18 | 21,62 | 49,15 | 65,78 |
| Substantivos-Nomes Próprios/16 | 22,94 | 19,77 | **19,01** | 38,82 | 66,58 |
| Substantivos-Adj-Nomes Próprios/8 | 21,39 | 18,00 | 17,55 | 33,29 | 56,46 |
| Substantivos-Adj-Nomes Próprios/16 | 22,44 | **16,96** | 18,27 | 37,98 | 54,21 |
| Adjetivos-Nomes Próprios/8 | 49,45 | 38,25 | **35,57** | 45,17 | 72,77 |
| Adjetivos-Nomes Próprios/16 | 48,53 | 38,62 | 37,75 | 47,63 | 74,64 |

A combinação gramatical Adjetivos-Nomes próprios apresentaram os priores resultados se comparado as demais estruturas. A menor taxa de erro (35,57%) foi obtida utilizando os 90 termos mais relevantes e 8 neurônios na camada intermediária. Assim, observa-se que a utilização dos substantivos é de extrema importância para o processo de categorização dos documentos.

Quanto ao tempo de treinamento dos classificadores para a técnica de aprendizado adotada, o treinamento da Rede MLP-BP resultou em aproximadamente 1068 segundos para a menor taxa de erro da seleção dos termos Substantivos-Adj-Nomes Próprios.

Um comparativo entre os melhores resultados das técnicas simbólicas e conexionistas em relação aos pré-processamentos adotados é mostrado na Tabela 6.

**Tabela 6. Comparativo do Menor Erro para as técnicas de aprendizado**

| Técnicas | Taxa | Combinação Gramaticcal | Nº Termos |
|---|---|---|---|
| Árvores de Decisão | 18,01 | Substantivos-Adjetivos | 90 |
| | 20,82 | Substantivos-Adj-Nomes Próprios | 30 |
| RNA-BP (16 neurônios) | 16,96 | Substantivos-Adj-Nomes Próprios | 60 |

## 4.2 Agrupamento de Textos

Para os experimentos realizados com os termos extraídos do pré-processamento usual e em informações lingüísticas no processo de agrupamento de textos foram consideradas

variações nos seguintes parâmetros: número de termos relevantes, variação do corpus, o tipo de codificação e o algoritmo de agrupamento *K-means*. Diversos experimentos foram feitos adotando-se um vetor global representado pelo modelo de espaço vetorial, utilizando a codificação por freqüência relativa, e a seleção dos 30, 90 e 150 termos relevantes para o conjunto de treinamento.

### 4.2.1 Pré-processamento baseado em Métodos Usuais

Foram realizados 9 experimentos de agrupamento baseado no pré-processamento usual com os parâmetros: semente randômica igual a 10 e o número grupos igual a 5. Desses 9 experimentos em apenas um dos casos foi possível identificar dois grupos com mais clareza. A Tabela 7 apresenta a matriz de confusão para esse resultado obtido com os 150 termos relevantes e versão V2 do corpus.

**Tabela 7. Matriz de Confusão para a versão V2 do corpus e 150 termos**

| | Cluster 0 | Cluster 1 | Cluster2 | Cluster 3 | Cluster 4 |
|---|---|---|---|---|---|
| Esporte | 1 | **31** | 2 | 0 | 23 |
| Imóveis | 2 | 0 | 4 | 0 | 51 |
| Informática | 0 | 0 | 1 | 0 | 55 |
| Política | 0 | 0 | 2 | **39** | 16 |
| Turismo | 5 | 0 | 17 | 0 | 33 |

Com base nos resultados obtidos para o agrupamento utilizando os métodos tradicionais de pré-processamento, pode-se verificar que, adotando o algoritmo *k-means* com 5 grupos, apenas 2 foram identificados. Devido aos resultados dos experimentos de categorização, onde as categorias Turismo e Imóveis apresentaram sempre as piores medidas, optou-se por eliminar as seções Imóveis e Turismo do conjunto de treino e teste, refazendo os experimentos com os mesmos parâmetros adotados anteriormente. Para esses novos experimentos foi possível identificar três grupos distintos, para as versões V1 e V2 do corpus. Como são variados os documentos que pertencem às versões do corpus, pode-se verificar que os termos selecionados para o agrupamento na versão V3 não foram significativos a ponto de conseguir identificar os grupos Esporte e Política.

Com as etapas usuais de pré-processamento, não foi possível identificar os 5 grupos correspondentes às seções do corpus jornalístico. Isso pode ser decorrente da similaridade dos termos que representam o conteúdo dos assuntos referentes a imóveis e turismo, pois, com a eliminação destas seções, foi possível identificar 3 grupos distintos utilizando os mesmos parâmetros adotados nos primeiros experimentos.

### 4.2.2 Pré-processamento baseado em Informações Lingüísticas

Para os experimentos com informações lingüísticas foram realizados 9 experimentos, objetivando encontrar cinco grupos de documentos correspondentes aos assuntos de Esporte, Imóveis, Informática, Política e Turismo. Utilizando a categoria gramatical substantivos apenas dois grupos referentes aos assuntos de Política e Turismo foram identificados nas versões V1 e V2 do corpus para o número de 150 termos selecionados.

Para a combinação Substantivos–Nomes próprios. foi possível identificar 3 grupos distintos para os 90 e 150 termos mais relevantes. A Tabela 8 apresenta a matriz de confusão para os 90 termos mais relevantes.

**Tabela 8. Matriz de Confusão da versão V2 do Corpus para os 90 termos**

| | Grupo 0 | Grupo 1 | Grupo 2 | Grupo 3 | Grupo 4 |
|---|---|---|---|---|---|
| Esporte | 26 | **31** | 0 | 0 | 0 |
| Imóveis | 43 | 0 | 13 | 0 | 1 |
| Informática | 25 | 0 | 0 | 0 | **32** |
| Política | 21 | 0 | 0 | **36** | 0 |
| Turismo | 57 | 0 | 0 | | 0 |

As combinações gramaticais Substantivos-Adjetivos e Substantivos-Adjetivos-Nomes próprios também possibilitaram a identificação de três grupos distintos para os assuntos referentes a Esporte, Informática e Política. Assim como nos experimentos do pré-processamento usual as seções Imóveis e Turismo foram eliminadas e novos experimentos com os 150 termos mais relevantes para a categoria gramatical Substantivos-Adjetivos-Nomes próprios foram realizados.

Pode-se observar que os termos substantivos-nomespróprios-adjetivos, foram os mais representativos para a formação dos grupos nas versões V1 e V3 do corpus. Por fim, pode-se verificar que a seleção de informações lingüísticas resultou na identificação de um numero maior de grupos no conjunto de experimentos.

## 5. Conclusões

O foco principal deste trabalho foi avaliar um método alternativo considerando a seleção de várias combinações gramaticais para as tarefas de categorização e agrupamento no processo de Mineração de Textos.

Analisando os resultados obtidos, pode-se verificar que o uso de informações lingüísticas na fase de pré-processamento apresentou-se como uma alternativa equivalente em ambas tarefas de categorização e agrupamento de textos. Os algoritmos de aprendizado simbólico mostraram um menor erro se comparado aos métodos usuais de pré-processamento, com uma taxa de erro de 18,01% para a seleção de substantivos e adjetivos, enquanto que o melhor resultado para os métodos usuais foi de 19,18%. Para o aprendizado utilizando RNA MLP-BP, no entanto, os métodos usuais apresentaram um melhor desempenho 14,64% para os 90 termos mais relevantes contra a taxa de erro de 16,96% para a combinação Substantivos-Adjetivos-Nomes próprios utilizando 60 termos relevantes. As RNAs apresentaram resultados com uma grande variação no erro de generalização à medida que o número de exemplos de entrada da rede aumentava. Considerando-se a média global, as RNAs apresentaram um erro maior do que as Árvores de Decisão.

O estudo realizado possibilitou verificar a importância da categoria gramatical substantivos na seleção dos termos relevantes, bem como na combinação dessa com os complementos adjetivos e nomes próprios. Nos experimentos realizados com o agrupamento, as combinações gramaticais permitiram a identificação de um número maior de grupos de documentos, similares às classes conhecidas de antemão, se

comparado aos métodos usuais de pré-processamento utilizando todas as seções do corpus.

Como trabalho futuro pretende-se comparar a metodologia proposta com a extração de sintagmas nominais simples (2 ou 3 palavras) dos textos, bem como a extração de n-gramas. Também queremos comparar os resultados obtidos com outros algoritmos de aprendizado, por exemplo SVM, devido a sua robustez em grandes números de termos. SVM tem sido aplicado para documentos utilizando o português Europeu [Gonçalves and Quaresma, 2003], porém o pré-processamento baseado em combinações gramaticais ainda não foi testado para o processo de categorização. Outros estudos poderiam ser a aplicação de outras línguas ou outros domínios.

## References


Bick, E. (2000) "The Parsing System PALAVRAS: Automatic Gramatical Analysis of Porutugese in a Constraint Grammar Framework", Århus University. Århus: Århus University Press.

Cole, R. M. (1998) "Clustering with Genetic Algorithms", Disponível em: http://www.cs.uwa.edu.au/pub/robvis/theses/RowenaCole, acessado em abril/2003.

Fayyad, U. M., et al. (1996) "From Data Mining to Knowledge Discovery:An Overview", In Advances in Knowledge Discovery and Data Mining, Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R., editors, 1--36. AAAI Press, Menlo Park, CA.

Gasperin, C.; Vieira, R.; Goulart, R. and Quaresma, P. (2003) "Extracting XML Syntactic Chunks from Portuguese Corpora", Proceedings of the TALN 2003 Workshop Natural Language Processing of Minority Languages and Small Languages - Batz-sur-Mer France June 11 – 14.

Gonçalves, T.; Quaresma, P. (2003) "A prelimary approach classification problem of Portuguese juridical documents", In: 11º Portuguese Conference on Artificial Intelligence Lectures Notes In: 11th Portuguese Conference on Artificial Intelligence, Béja. - Lecture Notes in Artificial Intelligence. Berlim: Springer Verlag

Tan, A.H. (1999) "Text Mining: The state of the art and the challenges", In Proc. of the PAKDD'99 workshop on Knowledge Discovery from Advanced Databases, pages 65-70, Beijing.

Theodoridas, S.; Koutroumbas, K. (1998) "Pattern Recognition", Academic Press.

Yang, Y.; Pederson, J. P. (1997) "A comparative study on feature selection in text categorization," in Fourteenth International Conference on Machine Learning (ICML'97), pp. 412--420.