Uso de Informações Lingüísticas em Categorização de Textos utilizando Redes Neurais Artificiais

Cassiana Fagundes da Silva, Renata Vieira, Fernando Santos Osório
PIPCA – UNISINOS/RS
Av. Unisinos, 950 – 93.022-000 – São Leopoldo, RS – Brasil
{cassiana, renata, osorio}@ exatas.unisnos.br

Abstract

This paper evaluates the use of linguistic information for categorization of Portuguese texts using neural networks. We present several experiments with different feature selection strategies, based on distinct grammatical groups. The results show clearly which grammatical categories are more relevant for text categorization using neural networks.

1. Introdução

A quantidade de documentos eletronicamente armazenados e a sobrecarga de informações são problemas que tem motivado o crescimento de áreas como Recuperação de Informação (RI) e Mineração de Textos (MT).

Um dos objetivos da MT é descobrir conhecimento útil nos textos e classificar estas informações automaticamente em categorias de documentos, facilitando sua visualização, manipulação e análise, está tarefa é denominada categorização de documentos.

Usualmente, o pré-processamento utilizado nesta tarefa consiste na remoção de termos irrelevantes (tais como, preposição, artigos, pronomes, entre outros), redução de afixos (termo reduzido ao seu radical) e seleção dos termos (palavras mais relevantes para representar cada documento).

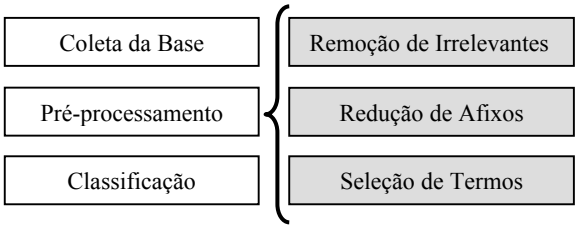
Este trabalho tem como objetivo adotar um pré-processamento alternativo, baseado na seleção de combinações gramaticais, visando avaliar o efeito do uso dessas informações nos resultados do processo de Categorização de Textos. Vários experimentos foram realizados utilizando a técnica de aprendizado de máquina Redes Neurais Artificiais (RNA) do tipo *Multi-layer Perceptron* (MLP) utilizando o algoritmo *Backpropagation* (BP).

O presente trabalho esta organizado conforme segue. Seção 2 apresenta uma visão geral do processo de Categorização de Textos. A seção 3 apresenta as ferramentas adotadas para adquirir conhecimentos lingüísticos. A metodologia aplicada aos experimentos e os resultados obtidos com a abordagem proposta são apresentados nas seções 4 e 5, respectivamente. Por fim, na seção 6 são feitas as considerações finais e relacionadas as propostas para trabalhos futuros.

2. Processo de Categorização de Textos

Categorizar documentos é classificá-los em uma ou mais categorias pré-existentes [2]. O processo de categorização, é formado por um conjunto de etapas fundamentais, conforme ilustra a Figura 1 e descritas a seguir.

Figura 1. Etapas do Processo de Categorização



A Coleta da Base consiste na busca de documentos relevantes ao domínio de aplicação do conhecimento a ser extraído. As etapas de pré-processamento contemplam uma seqüência de ações que transformam o conjunto de documentos em linguagem natural em uma lista de termos úteis. Por fim, é escolhida uma técnica de aprendizado de máquina e a codificação de cada documento é realizada. A codificação geralmente adota uma representação vetorial, onde as palavras selecionadas como relevantes tornam-se índices de vetores e os respectivos valores indicam a importância do termo no documento correspondente.

O pré-processamento no processo de categorização de textos é considerado uma etapa essencial e muito custosa. Os textos são originalmente não-estruturados

e uma série de passos são necessários para transformar os textos em um formato compatível para a extração de conhecimento. Os passos mais utilizados são: remoção de termos irrelevantes (*stopwords*), redução de afixos e seleção dos termos. Neste trabalho alternativamente investigamos o uso de informações lingüísticas na fase de pré-processamento verificando a repercussão das diferentes categorias gramaticais nos resultados de categorização.

Na seção que segue são descritos as ferramentas utilizadas para extração de conhecimento lingüístico para os experimentos realizados.

3. Ferramentas para extrair Conhecimento Lingüístico

O conhecimento lingüístico utilizado nos experimentos é baseado nos resultados de um analisador sintático denominado PALAVRAS [3]. Este analisador sintático é bastante robusto, pois possibilita a análise sintática de sentenças incorretas ou até mesmo incompletas. Uma vez analisados gramaticalmente os textos, é possível selecionar termos de acordo com diferentes combinações gramaticais e extrair esses termos em sua forma canônica[1]. A Figura 1 mostra a saída do analisador sintático para a sentença "*Janeiro começa com grandes liquidações"*.

Figura 1. Saída do analisador sintático PALAVRAS

```
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
=.
```

Outra ferramenta denominada Palavras Xtractor [4] foi utilizada para extrair facilmente características dos textos analisados. Esta ferramenta converte a saída do analisador sintático em três arquivos XML contendo: i) uma lista de palavras do texto e seu identificador (Figura 1), ii) informações morfo-sintática para cada palavra listada (Figura 2) e iii) a estrutura das sentenças (Figura 3).

Figura 2. *Words*

```
<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>
```

Figure 3. POS *(Part-of-Speech)*

```
<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"/>
</v>
</word>
...
```

Figure 4. *Chunks*

```
<text>
<paragraph id="paragraph_1">
<sentence id="sentence_1"
span="word_1..word_6">
  <chunk id="chunk_1" ext="sta"
  form="fcl" span="word_1..word_5">
  <chunk id="chunk_2" ext="subj"
  form="n" span="word_1">
</chunk>
...
```

A extração de combinações gramaticais é realizada utilizando folhas de estilos XSL[2] (*eXtensible Stylesheet Language*). As combinações gramaticais (cada combinação gramatical corresponde a um experimento) extraídas neste estudo são: substantivos; substantivos e adjetivos; substantivos e nomes próprios; substantivos, nomes próprios e adjetivos; nomes próprios e adjetivos. As listas de termos resultantes de acordo com cada combinação gramatical são passadas para a fase de categorização para serem submetidas a técnica de aprendizado de máquina adotada.

A próxima seção descreve a metodologia utilizada nos experimentos utilizando Redes Neurais Artificiais (RNA).

[1] A forma canônica de uma palavra significa a forma base de tal palavra, sem flexões de gênero, número ou grau.

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

4. Metodologia

Foram realizados vários experimentos para avaliar o efeito da seleção por informações lingüísticas no resultado do processo de categorização.

O corpus utilizado para a realização dos experimentos pertence a um extrato do conjunto de textos NILC (Núcleo Interinstitucional de Lingüística Computacional[3]) contendo 855 documentos correspondendo a artigos jornalísticos das seções: esporte, imóveis, informática, política e turismo. Três versões do mesmo corpus foram preparadas (V1, V2 e V3) buscando verificar a variação dos resultados em diferentes distribuições, sendo que cada versão é particionada em diferentes conjuntos de treino e teste, contendo 2/3 e 1/3 dos documentos, respectivamente.

A Tabela 1 mostra a distribuição dos documentos do conjunto de treino e teste para cada seção. Onde 'Esp', 'Im', 'Inf', 'Pol' e 'Tur' representam as categorias Esporte, Imóveis, Informática, Política e Turismo, respectivamente.

Tabela 1 . Distribuição dos Documentos

Categoria	Esp	Im	Inf	Pol	Tur	Total
B. Treino	114	114	114	114	114	570
B. Teste	57	57	57	57	57	285

Os 855 documentos foram pré-processados da seguinte maneira: análise sintática dos textos, geração de arquivos XML correspondentes e extração dos termos de acordo com a categoria gramatical usando XSL.

Os termos mais relevantes foram identificados no conjunto de treino com base no cálculo de freqüência relativa. Adotou-se como seleção de termos a técnica de *truncagem*, já que, além de ser implementada mais facilmente, não influi negativamente nos resultados, conforme testes realizados por Schütze [7]. Uma vez selecionados os termos que melhor representam cada categoria da coleção de documentos, estes foram numerados em ordem crescente e dispostos de forma seqüencial. Esta relação de termos resultante constituiu o *vetor local* de cada categoria.

Baseado nos *vetores locais* das categorias, foi construído um *vetor global* adotando a representação de documentos baseada no espaço vetorial baseado na categorização múltipla dos exemplos. O *vetor local* é consultado para que sejam obtidas as posições correspondentes que representam as características desejadas. Sendo que estas posições já definem o número de neurônios na camada de entrada da rede. Por exemplo, neste trabalho em que existem 5 categorias, decidiu-se por representar cada categoria com os 6, 12, 18, 24 e 30 termos mais freqüentes, totalizando 30, 60, 90, 120 e 150 termos no vetor global, respectivamente.

Uma vez concluída a etapa de construção dos *vetores globais*, estes foram submetidos à ferramenta *Weka*[4] [8] para o treinamento da rede neural *Multi-layer Perceptron* (MLP), sendo o algoritmo de aprendizado *Backpropagation* (BP).

As topologias das redes neurais são formadas por 3 ou mais camadas. O número de neurônios na camada de entrada corresponde ao número de termos no vetor global (dependente do número de termos relevantes selecionadas); o número de neurônios da camada intermediária varia conforme o algoritmo de aprendizado utilizado, no algoritmo BP as camadas intermediarias são definidas manualmente durante o treinamento da rede; e o número de neurônios na camada de saída (5) corresponde às categorias: Informática, Imóveis, Esporte, Política e Turismo. A condição de parada do aprendizado foi o número máximo de 3000 épocas, baseado em experimentos anteriores.

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 em 2, 4, 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 avaliação dos resultados no processo de categorização é baseada no erro de classificação de diferentes grupos gramaticais.

A seguinte seção apresenta os resultados obtidos para os experimentos realizados.

5. Resultados

Os experimentos realizados apresentam diferentes resultados para diversos grupos gramaticais: substantivos (Sub), substantivos e adjetivos (Sub-Adj), substantivos, adjetivos e nomes próprios (S-A-Npp), substantivos e nomes próprios (Sub-Npp) e adjetivos e nomes próprios (Adj-Npp).

Os resultados obtidos na Tabela 2 são baseados nos valores médios (para as 10 simulações) dos menores

[3] Disponível em http://www.nilc.org

[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/

erros de generalização obtidos no processo de aprendizagem para as três variações do corpus (3-*fold-cross-validation*: V1, V2 e V3) de acordo com as estruturas gramaticais. Na Tabela 2, Categ. equivale a categoria gramatical testada e N.N. corresponde ao número de neurônios na camada intermediária e.

Tabela 2. Média do Erro de Generalização

	Número de Termos				
Categ./N.N.	30	60	90	120	150
Sub/2	32,06	43,27	49,84	63,84	68,40
Sub/4	21,26	21,53	30,96	57,30	69,00
Sub/8	22,85	19,49	21,54	48,01	71,13
Sub/16	23,09	**19,22**	20,27	54,38	67,54
Sub-Adj/2	34,33	44,49	49,93	54,31	66,43
Sub-Adj/4	22,01	20,18	28,19	46,55	59,03
Sub-Adj/8	21,86	19,66	21,59	34,82	56,38
Sub-Adj/16	22,56	19,18	**19,05**	32,35	56,52
Sub-Npp/2	35,06	42,36	46,32	60,49	70,06
Sub-Npp/4	23,15	19,91	28,02	47,51	67,42
Sub-Npp/8	22,27	19,18	21,62	49,15	65,78
Sub-Npp/16	22,94	19,77	**19,01**	38,82	66,58
S-A-Npp/2	33,22	45,06	45,42	53,03	67,17
S-A-Npp/4	21,91	19,51	28,60	40,63	59,53
S-A-Npp/8	21,39	18,00	17,55	33,29	56,46
S-A-Npp/16	22,44	**16,96**	18,27	37,98	54,21
Adj-Npp/2	56,72	55,40	55,74	59,35	74,04
Adj-Npp/4	48,80	39,45	40,65	51,34	69,29
Adj-Npp/8	49,45	38,25	**35,57**	45,17	72,77
Adj-Npp/16	48,53	38,62	37,75	47,63	74,64

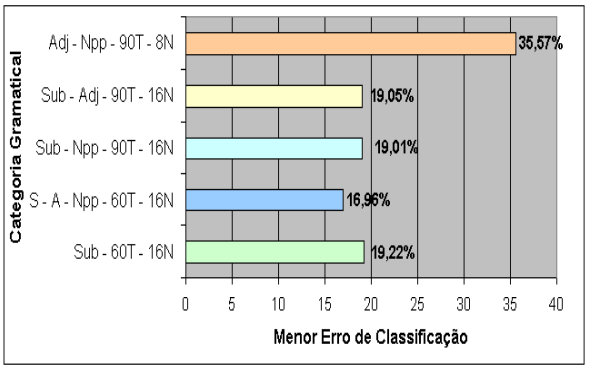
Os resultados obtidos na Tabela anterior mostram que a seleção gramatical que apresenta o menor erro de generalização é a combinação de substantivos com os complementos adjetivos e nomes próprios.

A combinação gramatical Adj-Npp apresentou 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.

Analisando os resultados obtidos na Tabela 2, pode-se observar 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. Essa variação não interfere no resultado da melhor categoria gramatical, como pode ser visto na Figura 6.

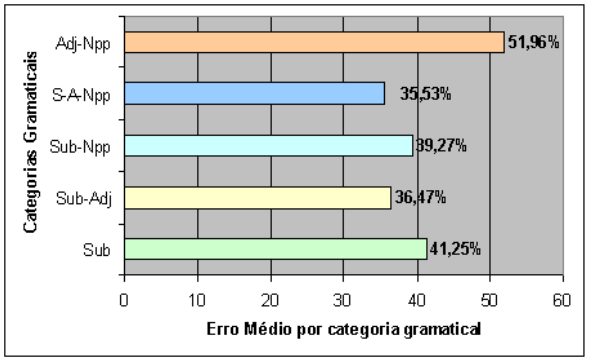
A Figura 5 mostra o menor erro de generalização para cada combinação gramatical selecionada, bem como o número de neurônios na camada intermediária.

Figura 5. Menor Erro para cada categoria gramatical



A Figura 6 apresenta a média de todos os experimentos realizados por categoria. O grupo gramatical que apresenta melhor resultado continua sendo a junção de substantivos, adjetivos e nomes próprios (S-A-Npp).

Figura 6. Média do Erro para cada categoria gramatical



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 S-A-Npp.

Outros experimentos foram realizados com base na metodologia adotada por Corrêa [2] na etapa de pré-processamento que consiste nas seguintes etapas: remoção de irrelevantes, redução de radicais e seleção dos termos. Os termos irrelevantes foram eliminados

dos documentos utilizando uma lista de *stopwords*[5] contendo 476 termos do português Europeu[6], adaptado para o português do Brasil. Enquanto que a normalização morfológica das palavras é realizada utilizando o algoritmo de Martin Porter[7]. A seleção dos termos foi a mesmo utilizada nas combinações gramaticais, baseada na truncagem, utilizando o calculo de freqüência relativa.

A Tabela 3 mostra o resultado desses experimentos utilizando as etapas fundamentais de pré-processamento e N.N. corresponde ao número de neurônios na camada intermediária.

Tabela 3. Média do Erro de Generalização

	Número de Termos				
N.N.	30	60	90	120	150
2	32,85	40,22	40,35	49,68	64,32
4	20,81	16,36	21,42	35,43	49,82
8	20,26	16,56	**14,64**	32,32	55,37
16	20,68	16,74	16,19	24,28	54,23

Como pode ser observado na Tabela 3 o menor erro (14,64%) foi obtido utilizando-se 8 neurônios na camada intermediária, e um número de 90 termos relevantes. Também a média do erro de todos os experimentos (32,13%) manteve-se inferior ao menor erro da combinação gramatical S-A-Npp.

No entanto a categorização com base em S-A-Npp apresenta seu melhor resultado com um número menor de termos (60).

Comparando os resultados obtidos com a utilização de diferentes categorias gramaticais aos obtidos com os métodos usuais de pré-processamento, pode-se verificar que esse último apresentou melhores resultados de categorização.

Corrêa (2002) apresenta resultados, obtidos com o pré-processamento usual para a língua inglesa, semelhantes para a coleção Metais[8] cujos documentos são extraídos da coleção "Reuters-21578, Distribution 1.0", e o médio de classificação para as 10 simulações foi de 15,09% com 16 neurônios na camada intermediária. As coleções de documentos PubsFinder[9] e K1[10] apresentaram erros médios de classificação de inferiores ou iguais a 14,34% com dimensão nos vetores globais de 200 termos e de 25,90% com 32 neurônios na camada intermediária, respectivamente.

6. Considerações Finais

Categorização de textos é uma técnica empregada para identificar a categoria a que determinado documento pertence, utilizando como base o seu conteúdo [9], [10]. Este processo pode ser utilizado em muitos contextos, desde a indexação automática de documentos baseada em um vocabulário de controle, filtragem de documentos, catálogos de recursos na Web e, em geral, qualquer aplicação que requeira organização de documentos ou seleção de mensagens [11] e contempla as seguintes etapas: coleta de documentos, pré-processamento e classificação.

O foco principal deste trabalho é a etapa de pré-processamento que geralmente consiste nas fases: remoção de termos irrelevantes, redução de afixos e seleção dos termos. Com base nessas fases foi proposto um método alternativo considerando a seleção de várias combinações gramaticais para avaliar o processo de categorização de textos utilizando RNAs.

A metodologia proposta apresentou-se eficiente para o problema de categorização abordado. Os parâmetros adotados, tais como o tipo de codificação, o número de palavras selecionadas e as topologias de rede, permitiram o acompanhamento do comportamento dos algoritmos, bem como a verificação da melhor configuração a ser adotada para a melhor combinação gramatical selecionada.

As taxas de erro obtidas, com o pré-processamento baseado em diferentes categorias gramaticais foram positivas para o resultado do categorizador. Dentre as classes gramaticais testadas, a estrutura substantivos, adjetivos e nomes próprios apresentaram-se suficientes para categorizar uma coleção de documentos utilizando RNAs com 16 neurônios na camada intermediária. Porém, cabe destacar que o algoritmo BP dependendo do número de entradas para o treinamento da rede apresenta um aumento no erro de generalização por não conseguir identificar em classes distintas a quantidade de padrões semelhantes que compõem a entrada da rede.

No entanto, os experimentos realizados com o pré-processamento baseado em etapas usuais (remoção de termos irrelevantes, redução de afixos e seleção dos termos), obtiveram melhores resultados se comparado as estruturas gramaticais.

---

[5] Tais como, artigos, preposições, pronomes, verbos auxiliares, entre outros.

[6] Criada por Paulo Quaresma da Universidade de Évora

[7] Disponível para várias línguas em http://snowball.sourceforge.net

[8] Disponível em http://www.research.att.com/~lewis

[9] Disponível em http://www.cin.ufpe.br/~mln/

[10] Disponível em ftp://ftp.cs.umn.edu/dept/users/boley/PDDPdata/K1

O presente trabalho pode ser estendido em várias direções visando melhorar o desempenho do processo de categorização de textos, os principais são:

– Experimentos com outros domínios ou corpora de outras línguas;

– Utilização de outras técnicas de aprendizado supervisionado tais como Support Vector Machines, devido a sua robustez em permitir a utilização de grandes números de termos. Support Vector Machines tem sido aplicado para documentos utilizando o português Europeu [12], porém o pré-processamento baseado em combinações gramaticais ainda não foi testado para o processo de categorização;

– Investigação de outros métodos para cálculo de relevância dos termos, a fim de verificar como eles podem afetar o desempenho de técnicas de aprendizado. O cálculo de relevância utilizado neste trabalho é freqüência relativa, outros como: freqüência absoluta, ganho de informação, Qui-quadrado podem ser testados;

– Investigação de outras técnicas para redução de termos para a representação dos documentos. Normalmente, a notação vetorial de documentos possui uma quantidade muito grande de atributos relevante, chegando a dezenas de milhares de termos disponíveis para a categorização dos textos, tornando-se computacionalmente intratável para a representação dos documentos.

– Aplicar conhecimentos lingüísticos mais sofisticados como categorias gramaticais, por exemplo, o uso de sintagmas nominais (com 2 ou 3 palavras) na etapa de pré-processamento.

7. References

[1] Y. Yang, J. Pederson, J. **A Comparative Study on Feature Selection in Text Categorization.** In Proceedings of the 14th International Conference on Machine Learning, Morgan Kaufmann Publishers, San Francisco, US, pp. 412-420, 1997.

[2] R. Correa, T. Ludermir. **Categorização Automática de Documentos: Estudo de Caso.** In: XVI Brazilian Symposium on Neural Networks, Porto de Galinhas, 2002.

[3] E. Bick. **The Parsing System PALAVRAS: Automatic Gramatical Analysis of Porutugese in a Constraint Grammar Framework.** Århus University. Århus: Århus University Press, 2000.

[4] C. Gasperin et. al.. **Extracting XML Syntactic Chunks from Portuguese Corpora.** In Proceedings of the Workshop TALN 2003 Natural Language Processing of Minority Languages and Small Languages - Batz-sur-Mer France June 11 – 14, 2003.

[5] J.A. Freeman, D.M. Skapura. **Neural Networks, Algorithms, Applications, and Programming Techniques**. Massachusetts: Addison-Wesley Publishing Company, Inc., 1991.

[6] S. Haykin. **Neural Netwoks - A Comprehensive Foundation**. New York: Macmillan College Publishing Company, 1994.

[7] K. Sparck-Jones, P. Willet (eds). **Readings in Information Retrieval.** San Francisco: Morgan Kaufmann, 1997.

[8] I.H. Witten et. al.. **Data mining: Pratical Machine Learning tools and techniques with Java implementations.** Academic Press, 2000.

[9] D. D. Lewis. **Evaluating Text Categorization.** In Proceedings Speech and Natural Language Workshop, pp. 312-318, 1991.

[10]Y. Yang, X. Liu. **An evaluation of statistical approaches to text categorization.** Journal of Information Retrieval, v.1, n.1/2, p.67-88, 1999.

[11]F. Sebastiani. **Machine learning in automated text categorization.** ACM Computing Surveys, vol. 34, no. 1, pp.1-47, 2002.

[12]T. Gonçalves, P. Quaresma. **A prelimary approach classification problem of Portuguese juridical documents.** In: 11th Portuguese Conference on Artificial Intelligence. Lectures Notes in Artificial Intelligence. Berlim: Springer Verlag, 2003.