MAC0499 – TRABALHO DE FORMATURA SUPERVISIONADO

Desenvolvimento de algoritmos para para Discriminar Português Brasileiro e Europeu

Orientadores: Arnaldo Mandel e Antonio Galves
Aluno: Denis Antonio Lacerda

1 de dezembro de 2008

Sumário

Parte I
Resumo

Na história do Português Europeu ocorreram duas mudanças importantes entre os séculos 17 e 19, uma afetando a sintaxe e outra a prosódia da língua. A Hipótese colocada em Galves & Galves (1995) é que as duas mudanças estão relacionadas e que a mudança prosódica precedeu e desencadeou a mudança sintática. Outra hipótese é de que o português brasileiro seja muito próximo do português clássico em termos rítmicos. Um Grande problema é que os únicos registros que dispomos da época são os escritos. Seria possível extrair padrões rítmicos de textos escritos? Como marcar os elementos pertinentes do ritmo no texto?

O objetivo do projeto desenvolvido foi obter evidências estatísticas para dar suporte as conjecturas citadas. Em particular, o objetivo do meu trabalho foi desenvolver ferramentas computacionais que implementassem os conceitos matemáticos, estatísticos e línguisticos colaborando com o projeto como um todo.

1 Introdução

Seqüências genéticas, cadeias de amino-ácidos, seqüências rítmicas na fala, seqüências de dados econômicos, parecem ter em comum um comportamento que, apesar de não ser determinístico,contém informações precisas a respeito do sistema que as produziu. No caso de cadeias linguísticas uma dessas características parece estar codificada no ritmo. Mas, o que é o ritmo de uma língua?

1.1 Conjectura das classes rítmicas

Conjectura-se que que as línguas se agrupam em classes rítmicas(Lloyd James, anos 40 e Abercrombie, anos 50), e que as classes rítmicas são caracterizadas pelo fato de certos domínios prosódicos e/ou certa informação fonético-fonológica serem ou não relevantes(Kleinhez, 1995).

Classes rítmicas conjecturadas:

- **línguas acentuais:** Holandês, Inglês, Polonês, Português Europeu,...
- **línguas silábicas:** Catalão, Espanhol, Francês, Italiano, Português Brasileiro, ...
- **línguas moraicas:** Japonês, ...

Divididas em função em função da unidade organizadora (sílaba, intervalo acentual ou mora), onde as línguas acentuais tendem a possuir uma estrutura silábica mais variável e um efeito duracional do acento enquanto as línguas silábicas apresentam uma estrutura silábica menos variável, e o efeito duracional do acento é menor ou nulo[1].

1.2 Rítmo em textos escritos

Como extrair padrões rítmicos de textos escritos? Possivelmente o texto completo com todas as letras tem informação demais e o ritmo fica escondido numa cadeia subjacente mais simples.

Uma tentativa foi marcar os elementos pertinentes do ritmo indicando para cada sílaba:

- se ela carrega ou não o acento principal da palavra Português Brasileiro, ...
- se ela é ou não começo de palavra

Isso nos levou a usar:

$$\{0,1\}^2 = \{(0,0),(0,1),(1,0),(1,1)\} \tag{1}$$

como conjunto de símbolos onde:

- o primeiro dígito indica se a sílaba é começo de palavra prosódica (0= não, 1=sim)
- o segundo dígito indica se a síliba carrega o acento principal da palavra (0=não, 1=sim).
- Em representação binária: (0, 0) = 0, (0, 1) = 1, (1, 0) = 2 e (1, 1) = 3.
- Acrescentamos o símbolo 4 para indicar o começo de frase.

Com isso chegamos no seguinte modelo

- 0 - Sílaba não acentuada
- 1 - Sílaba acentuada
- 2 - Sílaba não acentuada no início de palavra prosódica
- 3 - Sílaba acentuada no início de palavra prosódica
- 4 - Início de sentença

Onde Palavra prosódica é uma palavra lexical junto com uma palavra funcional não acentuada.

Exemplo:

	O menino	já	comeu	o doce
4	2 0 1 0	3	2 1	2 1 0

2 Conceitos e tecnologias estudadas

2.1 Cadeia de Markov de Alcance Variável(VLMC)

A sequência $\mathcal{A}$ = {1,2,3,4}não apresentam padrões identificáveis a olho nu sendo necessário identificar padrões na classe de modelos probabilístico capazes de gerar seqüências desse tipo. Dentre os modelos estatísticos analisados, Cadeias de Markov de Alcance Variável(VLMC) é o modelo que mais de adapta ao estudo por considerar a dependência de um estado em relação a uma porção relevante do passado chamado de contexto, e o comprimento dessa porção relevante varia de um passado para outro.

Uma VLMC é uma Cadeia de Markov de ordem K finita, com K fixado, para o qual é possível expressar as probabilidades de transição da seguinte forma:

$$P(X_0 = x_0 | X_{-K}^{-1} = x_{-K}^{-1}) = P(X_0 = x_0 | X_{-K}^{-1} = x_{-l(x_{-K}^{-1})}^{-1}) \tag{2}$$

onde $x_b^a = (x_a, x_{a-1}, ..., x_b), \forall a > b$ é chamada *função comprimento* e representa o número de passos que a cadeia precisa olhar para o passado para escolher o próximo estado.

O caso $l = 0$ corresponde a independência.

A função $c : \mathcal{A}^K -> U_{m=0}^K \mathcal{A}^m$

$$c : x_{-K}^{-1} | -> x_{-l(x_{-K}^{-1})}^{-1} \tag{3}$$

é chamada de *função contexto da cadeia*.

É muito comum representar uma VLMC através de uma árvore probabilística, onde cada galho da árvore representa um contexto.

3 Atividades realizadas

Durante o projeto de iniciação científica eu colaborei desenvolvendo e/ou aperfeiçoando programas para duas vertentes distintas do projeto, uma usando o "'Algoritmo do Contexto"' e outra usando "'Misturas de VLMC"'.

3.1 Algoritmo do Contexto

Os primeiros testes para detecção de padrões rítmicos em textos escritos usando VLMC foi utilizando o Algoritmo do Contexto, introduzido por Rissanen na década do 80 e implementado no pacote estatístico R. E minha primeira atividade no projeto foi analisar as estruturas do pacote VLMC e desenvolver rotinas para permitir o uso dos textos codificados em arquivos como entrada e melhorar a saida do programa que pode ser visualizada a seguir.

```
LitBras015.txt
n:  7072    K: 38.55215    AIC: 11468.55
```

Cont.	Probabilidades						
	0	1	2	3	4	n	freq
1 0 0	0	0	0.606	0.242	0.151	66	0.009
2 0 0	0.245	0.754	0	0	0	306	0.043
0 0 0	0.234	0.765	0	0	0	98	0.013
3 0 0	0	0	0.454	0.363	0.181	22	0.003
1 0	0.056	0	0.63	0.196	0.117	1177	0.166

2 0	0.366	0.633	0	0	0	835	0.118
3 0	0.048	0	0.643	0.271	0.037	457	0.064
1	0.727	0	0.184	0.069	0.018	1617	0.228
2	0.516	0.482	0	0	0	1617	0.228
3	0.68	0	0.184	0.127	0.007	672	0.095
4	0	0.004	0.529	0.465	0	204	0.028

Ao longo do estudo percebeu-se que o estimador proposto por Bulhman e Wynner, apesar de ser consistente, é muito pouco robusto, isto é, a contaminação com pequenas seqüências espúrias muda dramaticamente o resultado. Além disso, não é possível determinar a distribuição do erro do estimador. Com o objetivo de avaliar a distribuição do erro do estimador decidiu-se fazer 500 simulações para determinados tamanhos de amostra e utilizar a distância entre árvores sugerida por Fraiman. Para isso foi necessário desenvolver um simulador de VLMC e rotinas para processar e analisar diversas saídas

Esse estudo agrupou os textos de cada língua e obteve as probabilidades de transição pela estimativa de máxima verossimilhança e, com base nessas probabilidades, aplicou-se a técnica de bootstrap, obtendo uma amostra de 500 textos de tamanho 10.000. Em seguida, foi calculada a árvore mediana de cada língua, a qual é composta pelos galhos que aparecem em pelo menos 50% das amostras.

3.2 Misturas de VLMC

Uma mistura de VLMCs é o conjunto de modelos VLMC, isto é, um conjunto de árvores e suas respectivas *Probabilidades de Transição* neste conjunto.

Uma mistura de modelos pode ser interpretada como se em cada tempo 't', escolhêssemos um modelo de acordo com a distribuição, e com esse modelo gerássemos o próximo símbolo x_t do passado $x_0, ..., x_{t-1}$.

3.3 Como Ponderar as Árvores

O algoritmo para ponderar a mistura é o seguinte:

- primeiro, nós inicializamos os pesos da mistura com uma probabilidade *a priori* para as árvores($w^0(\tau)$).
- Então nós atualizamos os pesos das Árvore para cada símbolo do texto de entrada.

$$w^{t+1}(\tau) = w^t(\tau) P_\tau^t(x_t | c_\tau(x_0, ..., x_{t-1})). \quad (4)$$

onde $P_\tau^t(x_t | c_\tau(x_0, ..., x_{t-1}))$ é o estimador de Máxima Verossimilhança com a amostra até o t-ésimo símbolo.

3.4 O Peso Inicial

Como a Verossimilhança tende a ser maior nas árvores maiores(VLMCs com mais parâmetros), então esse peso inicial deve penalizar as árvores maiores.

Essa penalização também deve considerar o tamanho da amostra. Quanto maior a amostra, mais as árvores grandes serão beneficiadas pela verossimilhança. Portanto, quanto maior a amostra, mais as árvores grandes devem ser penalizadas.

Um exemplo desse tipo de penalização é $w^0(\tau) = (ct)^{-n}$, onde 'c' é uma constante, 't' é o tamanho do texto de entrada, e 'n' é o número de nós(parâmetros) da árvore τ.

3.5 A implementação

Foi feito um programa que implementa o modelo e o algoritmo, possibilitando a classificação automática dos textos, embora o programa possa potencialmente ser usado para classificação de outras sequências como por exemplo cadeias de aminoácidos, dados financeiros, etc.

Um problema no desenvolvimento desse programa é a complexidade do algoritmo e a grande quantidade de memória usada. Isso ocorre devido ao número exponencial de árvores geradas. Para resolver esse problema, os dados devem ser guardados numa estrutura bastante compacta e as atualizações são feitas numa árvore molde que contém todos os galhos possíveis.

Outro problema é a precisão limitada dos computadores para os cálculos. O algoritmo faz multiplicações com números cada vez menores e para textos grandes há o risco de perder dígitos significativos. Para resolver esse outro problema, todos os cálculos são realizados guardando apenas os digitos mais significativos dos resultados como em notação científica.

O que mais motivou a implementação desse software foi a necessidde de inserir restrições nas árvores geradas. Essa nova implementação elimina as árvores impossíveis, e para isso considera algumas restrições lingüísticas:

1. Toda palavra prosódica precisa conter uma, e somente uma, sílaba tônica.
2. Uma sílaba tônica pode ser seguida por no máximo 3 sílabas não tônicas dentro de uma palavra prosódica.
3. Finalmente, toda sentença precisa começar com uma palavra prosódica.

No programa desenvolvido, essas restrições foram facilmente implementadas usando Expressões Regulares para representar os contextos possíveis ou impossíveis.

3.6 Conjunto total de árvores

Para calcular o número total de árvores para uma Cadeia de Markov de Alcance Variável de profundidade máxima "n" dada podemos começar pensando nas duas árvores mais simples que são a de profundidade "zero" e a de profundidade "um".

Para saber o número de árvores máximas de profundidade máxima "n" com n superior a um, podemos imaginar essas árvores como sendo uma árvore de profundidade "um" e todas as possíveis combinações onde seus nós terminais são ramificados com árvores de profundidade "n-1'. Com isso podemos calcular de forma progressiva o número máximo de árvores de profundidade máxima "n".

4 Resultados e produtos obtidos

Foram feitos inúmeros testes usando os textos da nossa Base de Dados. A base de dados é composta por textos do Século XX, 35 escritos por Escritores Brasileiros e 31 escritos por Escritores Europeus. A base de dados pode ser consultada no endereço http://www.ime.usp.br/~tycho.

Os resultados foram surpreendentes, mostrando que esse é um excelente modelo para classificação sequüências simbólicas, no caso nossos textos.

Nos testes, dentre o enorme conjunto de todas as possíveis árvores, somente um pequeno número de árvores se destacaram com probabilidade significativas, e nesse pequeno conjunto de árvores que receberam probabilidades significativas, foram encontrados padrões que diferem entre o Portugüês Brasileiro, e o Portugüês Europeu, permitindo assim sua discriminação.

A figura 1 mostra a árvores com probabilidade(peso final normalizado) mais significativa obtida através dos testes feitos com uma amostra de 619282 símbolos em textos de autores brasileiros. O peso a priori foi $(500t)^N$. Já a figura 2 mostra a árvores com probabilidade(peso final normalizado) mais significativa obtida através dos testes feitos com uma amostra de 148887 símbolos em textos de autores europeus. O peso a priori foi $(500t)^N$.

Como é fácil notar, as duas árvores seguem um padrão diferente. Enquanto a árvore Européia ramifica apenas o galho '0', a árvore brasileira ramifica também o galho '1'. Fazendo uma breve análise lingüística temos que tanto no Português Brasileiro quanto no Português Europeu as frases são independentes. A diferença está justamente no acento, pois no Padrão rítmico Europeu o acento principal é um estado determinante, diferente do Padrão rítmico Brasileiro.

O programa encontra-se disponível no site http://www.ime.usp.br/~tycho/
prosody/vlmc/tools e pode ser baixado gratuItamente para fins acadêmicos.

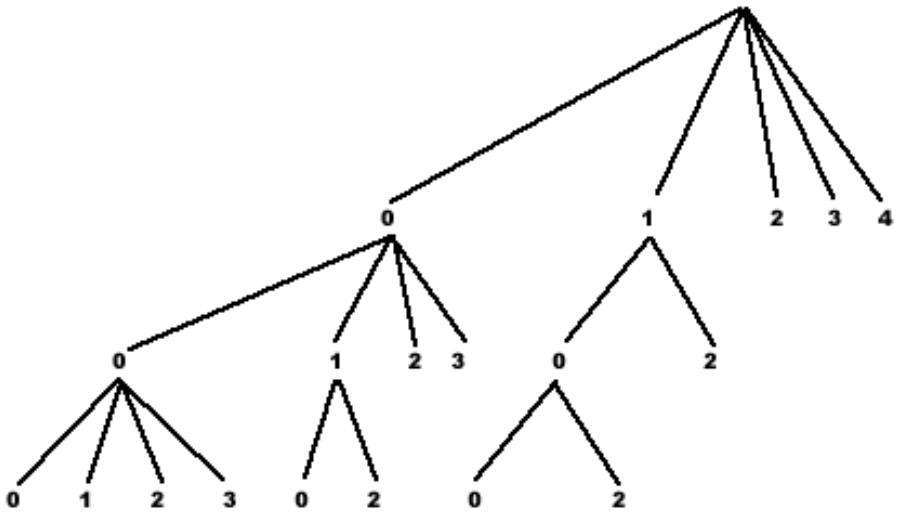

Figura 1: Padrão de Árvore Brasileira. Probabilidade: 0,80.

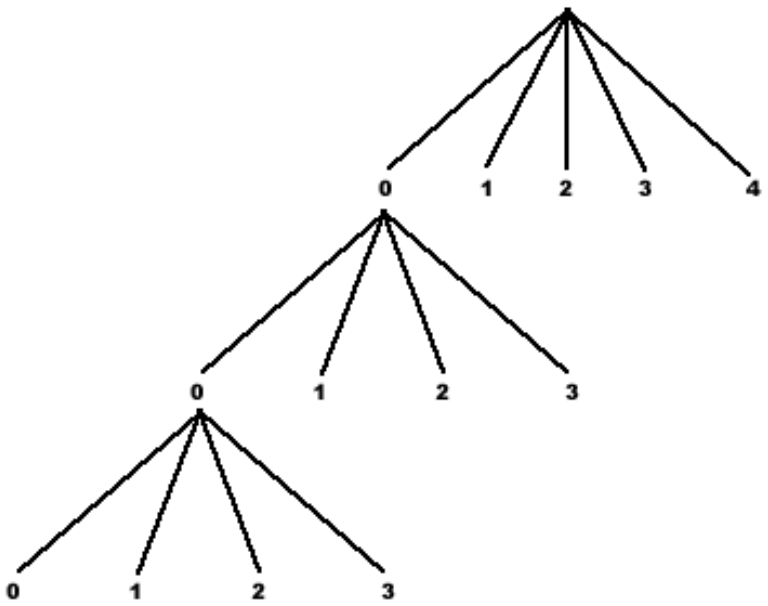

Figura 2: Padrão de Árvore Europeia. Probabilidade: 0,99.

5 Conclusões

É surprendente que num conjunto com cerca de 6000 árvores o procedimento leve a uma medida de probabilidade concentrada numa única árvore. A conclusão óbvia é que é os textos contém informações claras e precisas do ritmo do autor que o produziu. É possível concluir ainda que com o algoritmo apresentado é possível detectar de forma clara padrões em sequências corretamente modeladas sejam elas textos, cadeias de aminoácidos, etc.

Referências

[1] Sónia Frota, Marina Vigário & Fernando Martins, *Discriminação entre línguas: Evidência para classes rítmicas.*

[2] A. Galves, C. Galves, N. L. Garcia, C. Peixoto, *Correlates of rhythm in written texts of Brasilian and Modern European Portuguese.*

[3] Garcia, N. and Peixoto, C.,*Statistical Analysis of Written Texts Modern European Portuguese vs. Brazilian Portuguese*, (2001).
http://www.physik.uni-bielefeld.de/complexity/

[4] Buhlmann and Wyner

[5] Projeto Temático FAPESP *Padrões Rítmicos, Fixação de Parâmetros e Mudança Lingüística.*
http://www.ime.usp.br /t̃ycho/

Parte II
Subjetiva

6 Desafios e frustrações encontrados

Quando decidi participar do programa de iniciação científica, o primeiro desafio foi participar de um grande projeto com inúmeros pesquisadores de diversas universidades, o que deixou com a responsabilidade de escolher uma área dentro do grupo de pesquisa e estudar diversas matérias que estão fora da grade curricular do BCC o que exigiu que eu me tornasse um tanto autodidata.

Outro grande desafio foi ter que desenvolver softwares complexos para trabalhar com um grande volume de dados, o que me exigiu uma grande preocupação com a eficiência dos algoritmos criados e estruturas usadas.

Por fim, um último desafio foi ter que escrever essa monografia sem ter experiência com redação de textos científicos. Desafio que foi minimisado com a ajuda de meus orientadores e com o pouco de experiência que adquiri nos relatórios enviados a CNPQ e nas apresentações criadas para o Siicusp e Escola Brasileira de Probabilidades.

Uma grande frustação foi ter que deixar o projeto por motivos pessoais não podendo dar continuidade ao projeto. E mesmo durante o período que me dediquei ao projeto destaco a frustação de não ter conseguido emcontrar um meio de encontrar as árvores mais significativas e suas respectivas probabilidades sem ter que calcular a probabilidade de todas as árvores possíveis. Isso permitiria ao algoritmo ganhar muita eficiência possibilitanto o cálculo em árvores maiores.

Uma pequena frustação for não ter tido tempo para desenvolver uma interface gráfica mais amistosa para o software das Florestas Probabilísticas, porém creio que conseguirei fazer isso até o final do Ano.

7 Disciplinas relevantes para o desenvolvimento do projeto

Cito as disciplinas relevantes para o bom desenvolvimento do projeto de iniciação científica, e aproveito para cumprimentar e elogiar os professores que a ministraram sem os quais o processo de aprendizado se tornaria mais árduo e confuso.

- **MAC0122 - Princípios de Desenvolvimento de Algoritmos:** Disciplina onde aprendi a desenvolver algoritmos mais complexos me preocupando com a eficiência.
- **MAC0323 - Estruturas de Dados:** Disciplina onde aprendi diversas estruturas de dados assim como sua manipulação e aplicações. Aprendi ainda como as estruturas de dados usadas podem ser determinantes na eficiência dos programas e no consumo de recursos.
- **MAE0228 - Noções de Probabilidade e Processos Estocásticos:** Disciplina onde aprendi noções básicas de probabilidade condicional e cadeias de Markov. Matérias extremamente importantes para o projeto e que eu pude aprofundar durante a iniciação científica.
- **MAC0338 - Análise de Algoritmos:** Disciplina onde aprendi a analisar o desempenho de algoritmos. Isso foi extremamente útil pois trabalhei principalmente no desenvolvimento de algoritmos de alta complexidade.
- **MAC0441 - Programação Orientada a Objetos:** Disciplina onde pude solidificar meus conhecimentos num conceito de programação que utilizei nos principais programas desenvolvidos, e onde aprendi diversos padrões de desenvolvimento de software.
- **FLC0474 - Língua Portuguesa:** Quero acreditar que essa disciplina me ajudou na redação dessa monografia.

8 Próximos passos

O projeto é longo e ainda há muito o que estudar e desenvolver, afinal participei apenas das primeiras, satisfatórias e produtivas, fases. Tive experiências muito boas durante o período de iniciação científica, e pretendo num futuro próximo dar prosseguimento aos meus estudos na área de bioinformática.

Por enquanto deixo tudo que desenvolvi durante os anos de iniciação científica para que possa ser usado pelos pesquisadores do projeto ao qual pertenci e até mesmo por pesquisadores de outras áreas ou projetos. Espero poder dar prosseguimento ao projeto contribuindo com o desenvolvimento de novas ferramentas e aperfeiçoamento das atuais.

9 Agradecimentos

Agradeço a todos os professores com os quais apredi muito e adiquiri muita experiência durante o curso, e em particular meus orientadores pela atenção e dedicação que tiveram comigo. Agradeço também a minha mãe pela paciência, apoio e dedicação durante toda minha vida e em particular nesses últimos anos durante minha graduação. E por fim, agradeço a você que teve a paciência de ler essa monografia até essa última linha.