MAC 0499 – TRABALHO DE FORMATURA SUPERVISIONADO

Florestas Probabilísticas para Discriminar Português Brasileiro e Europeu

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

Textos escritos do Português Brasileiro e Europeu foram modificados e modelados como Cadeias De Markov de Alcance Variável(VLMC) representadas por árvores probabilísticas. Foi atribuído uma distribuição de probabilidade ao conjunto de todas as possíveis árvores de acordo com o texto de entrada.

A forma usada para ponderar essas árvores resultou num pequeno número de árvores com probabilidades significantes, e nessas árvores significantes foram encontrados padrões que diferem entre Português Brasileiro e Europeu, permitindo assim sua discriminação.

O objetivo é apresentar o modelo usado na codificação dos textos, e o método usado para ponderar as árvores.

A Codificação

O primeiro passo para a construção do modelo é a codificação dos textos.

Os textos foram codificados sobre o alfabeto A={0,1,2,3,4} de acordo com a localização da sílaba tônica de cada palavra, como descrito abaixo

- **0 – Sílaba átona**
- **1 – Sílaba tônica**
- **2 – Sílaba átona e início de palavra prosódica**
- **3 - Sílaba tônica e início de palavra prosódica**
- **4 – Início de sentença**

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

VLMC

Uma VLMC é uma cadeia de Markov de ordem finita cujas probabilidades de transição tem a seguinte propriedade

$$\mathbb{P}(X_0 = x_0 \mid X_{-K}^{-1} = x_{-K}^{-1}) = \mathbb{P}(X_0 = x_0 \mid X_{-K}^{-1} = x_{-\ell(x_{-K}^{-1})}^{-1})$$

Onde $\ell : \mathcal{A}^K \to \{1, \ldots, K\}$ **é uma função do passado que indica o número de passos que devemos olhar para trás para escolher o próximo estado da cadeia.**

Uma VLMC é convenientemente representada por uma árvore probabilística.

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*, com uma distribuição de probabilidades nesse conjunto.

Uma mistura de modelos pode ser interpretada como se em cada tempo t, procurássemos o modelo que melhor se ajusta ao texto até o presente momento, e gerássemos o próximo símbolo de acordo com o modelo escolhido.

Como ponderar as árvores?

O algoritmo para ponderar as árvores é como descrito abaixo:

1. **Primeiro, nós inicializamos os pesos da mistura com as probabilidades *a priori* das árvores (** $\omega^0(\tau)$**).**
2. **Então, nós atualizamos os pesos para cada símbolo do texto de entrada.**

$$\omega^{n+1}(\tau) = \omega^n(\tau) P_\tau^n(x_n | c_\tau(x_0, \ldots, x_{n-1}))$$

Onde $P_\tau^n(x_n | c_\tau(x_0, \ldots, x_{n-1}))$ **é o estimador de Máxima Verossimilhança com a amostra até o n-ésimo símbolo.**

O Peso *A Priori*

- **A Verossimilhança da amostra aumenta quando o tamanho da amostra aumenta, portanto as árvores maiores recebem pesos menores.**
- **O peso a priori também precisa considerar o tamanho da amostra. Quanto maior a amostra, menor o peso.**

Ex: w(T) = (ct)$^{-n}$, onde 'n' é o número de nós terminais da árvore 'T' e 't' é o tamanho do texto de entrada.

Implementação

Foi desenvolvido um software que implementa esse algoritmo.

Um problema no desenvolvimento desse software é a complexidade do algoritmo a a grande quantidade de memória usada devido ao exponencial número de árvores geradas. Para minimizar esse problema, o software trabalha e faz atualizações numa árvore molde.

O software também precisa eliminar as árvores impossíveis. Para isso são considerados algumas restrições lingüísticas:

- **Uma palavra prosódica deve conter uma, e somente uma, sílaba tônica.**
- **Uma sílaba tônica pode ser seguida por no máximo 3 sílabas átonas numa mesma palavra prosódica.**
- **Finalmente, uma sentença deve começar com o início de uma palavra prosódica.**

Uma versão desse software está disponível para todas as plataformas no site: http://www.ime.usp.br/~tycho/prosody/vlmc/tools.

Resultados

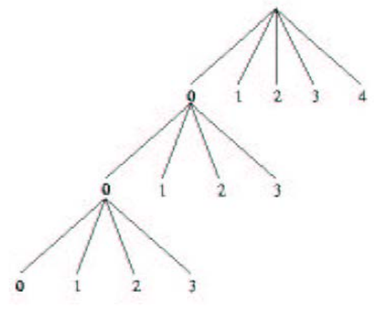
Testes usando o programa foram realizados com os textos de nossa base de dados. A Base de dados é composta por textos literários do século XX, 35 escritos em Português Brasileiro e 31 escritos em Português Europeu.

Os resultados impressionaram. Apenas um pequeno número de árvores receberam probabilidades significativas. Padrões foram encontrados nas árvores significativas que diferem entre o Português Brasileiro e Europeu.

Esses padrões podem ser verificados nas figuras abaixo.

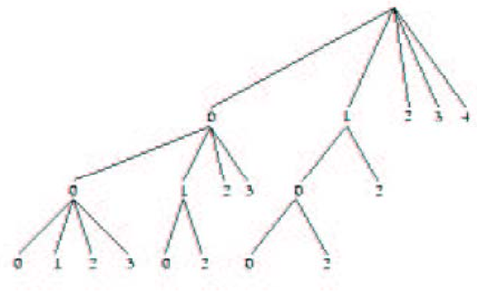
Conclusão

"Uma árvore resume toda a Floresta"



Padrão Europeu

Probabilidade: 0,99



Padrão Brasileiro

Probabilidade: 0,80

Apoio: CNPq Conselho Nacional de Desenvolvimento Científico e Tecnológico
IME - Instituto de Matemática e Estatística
Universidade de São Paulo