Aplicações de Treinamento Semi-Supervisionado em Casos Bipolares de Linguagem Natural

Trabalho final da da Matéria SCC5882,

Professores: Dr. Zhao Liang e Dr. Alneu de Andrade Lopes

Aluno: Renato Fabbri

renato.fabbri@gmail.com

IFSC-USP

6 de julho de 2010

Definição e suas Variações

- Linguagem Natural: linguagem utilizadas normalmente por seres humando para se comunicarem.
- Em geral $\rightarrow$ Somente Conteúdo Textual.
- Processamento de Fala.

PLN

- Fonética e Fonologia - O estudo dos sons linguísticos.
- Morfologia - O estudo dos componentes das palavras (e seus significados).
- Sintáxe - O estudo da relação estrutural entre palavras.
- Semântica - O estudo dos significados.
- Pragmática - O estudo do uso da linguagem para fins específicos.
- Discurso - O estudo de unidades linguísticas que envolvem conjuntos de colocações, frases, etc.

Processamento de Fala

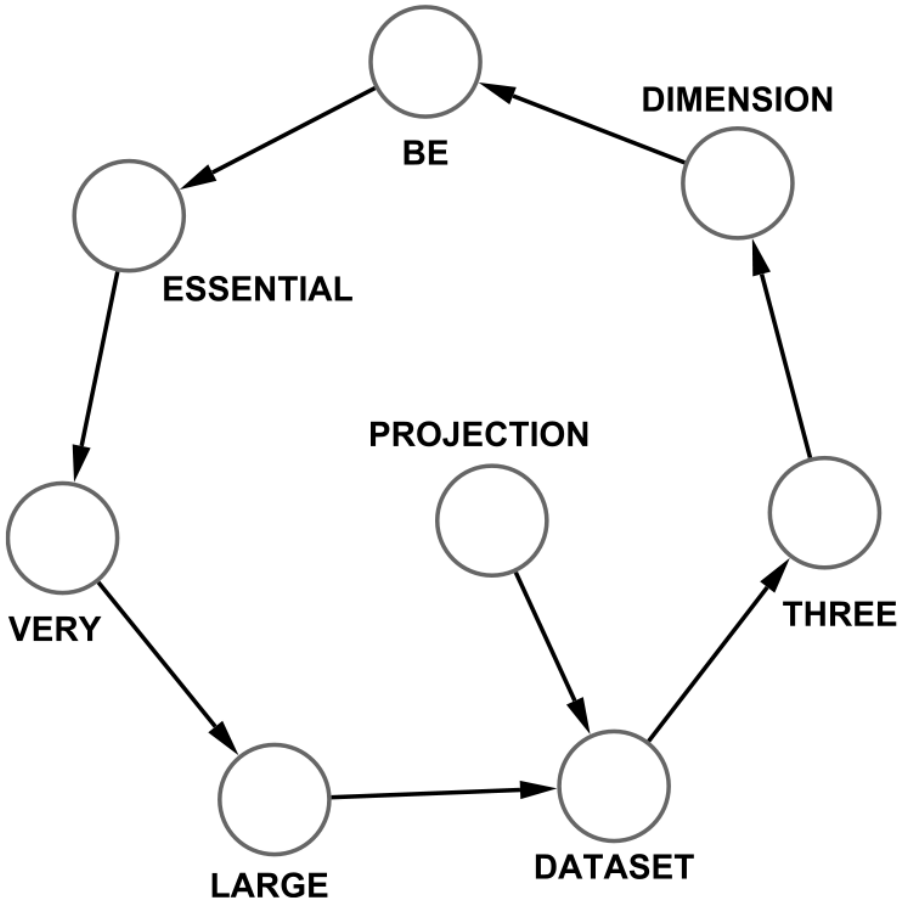
- Reconhecimento de Fala - lida com o a análise do conteúdo linguístico do sinal.
- Reconhecimento de Voz/Locutor - visa identificar o individuo que produz a fala.
- Codificação de voz - compactação de dados especializada.
- Análise de voz - fins médicos, estudos cognitivos, etc.
- Síntese de voz - fala artificial, geralmente gerada por computador.
- Melhora de fala - visa recuperar ou aumentar a inteligibilidade do sinal.

Formação (1)

Original sentence	Pre-processed sentence
The projection of the dataset into three dimensions is essential for very large datasets	projection dataset three dimension be essential very large dataset

Figura: Processamento de texto para formação de redes complexas.

Formação (2)


BE
DIMENSION
ESSENTIAL
PROJECTION
VERY
THREE
LARGE
DATASET

Formação - Caso Específico

- Textos da Folha de São Paulo, coletados por 10 anos.
- Opiniões negativas e positivas.
- Textos aglomerados até atingirem 1200 vértices.
- Extraímos as medidas de cada aglomerado: graus, coeficientes de clusterização, caminhos mais curtos, eficiência global, proximidade(closeness) e acessibilidade.

Atributos - Método Geral

- Medidas estatísticas
- Frequência Fundamental.
- Amplitude/Intensidade/Energia.
- Harmônicos.
- Articulação.
- Pausas.

Atributos

- Frequência fundamental a cada 0.01 segundo.
- Média, desvio padrão, âmbito, mediana, abaixo do limiar, acima do limiar.
- Praat e Python.

Normalização e PCA

- (medida - média) / desvio padrão
- PCA

Métodos Utilizados (Confronto)

- Bayesian Decision
- Decision Tree
- Decision Rules
- Naive Bayes

Propagação de Rótulo - Concepção

- Vizinho mais próximo $\rightarrow$ propaga mais fácil.
- Rótulos fixos nos rotulados de antemão.
- Estes "emanam" rótulos.

Convenções e Exposição do Problema

- $(x_1, y_1)...(x_l, y_l)$ os dados rotulados.
- $y \in 1...C$ com C o número de classes.
- $x_{l+1}...x_{l+u}$.
- Seja $n = l + u$
- L e U denotam os dados rotulados e não rotulados, respectivamente

Obtenção do Grafo

- Completo.
- $w_{ij} = exp(-\frac{\|x_i - x_j\|^2}{\alpha^2})$
- α é um *hiperparâmetro*

Matriz de Transição

- P_{ij} é a probabilidade de transição do nó i para o nó j (passagem do rótulo)
- $P_{ij} = P(i \rightarrow j) = \frac{w_{ij}}{\sum_{k=1}^{n} w_{ik}}$

Últimas Definições

- Matriz de rótulos Y_L, $l \times C$, cuja *iésima* linha possui 1 na coluna correspondente à classe do dado x_i.
- f uma matriz $n \times C$ em que cada linha pode ser interpretada como uma distribuição de probabilidade sobre os rótulos.

Implementação do Algoritmo de Propagação de Rótulo

$$f_u = (I - P_{UU})^{-1} P_{UL} Y_L \tag{1}$$

implementação $\rightarrow$ algoritmos/ssl-propagacao-de-rotulo.py

Mincut - Concepção e Implementação

- Objetos parecidos devem ficar juntos.
- Seccionamento por separações de custos mínimos.

implementação → algoritmos/ssl-mincut.py

Métodos Tradicionais

Method	Precision Pos.	Recall Pos.	Precision Neg.	Recall Neg.
Bayesian Decision	70.2 %	49.35 %	77.3 %	71.42 %
Decision Tree	71.1 %	38.0 %	57.7 %	84.5 %
Decision Rules	70.0 %	49.3 %	60.9 %	78.9 %
Naive Bayes	70.6 %	50.7 %	61.5 %	78.9 %

Figura: Reconhecimento de polaridades via métodos tradicionais.

Propagação de Rótulo

NA MONOGRAFIA (era texto/resultados/*)

Métodos Tradicionais

NA MONOGRAFIA (era audio/trad/*)

Propagação de Rótulo

NA MONOGRAFIA (era audio/resultados/*)

Discussão dos Resultados

- Resultados melhores para os casos que estudamos (se comparado aos métodos de treinamento supervisionado tradicionais)
- Maior variedade de taxas de acertos dependendo do conjunto de objetos escolhidos.
- Destaque para a utilização de várias amostras (em contraste com o que a literatura salienta).

Desenvolvimentos Futuros

- Afinar a extração de características (limiar utilizado só recentemente).
- Redes de Prosódia, visibilidade e seccionamento do espectro.
- Em conjunto com os métodos já utilizados, empregar a *Propagação de Rótulo* nos trabalhos.

Na Rede

- svn co http://svn.assembla.com/svn/audioexperiments/NinjaML

Referências Principais

- Zhu, X. and Lafferty, J. and Rosenfeld, R. "Semi-supervised learning with graphs.", 2005
- Zhu, X. and Ghahramani, Z., "Learning from labeled and unlabeled data with label propagation.", 2002
- Blum, A. and Chawla, S., "Learning from labeled and unlabeled data using graph mincuts", 2001
- (Mais detalhes na monografia)