Universidad del Cauca

Facultad de Ciencias Naturales, Exactas y de la Educación

Departamento de Matemáticas

Tipo de actividad: Asignatura(MAT662)	Créditos: 0
Nombre: MODELOS DEL LENGUAJE.	Intensidad Horaria: 4 Horas semanales.
Requisitos: MAT221,MAT261	Correquisitos:

Introducción

La construcción de modelos del lenguaje (ML) es una tarea central en los sistemas de reconocimiento automático del habla, procesamiento del lenguaje natural y otras aplicaciones. Su propósito es restringir y guiar la búsqueda de secuencias de unidades lingüísticas a través de un conjunto de hipótesis y contribuir a la determinación de la transcripción final. Un modelo del lenguaje no es otra cosa que una descripción del lenguaje. En su forma más simple puede ser una representación de las palabras que pertenecen al lenguaje, en los modelos mas complejos también se trata de describir la estructura y significado de las frases pertenecientes al lenguaje.

Los modelos estocásticos han ganado considerable aceptación en los últimos años debido a la eficiencia demostrada en áreas como reconocimiento automático del habla, traducción automática, desambiguación de fronteras, etiquetado automático con clases, corrección ortográfica y en otras aplicaciones en las que se procesan elementos del lenguaje en situaciones de conocimiento incompleto.

En términos simples un modelo de lenguaje estocástico es una distribución de probabilidades que permite modelar la probabilidad de que una frase sea dicha en un lenguaje específico.

Objetivo General

Dar a conocer las diferentes propuestas de formulación de modelos matemáticos que representan la distribución probabilística de las frases en un lenguaje dado.

Objetivos especificos

1. Adquirir los conceptos básicos del modelado del lenguaje.
2. Conocer las diferentes propuestas de modelos del lenguaje.
3. Implementar computacionalmente las técnicas para el entrenamiento y aplicación de los modelos del lenguaje.
4. Desarrollar aplicaciones computacionales en las que se utilicen los modelos del lenguaje

Contenido

CAPITULO I. Introducción.
1. Que es un modelo de lenguaje?
2. Diferentes enfoques para la construcción de modelos del lenguaje.
3. modelos de lenguaje probabilísticos
4. Planteamiento del modelo desde la perspectiva de la teora de la información.
5. Modelos condicionales.
6. Modelos Mixtos.

CAPITULO II. El modelo de n-gramas.
1. Definición del modelo de n-gramas
2. Modelo de bigramas y trigramas
3. Estimación de probabilidades para el modelo de bigramas y trigramas
4. Suavizado en el modelo de n-gramas
5. Interpolación para el modelo de trigramas.
6. Ventajas y desventajas del modelo de n-gramas.
7. Diferentes funciones discriminante.

CAPITULO III. Modelos basados en gramáticas incontextuales.
1. Introducción.
2. Gramáticas formales.
3. Gramáticas libres de contexto.
4. Aprendizaje de las reglas en una gramática libre de contexto.
5. Gramáticas probabilísticas.
6. Estimación de las probabilidades en las gramáticas libres de contexto.
7. Las gramáticas como modelo de lenguaje.
8. Otras alternativas basadas en gramáticas.

CAPITULO IV. Modelos basados en grafos.
1. Modelos basados en árboles de decisión.
2. Modelo VNSA.
3. Modelos basados en analizadores léxicos.
4. Otros modelos de árbol.

CAPITULO V. Modelo de máxima entropía.
1. Introducción.
2. El principio de máxima entropía.
3. Modelo del lenguaje basado en máxima entropía.
4. Estimación de los parámetros en el modelo de máxima entropía.
5. El algoritmo ”Generalized Iterative Scaling”.
6. El algoritmo ”Improved Iterative Scaling”
7. Entrenamiento y aplicación de los modelos de Máxima entropía.

CAPITULO VI. Modelo de mínima divergencia
1. La divergencia de Kullback.
2. El modelo de lenguaje de mínima divergencia.
3. Definición de las características en el modelo de mínima divergencia.
4. Estimación de los parámetros en el modelo de mínima divergencia
5. Aplicaciones del modelo de mínima divergencia.

CAPITULO VII. Modelos basados en análisis semántico latente (LSA).
1. Definición de LSA.
2. Aspectos generales del modelo
3. Métricas en el modelo LSA.
4. Espacios vectoriales y el modelo LSA.
5. Estimación de los parámetros en los modelos LSA.
6. Aplicación del modelo.

Bibliografía

1. F. Amaya. Modelos del lenguaje mixtos. Technical report, Departamento de Sistemas Informáticos y Computación, Universidad Politécnica de Valencia, 2000.
2. J. R. Bellegarda. A multispan language modeling framework for large vocabulary speech recognition. IEEE Transactions on Speech and Audio Processing, 6 (5):456–467, 1998.
3. J.R. Bellegarda. A latent semantic analysis framework for large-span language modeling. 3:1451–1454, 1997.
4. A. Borthwick. Survey paper on statistical language modeling. Technical report, New York University, 1997.
5. S. Chen and J. Goodman. An Empirical Study of Smoothing Techniques for Language Modeling. Technical Report TR-10-98, Center for Research in Computing Technology, Harvard University, 1998.
6. F. Jelinek. Statistical Methods for Speech Recognition. The MIT Press, Massachusetts Institut of Technology. Cambridge, Massachusetts, 1997.
7. R. Rosenfeld. Adaptive statistical language modeling: a Maximum Entropy aproach. Technical Report CMUCS-94-138, Carnegie Mellon University, 1994