

## Curso

### Lingüística Aplicada III (Procesamiento inteligente de textos)

#### **DURACIÓN:**

80 horas

#### **AUTOR DEL PROGRAMA:**

Dr. Alexander Gelbukh, Dr. Grigori Sidorov

#### **OBJETIVO GENERAL**

Describir que tipos de sistemas de procesamiento inteligente de textos existen y como se usan. Analizar a nivel detalle los sistemas de análisis automático morfológico y semántico.

#### **BIBLIOGRAFÍA**

1. I. Bolshakov, A. Gelbukh. Computational linguistics. CIC-IPN, 2004, 187 p.
2. J. Allen. Natural Language Understanding. The Benjamin/Cummings Publishing Company, 1995, 654 p.
3. R. Grishman. Computational Linguistics. Cambridge University Press, 1986.
4. Oxford Handbook on computational linguistics. Oxford University Press, 2003, 808 p.
5. Manning C. and H. Schütze. Foundations of statistical natural language processing. MIT Press, 1999, 680 p.
6. Jurafsky, D. and J. Martin. Speech and language processing. Prentice Hall, 2000, 934 p.

#### **CONTENIDO DEL CURSO**

##### **1. Aplicaciones y Fuentes de la Lingüística Aplicada**

- 1.1. Resumen de las aplicaciones del software lingüístico
  - 1.1.1. Sistemas de preparación de textos. Correctores
    - 1.1.1.1. Formateadores. Sistemas de guiones
    - 1.1.1.2. Correctores de ortografía
    - 1.1.1.3. Correctores de gramática
    - 1.1.1.4. Correctores de estilo
  - 1.1.2. Extracción de información y sistemas de minería de datos
    - 1.1.2.1. Sistemas no inteligentes de extracción de información
    - 1.1.2.2. Sistemas inteligentes de extracción de información
  - 1.1.3. Sistemas de traducción

- 1.1.3.1. Generaciones de sistemas de traducción
- 1.1.4. Interfaces de lenguaje natural
- 1.1.5. Otros sistemas lingüísticos y áreas vecinas
  - 1.1.5.1. Sistemas de generación de textos
  - 1.1.5.2. Sistemas de compresión de textos
  - 1.1.5.3. Sistemas de reconocimiento
    - 1.1.5.3.1. Reconocimiento de caracteres ópticos
    - 1.1.5.3.2. Reconocimiento del habla
      - 1.1.5.3.2.1. Sistemas de control de la voz
      - 1.1.5.3.2.2. Sistemas de dictado
      - 1.1.5.3.2.3. Compresión de voz
    - 1.1.5.3.4. Reconocimiento de escritura manuscrita
  - 1.1.5.4. Sistemas de referencia, diccionarios y thesauri
- 1.2. Lenguajes artificial y natural
  - 1.2.1. Principales problemas en el procesamiento del lenguaje natural
    - 1.2.1.1. Problemas de ambigüedad
    - 1.2.1.2. Información incompleta e implícita
    - 1.2.1.3. Conocimiento de papeles
  - 1.2.2. Conocimiento lingüístico y extralingüístico
- 1.3. Resumen de las fuentes utilizadas en los sistemas de lingüística aplicada
  - 1.3.1. Diferencias entre los diccionarios orientados para el uso humano o de la computadora
  - 1.3.2. Diccionarios tradicionales
  - 1.3.3. Diccionarios combinatorios
  - 1.3.4. Diccionarios de redes semánticas y ontologías
  - 1.3.5. Gramáticas

## **2. Análisis morfológico y homonimia**

- 2.1. Resultados de análisis morfológico
- 2.2. Homonimia morfológica
- 2.3. Etiquetación (tagging) y los corpus textuales
- 2.4. La etapa presintáctica de procesamiento

## **3. Synonymia morfológica**

- 3.1. Sinonimia morfológica en general
- 3.2. Sinonimia morfológica de formas de uno o varios lexemas

## **4. Sistemas de análisis y síntesis morfológico**

- 4.1. Funciones del procesador morfológico
  - 4.1.1. Análisis y Síntesis morfológico
  - 4.1.2. Normalización
  - 4.1.3. Corrector de ortografía
    - 4.1.3.1. Corrección de errores
      - 4.1.3.1.1. Errores tipográficos
      - 4.1.3.1.2. Errores gramaticales
    - 4.1.3.2. Corrección de palabras
  - 4.1.4. El papel del módulo morfológico en un procesador lingüístico
- 4.2. Análisis premorfológico
- 4.3. Representación de los elementos morfológicos
  - 4.3.1. Palabras. Lexemas
  - 4.3.2. Morfos. Morfemas

- 4.3.3. Paradigmas
- 4.3.4. Léxico: raíces, supletivismo, clases de palabras
- 4.3.5. Asuntos de codificación. Variantes gráficas. Capitalización
- 4.4. Análisis de derecha a izquierda y de izquierda a derecha
- 4.5. Organización del diccionario morfológico
  - 4.5.1. Requerimientos de un diccionario morfológico
    - 4.5.1.1. Cómo tratar la homonimia
    - 4.5.1.2. Requerimientos de memoria y velocidad
    - 4.5.1.3. *Caching of dictionary accesses*
  - 4.5.2. Métodos de derecha a izquierda
    - 4.5.2.1. Tablas *Hash*
    - 4.5.2.2. Cómo usar una base de datos estándar
  - 4.5.3. Métodos de izquierda a derecha
    - 4.5.3.1. Árboles
    - 4.5.3.2. Búsqueda por orden alfabético
      - 4.5.3.2.1. Principio de acceso a la localidad de la memoria
      - 4.5.3.2.2. Definición de un diccionario morfológico
      - 4.5.3.2.3. Búsqueda mejorada por orden alfabético
    - 4.5.3.3. *Caching of dictionary accesses*
  - 4.5.4. Diccionario universal para el procesamiento de textos
- 4.6. Síntesis morfológica
  - 4.6.1. Representación descriptiva de datos. Conjunto de datos común para el análisis y la síntesis
- 4.7. Enfoques de análisis morfológico automático
  - 4.7.1. Conjuntos de palabras.
  - 4.7.2. Transductores. PC-KIMMO.
  - 4.7.3. Generación de raíces durante el análisis.
  - 4.7.4. Generación de raíces de antemano.
  - 4.7.5. Las tablas gramaticales.
  - 4.7.6. Análisis a través de generación.
- 4.8. Ejemplo: AGME
- 4.9. Ejemplo: RMORPH
- 4.10. Ejemplo: AMORE™
  - 4.10.1. Representación de datos
    - 4.10.1.1. Tablas uniformes morfológicas
    - 4.10.1.2. Partes técnicas del discurso
    - 4.10.1.3. Tablas de morfos
    - 4.10.1.4. Listas de gramatemas
    - 4.10.1.5. Listas de máscaras
      - 4.10.1.5.1. Raíces supletivas
      - 4.10.1.5.2. Alternancias en las raíces
      - 4.10.1.5.3. Variedad
    - 4.10.1.6. Diccionario
  - 4.10.2. Algoritmos
    - 4.10.2.1. Análisis morfológico
    - 4.10.2.2. Síntesis morfológica
    - 4.10.2.3. Normalización

- 4.10.2.4. Corrección de errores
  - 4.10.2.4.1. Corrección errores gramaticales
  - 4.10.2.4.2. Corrección errores de una letra

## **5. Representación semántica. Pragmática**

- 5.1. Tareas de análisis semántico
  - 5.1.1. Representación semántica
  - 5.1.2. Redes y predicados semánticos
  - 5.1.3. El problema de la ambigüedad semántica
- 5.2. Estructuras asociativas en el lenguaje
  - 5.2.1. Resolución de anáforas
    - 5.2.1.1. Pronombres
    - 5.2.1.2. Artículos españoles con funciones anafóricas
  - 5.2.2. Co-referencia
    - 5.2.2.1. Nombres y referentes
    - 5.2.2.2. Uso referencial y no referencial
    - 5.2.2.3. Resolución co-referencial
      - 5.2.2.3.1. Articulación real
        - 5.2.2.3.1.1. El papel de los artículos
      - 5.2.2.3.2. Co-referencia de verbos
  - 5.2.3. Deixis
  - 5.2.4. *Scoping*
- 5.3. Restricciones semánticas
  - 5.3.1. Clases semánticas de palabras
  - 5.3.2. Otros métodos para imponer restricciones semánticas
- 5.4. Diccionarios utilizados para el análisis semántico
  - 5.4.1. Diccionarios combinatorios
  - 5.4.2. Diccionarios de redes semánticas
    - 5.4.2.1. Ejemplo: WordNet
    - 5.4.2.2. FACTOTUM® SemNet
- 5.5. Modelos de procedimiento de parsing basados en la semántica
  - 5.5.1. Ejemplo: *Parser* experto de palabras
- 5.6. Síntesis semántica
- 5.7. Análisis de discurso
  - 5.7.1. Estructura del discurso
  - 5.7.2. El papel de la pragmática

## **6. Aplicaciones del procesamiento inteligente de textos**

- 6.1. Sistemas de minería de datos como sistemas de Significado  $\Leftrightarrow$  Texto
- 6.2. Sistemas de traducción por computadora como sistemas de Significado  $\Leftrightarrow$  Texto
  - 6.2.1. Características de los sistemas de traducción
- 6.3. Sistemas de comprensión de textos y extracción de conocimiento como sistemas de Significado  $\Leftrightarrow$  Texto

## **7. Appendix. Clases prácticas y ejercicios**

- 7.1. Análisis y síntesis morfológica
  - 7.1.1. Cómo determinar la estructura morfológica de una oración en español
  - 7.1.2. Implementación del software de un sencillo analizador morfológico para el español

- 7.1.3. Compilación de un diccionario y de tablas morfológicas en español
- 7.2. Cómo determinar la estructura semántica de un texto corto en español.