Curso

Tecnologías de lenguaje natural

DURACIÓN:

80 horas.

AUTOR DEL PROGRAMA:

Dr. Alexander Gelbukh, Dr. Igor Bolshakov

OBJETIVO GENERAL

Presentar las diferentes tipos de lenguas y que consecuencias se generan según su estructura para el procesamiento automático. Presentar la estructura general del procesador lingüístico y como construirlo. Describir que fuentes de conocimiento lingüístico se usa en el procesador. Definir el concepto de la función léxica y sus tipos. Describir la teoría Significado-Texto-

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. D. Crystal. A dictionary of Linguistics and Phonetics. Blackwell Publishers, 1991.
5. The Handbook of Linguistics. Blackwell Publishers, 2001, 824 p.
6. Oxford Handbook on computational linguistics. Oxford University Press, 2003, 808 p.
7. Manning C. and H. Schutze. Foundations of statistical natural language procesing. MIT Press, 1999, 680 p.
8. Jurafsky, D. and J.Martin. Speeach and language processing. Prentice Hall, 2000, 934 p.
9. J. Steele, (Ed.) Meaning-Text Theory. University of Ottawa Press, 1990.

CONTENIDO DEL CURSO

1. **Taxonomía de las lenguas**
   - 1.1. Lenguas analíticas
   - 1.2. Lenguas sintéticas
     - 1.2.1. Lenguas inflectivas
     - 1.2.2. Lenguas intraflectivas
     - 1.2.3. Lenguas aglutinativas

1.2.4. Lenguas polisintéticas
1.3. Características de lenguas específicas
1.3.1. Lenguas utilizadas en México
1.3.1.1. Español
1.3.1.2. Náhuatl, maya, y otras lenguas indígenas
1.3.2. Otras lenguas
1.3.2.1. Inglés
1.3.2.2. Principales características del francés, alemán, ruso
1.4. Implicaciones para el diseño de sistemas de software para el procesamiento de distintas lenguas
1.4.1. Comparación de los métodos útiles para el procesamiento de textos en español, inglés, francés y ruso
1.4.2. Procesamiento del español: Lo que se puede adoptar de la tradición Norteamericana, y lo que no se puede
1.5. Necesidad para desarrollar sistemas orientados hacia el procesamiento de textos en Español

2. Procesadores lingüísticos
2.1. Estructura general de un sistema inteligente dotado para el procesamiento de un lenguaje natural
2.1.1. Módulo de razonamiento
2.1.2. Procesador lingüístico
2.2. Procesadores lingüísticos
2.2.1. Procesadores lingüísticos multipropósitos
2.2.2. Procesadores lingüísticos especializados
2.3. Estructura de un procesador lingüístico multipropósitos
2.3.1. Analizador y sintetizador de textos
2.3.2. Módulo morfológico
2.3.3. Módulo sintáctico
2.3.4. Módulo semántico

3. Desarrollo de sistemas lingüísticos
3.1. Problemas de los proyectos interdisciplinarios
3.1.1. El papel de los lenguajes formales en la codificación de datos
3.1.2. El papel del conocimiento lingüístico para los programadores
3.2. Problemas generales de los proyectos de desarrollo de grandes cantidades de software
3.2.1. Mantenimiento de los datos
3.2.2. Mantenimiento del código
3.2.3. Elección del (los) lenguaje(s) de programación
3.2.3.1. Lenguajes de “inicio rápido, alto inmediato”
3.2.3.2. Lenguajes de “inicio lento, desarrollo constante”
3.2.3.3. Lenguajes de inteligencia artificial
3.3. Fuentes lingüísticas
3.3.1. Representación de las fuentes lingüísticas
3.3.2. Mantenimiento de las fuentes lingüísticas
3.3.3. Especificaciones para la compilación y el mantenimiento de grandes series de datos
3.3.4. Lenguajes de programación para el trabajo lexicográfico

4. Fuentes lingüísticas y lenguajes para su representación

4.1. Conocimiento descriptivo y de procedimiento
  4.1.1. Principio de Yngve
  4.1.2. El papel de las fuentes lingüísticas
  4.1.3. Independencia de las lenguas
  4.1.4. Fuentes comunes para el análisis y la síntesis
  4.1.5. Limitaciones del principio de Yngve
    4.1.5.1. Ejemplo: Modelo de *parser* experto de palabra
4.2. Tipos de fuentes lingüísticas
  4.2.1. Gramáticas
  4.2.2. Diccionarios
    4.2.2.1. Diccionarios morfológicos
    4.2.2.2. Diccionarios de combinaciones de palabras
      4.2.2.2.1. Ejemplo: CrossLexica™
    4.2.2.3. Diccionario de patrones de gobierno
      4.2.2.3.1. Ejemplo: Diccionario de español
    4.2.2.4. Diccionarios de redes semánticas
      4.2.2.4.1. Ejemplo: WordNet
      4.2.2.4.2. Ejemplo: FACTOTUM® SemNet
  4.2.3. Corpus de textos
4.3. Compilación de fuentes lingüísticas
  4.3.1. Métodos de compilación de fuentes lingüísticas
    4.3.1.1. El papel del trabajo automático
    4.3.1.2. El papel del trabajo manual
      4.3.1.2.1. Introspección. Intuición lingüística
    4.3.1.3. Trabajo manual ayudado por una computadora
      4.3.1.3.1. Investigación de datos
      4.3.1.3.2. Uso del corpus y ejemplos de bases de datos
    4.3.1.4. Extracción de información del corpus de textos
      4.3.1.4.1. Nociones básicas de la teoría de la probabilidad
        4.3.1.4.1.1.Probabilidad de un evento complejo
        4.3.1.4.1.2.Leyes de distribución
      4.3.1.4.2. Ley de Zipf
      4.3.1.4.3. Métodos estadísticos para la compilación de diccionarios
      4.3.1.4.4. Gramáticas y diccionarios estadísticos y su utilización en el *parsing*
        4.3.1.4.4.1.El primero de los *parsings* que resulte el mejor
        4.3.1.4.4.2.Resolución de la ambigüedad
    4.3.1.5. Extracción de información de los diccionarios orientados hacia el uso humano
    4.3.1.6. Traducción de fuentes léxicas
      4.3.1.6.1. Importancia y limitaciones de la traducción de diccionarios
      4.3.1.6.2. Traducción de los diccionarios de combinaciones de palabras

4.3.1.6.3. Traducción de los diccionarios de redes semánticas

4.3.2. Mantenimiento de las fuentes

4.3.2.1. Fuentes lingüísticas como grandes series de datos

4.3.2.2. El papel de la documentación

4.3.2.3. Importancia de la consistencia y uniformidad

5. Funciones léxicas

5.1. Funciones léxicas tradicionales

5.1.1. Hipónimos e hipérnimos

5.1.2. Sinónimos y antónimos. Parónimos

5.1.3. Holónimos

5.1.4. Derivadas

5.2. Funciones léxicas avanzadas

5.2.1. Propiedades: *Magn*, *Bon*, etc.

5.2.2. Acciones: *Oper*, *Func*, *Labor*, etc.

5.2.3. Derivadas: *S*, *A*, etc. *Conv*

5.3. El uso de las funciones léxicas en la representación semántica

5.3.1. Normalización de la estructura semántica

5.3.2. Valencias sintácticas y semánticas

6. La teoría Sentido – Texto

6.1. Transformador de etapas múltiples y patrones de manejo

6.2. Árboles de dependencias

6.3. Vínculos semánticos

6.4. Puntos de vista posibles sobre los lenguajes naturales

6.5. El lenguaje como un transformador bi-direccional

6.6. Texto, ¿qué es?

6.7. Significado, ¿qué es?

6.8. Dos maneras de representar el sentido

6.9. Descomposición y atomización del sentido

6.10. Carencia de unicidad del Mapeo de Sentido a Texto: Sinonimia

6.11. Carencia de unicidad de Mapeo de Texto a Sentido : Homonimia

6.12. Carácter multietapa del Transformador Significado a Texto

6.13. Traducción como una transformación multietapa

6.14. Los dos lados de un signo

6.15. Signo lingüístico

6.16. El signo lingüístico en la MTT

6.17. El signo lingüístico en HPSG

6.18. ¿Son los significantes dados por naturaleza o por convención?

6.19. Ideas generativas, de la MTT y de constricciones en comparación

6.20. Los rasgos principales de la MTT: dinamismo, formalidad, transformacionalidad, la adopción total de las nociones y tradiciones de la lingüística estructural.

7. Appendix. Clases prácticas y ejercicios

7.1. Herramientas de software para el trabajo lexicográfico

7.1.1. Utilidades del procesamiento de textos

7.1.1.1. *grep*, herramienta de búsqueda

7.1.1.2. *tr*, herramienta de transliteración

7.1.1.3. *sort*, herramienta de arreglo