UNIVERSIDAD DE LOS ANDES
FACULTAD DE INGENIERÍA
POSTGRADO EN COMPUTACIÓN

Curso de Procesamiento del Lenguaje Natural
Facilitadores: Jacinto Dávila Q., Hilda Contreras H., Luis Astorga J., y Melva Márquez R. (Grupo de Ingieniería Lingüística, CESIMO, ULA)
Fecha: 25-29 de enero de 2010
Duración: 20 horas

Introducción y objetivos

La lingüística computacional es un área de conocimiento interdisciplinario en la que confluyen la lingüística teórica y aplicada, la informática, la inteligencia artificial y la ciencia cognitiva. Su principal objetivo es la modelización a través de programas informáticos de los comportamientos lingüísticos de los pares hablante-oyente y escritor-lector, en los cuales se resumen las cuatro grandes habilidades lingüísticas humanas: comprensión/producción oral y escrita. Estas habilidades en el ámbito computacional llevan por nombre generación y síntesis (producción) y análisis o interpretación (comprensión).

El curso de procesamiento de lenguaje natural diseñado para este programa de postgrado tiene dos grandes objetivos:

1. Iniciar al estudiante en el conocimiento de la lingüística computacional y el procesamiento del lenguaje natural (humano) - PLN, a través de la comprensión de conceptos centrales, de relaciones con otras disciplinas de estudios, alcances, limitaciones y posibilidades de expansión dentro del interdisciplinariedad predominante en su naturaleza.
2. Diseñar y elaborar programas de PLN empleando arquitecturas simbólicas, probabilísticas o híbridas a partir de recursos lógicos y matemáticos.

El programa de esta asignatura posee cuatro grandes secciones, a saber: La primera sección es introductoria, dentro de la cual se ofrecen fundamentos de la lingüística formalizables en el PLN; la segunda sección muestra algunas aplicaciones y las diferentes arquitecturas conocidas en el PLN, así como también, los niveles de análisis que se aplican en un sistema de PLN; la tercera sección, más orientada hacia el PLN de textos escritos, se enfoca en los diferentes modelos empleados, así como también en los conceptos y recursos más vistosos de la aproximación computacional hacia el texto escrito, por lo cual tiene un subapartado que explica el resumidor de textos especializados diseñado por miembros del GIL-CESIMO; la cuarta y última sección está más orientada hacia la programación matemática de la lingüística. Dentro de esta sección, se ofrece una introducción a la Teoría de Categorías, así como a la programación funcional con Haskell y a los analizadores monádicos en sintaxis y semántica, principalmente. .

Temario

Sección 1: Introducción: La lingüística en el PLN
Facilitadora: Melva J. Márquez Rojas

1.1 Definición, objetivos, marco interdisciplinar, pinceladas históricas.
1.2 Representación y niveles del conocimiento lingüístico.
1.3 Las gramáticas formales.
1.4 Corpus lingüísticos.
1.5 Fundamentos lingüísticos para el desarrollo de un anglicizador.

Sección 2: Aplicaciones, arquitectura y herramientas de PLN

Facilitadora: Hilda Y. Contreras Hernández

2.1 Aplicaciones y arquitectura de PLN
- 2.1.1 Niveles linguísticos
- 2.1.2 Arquitecturas simbólicas, probabilísticas e híbridas

2.2 Herramientas de PLN
- 2.2.1 Preprocesos
- 2.2.2 Análisis morfológico (*tagging*)
- 2.2.3. Análisis sintáctico superficial (*chunking*)
- 2.2.4. Análisis sintáctico (*parsing*)
- 2.2.5. Análisis semántico
- 2.2.6. Directorios de herramientas, recursos y documentación, aplicaciones

Sección 3: Extracción de conocimiento a partir de textos en lenguaje natural.
Facilitador: Jacinto Dávila Quintero

A Objetivo general de la investigación

3.1 Lingüística textual
3.2 Procesamiento humano del lenguaje textual
3.3 Procesamiento del lenguaje natural por parte del computador
- 3.3.1 Conocimiento Lingüístico
- 3.3.2 Estado del arte de las técnicas de resumen automático

3.4 El estilo en el lenguaje escrito
- 3.4.1 Reglas de estilo de Williams
- 3.4.2 Claridad
- 3.4.3 Cohesión y Coherencia

3.5 Problema de la investigación
- 3.5.1 El idioma y la gramática
- 3.5.2 Tipos de textos y estilo
- 3.5.3 Semántica y Pragmática
- 3.5.4 Problema específico

B Un resumidor simbólico: un experimento para evaluar. Las gramáticas basadas en estilo

3.6.1 Descripción general de la estrategia del resumidor simbólico
3.6.2 Descripción detallada e implementación del resumidor
- 3.6.2.1 Metodología y herramientas empleadas en la implementación
- 3.6.2.2 Módulos del resumidor
  - 3.6.2.2.1 Tokenizador
  - 3.6.2.2.2 Gramática
  - 3.6.2.2.3 Claridad

3.6.2.2.4 Cohesión / Coherencia
3.6.2.2.5 Tópico común
3.6.2.2.6 Salida
3.6.2.3 Diccionarios empleados
3.6.2.3.1 Diccionario de verbos
3.6.2.3.2 Diccionarios de conectores
3.7Modelos textuales
3.7.1.Noción de texto
3.7.2.Definición de Modelos textuales
3.7.3.Intención comunicativa
3.7.4.Carácter cultural
3.7.5.Base textual y base cognitiva de los textos de especialidad

3.8. Artículos de investigación científica (AIC). Modelo Textual
3.8.1. Base textual de los AIC
3.8.1.1 Dimensión del discurso según su contenido
3.8.1.2 Organización del discurso
3.8.1.3 Estilo del discurso
3.8.2 Base cognitiva de los AIC
3.8.2.1 Modelo IMRD
3.8.2.2 Las introducciones en los AIC
3.8.2.2.1 Modelo Swales
3.8.2.3 Las discusión en los AIC

3.9Procesamiento computacional del lenguaje natural
3.9.1. Lingüística computacional
3.9.2. Lenguajes formales
3.9.2.1. Gramáticas formales
3.9.2.2.Gramáticas de estructura de frase (DCG)
3.9.3 Lenguajes de programación lógica
3.9.4 PROLOG

3.10 Resumidor automático

3.11 Un repositorio para procesamiento del lenguaje natural y extracción de conocimiento textual.
http://resumidor.sourceforge.net/
http://resumidor.svn.sourceforge.net/viewvc/resumidor/

Sección 4: Lingüística Matemática
Facilitador: Luis Astorga Junquera

4.1Introducción a la Teoría de Categorías
4.1.1.Categorías, funtores, transformaciones naturales y dualidad. Ejemplos: algebra de términos y autómatas (co-algebras).
4.1.2.Construcciones básicas. Ejemplos: circuitos lógicos y diagramas de flujo.
4.1.3.Adjunción y mónadas. Ejemplos: algebras libres y categorías de Kleisi.
4.2Rudimentos de programación funcional con Haskell
4.2.1.Sesiones, declaraciones, reducción de expresiones y evaluación

perezosa.
4.2.2.Polimorfismo, constructores de tipos y sistema de clases. Ejemplo: recursión vs. iteración en listas y árboles.
4.2.3.Especificaciones algebraicas y coalgebraicas de estructuras de datos. Ejemplos: co-inducción y co-recursión en listas y árboles infinitos.
4.2.4.Programación con mónadas, mónadas básicas, combinación de mónadas y listas por comprensión. Ejemplo: la mónada de las expresiones regulares y los analizadores monádicos.
4.3 Analizadores monádicos para lenguajes naturales
4.3.1. El cálculo Lambda con tipos y las categorías cartesianas cerradas.
4.3.2.Gramáticas categoriales combinatorias y mónadas sintácticas.
4.3.3.Gramáticas de Montague y mónadas semánticas.

Referencias

[1] Contreras, H. Yelitza. Una técnica para la extracción automática de resúmenes
basada en una gramática de estilo. Tesis de Maestría. Postgrado en Computación. Universidad de Los Andes. 2002.

[2] Parra, M. Marilú. Un modelo computacional para la generación de resúmenes automáticos de artículos científicos en español. Tesis de Maestría. Postgrado en Modelado y Simulación de la ULA. 2004.

[3] Arteaga Matute, Leonardo José. Estrategia para el procesamiento de lenguaje natural español escrito en páginas web con un resumidor de textos simbólicos basado en estilos. Tesis de Licenciatura en Computación. La Universidad del Zulia. 2008.

[4] María Marilú Parra y Jacinto Dávila. Un Modelo Computacional para la Generacion de Resumenes Automaticos de Articulos Cientificos en Español. ACTAS del IX Simposio Internacional de Comunicación Social. Santiago de Cuba, 24 al 28 de Enero, CUBA, 2005.

[5] Melva Márquez y Jacinto Dávila. La terminología y la ingeniería lingüística en la Universidad de Los Andes: historia y situación actual. Terminometro. Ed. Unión Latina. Número 6. Paris. 2002.

[6] Jacinto A. Dávila, Luis Astorga, Melva Márquez, H. Yelitza Contreras, Jacobo Myerston y M. Marilú Parra . Introducción a la lingüística computacional con una perspectiva interdisciplinaria. Terminometro. Ed. Unión Latina. Número. 6. Paris. 2002.

[7] Jacinto Dávila y H. Yelitza Contreras. Una gramática de estilos para resumir textos en español. Revista de La Sociedad Español para el Procesamiento del Lenguaje Natural (SEPLN). Valladolid 11 al 13 de Septiembre de 2002.
http://webdelprofesor.ula.ve/ingenieria/jacinto/indices/index-sepln-2002.htm

[18] Sergi Balari Ravera. Formalismos gramaticales de unificación y procesamiento basado en restricciones. Volumen Monográfico, 1999.

Disponible en: http://dialnet.unirioja.es/servlet/fichero_articulo?codigo=227028&orden=67517

[19] Sofía Galicia Haro y Alexander Gelbukh. Investigaciones en análisis sintáctico para el español. 2007. Disponible en: http://www.gelbukh.com/libro-investigaciones/

[20] Melva Márquez. El anglicismo terminológico integral en los textos especializados: pautas para su tratamiento automatizado. Barcelona: Universitat Pompeu Fabra Tesis de doctorado. 2005. Disponible en http://www.tesisenxarxa.net/TDX-0307105-134028/index.html

[21] Carme Bach i Martorell. Los marcadores de reformulación como localizadores de zonas discursivas relevantes en el discurso especializado. Revista electrónica Debate terminológico, Número 1. 2005. Disponible en: http://www.riterm.net/revista/n_1/bach.pdf.

[22] Ruslan Mitkov (editor) The Oxford Handbook of Computational Linguistics. Oxford University Press. 2004. Disponible en: http://books.google.co.ve/books?id=OaClhre-vW4C&dq=The+Oxford+handbook+of+computational+linguistics&printsec=frontcover&source=bn&hl=es&ei=o5JMS4HXBsqWtgfcx9DhDA&sa=X&oi=book_result&ct=result&resnum=5&ved=0CCQQ6AEwBA#v=onepage&q=&f=false

[23] María Antonia Martí y Joaquim Llisterri. La ingeniería lingüística en la sociedad de la información. Digithum, Revista digital d'humanitats, Número 3, Universitat Oberta de Catalunya- . Publicación electrónica. 2001. Disponible en: http://www.uoc.edu/humfil/articles/esp/llisterri-marti/llisterri-marti.html