ASIGNATURA DE MÁSTER:

ACCESO INTELIGENTE A LA INFORMACIÓN

Curso 2011/2012

(Código:31101019)

1.PRESENTACIÓN

Tipo	Optativa
Cuatrimestre	Primero
Créditos/horas totales	6/150
Horas de estudio teórico	100
Horas de prácticas	50
Horas complementarias	0

2.CONTEXTUALIZACIÓN

Esta asignatura se encuadra en el módulo "ESP-LSI-1 Tecnologías del Lenguaje en la Web" dentro de la especialidad con el mismo nombre de la titulación de posgrado "Master en Lenguajes y Sistemas Informáticos". Dentro de esta especialidad, "Acceso Inteligente a la Información" revisa los avances más recientes y las áreas de investigación más activas dentro de las tecnologías de la lengua y la recuperación de información.

3.REQUISITOS PREVIOS RECOMENDABLES

Lectura fluida del inglés y conexión a Internet.

4.RESULTADOS DE APRENDIZAJE

El curso pretende introducir a los alumnos en los temas de investigación más recientes relacionados con el Procesamiento del Lenguaje Natural aplicado a la exploración y manipulación de grandes volúmenes de información textual (como la disponible en la Web). El temario inicial cubre la Recuperación de Información Multilingüe, la extracción de resúmenes, la búsqueda automática de respuestas y la extracción de información, todos ellos temas en los que se puede iniciar, con posterioridad al curso, un trabajo de iniciación a la investigación en la especialidad de Tecnologías del Lenguaje en la Web. El temario se actualizará anualmente para adecuarse a las nuevas aplicaciones que surjan en este campo.

El alumno adquirirá las siguientes destrezas y competencias:

Debe tener una visión de conjunto de la investigación en sistemas de acceso a la información, como punto de confluencia entre los campos de Recuperación de Información y Procesamiento del Lenguaje Natural.

Debe ser capaz de realizar una lectura crítica de artículos científicos sobre el tema, de localizar y discriminar información bibliográfica relevante, y de sintetizar información de distintas fuentes.

Debe ser capaz de redactar con rigor científico y de comunicar y debatir con pares (en este caso, sus compañeros) sus

análisis y opiniones en torno a los temas de la asignatura.

Debe ser capaz de diagnosticar los puntos débiles del estado del arte actual en el campo y de esbozar temas y líneas de investigación potenciales.

5.CONTENIDOS DE LA ASIGNATURA

Estructura y contenido teórico

Tema 1: Recuperación de Información Multilingüe.

-1.1 Recursos para la Recuperación de Información dependientes del idioma:

-Lematizadores.

-Stemmers.

-Segmentación de compuestos y palabras.

-1.2 Traducción de Consultas, el problema de la Fusión Documental:

-Diccionarios bilingües, diccionarios con información gramatical.

-Estructuración de la consulta según la traducción.

-Idiomas pivote.

-Utilización de corpora paralelos/comparables.

-Programas de traducción automática.

-Tesauros.

-Múltiples idiomas: el problema de la fusión documental.

-1.3 Traducción de Documentos, otros enfoques al problema:

-Traducción de los documentos vs. traducción de las consultas.

-Traducciones bidireccionales.

-Representación conceptual.

-1.4 Sistemas interactivos de búsqueda de información multilingüe.

Tema 2: Extracción de Información.

-2.1 Perspectiva histórica y objetivos:

-Reconocimiento de información relevante

-Las conferencias MUC.

-2.2 Arquitectura de los sistemas de Extracción de Información:

-Preprocesado y análisis de los documentos.

-Reconocimiento de patrones y entidades

-Resolución de co-referencias.

-Generación de la salida esperada.

-2.3 Extracción de Información Multilingüe:

-Reconocedores del idioma.

-Traducción de la información extraída vs. traducción de los documentos:

-Sistemas de extracción de información translingües.

2.4 Aprendizaje Máquina aplicado a la Extracción de Información:

-Aprendizaje de reglas para la extracción de información.

-Aprendizaje estadístico.

2.5 Ejemplos de Sistemas de Extracción de Información:

-ARMADILLO

-LaSIE

-PROTEUS

-TURBIO

2.Evaluación:

Conferencias MUC.

Tema 3: Extracción Automática de Resúmenes y Síntesis de Información.

-3.1 Tipos de resumen: consideraciones sobre el texto a procesar y los objetivos del resumen:

-Resumen mono/multi-documento

-Resumen genérico vs. orientado a consulta

-Resumen informativo vs. indicativo

-Resumen multi-evento vs. mono-evento

-3.2 Caracterización de fragmentos relevantes:

-Localización y longitud del fragmento.

-Presencia de términos relevantes

-Expresiones indicativas de relevancia

-Nombres propios.

-3.3 Técnicas de resumen basadas en coherencia y cohesión:

-Conceptos de cohesión y coherencia.

-Aplicación de la cohesión y la coherencia en la generación automática de resúmenes.

-Aplicación combinada de cohesión y coherencia.

-3.4 Resumen multidocumento y síntesis de información:

-Características del resumen multi-documento frente a mono-documento.

-Síntesis de Información frente a resumen multi-documento.

-3.5 Resúmenes multilingües:

-Diversas aproximaciones al problema.

-3.6 Evaluación de resúmenes

-Evaluación basada en la coherencia o en la información contenida.

-Evaluación mediante resúmenes de referencia.

-Evaluación en relación a los documentos de partida.

-Conferencias DUC.

Tema 4: Sistemas de Búsqueda Automática de Respuestas.

-4.1 Búsqueda de Respuestas vs. Recuperación de Información:

-Evolución de la Recuperación de Información hacia la Búsqueda de Respuestas.

-4.2 Arquitectura básica de un sistema de Búsqueda de Respuestas:

-Análisis de la pregunta.

-Selección de documentos.

-Extracción de respuestas.

-4.3 Clasificación de Sistemas de búsqueda automática de respuestas:

-Nivel de utilización de técnicas de PLN.

-Taxonomía de Moldovan.

-Situación actual de la investigación en este campo.

-4.4 Tipos de preguntas y respuestas:

-Clasificación de los diferentes tipos y subtipos de preguntas.

-4.5 La barrera del idioma en la Búsqueda de Respuestas:

-El track QA@CLEF: búsqueda translingüe de respuestas.

-Diferentes enfoques al problema.

-4.6 Interacción con el usuario:

-Sistemas de ayuda para la búsqueda de respuestas.

-Comparación entre sistemas automáticos de búsqueda de respuestas y asistentes interactivos de ayuda a la búsqueda de respuestas.

-4.7 Evaluación:

-Conferencias TREC, CLEF y NTCIR.

Objetivos por tema y orientaciones breves

Tema 1: Recuperación de Información Multilingüe.

Objetivo:

El objetivo global del tema es introducir al alumno en el tema de investigación sobre Recuperación de Información Multilingüe, introduciendo en primer lugar los aspectos monolingües del problema y extendiendo éstos a entornos multilingües e interactivos.

Este objetivo global puede descomponerse en los siguientes objetivos más concretos:

O.1.1: Conocer el problema de la Recuperación de Información e identificar los problemas específicos que se presentan en un entorno multilingüe.

O.1.2: Conocer las diferentes aproximaciones para resolver el problema de la Recuperación de Información Multilingüe e identificar las ventajas e inconvenientes de cada uno de ellos.

O.1.3: Identificar los problemas adicionales que se presentan cuando el problema se aborda desde un punto de vista interactivo y conocer las principales aproximaciones realizadas para resolverlos.

Dentro de las actividades de aprendizaje se especifican las lecturas más adecuadas para cada uno de los objetivos del tema.

Tema 2: Extracción de Información.

Objetivo:

El objetivo global del tema es introducir al alumno en el problema de la Extracción de Información, estudiando su perspectiva histórica a través de las Conferencias MUC y analizando las diferentes partes del problema tanto en entornos monolingües como multilingües.

Este objetivo global puede descomponerse en los siguientes objetivos más concretos:

O.2.1: Conocer el problema de la Extracción de Información Recuperación de Información desde la perspectiva histórica de las conferencias MUC (Message Understanding Conference).

O.2.2: Identificar las diferentes fases del proceso de extracción de información y conocer las principales aproximaciones para afrontarlos.

O.2.3: Identificar los problemas específicos de la Extracción de Información cuando se extiende a un entorno multilingüe y conocer las aproximaciones para resolverlos.

O.2.4: Examinar las técnicas de Aprendizaje Máquina que se han empleado para resolver el problema de la Extracción de Información.

Dentro de las actividades de aprendizaje se especifican las lecturas más adecuadas para cada uno de los objetivos del tema.

Tema 3: Extracción Automática de Resúmenes y Síntesis de Información.

Objetivo:

De la Revista Iberoamericana de Inteligencia artificial, n. 22, vol. 8, 2004 (número especial sobre Acceso a Información Multilingüe), los siguientes textos de referencia:

-López-Ostenero, F., Gonzalo, J. y Verdejo, M. F. Búsqueda de información multilingüe: estado del arte.

-Vicedo, José Luis. La búsqueda de respuestas: estado actual y perspectivas de futuro.

-Turmo, Jordi. Information Extraction: Multilinguality and Portability.

-Alonso, Laura, Castellon, Irene, Climent, Salvador, Fuentes, María, Padró, Lluís y Rodríguez, Horacio. Approaches to Text Summarization: Questions and Answers.

El alumno deberá aprender las principales técnicas empleadas para la Extracción Automática de Resúmenes, así como diferenciar este problema del de la Síntesis de Información.

Este objetivo global puede descomponerse en los siguientes objetivos más concretos:

O.3.2: Estudiar diferentes técnicas para la localización de fragmentos relevantes del documento y conocer cómo se mide la relevancia de los mismos.

O.3.3: Conocer los conceptos de cohesión y coherencia aplicados al problema. Estudiar las diversas técnicas que aplican dichos conceptos.

O.3.4: Saber identificar las diferencias entre la Síntesis de Información y los Resúmenes Multi-Documento.

O.3.5: Conocer los problemas que se plantean cuando el problema estudiado se lleva a un entorno multilingüe.

O.3.6: Estudiar las medidas que se han probado para evaluar la calidad de los resúmenes automáticos, principalmente en las Conferencias DUC (Document Understanding Conference).

Dentro de las actividades de aprendizaje se especifican las lecturas más adecuadas para cada uno de los objetivos del tema.

Tema 4: Sistemas de Búsqueda Automática de Respuestas.

Objetivo:

Introducir al alumno en los Sistemas Autmáticos de Búsqueda de Respuesta, conociendo la problemática que se plantea tanto en entornos multilingües como interactivos.

Este objetivo global puede descomponerse en los siguientes objetivos más concretos:

O.4.1: Conocer cómo se produjo la evolución de la Recuperación de Información hacia la Búsqueda de Respuestas.

O.4.2: Conocer la arquitectura básica de un Sistema Automático de Búsqueda de Respuestas, estudiar las diferentes clasificaciones de éstos.

O.4.3: Estudiar los problemas específicos de la Búsqueda de Respuestas en entornos multilingües y conocer los diferentes enfoques empleados para resolverlos.

O.4.4: Considerar los Sistemas de Búsqueda de Respuestas como asistentes interactivos para el usuario. Comparar dichos asistentes con los Sistemas Automáticos.

Actividades y plan de trabajo

Actividades prácticas programadas

Las tareas que se asignan en esta asignatura tienen tanto que ver con la asimilación de los conocimientos propios de la materia, como con el desarrollo de la capacidad para investigar.

Algunos de los tipos de tareas que se proponen son:

-Lectura y análisis de un artículo de investigación.

-Evaluación de un artículo, calificando de forma razonada su originalidad, su impacto potencial en el área, la pertinencia y completitud de las referencias bibliográficas, la calidad del trabajo, etc.

-Estudio del impacto de un artículo: ¿Cuáles son los aspectos del artículo por los que es referenciado? ¿Coinciden con los aspectos sobre los que los autores habían hecho énfasis, o son aspectos inicialmente marginales? ¿Se ha hecho algún avance sustancial respecto a las conclusiones del artículo? ¿Se han refutado las conclusiones del artículo, se han corroborado, se ha profundizado en ellas, se han propuesto vías alternativas?

-Actualización de un artículo de revisión del estado del arte, sintetizando los avances más significativos posteriores a la publicación de la revisión inicial.

-Propuesta de "lecturas recomendadas" para un tema, consensuando una lista razonada a partir del debate entre todos los alumnos de la asignatura.

-Evaluación comparada de servicios de búsqueda Web alternativos, utilizando tanto la revisión bibliográfica como la experimentación directa.

-Diseño e implementación de un servicio de búsqueda Web con algún componente novedoso, partiendo de herramientas de código abierto (como Lucene) o servicios Web (como las API de Google, Yahoo, etc).

Otras actividades prácticas programadas

Se irán anunciando de forma dinámica en el entorno virtual.

3.3 Plan de trabajo

Tema 1 (15 horas) Semanas 1-4. Estudio de materiales de referencia y ejercicios relacionados con la consulta bibliográfica.

Tema 2 (15 horas) Semanas 4-8. Estudio de materiales de referencia y ejercicios relacionados con la consulta bibliográfica.

Tema 3 (20 horas) Semanas 9-12. Estudio de materiales de referencia y ejercicios relacionados con la consulta bibliográfica.

Tema 4 (25 horas) Semanas 13-16. Estudio de materiales de referencia y ejercicios relacionados con la consulta bibliográfica.

Trabajo individual (50 horas). Semanas 16-23.

6.EQUIPO DOCENTE

- FERNANDO LOPEZ OSTENERO
- ALVARO RODRIGO YUSTE

7.METODOLOGÍA

La general del programa de posgrado. En particular, el alumno realiza dos tipos de actividades en esta asignatura: las relacionadas con la consulta bibliográfica y las de implementación y experimentación. Las

primeras son comunes a todos los alumnos y están fijadas dentro del material de estudio correspondiente a cada tema. En una segunda parte de la asignatura, cada alumno realiza un trabajo individual sobre un tema acordado con el equipo docente. Todo el material de estudio está disponible en el entorno virtual del posgrado, y toda la interacción entre profesores y alumnos se puede llevar a cabo en este entorno.

8.BIBLIOGRAFÍA BÁSICA

Comentarios y anexos:

Bibliografia general de consulta

De la Revista Iberoamericana de Inteligencia artificial, n. 22, vol. 8, 2004 (número especial sobre Acceso a Información Multilingüe), los siguientes textos de referencia:
-López-Ostenero, F., Gonzalo, J. y Verdejo, M. F. Búsqueda de información multilingüe: estado del arte.
-Vicedo, José Luis. La búsqueda de respuestas: estado actual y perspectivas de futuro.
-Turmo, Jordi. Information Extraction: Multilinguality and Portability.
-Alonso, Laura, Castellon, Irene, Climent, Salvador, Fuentes, María, Padró, Lluís y Rodríguez, Horacio. Approaches to Text Summarization: Questions and Answers.
De la Revista Iberoamericana de Inteligencia artificial, n. 22, vol. 8, 2004 (número especial sobre Acceso a Información Multilingüe), los siguientes textos de referencia:
-López-Ostenero, F., Gonzalo, J. y Verdejo, M. F. Búsqueda de información multilingüe: estado del arte.
-Vicedo, José Luis. La búsqueda de respuestas: estado actual y perspectivas de futuro.
-Turmo, Jordi. Information Extraction: Multilinguality and Portability.
-Alonso, Laura, Castellon, Irene, Climent, Salvador, Fuentes, María, Padró, Lluís y Rodríguez, Horacio. Approaches to Text Summarization: Questions and Answers.

9.BIBLIOGRAFÍA COMPLEMENTARIA

10.RECURSOS DE APOYO AL ESTUDIO

El campus virtual de posgrados de la UNED, proporcionará interfaz de interacción entre el alumno y sus profesores. La plataforma UNED (Alf) permite impartir y recibir formación, gestionar y compartir documentos, crear y participar en comunidades temáticas, así como realizar proyectos online.

11.TUTORIZACIÓN Y SEGUIMIENTO

La tutorización de los alumnos se llevará a cabo a través de la plataforma de enseñanza virtual de posgrado de la UNED.

12.EVALUACIÓN DE LOS APRENDIZAJES

La evaluación se realizará a partir de las actividades realizadas en cada tema y el trabajo individual de cada alumno.

13.COLABORADORES DOCENTES

Véase equipo docente.