Procesamiento del Lenguaje Natural, Revista nº 34, marzo de 2005

sociedad española para el Procesamiento del Lenguaje Natural

ISSN: 1135-5948

Artículos

Análisis de Metodologías de Evaluación de Sistemas de Diálogo Multimodal
Ramón López-Cózar, Zoraida Callejas, Miguel Gea 9

Aplicación del Procesamiento de Lenguaje Natural en la Recuperación de Información
Yenory Rojas, Antonio Ferrández, Jesús Peral 17

Búsqueda de respuestas multilingüe Clasificación de preguntas en español basada en aprendizaje
Miguel Ángel García, Luis Alfonso Ureña, Fernando Martínez, Arturo Montejo 31

Construcción de sistemas de recuperación de información sobre córpora textuales estructurados de grandes dimensiones
Fco. Mario Barcala, Miguel A. Molinero, Eva Domínguez 41

Reentrenamiento: Aprendizaje Semisupervisado de los Sentidos de las Palabras
Armando Suárez, Manuel Palomar, German Rigau 49

The Spanish pronominal clitic system
Luis Pineda, Ivan Meza 67

Un sistema de diálogo multilingüe dirigido por la semántica
Marta Gatius, Meritxell González 105

Tesis

Adquisició d'informació lèxica i morfosintàctica a partir de corpus sense anotar: aplicació al rus i al croat
Antoni Oliver 115

Algoritmo LVQ aplicado a tareas de procesamiento del lenguaje natural
María Teresa Martín Valdivia 117

El problema de la fusión de colecciones en la recuperación de información multilingüe y distribuida: cálculo de la relevancia de documental en dos pasos
Fernando Martínez Santiago 119

Resolución de la Ambigüedad Semántica de las palabras mediante Modelos de Probabilidad de Máxima Entropía
Armando Suárez 121

Selección de contenidos a partir de un corpus enriquecido y especializado para la generación personalizada de documentos
Guillermo Barrutieta 123

Supervised Word Sense Disambiguation: Facing Current Challenges
David Martínez 125

Una arquitectura de integración de recursos léxicos de naturaleza heterogénea. Una aportación desde la perspectiva de la integración de datos
Aitor Soroa 127

Información General
XXI Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural 131

Impresos de Inscripción

Procesamiento del Lenguaje Natural, núm. 34 (2005), pp. 31-39 recibido 10-03-2005, aceptado 30-03-2005

Búsqueda de respuestas multilingüe: Clasificación de preguntas en español basada en aprendizaje.*

Miguel Ángel García Cumbreras
Fernando Martínez Santiago
L. Alfonso Ureña López
Arturo Montejo Raez
Grupo Sistemas Inteligentes de Acceso a la Información
Departamento de Informática
Universidad de Jaén
http://sinai.ujaen.es
e-mail:{magc,laurena,dofer,amontejo}@ujaen.es

Resumen: Este artículo presenta un sistema de clasificación de preguntas en español basado en aprendizaje, utilizando colecciones en inglés y diversas *máquinas de traducción* online como recursos de traducción de la pregunta original en español al inglés. En este estudio se mide de forma cuantitativa la bondad de cuatro traductores automáticos para la pareja de idiomas español-inglés, comparando los resultados obtenidos para las preguntas originales en inglés con los obtenidos de cada una de las traducciones. El sistema se ha implementado de forma modular utilizando varios métodos de aprendizaje tales como LibSVM, Bayesian Logistic Regression o PLAUM. En la tarea de clasificación de preguntas se demuestra que la pérdida de precisión debida a la traducción automática es moderada, situandose entorno a un 5 %.
Palabras clave: Clasificación de Preguntas, Sistemas de Búsqueda de Respuestas, aprendizaje automático, traductores automáticos

Abstract: This paper presents an Spanish question classification system based on machine learning, that uses English collections, different online machine translators and other NLP English resources. The original Spanish questions are translated into English. Four machine translators are evaluated in terms of precision and the results are compared with the result obtained by using original English questions. Our system has been developed into separated modules and we have tested several machine learning methods, such as LibSVM, Bayesian Logistic Regression or PLAUM. The obtained results show that these online machine translators, used for the language pair Spanish-English, and for the query translation task in a multilingual question answering system, work well. It is showed that the loss of precision because of the machine translation, in a question classification task, is reasonable, around 5 %.
Keywords: Question Classification, Question Answering Systems, machine learning, machine translation

1. *Introducción*

La cantidad de información digital ha experimentado un fuerte crecimiento, así como el número de usuarios finales que a través de ordenadores personales interactuan con esta información. Esto implica el creciente interés por los sistemas de recuperación de información multilingüe (CLIR, del inglés Cross Language Information Retrieval) así como por los sistemas de búsqueda de respuestas (BR) tanto monolingües como multilingües.

Un sistema CLIR es un sistema de recuperación de información que tiene capacidad para operar sobre una colección de documentos o pasajes multilingüe, esto

* Este trabajo ha sido financiado por el Ministerio de Ciencia y Tecnología mediante el proyecto TIC2003-07158-C04-04

ISSN 1135-5948