Procesamiento del Lenguaje Natural, nº 31, septiembre de 2003 ISSN 1135-5948

SEPLN

Sociedad Española para el Procesamiento del Lenguaje Natural

Artículos

Análisis automático del contenido textual 4

CARPANTA eats words you don't need from e-mail.
Laura Alonso, Bernardino Casas, Irene Castellón, Salvador Climent y Lluís Padró 5

Categorización de texto sensible al coste para el filtrado de contenidos inapropiados en Internet.
José María Gómez Hidalgo, Enrique Puertas Sanz, Francisco Carrero García y Manuel de Buenaga Rodríguez 13

Extracción automática de respuestas para documentación técnica.
Fabio Rinaldi y Elia Yuste 21

Un sistema para resumen automático de textos en castellano.
Pedro Luis Mateo, José Carlos González, Julio Villena y José Luis Martínez 29

Semántica, pragmática y discurso 37

Exploring Large-scale Acquisition of Multilingual Semantic Models for Predicates.
Jordi Atserias, Mauro Castillo, Francis Real, Horacio Rodríguez y Germán Rigau 39

Making Wordnet Mappings Robust.
J. Daudé, L. Padró y G. Rigau 47

Aplicaciones industriales del PLN 55

Algoritmo de Clustering On-Line utilizando metaheurísticas y técnicas de muestreo.
Arantza Casillas, Mª Teresa González de Lena y Raquel Martínez 57

Desarrollo de un corrector ortográfico para aplicaciones de conversión texto-voz.
Ana Armenta, Gregorio Escalada, Juan María Garrido y Miguel Ángel Rodríguez 65

Reconocimiento y síntesis de voz 73

Ajuste subjetivo de pesos para selección de unidades a través de algoritmos genéticos interactivos.
Francesc Alías, Xavier Llorà, Ignasi Iriondo y Lluís Formiga 75

Arquitectura para conversión texto-habla multidominio.
Francesc Alías, Xavier Sevillano, Pere Barnola y Joan Claudi Socoró 83

Estrategias de generación y reducción de variantes de pronunciación en sistemas de reconocimiento automático de habla: consideraciones arquitecturales.
Javier Macías Guarasa, Javier Ferreiros, Ricardo de Córdoba, Juan Manuel Montero, José David Romeral y José M. Pardo .. 91

Modelos específicos de comprensión en un sistema de diálogo.
Fernando García, Emilio Sanchís, Lluís Hurtado y Encarna Segarra 99

Phrase break prediction: a comparative study.
Pablo Daniel Agüero,y Antonio Bonafonte 107

Selección de unidades léxicas para reconocimiento automático del habla continua en euskera.
Karmele López de Ipiña, Manuel Graña, Ekaitz Zulueta y Aitzol Ezeiza 115

Resolución de la ambigüedad léxica 123

Aprendizaje competitivo LVQ para la desambiguación léxica.
Manuel García Vega, María Teresa Martín Valdivia y Luis Alfonso Ureña López 125

Colaboración entre información paradigmática y sintagmática en la Desambiguación Semántica Automática.
Iulia Nica, Mª. Antònia Martí Antonín y Andrés Montoyo Guijarro 133

Método de desambiguación léxica basada en el recurso léxico Dominios Relevantes.
Sonia Vázquez, Andrés Montoyo y Germán Rigau 141

Gramáticas y formalismos para el análisis morfológico y sintáctico 149

Análisis ascendente bidireccional de TAG dirigido por el núcleo TIG.
Vicente Carrillo Montero, Víctor J. Díaz Madrigal y Miguel A. Alonso Pardo 151

Análisis morfosintáctico estadístico en lengua gallega.
Francisco Méndez Pazó, Francisco Campillo Díaz, Eduardo Rodríguez Banga y Elisa Fernández Rei 159

Análisis sintáctico ascendente con un algoritmo evolutivo.
Lourdes Araujo 167

Aprendizaje de gramáticas probabilísticas a partir de árboles sintácticos.
José Luis Verdú-Mas 175

Earley-based stochastic context-free grammar estimation from bracketed corpora and its use in a hybrid language model.
Diego Linares, José-Miguel Benedí y Joan-Andreu Sánchez 183

Lingüística de corpus 191

3LB-SAT: Una herramienta de anotación semántica.
Empar Bisbal, Antonio Molina, Lidia Moreno, Ferrán Pla, Maximiliano Saiz-Noeda y Emilio Sanchís 193

Análisis cualitativo y cuantitativo de acuerdo entre anotadores en el desarrollo de corpus interpretados lingüísticamente.
M. Civit, A. Ageno, B. Navarro, N. Bufi y M. A. Martí 201

Asignación automática de etiquetas de dominios en WordNet.
Mauro Castillo, Francis Real y Germán Rigau 209

NATools — A Statistical Word Aligner Workbench.
Alberto Simões y José João Almeida 217

Extracción y recuperación de información monolingüe y multilingüe 225

Aprendizaje neuronal aplicado a la fusión de colecciones multilingües en CLIR.
Mª Teresa Martín Valdivia, L. Alfonso Ureña López y Fernando Martínez Santiago 227

Identificación de entidades con nombre basada en modelos de Markov y árboles de decisión.
José A. Troyano, Víctor J. Díaz, Fernando Enríquez, Javier Barroso y Vicente Carrillo 235

Un enfoque gramatical para la extracción de términos índice.
Jesús Vilares Ferro y Miguel A. Alonso Pardo 243

Uso de información contextual en la interfaz de entrada de un sistema de diálogo.
R. López-Cózar 251

Lexicografía computacional 259

Starting up the Multilingual Central Repository.
Jordi Atserias, Germán Rigau, Luís Villarejo 261

Uso de Internet para aumentar la cobertura de un sistema de adquisición léxica del ruso.
Antoni Oliver, Irene Castellón y Lluís Màrquez 269

Traducción automática 277

Construcción rápida de un sistema de traducción automática español-portugués partiendo de un sistema español-catalán.
Patricia Gilabert-Zarco, Javier Herrero-Vicente, Sergio Ortiz-Rojas, Antonio Pertusa-Ibáñez, Gema Ramírez-Sánchez,
Felipe Sánchez-Martínez, Marcial Samper-Asensio, Miriam A. Scalco y Mikel L. Forcada 279

Natural Language Interface Framework for Spatial Object Composition Systems.
Hiram Calvo y Alexander Gelbukh 285

Proyectos 293

SAÓ - Sistema de ayuda ortoépica para la lectura en voz alta del valenciano.
Mikel L. Forcada, Vicent Beltrán, Carles Segura y Jordi Colomina 295

3LB: Construcción de una base de datos de árboles sintáctico semánticos.
UA, UPC, EHU, FBiG y UPV 297

IMAGINE: Interfacing Mobile Application with Voice Natural Language Interactivity.
C. Arana Ferrandiz, I. Recio Sánchez, Mª J. Carrión, J. de Frutos, I. Dattani, F. Choi, P. Wilken, S. Hefeez
M. Cecil-Wrigh, P. Schmidt, M. Marimom, C. Pease, V. Hinz, J. Caminero, D. Castell, L. Hernández, J. Relaño,
K, Gladstone, R. Pick, E. Ramos, J. Alicarte, A. Pons 299

C-ORAL-ROM. Corpus integrado de referencia en lenguas romances.
Manuel Alcántara Plá, Antonio Moreno Sandoval, Guillermo De la Madrid Heitzmann,
Ana González Ledesma y Fernando Ares Chicote 301

Adquisición de recursos básicos de lingüística computacional del gallego
para aplicaciones informáticas de tecnología lingüística.
Luz Castro Pena, Ángel López López, José Ramom Pichel Campos, José Luis Aguirre Moreno,
Alberto Álvarez Lugrís, Xavier Gómez Guinovart, Elena Sacau Fontenla y Lara Santos Suárez 303

Proyecto ALIADO: Tecnologías del habla y el lenguaje para un asistente persona.
José B. Mariño y Horacio Rodríguez 305

The MEANING Project.
German Rigau, Eneko Agirre y Jordi Atserias 307

A corpus-based approach to generalising a chatbot system.
Bayan Abu Shawar y Eric Atwell 309

Desarrollo de un analizador morfológico de catalán antiguo basado en corpus textuales.
Mikel L. Forcada, Alicia Garrido-Alenda, Patricia Gilabert-Zarco, Marinela García-Sempere,
Sandra Montserrat-Buendia y Amaia Iturraspe-Bellver 311

Sistema de comprensión de comunicaciones habladas para el control de tráfico aéreo del proyecto INVOCA.
V. Sama Rojo, F. Fernández Martínez, J. Ferreiros López, J. Macias-Guarasa,
R. De Córdoba, J. M. Montero Martínez, J. Colas Pasamontes y J. M. Pardo Muñoz 313

Información colocacional y recuperación de la información.
Margarita Alonso Ramos y Leo Wanner 315

Lexicon and Corpora for Speech to Speech Translation (LC-STAR).
Maximilian Bisani, Antonio Bonafonte, Nuria Castell, Elviira Hartikainen,
Giulio Maltese, Asunción Moreno, Shaunie Shammass y Ute Ziegenhain 317

X-Flow: Gestión de flujo de contenidos multilingües sobre XLIFF y TMX.
Inés Jacob, Joseba Abaitua, Josuka Díaz y Fernando Quintana 319

Construcción de un sistema de recuperación multilingüe en la Web.
Antonio Ferrández, Alfonso Ureña y Víctor J. Díaz 321

Demostraciones 323

ÁGORA. Multilingual Multiplatform Architecture for the development of Natural Language Voice Services.
José Relaño, Luis Villarrubia, Mari Carmen R. Gancedo y Luis Hernández 325

First release of the Multilingual Central Repository of MEANING.
Luis Villarejo, Jordi Atserias, Gerard Escudero y Germán Rigau 327

Programa deductor de elementos morfolgicos en contextos de oraciones infinitas iterativas.
Carlos Alonso Hidalgo Alfageme 311

Conversor texto a voz multilingüe de Telefonica I+D.
Ana Armenta, Gregorio Escalada, Juan María Garrido y Miguel Ángel Rodríguez 331

Lexicometría de corpus.
Jordi Porta Zamorano y Rafael J. Ureña Ruiz 333

X-Not@rial: Sistema de Recuperación y Extracción de Información Notarial.
R. Muñoz, F. Llopis, R. Izquierdo y M. C. Calle 335

Demostración del sistema de comprensión de comunicaciones habladas para control de tráfico aéreo del proyecto INVOCA.
F. Fernández Martínez, V. Sama Rojo, J. Ferreiros López, J. Macias-Guarasa,
R. De Córdoba, J. M. Montero Martínez, J. Colas Pasamontes, E. Campos Palarea y J. M. Pardo Muñoz 337

The LinguaStream Platform.
Frédérik Bilhaut 339

La plataforma de adquisición de diálogos en el proyecto Dihana.
M. I. Galiano, R. Granell, Ll. F. Hurtado, A. Miguel, J. A. Sánchez y E. Sanchís 341

A new approach to the analysis and annotation of speech and prosody based on computerized cross-linguistic corpora.
Dolores Ramírez Verdugo 343

TransType2. Un sistema de ayuda a la traducción.
Antonio L. Lagarda, Luis Rodríguez, Elsa Cubel, Enrique Vidal y Francisco Casacuberta 345

Iuriservices: Un FAQ inteligente para los jueces en su primer destino.
Jesús Contreras, V. Richard Benjamins, Pompeu Casanovas, Lissette Lemus y Cristina Urios 347

Talleres 349

La Búsqueda de Respuestas: estado actual de la tecnología, aplicaciones y líneas de futuro

Los sistemas de búsqueda de respuestas desde una perspectiva actual.
José Luis Vicedo, Horacio Rodríguez, Anselmo Peñas y Marc Massot 351

Tecnología del Habla: pasado presente y futuro. Particularización sobre tecnología del español

Modelo de evolución de la Tecnología del Habla, y tendencias futuras.
Luis Hernández Gómez 369

¿Qué queremos que sea Tecnología del Habla?
Javier Ferreiros López 375

Tecnologías del habla y lenguas minoritarias.
Carmen García Mateo 381

©2003 Sociedad Española para el Procesamiento del Lenguaje Natural

XIX Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural

SEPLN

Universidad de Alcalá, 10, 11 y 12 de septiembre de 2003

Instituto Cervantes

UNIVERSIDAD DE ALCALÁ

Excmo. Ayuntamiento de ALCALA de HENARES
Concejalía de Turismo

CIUDAD PATRIMONIO de la Humanidad
ALCALÁ de Henares

MINISTERIO DE CIENCIA Y TECNOLOGIA

Telefónica IBERIA

FUNDACION MAPFRE ESTUDIOS

EDITADO POR:

Isabel Bermejo Rubio (Instituto Cervantes)
Maxilimiliano Saiz Noeda (Universitat d´Alacant)

COMITÉ CIENTÍFICO

Presidente:

Prof. Maximiliano Saiz Noeda (Universitat d' Alacant)

Miembros:

Prof. José Gabriel Amores Carredano (Universidad de Sevilla)
Prof. Toni Badia i Cardús (Universitat Pompeu Fabra)
Prof. Manuel de Buenaga Rodríguez (Universidad Europea de Madrid)
Prof.ª Irene Castellón Masalles (Universitat de Barcelona)
Prof.ª Arantza Díaz de Ilarraza (Euskal Herriko Unibertsitatea)
Prof. Antonio Ferrández Rodríguez (Universitat d'Alacant)
Prof. Mikel Forcada Zubizarreta (Universitat d' Alacant)
Prof.ª Ana María García Serrano (Universidad Politécnica de Madrid)
Prof. Koldo Gojenola Galletebeitia (Euskal Herriko Unibertsitatea)
Prof. Xavier Gómez Guinovart (Universidade de Vigo)
Prof. Julio Gonzalo Arroyo (Universidad Nacional de Educación a Distancia)
Prof. José Miguel Goñi Menoyo (Universidad Politécnica de Madrid)
Prof. Joaquim Llisterri Boix (Universitat Autònoma de Barcelona)
Prof. Javier Macías Guarasa (Universidad Politécnica de Madrid)
Prof. José B. Mariño Acebal (Universitat Politècnica de Catalunya)
Prof.ª M. Antonia Martí Antonín (Universitat de Barcelona)
Prof.ª Lidia Ana Moreno Boronat (Universitat Politècnica de València)
Prof. Lluis Padró (Universitat Politècnica de Catalunya)
Prof. Manuel Palomar Sanz (Universitat d' Alacant)
Prof. José Manuel Pardo Muñoz (Universidad Politécnica de Madrid)
Prof.ª Natividad Prieto Sáez (Universitat Politècnica de València)
Prof. Germán Rigau Claramunt (Universitat Politècnica de Catalunya)
Prof. Horacio Rodríguez Hontoria (Universitat Politècnica de Catalunya)
Prof. Kepa Sarasola Gabiola (Euskal Herriko Unibertsitatea)
Prof. L. Alfonso Ureña López (Universidad de Jaén)
Prof.ª Mª Felisa Verdejo Maillo (Universidad Nacional de Educación a Distancia)
Prof. Manuel Vilares Ferro (Universidade de Vigo)

REVISORES EXTERNOS:

Alicia Ageno Pulido, Iñaki Alegría, Laura Alonso Alemany, Alberto Álvarez Lugrís, Aitziber Atutxa , Rafael C. Carrasco, María José Castro Bleda, Montserrat Civit Torruella, Alberto Díaz Esteban, Ana Mª Fernández Montraveta, Manuel García Vega, Pablo Gervás Gómez-Navarro, Ignacio Giráldez, Fernando Llopis Pascual, Pilar Manchón Portillo, Manuel J. Maña López, Montserrat Marimón Felipe, Lluís Màrquez, Patricio Martínez Barco, José Luis Martínez Fernández, Paloma Martínez Fernández, Aingeru Mayor Martínez, Louise McNally, Juan Andrés Montoyo Guijarro, Rafael Muñoz Guillena, Maite Oronoz , Anselmo Peñas Padilla, Jesús Peral Cortés, Juan Antonio Pérez Ortiz, Ferrán Plá Santamaría, Antonio Sánchez Valderrábanos, Mariona Taulé Delor, Enric Vallduví, Julio Villena Román

Comité organizador

Presidente:

D. Jesús Antonio Cid Martínez
Director Académico del Instituto Cervantes

Coordinadora:

Dña. Isabel Bermejo Rubio
Responsable de la Oficina del Español en la Sociedad de la Información (OESI) del Instituto Cervantes

Vocales:

Dña. Eva Mª García García
Técnico de la OESI del Instituto Cervantes

Dña. Raquel Tapias Aparicio
Técnico de la OESI del Instituto Cervantes

D. John Michael Urresti Graña
Técnico de la OESI del Instituto Cervantes

Colaboradores:

Dña. Rosario Guijarro Huerta
Colaboradora de la OESI del Instituto Cervantes

Asesores:

Dña. Esmeralda de Luis Martínez
Jefe del Dpto. de Relaciones Exteriores e Institucionales del Instituto Cervantes

Dña. Gloria Gamarra Alonso de Linaje
Técnico del Dpto. de Relaciones Exteriores e Institucionales del Instituto Cervantes

Procesamiento del Lenguaje Natural, nº 31 (2003), pp-227-234 recibido 16-05-2003; aceptado 20-06-2003

Aprendizaje neuronal aplicado a la fusión de colecciones multilingües en CLIR

Mª Teresa Martín Valdivia
Universidad de Jaén
Av. Madrid 37, 23071
maite@ujaen.es

Fernando Martínez-Santiago
Universidad de Jaén
Av. Madrid 37, 23071
dofer@ujaen.es

L. Alfonso Ureña López
Universidad de Jaén
Av. Madrid 37, 23071
laurena@ujaen.es

Resumen: Un problema común al trabajar con sistemas CLIR (Cross-Lingual Information Retrieval) basados en la traducción de consultas consiste en obtener una única lista de documentos relevantes a partir de los resultados locales obtenidos para cada colección monolingüe. En este trabajo se presenta un estudio comparativo de las estrategias tradicionalmente usadas para resolver este problema. Se incluyen en el estudio dos técnicas recientes: la regresión logística y el cálculo del RSV (Retrieve Status Value) en dos pasos. Además, se presenta e implementa una nueva técnica basada en redes neuronales artificiales que utiliza el algoritmo LVQ (Learning Vector Quantization) y con la que se obtienen resultados prometedores. Como muestran los experimentos realizados, los mejores resultados son obtenidos mediante el uso de la técnica denominada RSV en dos pasos. Sin embargo, este método requiere que las consultas estén alineadas a nivel de término. Esto es, para cada término de la consulta, debe conocerse cómo ha sido traducido al resto de los idiomas. Dado que tal información no siempre está disponible, es usual que las consultas cuenten con una parte alineada y otra no alineada. Es por ello que la segunda parte del artículo, estudia la forma de integrar la información obtenida a partir de la parte alineada y la no alineada en el método RSV en dos pasos mediante el uso de regresión logística y LVQ.
Palabras clave: Redes neuronales artificiales, Sistemas CLIR, LVQ, RSV en dos pasos, Regresión logística, Estrategias de fusión de documentos.

Abstract: A very common problem that arises when we deal with CLIR systems based on queries translation consists of obtaining an only relevant documents list from the local results of each monolingual collection. In this work a comparative study of the strategies traditionally used appears to solve this problem. Two recent techniques are included in the study: the logistic regression and the calculation of the 2-step RSV. We include two recent techniques: logistic regression and 2- step RSV. Moreover, we present and implement a new technique based on neural networks using LVQ algorithm with promising results obtained. The experiments show that best results are obtained by the 2-step RSV technique. Nevertheless, this method requires that the query must be aligned at term level. That is, it must be known how each term of the query has been translated into different languages. As such information is not always available, it is usual that the query has an aligned and not aligned part. So, we study the way of integrating the data obtained from the aligned and not aligned part in 2-step RSV method by means of the use of logistic regression and LVQ.

Keywords: Artificial neural networks, CLIR systems, LVQ, 2-step RSV, Logistic regresion, Merging document strategies

ISSN 1135-5948 ©2003 Sociedad Española para el Procesamiento del Lenguaje Natural

1 Introducción

En la última década, el interés por desarrollar sistemas de recuperación de información multilingüe (CLIR - Cross Lingual Information Retrieval) ha crecido de manera espectacular (Grefenstette, 1998). Un sistema CLIR es un sistema de recuperación de información capacitado para operar sobre una colección de documentos multilingüe. Esto es, supuesto que un usuario consulte un sistema CLIR, éste debe recuperar todos aquellos documentos relevantes de entre los que se encuentran en la colección, con independencia del idioma utilizado tanto en la consulta como en los documentos. Así, la salida de uno de estos sistemas será frecuentemente una lista heterogénea de documentos escritos en inglés, español, francés, alemán... y ordenada según la puntuación obtenida por cada documento para la consulta dada.

Para implementar sistemas CLIR existen varias aproximaciones. Por un lado, existen sistemas CLIR que traducen las consultas a los idiomas necesarios, mientras que otros crean una colección de documentos monolingüe mediante la traducción de la colección original multilingüe. También se han realizado sistemas que utilizan un enfoque mixto, traduciendo las consultas, pero manteniendo un único índice de documentos multilingüe. Si bien la opción de traducir únicamente las consultas parece que es la predominante actualmente, este enfoque dificulta la obtención de una única lista de documentos relevantes pues, en general, obtendremos tantas listas como idiomas estén presentes en la colección. La traducción documental, por su parte, presenta problemas de escalabilidad, además de resultar pesada la traducción de toda la colección, especialmente en un ambiente experimental, con frecuentes cambios y la consecuente reindexación de la colección.

En un sistema CLIR basado en traducción de consultas se realiza un proceso de recuperación de información monolingüe independiente para cada idioma. De esta manera, se obtienen tantas colecciones como lenguajes dividiendo los documentos por idiomas. A continuación, cada consulta se lanza contra su colección correspondiente obteniendo una lista de documentos relevantes por cada uno de los idiomas. El último paso, consiste en mezclar estas listas para proporcionar al usuario una única lista de documentos relevantes. Teniendo en cuenta que la relevancia de cada documento es obtenida con relación a la colección de documentos monolingüe a la cual pertenece, y no con relación a la colección original multilingüe, la obtención de una única lista de documentos relevantes a partir de las listas monolingües no es un problema trivial. Se trata de un problema abierto en el que se han experimentado diversas estrategias, desde aplicar un sencillo algoritmo estilo Round-Robin, hasta normalizar la puntuación obtenida por cada documento. Sin embargo, la pérdida de precisión con respecto a los esquemas recuperación de información monolingüe que trabajan con una única gran colección baja considerablemente (Voorhees, 1995, Savoy, 2001).

Recientemente, Martínez, Martín y Ureña, (2002) proponen un nuevo método de fusión de documentos para conseguir una única lista de artículos relevantes. Esta aproximación, denominada RSV en dos pasos (RSV: Retrieval Status Value), se basa en la reindexación de los documentos recuperados en un único nuevo espacio multilingüe donde cada término y sus traducciones son considerados sinónimos y, por consiguiente, con una frecuencia documental común. El problema que presenta este método es que para que funcione correctamente es necesario que el vocabulario esté alineado: dada una palabra en una consulta, se debe conocer su traducción a cada uno del resto de idiomas. Lamentablemente, esto no siempre es posible, por lo que se hace necesaria un estrategia alternativa que permita manejar consultas con un nivel de alineación de palabras parcial. Como consecuencia de ello, para cada documento obtendremos, al menos, dos puntuaciones: la alcanzada por la parte alineada de la consulta (calculada mediante RSV en dos pasos) y la puntuación lograda por la parte no alineada de la consulta (calculada de manera tradicional). En este trabajo se propone una nueva estrategia basada en redes neuronales que permite integrar tanto la puntuación de la consulta obtenida por la parte alineada como por la no alineada.

En primer lugar, se presenta un estudio comparativo entre las distintas estrategias tradicionales que se han venido utilizando en la fusión de documentos y métodos más recientes como el RSV en dos pasos y el uso de regresión logística. Como novedad, se incluye en la comparativa una técnica prometedora basada en

Redes Neuronales Artificiales (RNA) para la obtención de una única lista de documentos relevantes.

En segundo lugar, el artículo estudia la utilización de la regresión logística y las RNA para integrar la parte alineada y no alineada de una consulta en el cálculo del RSV en dos pasos. Como muestran los resultados, la combinación de estas nuevas técnicas para calcular el RSV en dos pasos supone un incremento considerable en la precisión.

El resto del artículo se organiza de la siguiente manera: En la siguiente sección, se presenta una breve introducción a las RNA. A continuación, se describen las técnicas tradicionales en la fusión de documentos en sistemas CLIR, así como la nueva técnica propuesta basada en el algoritmo LVQ. Después, se explica el uso de la regresión logística y el algoritmo LVQ para la integración de la parte no alineada en las consultas con el método RSV en dos pasos. Por último, se describen los experimentos realizados y los resultados obtenidos, así como las conclusión finales.

2 *Redes neuronales artificiales*

En este artículo se describe el uso de RNA aplicadas a la resolución del problema de la fusión de documentos en sistemas CLIR.

Las RNA son modelos estadísticos de procesamiento de información que están inspirados en un sistema nervioso biológico (McClelland y Rumelhart, 1986). Una RNA se compone de un conjunto de elementos de procesamiento denominados neuronas e interconectados entre sí a través de unos pesos de conexión. Mediante un proceso de aprendizaje (fase de entrenamiento), la RNA ajusta sus pesos de conexión para generar un modelo capaz de incorporar la información de los datos de ejemplo.

Las RNA se han aplicado a un gran número de problemas reales de complejidad considerable (Fiesler y Beale, 1997). Estos problemas incluyen reconocimiento de patrones y pronósticos, clasificación de datos, aproximación de funciones...

Existen muchos tipos de redes dependiendo de su arquitectura, tipo de aprendizaje utilizado o algoritmo de entrenamiento aplicado a la red. Una de las redes más utilizadas en clasificación y que se puede entender como un método de regresión es el modelo LVQ (Learning Vector Quatization) (Kohonen, 1995). Se trata de una red basada en el modelo de Kohonen que utiliza aprendizaje competitivo supervisado para ajustar los pesos de conexión. Una red competitiva permite agrupar y representar los datos que están situados en un mismo espacio de entradas. Los pesos de cada neurona representan puntos en el espacio de entrada llamados vectores prototipo o codebooks. En el aprendizaje competitivo, las neuronas de la capa de salida compiten entre sí para conseguir proclamarse ganadora. Sólo la unidad de salida que gane la competición será la que modifique sus pesos de conexión.

Por otra parte, puesto que se trata de un aprendizaje supervisado, la red necesita un conjunto de datos de entrenamiento que incluyan tanto las características que se desean aprender como la salida correcta que debería dar la red en respuesta a esa entrada.

En la sección 4 se describe con detalle el funcionamiento de la red LVQ aplicada a la fusión de documentos en colecciones CLIR.

3 *Estrategias de fusión de documentos*

Existen varias aproximaciones para mezclar colecciones monolingües pero, como se muestra en (Savoy, 2001), la pérdida de precisión se encuentra entre un 20% y un 40% dependiendo de la colección.

A pesar de su simplicidad, una de las técnicas más utilizadas en esta tarea consiste en ordenar simplemente atendiendo a la puntuación alcanzada por cada documento (raw scoring). En un intento de hacer comparables las puntuaciones de las distintas colecciones monolingües, se ha propuesto en diversas ocasiones normalizar tales valores localmente (raw scoring normalizado). Las fórmulas 1 y 2 muestran distintas formas de llevar a cabo esta normalización (Powell et al., 2000).

$$RSV_i' = \frac{RSV_i}{\max(RSV)} \quad (1)$$

$$RSV_i' = \frac{RSV_i - \min(RSV)}{\max(RSV) - \min(RSV)} \quad (2)$$

El uso de algoritmos tipo round-robin es un enfoque igualmente simplista y con unos resultados similares, sólo que, en este caso, se ordena atendiendo no a la puntuación sino a la posición alcanzada por cada documento

localmente en relación a la colección monolingüe a la cual tal documento pertenece.

Una aproximación que ha dado buenos resultados hace uso de la regresión logística. Se trata de un método estadístico que permite calcular la probabilidad de relevancia de un documento d_i en base a su puntuación original y el logaritmo del ranking obtenido. Dependiendo de estas probabilidades de relevancia, estimadas a partir de la fórmula 3 para cada colección monolingüe, los documentos se intercalan creando una única lista ordenada por probabilidad (Le Calvé, 2000, Savoy, 2003).

$$\text{Prob}(d_i) = \frac{e^{\alpha+\beta_1 \cdot \ln(rank_i)+\beta_2 \cdot rsv_i}}{1+e^{\alpha+\beta_1 \cdot \ln(rank_i)+\beta_2 \cdot rsv_i}} \quad (3)$$

Los coeficientes α, β_1 y β_2 son parámetros desconocidos que deben ser calculados usando otros métodos como máxima probabilidad o métodos iterativos basados en el cálculo de mínimos cuadrados. Puesto que la regresión logística requiere ajustar el modelo, se debe disponer de un conjunto de datos de entrenamiento (las consultas y sus juicios de relevancia) para cada una de las colecciones monolingües.

Recientemente, Martínez, Martín y Ureña, (2002) proponen un nuevo método de fusión de documentos denominado RSV en dos pasos, obteniendo unos resultados prometedores. Un término y sus traducciones deben compartir la misma frecuencia documental, de tal manera que el factor *idf* (*inverse document frequency*) aplicado a la hora de calcular el peso de un término, no debe depender de la colección considerada, sino del conjunto de todas ellas. El método propuesto calcula el RSV en dos fases:

1. La fase de preselección de documentos se corresponde con la traducción y lanzamiento de la consulta sobre cada colección monolingüe, D_j, como se realiza de manera usual en los sistemas CLIR basados en traducción de consultas. Esta fase genera dos resultados. En primer lugar, se obtiene una única colección multilingüe de documentos preselecionados (D') con la unión de los 1.000 primeros documentos recuperados para cada idioma. Por lo tanto, esta colección multilingüe, D', tendrá un total de N*1.000 documentos, donde N es el número de idiomas. En segundo lugar, para cada término de la consulta original, se obtiene su traducción al resto de los idiomas. Al conjunto de términos que son traducciones de un término dado, se le llamará "concepto". Un concepto es independiente del idioma. Así, se obtiene un vocabulario T, formado por todos los conceptos presentes en la consulta.
2. La fase de reordenamiento consiste en reindexar la colección D', considerando el vocabulario T. Se crea un índice de conceptos, en lugar de uno de términos, ya que todos los términos de un mismo concepto se tratan como ocurrencias del mismo concepto. Por último, se lanza la consulta sobre este índice, consulta que estará formada por conceptos y no por términos, con lo que es independiente del lenguaje.

El problema principal que presenta esta aproximación es que requiere que todos los términos estén alineados. Sin embargo, esto no es siempre posible y se hace necesario utilizar una técnica que combine de alguna manera la parte de consulta alineada y la no alineada. En este trabajo se presentan dos nuevas aproximaciones para resolver este problema: el uso de regresión logística y el uso de una RNA basada en el modelo de Kohonen.

4 *Una nueva estrategia de fusión de documentos: LVQ*

En esta sección se presenta una nueva técnica para obtener una única lista de documentos para la colección multilingüe basada en la utilización de una RNA. Concretamente, se ha utilizado el algoritmo de aprendizaje competitivo LVQ. El algoritmo LVQ usa un conjunto de vectores de similitud denominados vectores de pesos, vectores de referencia o *codebooks*. A cada clase se le asocia un conjunto de vectores de pesos w_k, de manera que durante el proceso de aprendizaje, uno de ellos será seleccionado y la clase a la que pertenece será elegida como ganadora de la competición.

Aunque en principio el algoritmo LVQ se utiliza para clasificación, en este trabajo se ha ajustado la salida de la red para que dé una medida de proximidad a una clase en lugar de simplemente determinar si un ejemplar pertenece a una determinada clase. Concretamente, en este trabajo se utiliza el algoritmo LVQ para determinar la probabilidad de que un documento sea relevante. Para ello, se utilizan dos codebooks como representantes de cada clase, la clase 0 que representa a los

documentos no relevantes y la clase 1 que representa a los documentos relevantes.

Con este método, al igual que ocurre con la regresión logística, se requiere un conjunto de datos de entrenamiento que incluyan por cada ejemplar (por cada documento) las características que se desean aprender (el ranking y el RSV) junto con la respuesta correcta (1 ó 0 dependiendo de si el documento es o no relevante). Cada documento se representa como un vector x_i con tantas dimensiones como características se deseen aprender (en nuestro caso, dos). Con esto, se implementan tantas redes LVQ independientes como idiomas haya en nuestro sistema, de manera que, para cada colección monolingüe, se tiene una red con dos unidades de entrada y una única unidad de salida.

El proceso de entrenamiento tiene lugar de la siguiente manera. En cada iteración, el algoritmo selecciona un vector de entrada, x_i, y lo compara con cada vector de pesos, w_k, usando alguna medida de similitud (en nuestro caso, concretamente se ha utilizado la distancia euclídea $\|x_i-w_k\|$); el vector w_c será el ganador si es el más cercano a x_i, por lo que c será la clase asignada:

$$\|x_i - w_c\| = \min_k \{\|x_i - w_k\|\} \qquad (4)$$

Las clases compiten entre ellas para encontrar el vector de entrada más parecido, para que el ganador sea el que menor distancia euclídea tenga respecto al vector de entrada. Sólo la clase ganadora podrá modificar el vector de pesos usando un algoritmo de aprendizaje reforzado, o positivo o negativo, dependiendo de que la clasificación sea correcta o no. De este modo, si la clase ganadora pertenece a la misma clase que el vector de entrada (la clasificación ha sido correcta), el peso se reforzará positivamente, acercándose ligeramente al vector de entrada. Por el contrario, si la clase ganadora es diferente a la clase del vector de entrada (la clasificación no ha sido correcta), se penalizará el peso de manera negativa, alejándose ligeramente del vector de entrada.

Sea $x_i(t)$ un vector de entrada en el tiempo t, y $w_k(t)$ el vector de pesos para la clase k en el tiempo t. La siguiente ecuación define el proceso de aprendizaje básico para el algoritmo LVQ.

$$w_c = \begin{cases} w_c + \alpha(t)\cdot(x_i - w_c) \text{ si } c = d \\ w_c - \alpha(t)\cdot(x_i - w_c) \text{ si } c \neq d \end{cases} \qquad (5)$$

donde $\alpha(t)$ es el ratio de aprendizaje, siendo $0<\alpha(t)<1$, una función monótona decreciente del tiempo. Se recomienda que $\alpha(t)$ sea más bien pequeña inicialmente, es decir, menor de 0,5, y que decrezca hasta un umbral dado muy cercano a 0 (Kohonen, 1995).

Una vez concluido el entrenamiento comienza la fase de producción. Para cada documento, se suministra a la red el ranking y el RSV, y la red da como salida la distancia del documento a la clase relevante (clase 1). Este valor se toma como la puntuación dada por la red a dicho documento.

5 RSV en dos pasos con la consulta parcialmente alineada

El principal problema del método RSV en dos paso es que necesita que cada término de la consulta debe estar alineado con los términos de los demás idiomas. Sin embargo, esto no siempre es posible, bien por la estrategia de traducción elegida o bien por el uso técnicas de expansión de consultas locales para cada colección.

Para tratar las consultas parcialmente alineadas se deben integrar tanto la parte alineada como la no alineada. Para ello, el peso de los términos que se encuentran alineados en la consulta se calcula utilizando el método RSV en dos pasos, tal como queda descrito más arriba. De esta manera obtenemos un primer indicador de la relevancia para cada documento. Independientemente de este indicador, se calcula un segundo peso para cada documento considerando tan solo los términos no alineados. Esto es, creamos una subconsulta formada únicamente por términos no alienados y se calcula la similitud entre esta subconsulta y cada uno de los documentos recuperados en la fase de preselección. La similitud documento-subconsulta se calcula siguiendo algún modelo tradicional como el vectorial, Okapi u otro (Frakes y Baeza-Yates, 1992). De esta manera, para cada documento se calculan dos puntuaciones o índices de relevancia . En este trabajo se propone dos nuevos métodos para integrar la parte alineada y la no alineada:

- Uso de regresión logística: Al igual que en el método descrito con regresión logística (Fórmula 3) en el que se usó la puntuación y el ranking para la fusión de documentos,

se puede utilizar el mismo método para integrar el ranking (encontrado por el documento al final de la primera fase de preselección de documentos del método RSV en dos pasos), el RSV alineado (rsv_i^a) y el RSV no alineado (rsv_i^{na}) (Fórmula 6):

$$\text{Prob}(d_i) = \frac{e^{\alpha+\beta_1 \cdot \ln(rank_i)+\beta_2 \cdot rsv_i^a+\beta_3 \cdot rsv_i^{na}}}{1+e^{\alpha+\beta_1 \cdot \ln(rank_i)+\beta_2 \cdot rsv_i^a+\beta_3 \cdot rsv_i^{na}}} \quad (6)$$

- Uso de RNA: El algoritmo LVQ descrito en la sección anterior se utiliza para integrar la parte alineada y no alineada además del ranking original del documento. En este caso, la red utiliza 3 neuronas de entrada (se pretende aprender 3 características) y una neurona de salida. Así, tanto los vectores de entrada como los vectores prototipo tienen 3 dimensiones. El proceso de aprendizaje y proceso de producción posterior es el mismo que el descrito para el algoritmo LVQ en la sección anterior.

De nuevo, el inconveniente que presentan ambos métodos es que para ajustar los parámetros es necesario un conjunto de datos de entrenamiento. Sin embargo, con estas aproximaciones no sólo se tienen en cuenta la parte alineada y no alineada sino que además integran el ranking original obtenido por el documento. Además, estos métodos se pueden aplicar cuando se tienen consultas totalmente alineadas (RSV no alineado es igual a 0) como una manera de mejorar el método RSV en dos pasos puesto que permite el uso de información extra procedente del primer paso (fase de preselección de documentos): el ranking del documento obtenido mediante el proceso de recuperación monolingüe.

6 *Experimentos y resultados*

Para los experimentos se han utilizado las colecciones del CLEF[1] 2001 y 2002 en 5 idiomas (inglés, español, alemán, francés e italiano), así como los juicios de relevancia.

	Inglés	Alemán	Francés	Español	Italiano
Nºdocs	113.005	225.371	87.191	215.738	108.578
Tamaño (MB)	425	527	243	509	278
Nºconsultas CLEF 2001	47	49	48	49	47
Nºdocs rel CLEF 2001	856	2.238	1.193	2.694	1.246
Nºconsultas CLEF 2002	42	50	50	50	49
Nºdocs rel CLEF 2002	821	1.938	1.383	2.854	1.072

Tabla 1: Descripción de las colecciones CLEF (extraído de (Savoy, 2001, Savoy, 2002))

Cada colección ha sido preprocesada usando listas de parada y algoritmos de stemming disponibles en Internet[2] (Frakes y Baeza-Yates, 1992). Las listas de parada se han incrementado con algunas otras palabras de uso común. Además, ya que el alemán utiliza palabras compuestas, se ha utilizado el paquete MORPHIX (Neuman, 2003) para reducirlas a palabras simples. Una vez preprocesadas las colecciones, se han indexado con el sistema de recuperación de información Zprise usando el modelo probabilístico Okapi (Robertson, Walker y Beaulieu, 2000).

Para realizar los experimentos, el primer paso consiste en conseguir una lista de documentos relevantes para cada una de las colecciones monolingües mediante la traducción de consultas a cada uno de los idiomas disponibles. Para traducir la consulta se ha utilizado Babylon[3]. Para cada término de la consulta, este diccionario bilingüe propone varias traducciones. En los experimentos se ha utilizado la primera palabra traducida propuesta para cada término. Sólo se ha tenido en cuenta los campos *Title* y *Description* de la consulta. Por último, se ha utilizado pseudo-realimentación (PRF – pseudo-relevance feedback) para expandir las consultas adoptando la aproximación de Robertson-Croft (Harman, 1992). Los resultados de los experimentos bilingües se muestran en la Tabla 2.

[1] El CLEF (Cross Language Evaluation Forum) es una actividad de carácter anual y de ámbito europeo que se celebra desde el año 2000 coordinado por DELOS Network of Excellence for Digital Libraries en colaboración con el NIST y el TREC.

[2] http://www.unine.ch/info/clef

[3] Babylon es un diccionario electrónico disponible a través de http://www.babylon.com

Idioma	Prec. Media CLEF 2001	Prec. Media CLEF 2002
Inglés	0,4582	0,5049
Inglés->Alemán	0,3232	0,3187
Inglés->Francés	0,4112	0,4677
Inglés->Español	0,4533	0,3867
Inglés->Italiano	0,3150	0,2817

Tabla 2: Experimentos bilingües

Para implementar la regresión logística se ha utilizado el paquete R (Ihaka y Gentleman, 1996).

Las pruebas realizadas con el algoritmo LVQ, fueron llevados a cabo usando la implementación descrita en la documentación de LVQ_PAK (Kohonen et al., 1996) con los parámetros por defecto. Así, para cada experimento se utilizan 2 codebooks o vectores de pesos, uno para la clase 0 y otro para la clase 1. El ratio de aprendizaje α se inicializa a 0,3.

La Tabla 3 muestra los resultados obtenidos mediante la utilización de las distintas técnicas descritas para la fusión de documentos con la consulta parcialmente alineada. Concretamente, se han utilizado la siguientes aproximaciones: Round-Robin, Raw Scoring, Raw Scoring Normalizado con las dos ecuaciones propuestas (Fórmulas 2 y 3), Regresión Logística, LVQ y RSV en dos pasos. En esta tabla también se muestra la precisión óptima teórica calculada mediante el algoritmo propuesto por Chen (2002). Este algoritmo mezcla las listas monolingües de forma óptima bajo la suposición de que se conserva el orden relativo de los documentos. Requiere para su cálculo el uso de los juicios de relevancia de las consultas por lo que su utilidad es puramente teórica.

Estrategia de fusión	Prec. Media CLEF 2001	Prec. Media CLEF 2002
Round-Robin (caso base)	0,273 (0%)	0,251 (0%)
Raw Scoring	0,291 (6,6%)	0,281 (11,9%)
RS N1	0,271 (-0,7%)	0,235 (-6,4%)
RS N2	0,297 (8,8%)	0,272 (8,4%)
Regresión Logística	Entrenamiento	0,289 (15,1%)
LVQ	Entrenamiento	0,293 (16,4%)
RSV en 2 pasos	0,327 (19,8%)	0,308 (22,7%)
Prec. Óptima	*0,420 (53,8%)*	*0,367 (46,2%)*

Tabla 3. Experimentos multilingües con consultas parcialmente alineadas

Como se observa, los mejores resultados se obtienen con el cálculo del RSV en dos pasos. La mejora obtenida es de un 22,7%. Por otra parte, el uso del algoritmo LVQ de manera independiente supera levemente los resultados obtenidos con la regresión logística. Sin embargo, ambas aproximaciones requieren la utilización de un conjunto de datos de entrenamiento. Para ello, se ha utilizado la colección CLEF 2001 con los juicios de relevancia para el entrenamiento mientras que la colección CLEF 2002 se usa para evaluación.

En un intento por mejorar los resultados obtenidos con el RSV en dos pasos, se han utilizado las dos técnicas que presentan la siguiente mejor precisión (regresión logística y LVQ) para integrar la parte alineada y la no alineada de la consulta. Los resultados se muestran en la Tabla 4.

Los resultados obtenidos mejoran considerablemente la precisión obtenida cuando se aplican los métodos por separado. El incremento de precisión con respecto al método RSV en dos pasos original es de un 4,6% cuando se utiliza regresión logística para integrar la parte alineada y no alineada, mientras que con el algoritmo LVQ la mejora es de un 9,0%.

	Prec. Media CLEF 2001	Prec. Media CLEF 2002
RSV en 2 pasos con RL	Entrenamiento	0,323 (28,7%)
RSV en 2 pasos con LVQ	Entrenamiento	0,337 (34,3%)
Prec. Óptima	0,420 (53,8%)	0,367 (46,2%)

Tabla 4. Integración de la parte alineada y no alineada con RSV en dos pasos

7 *Conclusiones*

En este artículo se presenta un estudio comparativo de distintas estrategias de fusión de documentos. En el estudio se incorpora una nueva técnica basada en el uso de aprendizaje neuronal. Concretamente, se aplica el algoritmo competitivo LVQ para generar una única lista de documentos en un sistema CLIR.

Aunque como muestran los resultados, la mejor opción es el método RSV en dos pasos, la red neuronal LVQ se presenta como una nueva alternativa con resultados prometedores que superan ligeramente a la regresión logística.

Por otra parte, el problema fundamental del método RSV en dos pasos es que requiere que los términos de la consulta estén totalmente alineados. Para resolver el problema, se ha integrado la parte alineada y la no alineada utilizando regresión logística y el algoritmo LVQ. Los resultados obtenidos muestran que el uso de ambas técnicas aumentan la precisión considerablemente.

Como trabajos futuros, se pretende utilizar otras redes neuronales aplicadas a la fusión de documentos CLIR como, por ejemplo, la red de contrapropagación (CPN – Counter Propagation Network) o la red de propagación hacia atrás (BPN – Backpropagation Network).

Otro aspecto de interés es el uso del método RSV en dos pasos con otras técnicas de traducción no basadas en diccionarios electrónicos, tales como máquinas de traducción automáticas o tesauros de similitud multilingües.

Agradecimientos

Este trabajo ha sido financiado con el proyecto (MCYT) FIT-150500-2003-412.

Bibliografía

Chen, A. 2002. Cross-language retrieval experiments at CLEFF-2002. *Proc. CLEF 2002*. pp. 5-20.

Fiesler, E., R. Beale. 1997. Handbook of Neural Computation. Oxford University Press.

Frakes, W.B., R. Baeza-Yates. 1992. Information retrieval: Data, structures and algorithms. Pretice Hall.

Grefenstette, G. 1998. Cross-Language Information Retrieval. Kluwer academic publishers. Boston.

Harman, D. 1992. Relevance feeback revisited. *Proc. ACM-SIGIR'92*. pp. 1-10.

Ihaka, R. y R. Gentleman. 1996. R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics*. Vol. 5. pp. 299-314.

Kohonen, T. 1995. Self-organization and associative memory. 2ª Edición, Springer.Verlag, Berlín.

Kohonen, T., J. Hynninen, J. Kangas, J. Laaksonen, K. Torkkola. 1996. Informe técnico, LVQ_PAK: The Learning Vector Quantization Program Package. Helsinki University of Technology, Laboratory of Computer and Information Science, FIN-02150 Espoo, Finlandia.

Le Calvé, A. 2000. Database merging strategy based on logistic regresión. Information Processing & Management. Vól. 36:341-359

Martínez, F., M.T. Martín, L.A. Ureña. 2002. SINAI on CLEF 2002: experiments with merging strategies. *Proc. CLEF'2002*. pp. 103-110.

McClelland, J., D. Rumelhart. 1986. Parallel Distributed Processing. Volúmenes I y II. MIT Press. Cambridge, MA.

Neumann, G. 2003. Morphix software package. http://www.dfki.de/~neumann/morphix/morphix.html. Disponible en mayo de 2003.

Powell, A.L., J.C. French, J. Callan, M. Connell y C.L. Viles. 2000. The impact of database selection on distributed searching. *Proc. of ACM-SIGIR'2000*. pp. 232-239. New York.

Robertson, S.E., S. Walker y M. Beaulieu. 2000. Experimentation as a way of life: Okapi at TREC. *Information Processing and Management*. Vol, 1, pp. 95-108.

Savoy, J. 2001. Report on CLEF-2001 experiments. *Proc. CLEF'2001*. pp. 27-43.

Savoy, J. 2002. Report on CLEF-2002 experiments: combining multiple sources of evidence. *Proc. CLEF'2002*. pp. 31-46.

Savoy, J. 2003. Cross-language information retieval: experiments based on CLEF-2000 corpora. *Inforamtion Processing and Management*. Vol. 39. pp. 75-115.

Voorhees, E. 1995. The collection fusion problem. *Proc. TREC-3*. pp. 95-104. Gaithersburg.