XI Conferencia de la Asociación Española para la Inteligencia Artificial

CAEPIA 2005

Santiago de Compostela, 16-18 de Noviembre de 2005

Actas, Volumen I

Entidades Organizadoras:

UNIVERSIDADE DA CORUÑA

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Entidades Colaboradoras:

- Ministerio de Educación y Ciencia.
- Consellería de Innovación e Industria (Dirección Xeral de Investigación, Desenvolvemento e Innovación), Xunta de Galicia.
- Consellería Educación e Ordenación Universitaria, Xunta de Galicia.
- Universidad de A Coruña.
- Universidad de Santiago de Compostela.

Editores:

Roque Marín
Eva Onaindia
Alberto Bugarín
José Santos

Imprime: CopyNino, Santiago de Compostela.
ISBN: 84-96474-13-5
Depósito Legal: C-2534-05

Prefacio

Hace poco más de cincuenta años, el 31 de Agosto de 1955, Marvin Minsky, John McCarthy, Nathan Rochester y Claude Shannon propusieron la celebración de una reunión de dos meses de duración, que tuvo lugar en el Dartmouth College durante el verano de 1956. En la llamada a la participación indicaban que el objetivo era discutir "la conjetura de que todos los aspectos del aprendizaje o de cualquier otra característica de la inteligencia pueden, en principio, ser descritos de modo tan preciso que se pueda construir una máquina capaz de simularlos". El tema era tan novedoso que acuñaron un nuevo término para él: Inteligencia Artificial (IA).

Está, pues, a punto de cumplirse el 50 aniversario de lo que se considera como el nacimiento histórico del campo de la IA. Ello nos brinda la oportunidad de lanzar una mirada crítica hacia el estado actual de la IA en el mundo, y en España en particular. Es tiempo también de expresar nuestro reconocimiento hacia las sucesivas generaciones de investigadores que han contribuido a consolidar y hacer avanzar este campo.

La Asociación Española para la Inteligencia Artificial (AEPIA), creada en 1983, ha contribuido a guiar y aglutinar a los investigadores españoles en IA durante más de 20 años, casi la mitad de esta andadura. AEPIA es miembro del ECCAI (European coordinating committee for Artificial Intelligence) y miembro fundador de IBERAMIA. AEPIA organiza bienalmente la Conferencia de la Asociación, cuyas ediciones anteriores se han celebrado en Madrid, Alicante, Málaga, Murcia, Gijón y Donostia.

Este año 2005, la XI Conferencia de la Asociación Española para la Inteligencia Artificial (CAEPIA'05) es acogida por la ciudad de Santiago de Compostela, cuya hospitalidad agradecemos sinceramente, y está organizada de modo conjunto por las Universidades de A Coruña y Santiago de Compostela.

Como en las anteriores ediciones, los objetivos de esta Conferencia son: facilitar la diseminación de nuevas ideas y experiencias, fortalecer los lazos entre los distintos grupos de investigación implicados, promover el trasvase de conocimiento entre nuevos investigadores y grupos consolidados, y ayudar a difundir los nuevos desarrollos hacia la sociedad. En CAEPIA'05, la comunidad de investigadores que trabaja en temas relacionados con la Inteligencia Artificial se reúne para presentar y discutir los últimos avances científicos y tecnológicos en este campo. Se acompaña, además, de otros eventos, como Talleres, Paneles y Tutoriales.

Ha sido nuestra pretensión consolidar y aumentar el nivel científico alcanzado en ediciones anteriores. Hemos contado, en esta undécima edición, con la participación excepcional, como conferenciantes invitados, del profesor Stephen Muggleton, director del *Computational Bioinformatics Laboratory* en el *Imperial College*, Londres; del profesor Luc Steels de la *Free University of Brussels* y director del *Sony Computer Science Laboratory* en Paris; del profesor Gerhard Brewka, del *Intelligent Systems Department - Computer Science Institute* de la Universidad de Leipzig, Alemania; de Peter Lucas, del *Institute for Computer and Information Sciences* de la Universidad de Nijmegen, Holanda. Finalmente, el profesor Vicente Botti, de la

Universidad Politécnica de Valencia, ha recibido el premio AEPIA a una carrera científica en IA. A todos ellos agradecemos muy sinceramente su contribución al éxito de este evento.

Los datos de participación de esta edición son un reflejo de la solidez de la investigación actual en el campo de la IA y de los fuertes lazos de colaboración internacional establecidos entre investigadores españoles y extranjeros. Se han recibido 147 artículos, presentados por un total de 319 autores de 16 países distintos. De estos 319 autores, el 20 % proceden de fuera de nuestras fronteras. Un 13 % de los artículos recibidos corresponde a trabajos predoctorales, confirmando que el flujo de jóvenes investigadores hacia el campo de la IA mantiene su tendencia al alza, comparado con los datos de anteriores ediciones.

De los 147 artículos recibidos, el Comité de Programa ha aceptado, con una selección rigurosa, un total de 84, lo que supone un índice de aceptación global del 57 %. Cada artículo ha sido revisado por tres miembros diferentes del Comité de Programa. Éste se compuso de 76 miembros de 14 países distintos. En concreto, el 27 % son revisores de procedencia internacional.

Los editores queremos manifestar nuestro agradecimiento a todos aquellos que, con su contribución generosa y desinteresada, han hecho posible la realización de esta Conferencia. Reciban nuestra más especial gratitud los conferenciantes y ponentes invitados, así como los investigadores que han contribuido con sus valiosos trabajos a que este libro sea una realidad.

Asimismo, queremos expresar nuestro agradecimiento al Comité de Programa y a los revisores externos, por haber realizado las revisiones de los trabajos con extraordinaria eficacia, prontitud y rigor científico, y a los miembros del Comité Organizador por la dedicación y diligencia mostradas en su labor. Expresamos también un especial agradecimiento a los miembros del grupo de apoyo al Comité de Programa que, desde la Universidad de Murcia, han dado desinteresadamente soporte a todas las labores técnicas asociadas al proceso de revisión, respondiendo con prontitud a todas las incidencias que siempre se encuentran en un proceso como éste.

Nuestra gratitud se extiende a la Junta Directiva de la Asociación Española para la Inteligencia Artificial y a las entidades que han colaborado en la financiación de un evento de esta envergadura. Entre éstas queremos destacar, por la financiación concedida, al Ministerio de Educación y Ciencia, a la Xunta de Galicia, que nos ha financiado a través de sus Consellerías de Innovación e Industria (Dirección Xeral de Investigación, Desenvolvemento e Innovación) y Educación e Ordenación Universitaria y al Ayuntamiento de Santiago de Compostela, además de las Universidades de A Coruña y Santiago de Compostela que, como no podía ser de otro modo, nos han brindado además todo su soporte y apoyo.

Gracias a todos por vuestras aportaciones y esfuerzo.

Roque Marín
Eva Onaindia
Alberto Bugarín
José Santos

Preface

About 50 years ago, on 31st of August of 1955, Marvin Minsky, John McCarthy, Nathan Rochester and Claude Shannon proposed the celebration of a two-month meeting at Dartmouth College during the summer of 1956. In the call for participation they indicated that the objective was to discuss "the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it". The subject was so novel that they coined a new term: Artificial intelligence (AI).

We are about to celebrate the 50th anniversary of what it is considered to be the historical birth of the field of AI. This event offers us the opportunity to take a critical glance at the present state of AI in the world, and in Spain specifically. It is also time to recognize the efforts of the successive generations of researchers who have contributed to consolidating and making progress in this field.

The Spanish Association for Artificial Intelligence (AEPIA), created in 1983, has contributed to guiding and gathering the Spanish researchers on AI for more than 20 years, almost half of this path. AEPIA is a member of the ECCAI (European Coordinating Committee for Artificial Intelligence) and founder member of IBERAMIA. AEPIA holds the Conference of the Association every two years. The previous conferences have taken place in Madrid, Alicante, Malaga, Murcia, Gijón and Donostia.

This year, 2005, the XI Conference of the Spanish Association for Artificial Intelligence (CAEPIA'05) is welcomed by the city of Santiago de Compostela, whose hospitality we sincerely appreciate. CAEPIA'05 is jointly organized by the Universities of A Coruña and Santiago de Compostela.

As in previous conferences, the goals of CAEPIA'05 are to create the necessary conditions for researchers to disseminate their work, to strengthen the relationships among the AI research groups, to facilitate the interaction between new researchers and consolidated groups and to help in spreading the new developments to society. In CAEPIA'05, the AI research community will have the opportunity to present and discuss their work and developments. The main conference is organized together with other events like Workshops, Panels and Tutorials.

It has been our aim to consolidate and increase the scientific level achieved in previous editions. In this eleventh edition we have benefited from the exceptional participation as invited speakers of Professor Stephen Muggleton, Head of the *Computational Bioinformatics Laboratory* at the *Imperial College*, London; Professor Luc Steels from the *Free University of Brussels* and Head of the *Sony Computer Science Laboratory* in Paris; Professor Gerhard Brewka, from the *Intelligent Systems Department - Computer Science Institute* at Leipzig University, Germany; Peter Lucas, from the *Institute for Computer and Information Sciences* at the University of Nijmegen, Holanda; and, finally, Vicente Botti, from the Polytechnic University of Valencia, who has, in addition, received this year´s AEPIA Scientific Career Award. We would like to sincerely acknowledge their willingness to participate in this event.

The participation rate of CAEPIA'05 is a reflection of the maturity of the current research on AI in Spain and the strong links of international collaboration between Spanish and foreign researchers. CAEPIA'05 received 147 submissions from 16 different countries by 319 authors, being 20% of the authors from abroad. 13% of the received articles correspond to pre-doctoral work, confirming that the flow of young researchers maintains its rising trend, compared to the previous conferences.

Among the 147 submissions received, a total of 84 papers were accepted which means an acceptance rate of 57%. Each paper was sent to three different reviewers. The Programme Committee is made up of 76 members from 14 different countries. In particular, 27% of the PC composition was international.

We would like to warmly thank all those people that with their generous contribution have made it possible to hold this conference. Special thanks are also due to the invited lecturers and speakers and to the researchers who have contributed with their valuable work to make this book a reality.

Likewise, we would like to thank the PC members and external reviewers for their efficient, rigorous and prompt reviews, as well as the Organizing Committee for the dedication and efficiency. Also, we would like to give special thanks to the members of the supporting group of the PC from the University of Murcia for their selfless and timely support in all those incidences that occur in a reviewing process.

We would also like to express our gratitude to the Board of directors of the AEPIA and to the sponsors for their financial support for hosting an event of this scope. Among these, due to the amount of funding provided, we would like to mention the Ministry of Education and Science of Spain, the *Xunta de Galicia*, who has provided funding through their *Consellerías de Innovación e Industria (Dirección Xeral de Investigación, Desenvolvemento e Innovación)* and *Educación e Ordenación Universitaria* and the City Council of Santiago de Compostela, as well as the Universities of A Coruña and Santiago de Compostela who have wholeheartedly supported this effort.

Thanks you all for your contributions and effort.

Roque Marín
Eva Onaindía
Alberto Bugarín
José Santos

Organización

Presidente de la Asociación:

Federico Barber, Universitat Politècnica de València (España).

Presidente del Comité de Programa:

Roque Marín, Universidad de Murcia (España)

Vicepresidente del Comité de Programa:

Eva Onaindia, Universitat Politècnica de València (España)

Miembros del Comité de Programa:

Carlos Alonso, Universidad de Valladolid (España).
Luis Alvarez, Universidad de Las Palmas de Gran Canaria (España).
José Angel Bañares, Universidad de Zaragoza (España).
Emilia Barakova, Brain Science Institute, (Japón)
Federico Barber, Universitat Politècnica de València (España).
Alvaro Barreiro, Universidade da Coruña (España).
Senén Barro, Universidade de Santiago de Compostela (España).
Beatriz Barros, UNED (España).
Daniel Borrajo, Universidad Carlos III de Madrid (España).
Vicent Botti, Universitat Politècnica de València (España).
Alberto Bugarín, Universidade de Santiago de Compostela (España).
Nuria Castell, Universitat Politècnica de Catalunya (España).
Helder Coelho, University of Lisboa (Portugal)
Ricardo Conejo, Universidad de Málaga (España).
Juan M. Corchado, Universidad de Salamanca (España).
Ulises Cortés, Universitat Politècnica de Catalunya (España).
Juan José del Coz, Universidad de Oviedo (España).
Angel P. del Pobil, Universitat Jaume I (España).
Alvaro del Val, Universidad Autónoma de Madrid (España).
Miguel Delgado, Universidad de Granada (España).
Yves Demazeau, Leibniz, Institut IMAG (Francia)
Rose Dieng, INRIA Sophia-Antipolis Unit (France).
Ed Durfee, University of Michigan, (E.E.U.U.)
Francisco Escolano, Universitat d'Alacant (España).
Wolfgang Faber, University of Calabria (Italia)
Luis Fariñas del Cerro, Université Paul Sabatier (Francia)
Juan Pedro Febles Rodríguez, Centro Nacional de Bioinformática (Cuba).
Isabel Fernández de Castro, Euskal Herriko Unibertsitatea/Universidad del País Vasco (España).
José Antonio Gámez, Universidad de Castilla-La Mancha (España).

Ana García-Serrano, Universidad Politécnica de Madrid (España).
Francisco Garijo, Telefonica I+D (España).
Hector Geffner, Universitat Pompeu Fabra (España).
Lluís Godo, Institut d'Investigació en Intel.ligència Artificial (España).
Antonio Gómez Skarmeta, Universidad de Murcia (España).
Asunción Gómez-Pérez, Universidad Politécnica de Madrid (España).
Manuel Hermenegildo, Universidad Politécnica de Madrid (España).
José Hernández Orallo, Universitat Politècnica de València (España).
Miguel Angel Jaramillo, Universidad de Extremadura (España).
Elena Lazkano, Euskal Herriko Unibertsitatea/Universidad del País Vasco (España).
Peter Lucas, University of Nijmegen (Holanda).
Lawrence Mandow, Universidad de Málaga (España).
Roque Marín, Universidad de Murcia (España).
Fernando Martín, Universidad de Murcia (España).
Mark Maybury, MITRE Corporation (E.E.U.U.)
Gaspar Mayor, Universitat de les Illes Balears (España).
Erica Melis, Universität des Saarlandes (Alemania).
José Mira, UNED (España).
Serafín Moral, Universidad de Granada (España).
Eduardo Morales Manzanares, ITESM, Morelos (México).
Juan José Moreno Navarro, Universidad Politécnica de Madrid (España).
José A. Moreno Pérez, Universidad de La Laguna (España).
Pablo Noriega, Institut d'Investigació en Intel.ligència Artificial (España).
Nuria Oliver, Microsoft Corporation (E.E.U.U.).
Eva Onaindia, Universitat Politècnica de València (España).
Sascha Ossowski, Universidad Rey Juan Carlos (España).
Ramón P. Otero, Universidade da Coruña (España).
José Tomás Palma, Universidad de Murcia (España).
David Pearce, Universidad Rey Juan Carlos (España).
Luigi Portinale, Universita' del Piemonte Orientale (Italia).
Maria Cristina Riff, Universidad Técnica Federico Santa María (Chile).
José Cristóbal Riquelme, Universidad de Sevilla (España).
Ramón Rizo, Universitat d'Alacant (España).
Jesús María Rodríguez Presedo, Universidade de Santiago de Compostela (España).
Camino Rodríguez Vela, Universidad de Oviedo (España).
José Santos Reyes, Universidade da Coruña (España).
Abdul Sattar, Griffith University (Australia).
Humberto Sossa, Instituto Politécnico Nacional (IPN) (México).
Luis Enrique Sucar, ITESM, Morelos (México).
María Jesús Taboada, Universidade de Santiago de Compostela (España).
Miguel Toro, Universidad de Sevilla (España).
Maite Urretavizcaya, Euskal Herriko Unibertsitatea/Universidad del País Vasco (España).
Felisa Verdejo, UNED (España).
José M. Vidal, University of South Caroline (E.E.U.U.).
Enrique Vidal, Universitat Politècnica de València (España).
Gerson Zaverucha, Federal University of Rio de Janeiro (Brasil).
Changjiu Zhou, Singapore Polytechnic (Singapur).

Revisores Externos:

Eva Armengol Voltas
Roi Blanco González
Manuel Campos Martínez
Purificación Cariñena Amigo
Carlos Carrascosa Casamayor
José M. Casanova Crespo
José Luis Correa Pombo
Carmelo del Valle Sevillano
Félix Díaz Hermida
Vicent Estruch Gregori
Paulo Félix Lamas
Antonio Fernández-Caballero
Manuel Fernández-Delgado
Cèsar Ferri
Pablo García Tahoces
Attilio Giordana
Laura Giordano
Adriana Giret Boggino
Giovambattista Ianni
Roberto Iglesias Rodríguez
José Manuel Juárez Herrero
Vicente Julián Inglada
Manuel Lama Penín
Carlos Linares López
Jim Lipton
Pedro López García
David E. Losada Carril
Antonio Lova
Marco Maratea
Julio Mariño Carballo
Rafael Martínez Gasca
Jesús Medina Moreno
Andreas Meier
Stefania Montani
Manuel Mucientes Molina
Isabel Navarrete Sánchez
Hector Palacios Verdes
Terenziani Paolo
José Luis Pérez de la Cruz Molina
Simona Perri
Jorge Puente Peinador
Jose Miguel Puerta Callejon
María José Ramírez Quintana

Miguel A. Rodríguez González
Horacio Rodriguez Hontoria
Miguel A. Salido
Eduardo M. Sánchez Vila
Laura Sebastiá Tarín
Basilio Sierra Araujo
Giorgio Terracina
Alicia Troncoso Lora
Mercedes Valdés Vela
Carlos Vázquez Regueiro
Xose A. Vila Sobrino
Alicia Villanueva

Co-presidentes del Comité Organizador:

Alberto J. Bugarín Diz, Universidade de Santiago de Compostela (España)
José Santos Reyes, Universidade da Coruña (España)

Presidente Workshops:

Richard José Duro Fernández, Universidade da Coruña (España)

Co-presidentes Tutoriales:

Manuel Lama Penín, Universidade Santiago de Compostela (España)
Eduardo M. Sánchez Vila, Universidade Santiago de Compostela (España)

Secretario del Comité Organizador:

Ramón P. Otero, Universidade da Coruña (España)

Vocales del Comité Organizador:

Alvaro Barreiro García, Universidade da Coruña (España)
José Antonio Becerra Permuy, Universidade da Coruña (España)
Francisco Bellas Bouza, Universidade da Coruña (España)
María J. Carreira Nouche, Universidade Santiago de Compostela (España)
Juan Luis Crespo Mariño, Universidade da Coruña (España)
Paulo Félix Lamas, Universidade Santiago de Compostela (España)
Manuel Fernández Delgado, Universidade Santiago de Compostela (España)
Pablo García Tahoces, Universidade Santiago de Compostela (España)
Adolfo Lamas Rodríguez, Universidade da Coruña (España)
David E. Losada Carril, Universidade Santiago de Compostela (España)
Jesús M. Rodríguez Presedo, Universidade Santiago de Compostela (España)
María Jesús Taboada Iglesias, Universidade Santiago de Compostela (España)
Carlos Vázquez Regueiro, Universidade da Coruña (España)

Soporte Técnico al Comité de Programa:

Manuel Campos Martínez, Universidad de Murcia (España)
José Manuel Juárez Herrero, Universidad de Murcia (España)
Antonio Morales Nicolás, Universidad de Murcia (España)
José Tomás Palma Méndez, Universidad de Murcia (España)
José Salort Rodríguez-Navas, Universidad de Murcia (España)

Índice / Table of Contents

Volumen I

Índice / Table of Contents

Volumen II

Sistemas Multiagente /Multiagent Systems

Visión Artificial / Artificial Vision

Panel "Experiencias empresariales en el campo de Inteligencia Artificial"

and problematic issues about Asbru are not so meaningful as for MHB, which has not been too much explored yet.

From our approach we conclude that, although MHB covers mostly all the features of the guideline we have modelled, there are still some aspects on which it can be improved.

On the one hand, the methodological process followed in this practical approach, solves in the first stage most difficulties and problematic points coming from the medical field and from the lack of medical knowledge of the knowledge engineer. Although this approach does not turn into an automatic process, the Asbru formalisation of the guideline seems to become easier using MHB, as shown in our experiment. Just before the Asbru modelling, we are provided, by the MHB model, with a series of useful resources such as a hierarchy of plans and a list of data items. Consequently, using this pre-processed information obtained from a unique additional step, formal modelling is eased by reducing effort and time resources.

The ongoing work is (a) to demonstrate that MHB is a good IR language towards other guideline representation languages, (b) to add other aspects not covered by MHB and, (c) to define a methodology for MHB modellers and also Asbru modellers to work in a systematic way in order to obtain better results in less time.

Acknowledgments

We want to thank all Protocure II members for their contribution to the work presented in this paper.

References

1. Field, M.J., Lohr, K.H.: Clinical Practice Guidelines: Directions for a New Program. National Academy Press, Institute of Medicine, Washington DC (1990)
2. Hanka, R., O'Brien, C., Heathfield, H., Buchan, I.E.: WAX ActiveLibrary: A tool to manage information overhead. Topics in Health Informatics Management (1990) 69–82
3. Woolf, S., Grol, R., Hutchinson, A., Eccles, M., Grimshaw, J.: Potential benefits, limitations, and harms of clinical guidelines. In: BMJ. Volume 318. (1999) www.bmj.com.
4. Lobach, D.F., Hammond, W.Ed., Durham: Computerized decision support based on a clinical practice guideline improves compliance with care standards. Volume 102., North Carolina, Am J Med (1997) 89–98
5. Seyfang,A., Miksch, S., Polo-Conde, C., Wittenberg, J., Marcos, M. and Rosenbrand, K.: MHB - A Many-Headed Bridge between Informal and Formal Guideline Representations. In: Springer's Lecture Notes of Computer Science Series. (2005)
6. Shahar, Y., Young, O., Shalom, E., Mayaffit, A., Moskovitch, R., Hessing, A., Galperin, M.: Degel: A hybrid, multiple-ontology framework for specification and retrieval of clinical guidelines. In: AIME. (2003) 122–131
7. Votruba, P., Miksch, S., Seyfang, A., Kosara, R.: Tracing the formalization steps of textual guidelines. In Kaiser, K., Miksch, S., Tu, S.W., eds.: Computer-based Support for Clinical Guidelines and Protocols, IOS Press (2004) 172–176
8. Miksch, S., Shahar, Y., Horn, W., Popow, C., Paky, F. and Johnson, P.: Time-oriented skeletal plans: Support to design and execution. In: Springer. (1997) 24–26
9. Seyfang, A., Kosara, R. and Miksch, S.: Asbru 7.3 Reference Manual. Vienna University of Technology, Institut of Software Technology and Interactive Systems. (2002)

A Comparative Impact Study of Corpus Preprocessing for the Construction of Anti-Spam Filtering Software

J. R. Méndez[1], E. L. Iglesias[1], F. Fdez-Riverola[1], F. Díaz[2], J.M. Corchado[3]

[1] Dept. Informática, University of Vigo, Escuela Superior de Ingeniería Informática, Edificio Politécnico, Campus Universitario As Lagoas s/n, 32004, Ourense, Spain
{moncho.mendez | eva | riverola}@uvigo.es
[2] Dept. Informática, University of Valladolid, Escuela Universitaria de Informática, Plaza Santa Eulalia, 9-11, 40005, Segovia, Spain
fdiaz@infor.uva.es
[3] Dept. Informática y Automática, University of Salamanca, Plaza de la Merced s/n, 37008, Salamanca, Spain
corchado@usal.es

Abstract. Spam mail is producing a considerable damage to the whole Internet community. Among the technical mechanisms proposed to reduce this problem filtering is specially promising. In order to train and test filters, it is necessary to have a large spam and legitimate e-mail corpus. This paper analyzes the strengths and weaknesses of the preprocessing techniques used in traditional text retrieval when they are applied to an e-mail corpus. We show the results obtained by three well-known machine approaches (Naïve Bayes, boosting trees and Support Vector Machines) and three case-based systems for spam filtering that can learn dynamically when the preprocessing of the training corpus changes. From the experiments carried out useful information is provided in order to choose the best method for spam filtering.

1 Introduction

At the moment anti-spam filtering software seems to be the most viable solution to the spam problem. Spam filtering methods are often classified as *collaborative* or *content-based* [1]. In the collaborative filtering, the selection of messages is made according to annotations or judgements made by other users [2]. On the other hand, in the content-based the selection is made according to intrinsic properties of them (e.g. message subject or body contents, structure, etc.) [3]. Despite there is no doubt collaborative techniques help to spam filtering, we focus in this paper in content-based approaches.

The main types of content-based techniques are machine learning (ML) algorithms and case-based (memory-based) reasoning approaches. ML approaches use an algorithm to 'learn' the classification from a set of training messages. On the other hand, memory-based and case-based reasoning techniques store all training instances in a memory structure and try to classify new messages finding similar e-mails on it. Hence, the decision of how to solve a problem is deferred until the last moment.

In order to train and test content-based filters, it is necessary to build a large corpus with spam and legitimate e-mails or use a public corpus. Anyway, e-mails have to be preprocessed to extract their words (*features*). Features can be extracted of the message subject, the body as well as attachments. Also, since the number of features in a corpus can end up being very high, in general it will also be necessary to choose those features that better represent e-mails before carrying out the filter training to prevent the classifiers from over-fitting.

The effectiveness of content-based anti-spam filters relies on the appropriate choice of the features. If the features are chosen so that they may exist both in a spam and legitimate messages then, no matter how good learning algorithm is, it will make mistakes. Therefore, the preprocessing steps of e-mail features extraction and the later selection of the most representative are crucial for the performance of the filter.

In this paper we analyze what are strengths and weaknesses of different feature extraction techniques used in text categorization when they are applied to the spam problem domain. Therefore, we will show the results obtained by different well-known content-based techniques when the preprocessing of the training corpus changes. The selected models for the evaluation are Naïve Bayes [4], boosting trees [5], Support Vector Machines [6], and three case-based systems for spam filtering that can learn dynamically: a Cunningham *et al.* system which we call *Cunn Odds Rate* [7], its improved version named *ECUE* [8] and the *SpamHunting* system [9].

The rest of the paper is structured as follows: in Section 2 we outline machine learning and case-based e-mail filters mentioned above. In Section 3 we describe the some available public corpus for empirical model evaluation and we study several issues related with message representation and feature selection. In Section 4 we present the experiments carried out and, finally, in Section 5 we expose the main conclusions reached.

2 Spam Filtering Techniques

The most recent trends in anti-spam filtering are based on the use of Case-based reasoning systems. Case-based approaches outperform classical machine learning techniques in anti-spam filtering [8]. Case-based classification works well for disjoint concepts as spam (spam about *porn* has little in common with spam offering *rolex*) whereas classical ML techniques try to learn a unified concept description. Another important advantage of this approach is the ease with which it can be updated to tackle the *concept drift* problem in the anti-spam domain [10].

Cunningham *et al.* present in [8] a successful case-based system for anti-spam filtering that can learn dynamically. The system uses a similarity retrieval algorithm based on Case Retrieval Nets (CRN) [11]. CRN networks are equivalent to the *k*-nearest neighbourhood algorithm but are computationally more efficient in domains where there is feature-value redundancy and missing features in cases, as spam. This classifier use unanimous voting to determine whether a new e-mail is spam or not. All the returned neighbours need to be classified as spam e-mails in order to classify as spam the new message. Delany *et al.* present in [7] an evolution from this system called ECUE (*E-mail Classification Using Examples*). The former uses an odds ratio method for feature selection in opposite of the latest that uses Information Gain (IG) and includes a case base editing technique used for eliminate inconsistencies on the knowledge.

Included in the context of lazy learning, SpamHunting an hybrid system has been introduced in [9] to accurately solve the problem of spam labelling and filtering. The model follows an Instance-Based Reasoning (IBR) approach. According to this, SpamHunting uses an instance memory structure as primary way of manage knowledge. The retrieval stage is carried out using a novel dynamic *k*-NN Enhanced Instance Retrieval Network (EIRN). The EIRN network facilitates the indexation of instances and the selection of those that are most similar to the new e-mail. Similarity between two given e-mails is measured by the number of relevant features found in both messages. EIRN can quickly retrieve all stored e-mails having at least one shared feature with a target message. The reuse of similar messages is done by using a simple unanimous voting mechanism to determine whether the target case is spam or not. The revision stage is only carried out in the case of unclassified messages, where the system employs general knowledge in the form of meta-rules extracted from the e-mail headers to assign a final class.

By other side, the most popular classical filtering models are bayesian methods. Bayesian filtering is based on the principle that most of the events are conditioned. So, the probability that an event happens can be deduced from the previous appearances of that event. This technique can be used for spam filtering. If some feature is often in spam but not in legitimate e-mails, then it would be reasonable to assume that an e-mail including this feature will be probably spam. Although there are several approaches of the bayesian method, the most widely used to spam filtering is Naïve Bayes algorithm [4].

Besides bayesian models, Support Vector Machines (SVM) and boosting techniques are also well-known ML techniques used in this field.

SVMs [6] has become very popular in the ML and DM community due to its good generalization performance and its ability to handle high-dimensional data through the use of kernels. They are based on representing e-mails as points in an *n*-dimensional space and finding an hyperplane that generates the largest margin between the data points in the positive class and those in the negative class. Some implementations of SVM can be found in ML environments such as Waikato Environment for Knowledge Analysis[1] (WEKA) or Yet Another Learning Environment[2] (YALE). Particularly, WEKA includes the *Sequential Minimal Optimization* (SMO) algorithm which has demonstrated a good trade-off between accuracy and speed (see [12] for details).

Boosting algorithms [5] are techniques based on the use of weak learners; that is to say, algorithms that learn with a next error rate to 50%. The main idea of boosting is to combine the hypotheses to one final hypothesis, in order to achieve higher accuracy than the weak learner's hypothesis would have. Different boosting algorithms

[1] WEKA is available from http://www.cs.waikato.ac.nz/ml/weka/
[2] YALE is available from http://yale.sourceforge.net

have been developed for classification tasks, so much binary as multi-class. Among them we could highlight Adaboost [13].

Several new ML models has been introduced for e-mail classification such as Chung-Kwei [14], which is based on pattern-discovery. As well as it is faster than other ML approaches, it becomes better on performance.

3 Corpus Preprocessing

3.1 Analysing the Available Corpus

In order to carry out a benchmarking of the spam filter models, it is essential to have several collection of emails for training and testing purposes. Some research workers on the spam filtering domain, had built their own corpus and shared it with the scientific community.

Table 1. Comparative study of the most well-known corpus

Corpus	%Legitimate	%Spam	Format	Preprocessing steps applied
Ling Spam	83,3	16,6	Tokens	Tokenized
PU1	56,2	43,8	Token Ids	Tokenized with id representation for each token
PU2	80	20	Token Ids	Tokenized with id representation for each token
PU3	51	49	Token Ids	Tokenized with id representation for each token
PUA	50	50	Token Ids	Tokenized with id representation for each token
SpamAssassin	84,9	15,1	RFC 822	Not preprocessed
Spambase	39,4	60,6	Feature Vectors	Vectors made with a previous feature selection
Junk-Email	0	100	XML	Not preprocessed
Bruce Guenter	0	100	RFC 822	Not preprocessed
DivMod	0	100	RFC 822	Not preprocessed

Despite privacy issues, a large number of corpus like SpamAssassin[3], Ling-Spam[4], DivMod[5], SpamBase[6] or JunkEmail[7] can be downloaded from Internet. Table 1 shows a short description of the existent corpus focussing in the spam and legitimate ratio and the distribution form.

[3] Available at http://www.spamassassin.org/publiccorpus/
[4] Available at http://www.iit.demokritos.gr/
[5] Available at http://www.divmod.org/cvs/corpus/spam/
[6] Available at http://www.ics.uci.edu/~mlearn/MLRepository.html
[7] Available at http://clg.wlv.ac.uk/projects/junk-e-mail/

In our work, we use the SpamAssassin corpora. It contains 9332 different messages from January 2002 up to and including December 2003. Since this corpus has not been preprocessed by the author, it can successfully be used to analyze the impact of applying different preprocessing techniques.

3.2 Message Representation

A relevant issue in ML applied to spam filtering is the internal structure of the messages used by different models during the training and the classification stages. Each message is often converted into a reliable message descriptor which can be easily assembled. In learning algorithms, training messages are usually represented as a vector $\vec{t} = \langle t_1, t_2, ..., t_p \rangle$ of weighted terms, T_i, much as in the vector space model in information retrieval [15, 16].

Feature identification can be performed by using a variety of generic lexical tools, primarily by tokenising the e-mail into words. At first glance, it seems to be a simple tokenising task guided by spaces as word separators. However, at least the following particular cases have to be considered with care: hyphens, punctuation marks, and the case of the letters (lower and upper case).

In the spam domain punctuation marks and hyphenated words are among the best discriminating attributes in a corpora, because they are more common in spam messages than legitimate ones. The tokenising step should be done carefully because the use of some lexical analizers can delete this kind of symbols. On the subject of case, the lexical analyzer normally converts all the text to either lower or upper case.

When the tokenising step has been completed, *stopword removal* (which drop articles, connectives and other words without semantic content) and/or a *stemming* (which reduces distinct words to their common grammatical root) can be applied to identified tokens [16].

Once carried out the lexical analysis over the corpus, the weight of terms in each message e, need to be calculated. The measure of the weight can be (*i*) binary (1 if the term occurs in the message, 0 otherwise), (*ii*) the *term frequency* (TF) representing the number of times the term occurs in the message calculated by Expression (1) or (*iii*) the *inverse document frequency* (IDF) given by Expression (2) denoting those terms that are common across the messages of the training collection.

$$t_i(e) = \frac{n_i(e)}{N(e)} \tag{1}$$

$$t_i(e) = \frac{n_i(e)}{N(e)} \log_2 \frac{m}{df(T_i)} \tag{2}$$

In Equations (1) and (2), $n_i(e)$ is the number of occurrences of term T_i in e, $N(e)$ represents the total number of occurrences of terms in e, m is the number of training messages and $df(T_i)$ stands for the number of training messages where the term T_i occurs.

4 Evaluation

The main goal of our experiments is to analyze the impact of applying distinct preprocessing techniques over the corpus. Naïve Bayes, SVM, Adaboost, and the three previously commented CBR systems (ECUE, *Cunn Odds Rate* and SpamHunting) had been analyzed in three different preprocessing scenarios: (*i*) applying stopword removal and stemming analysis, (*ii*) applying stopword but without stemming and (*iii*) without applying neither stopword nor stemming.

Six well-known metrics [3] have been used in order to evaluate the performance (efficacy) of all the analyzed models: *total cost ratio* (TCR) with three different cost values, spam *recall*, spam *precision*, percentage of correct classifications (%OK), percentage of False Positives (%FP) and percentage of False Negatives (%FN).

All the experiments have been carried out using a 10-fold stratified cross-validation [17] in order to increase the confidence level of results obtained.

The most widely used feature selection method is based on computing the *Information Gain* (IG) [18] of each term t with respect to the category c, where $c \in \{l, s\}$ (legitimate and spam categories, respectively) using the Expression (3). Subsequently, those terms whose value of IG overcomes a certain threshold are selected.

$$IG(t,c) = \sum_{c \in \{l,s\}} P(t \wedge c) \cdot \log \frac{P(t \wedge c)}{P(t) \cdot P(c)} \qquad (3)$$

Analyzed models except from *Cunn Odds Rate* and SpamHunting systems use IG to select the most predictive features as it has been shown to be an effective technique in aggressive feature removal in text classification [18]. For our comparisons, we have selected the best performance model of each technique varying between 100 and 2000 features. For *Cunn Odds Rate* model, we have maintained the original technique of selecting 30 words for representing spam e-mails plus 30 words representing legitimate messages. The algorithm employed for sorting the vocabulary is based on the odds-ratio described in [8].

SpamHunting terms selection is not made using the vocabulary of the whole corpus. Instead of this, each message has its own relevant terms. The relevant feature list of each message is computed as the minimum set containing the most frequent features of the specified e-mail, which frequency amount is greater than a specified threshold in the range [0,1]. As the best results have been obtained using the 30% frequency amount, we computed the relevant feature list as the most repeated features whose frequency amount is greater than mentioned threshold.

4.1 Results

In order to compare the performance of the models taking into account the three predefined scenarios but with a cost-sensitive point of view, we calculate the TCR score in the mentioned different situations. TCR assumes than FP errors are λ times more costly than FN errors, where λ depends on the usage scenario (see [3] for more details). In the experiments carried out in this paper, the values for λ parameter were 1, 9 and 999.

Table 2. TCR scores over 10 fold-cross validation

		Na Bayes	Ada boost	SVM	Cunn Odds Rate	ECUE	Spam Hunting
Scenario 1	TCR λ=1	2.985	4.659	19.011	1.174	4.562	6.024
	TCR λ=9	0.568	1.689	3.353	1.153	3.071	4.823
	TCR λ=999	0.006	0.022	0.033	0.779	0.099	1.545
Scenario 2	TCR λ=1	2.863	5.003	17.990	1.364	6.020	7.498
	TCR λ=9	0.508	1.618	3.559	1.351	2.846	5.331
	TCR λ=999	0.005	0.019	0.037	1.130	0.046	0.874
Scenario 3	TCR λ=1	3.174	4.854	23.120	1.247	5.622	4.594
	TCR λ=9	0.560	1.432	4.920	1.242	2.498	2.982
	TCR λ=999	0.005	0.017	0.050	1.140	0.036	0.081

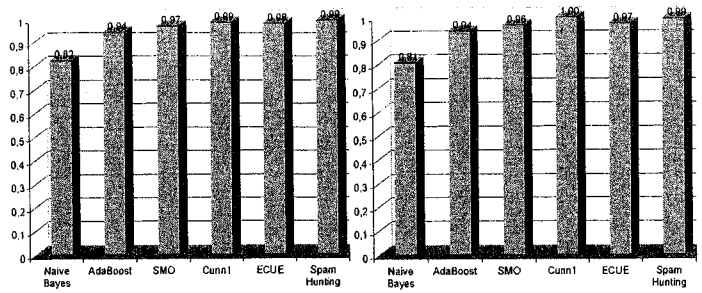
Table 2 shows the results taking into account the TCR score and varying the λ parameter. The best performance on classical ML approaches are obtained when no stemming and no stopword removal are applied (Scenario 3) except for Adaboost, that increases its TCR score when only stopword are applied (Scenario 2). By other side, CBR/IBR approaches work better if stopword removal is used (Scenarios 1 and 2).

Table 3. Recall scores using different scenarios

	Scenario 1	Scenario 2	Scenario 3
Naïve Bayes	0.842	0.853	0.869
Adaboost	0.836	0.854	0.858
SVM	0.978	0.977	0.916
Cunn Odds Rate	0.150	0.266	0.198
ECUE	0.793	0.857	0.846
Spam Hunting	0.837	0.871	0.795

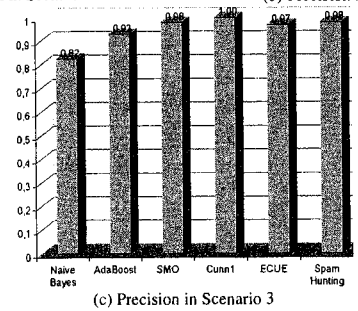
Table 3 shows a comparative study between the three proposed scenarios using recall score. Evaluation results on recall using classical ML models are better when no stopword removal and no stemming is used (Scenario 3). However, CBR/IBR approaches become better when only stopword removal is performed (Scenario 2) while the worst results are obtained when both stemming and stopword removal is applied (Scenario 1).

As we can see from Table 3, applying stemming can significantly reduce the number of selected features belonging to the corpus. According to this, it will also decrease the time needed to compute IG for all features and the spam recall score.



(a) Precision in Scenario 1

(b) Precision in Scenario 2



(c) Precision in Scenario 3

Fig. 1. Precision score graphics in the three different scenarios

Figure 1 shows a comparative study using the precision scores between the proposed scenarios. Analyzing in Figure 1 precision scores gathered from experiments, we can realize that it is possible to obtain better results when stopword removal and stemming is applied (Scenario 1) except for SVM and *Cunn Odds Rate* models. These techniques work better without any preprocessing step (Scenario 3).

Table 4 shows the amount of correct classifications, false positives and false negatives belonging to the experimental work with the six analyzed models over the defined scenarios. Analyzing Table 4 we can realize that Naïve Bayes and SVM techniques get better performance with no stemming and no stopword removal (Scenario 3). When stopword and stemming are used (Scenario 1) SpamHunting, ECUE and Adaboost report better accuracy but generally worst results. However, applying only stopword removal (Scenario 2) in those models that do not incorporate IG for feature selection leads to a significant accuracy increment.

Table 4 suggests that classical ML based models can get the best number of correctly classified messages by removing all preprocessing steps, but stopword removal and stemming is recommended if best accuracy is needed. Also, CBR/IBR models

can obtain better performance by stopword removal although stemming can improve accuracy.

Table 4. Mean value of correct clasifications, FPs and FNs with 10 fold-cross validation

		Scenario 1	Scenario 2	Scenario 3
Naïve Bayes	OK	852.5	849.6	857.9
	False Positives	43	48.6	44
	False Negatives	37.7	35	31.3
Adaboost	OK	881.9	885.1	883.4
	False Positives	12.2	13.4	16
	False Negatives	39.1	34.7	33.8
SVM	OK	920.2	919.1	922.2
	False Positives	7.8	8.7	5.3
	False Negatives	5.2	5.4	5.7
Cunn Odds Rate	OK	730.4	758.3	742.1
	False Positives	0.5	0.2	0.1
	False Negatives	202.3	174.7	191
ECUE	OK	880.2	893	889.8
	False Positives	3.6	6.1	6.8
	False Negatives	49.4	31.4	36.6
Spam Hunting	OK	892.9	900.8	880.3
	False Positives	1.5	1.8	4.2
	False Negatives	38.8	30.6	48.7

5 Conclusions

The final goal of our experiments is a comparative study of the above corpus preprocessing steps when they are applied in spam filtering. The experiments have been done using implementations of Naïve Bayes, Adaboost, SVM, and three case-based systems: a Cunningham *et al.* system which we call *Cunn Odds Rate*, its improved version named *ECUE* and the *SpamHunting* system.

We have explored the use of two text preprocessing techniques: the stopword removal and the stemming. Six well-known metrics have been used in order to evaluate the performance: percentage of correct classifications (OK), percentage of False Positives (FP), percentage of False Negatives (FN), spam *recall*, spam *precision* and *total cost ratio*. Lastly, all the experiments have been carried out using a 10-fold stratified cross-validation in order to increase the confidence level of results obtained.

From the analysis of these results, we can infer that stemming should be used for reducing the amount of the FP errors. Legitimate messages are better classified when stemming is used because the identification of the semantic roots can be successfully done. Nevertheless, stemming is unsuitable for the maximization of the correct classification rate.

Finally, as the experimental results show, the preprocessing techniques applied over the corpus need to be kept in mind for compare the accuracy obtained by different models.

References

1. Oard, D.W.: The state of the art in text filtering. User Modeling and User-Adapted Interaction, Vol.7, (1997) 141–178
2. Wittel, G.L., Wu, S.F.: On Attacking Statistical Spam Filters. Proc. of the First Conference on E-mail and Anti-Spam CEAS, (2004)
3. Androutsopoulos, I., Paliouras, G., Michelakis, E.: Learning to Filter Unsolicited Commercial E-Mail. Technical Report 2004/2, NCSR "Demokritos", (2004)
4. Sahami, M., Dumais, S., Heckerman, D., Horvitz, E.: A Bayesian approach to filtering junk e-mail. In Learning for Text Categorization – Papers from the AAAI Workshop, Technical Report WS-98-05, (1998) 55–62
5. Carreras, X., Màrquez, L.: Boosting trees for anti-spam e-mail filtering. Proc. of the 4th International Conference on Recent Advances in Natural Language Processing, (2001) 58–64
6. Vapnik, V.: The Nature of Statistical Learning Theory. 2nd Ed. Statistics for Engineering and Information Science, (1999)
7. Delany, S.J., Cunningham P., Coyle L.: An Assessment of Case-base Reasoning for Spam Filtering. Proc. of Fifteenth Irish Conference on Artificial Intelligence and Cognitive Science: AICS-04, (2004) 9–18
8. Cunningham, P., Nowlan, N., Delany, S.J., Haahr, M.: A Case-Based Approach to Spam Filtering that Can Track Concept Drift. Proc. of the ICCBR'03 Workshop on Long-Lived CBR Systems, (2003)
9. Fdez-Riverola, F., Méndez, J. R., Iglesias, E. L., Díaz, F. Representación Flexible de e-mails para la construcción de filtros antispam: un caso práctico. Proc. of the I Congreso Español de Informática CEDI'05 (2005) 109-116
10. Widmer, G., Kubat, M.: Learning in the presence of concept drift and hidden contexts. Machine Learning, Vol. 23 (1). (1996) 69–101
11. Lenz, M., Auriol, E., Manago, M.: Diagnosis and Decision Support. Case-Based Reasoning Technology. Lecture Notes in Artificial Intelligence, Vol. 1400, (1998) 51–90
12. Platt, J.: Fast training of Support Vector Machines using Sequential Minimal Optimization. In Sholkopf, B., Burges, C., Smola, A. (eds.). Advances in Kernel Methods – Support Vector Learning, MIT Press, (1999) 185–208
13. Schapire, R.E., Singer, Y.: BoosTexter: a boosting-based system for text categorization. Machine Learning, Vol. 39 (2/3). (2000) 135–168
14. Isidore Rigoutsos and Tien Huynh. Chung-Kwei: a Pattern-discovery-based System for the Automatic Identification of Unsolicited E-mail Messages (SPAM). Proc. of the First Conference on E-mail and Anti-Spam CEAS, (2004)
15. Salton, G., McGill, M.: Introduction to modern information retrieval, McGraw-Hill, (1983)
16. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. Addison Wesley, (1999)
17. Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the 14th International Joint Conference on Artificial Intelligence: IJCAI-95, (1995) 1137–1143
18. Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. Proc. of the Fourteenth International Conference on Machine Learning: ICML-97, (1997) 412–420

Solución basada en un Lenguaje de Representación y un Motor de Ejecución de Guías de Práctica Clínica para la ayuda a la decisión en Atención Primaria y Urgencias

Flórez S.H., Pikatza J.M., Larburu I.U., Sobrado F.J.

Departamento de Lenguajes y Sistemas Infonnáticos (UPV/EHU)[1]
{sh.florez, jm.pikatza, jiblaeni, fj.sobrado}@ehu.es

Abstract. En este artículo se presenta la creación de un nuevo lenguaje para la representación de guías denominado LAGRE y un nuevo motor de ejecución para dichas guías llamado TESEO. LAGRE está diseñado a partir de PRO*forma*, simplificando los componentes de modelado e integrándolos en el entorno Protégé-2000. La implementación del motor en Java está orientada a la construcción de sistemas Web. El rediseño del motor para ser desarrollado en CLIPS hace que TESEO sea ejecutable en dispositivos móviles. Ambas aportaciones han sido probadas en procesos de implementación de guías en la construcción de sistemas de ayuda a la decisión resultando satisfactorias para los expertos.

1 Introducción

Los servicios de urgencias (SU) son una parte fundamental de toda la estructura hospitalaria. Sus funciones son: 1) accesibilidad, debe estar capacitado para recibir pacientes urgentes durante las 24 horas del día, todos los días del año; 2) recepción, valoración y manejo inicial de los pacientes, incluso los no graves, realizando las actuaciones precisas para la rápida estabilización clínica; 3) derivación adecuada, organizar el acceso de los pacientes al lugar adecuado según su estado; y 4) docencia, investigación y actividades preventivas y educativas [15]. Son varios los problemas que derivan de dichas funciones de entre los cuales nosotros consideraremos dos: la clasificación de pacientes y su estabilización en urgencias, y la atención primaria en enfermedades crónicas como, por ejemplo, el asma. Las decisiones y acciones a tomar en dichos ámbitos dependen totalmente de la experiencia y conocimientos del personal sanitario que la realice, por lo que suele conllevar una falta de uniformidad en la práctica clínica.

Durante la última década se ha extendido el fenómeno de la creación y difusión de guías de práctica clínica (GPC) con el objetivo de reducir la variabilidad en la práctica médica, fomentado por la propagación de la Medicina Basada en la Evidencia (MBE) [17]. Su implantación pretendió ser inicialmente el puente entre la MBE y la práctica clínica. Actualmente ha logrado que también se promocione el control de costos en

[1] Fac. Informática (UPV/EHU), Dep. L.S.I, Apdo. 649, 20080 Donostia - San Sebastián.