

INTERNATIONAL SOCIETY FOR KNOWLEDGE ORGANIZATION [ISKO]

# Challenges in Knowledge Representation and Organization for the 21st Century. Integration of Knowledge across Boundaries

Proceedings of the  
Seventh International ISKO Conference  
10-13 July 2002  
Granada, Spain

Organized by the  
ISKO – Spanish Chapter  
and  
The University of Granada

Edited by

**MARÍA J. LÓPEZ-HUERTAS**

With the assistance of

**FRANCISCO J. MUÑOZ-FERNÁNDEZ**

MARÍA J. LÓPEZ-HUERTAS (Ed.)  
Challenges in Knowledge Representation and Organization for  
the 21st Century. Integration of Knowledge across Boundaries

ISBN 3-89913-247-5

Ergon

Ergon

Predocumentation

The volume contains:

Introduction - Keynote address - Theoretical models and universals in knowledge organization and representation - Epistemological foundations for knowledge structures and analysis - Models and methods for knowledge representation - Models and methods for knowledge organization. Tools and systems - Models and methods for knowledge organization and retrieval - Organization of integrated knowledge in the electronic environment. The internet - Models and methods for knowledge organization and conceptual relationships - Integration of knowledge in the Internet. Representing knowledge in Web sites - Models and methods for knowledge integration in information systems - Applications of Artificial Intelligence Techniques to Information Retrieval (Part I) - Integration of knowledge in multicultural domain-oriented and general systems (Part I) - Applications of Artificial Intelligence Techniques to Information Retrieval (Part II) - Epistemological approaches to classification principles, design and construction - Professional ethics. Users and information structures. Evaluation of systems - Integration of knowledge in multicultural domain-oriented and general systems (Part II) - Applications of Artificial Intelligence Techniques to Information Retrieval (Part III) - Subject Index - List of Contributors

Die Deutsche Bibliothek - CIP-Einheitsaufnahme  
Ein Titeldatenatz für diese Publikation ist bei  
Der Deutschen Bibliothek erhältlich

© 2002 ERGON Verlag · Dr. H.-J. Dietrich, D-97080 Würzburg

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in other ways and storage in databanks. Duplication of this publication or parts thereof is only permitted under the provisions of the German Copyright Law, a copyright fee must always be paid.  
Cover Design: Jan von Hugo

Printed in Germany

ISBN 3-89913-247-5

ISSN 0938-5495

Contents

Introduction 12

Key note address

Rebecca GREEN  
Conceptual Universals in Knowledge Organization and Representation 15

1. Theoretical Models and Universals in Knowledge Organization and Representation

Jack ANDERSEN  
Ascribing Cognitive Authority to Scholarly Documents on the (Possible) Role of Knowledge Organization in Scholarly Communication 28  
Elin K. JACOB  
Augmenting Human Capabilities: Classification as Cognitive Scaffolding 38  
Clare BEGHTOL  
Universal Concepts, Cultural Warrant and Cultural Hospitality 45  
Maria Néida GONZÁLEZ DE GÓMEZ  
Knowledge, Communication, Information: Intersubject Links Institutional and Technological Mediations in Information 50  
Joe TENNIS  
Subject Ontogeny: Subject Access Through Time and the Dimensionality of Classification 54

2. Epistemological Foundations for Knowledge Structures and Analysis

Nuno SILVA and João ROCHA  
Merging Ontologies Using a Bottom-Up Lexical and Structural Approach 60  
Giliola NEGRINI and Patrizia ZOZI  
Ontological Analysis of the Literary Work of Art Jarmo SAARTI 67  
The Analysis of the Information Process of Fiction: a Holistic Approach to Information Processing 74  
N.Y. KOBASHI, J.W. SMIT and M. de F.G.M. TÁLAMO  
Constitution of the Scientific Domain of Information Science 80

3. Models and Methods for Knowledge Representation

Anita COLEMAN  
A Classification of Models 86

- Gian Piero ZARRI  
Indexing and Querying of Narrative Documents, a Knowledge  
Representation Approach 93
- Jeremy J. SHAPIRO  
Interdisciplinary Knowledge Integration and Intellectual Creativity  
Rahmatollah FATAHI and Mehri PARJROKH  
Restructuring the Bibliographic Record for Better Organization,  
Management, and Representation of Knowledge in the Global Online  
Environment: a New Approach 100
- Devika P. MADALLI and A.R.D. PRASSAD  
Vyasa: a Knowledge Representation System for Automatic Maintenance  
of Analytico-Synthetic Scheme 107
- Catalina NAUMIS PEÑA  
Images and Words 113
- 120
- 4. Models and Methods for Knowledge Organization. Tools and  
Systems**
- María Inés CORDEIRO and Aida SLAVIC  
Data Models for Knowledge Organization Tools: Evolution and  
Perspectives 127
- Vanda BROUGHTON  
Facet Analytical Theory as a Basis for Knowledge Organization Tool in a  
Subject Portal 135
- Stella G DEXTRE CLARKE  
Planning Controlled Vocabularies for the Uk Public Sector 142
- Widad MUSTAFA el HADI  
Terminology & Information Retrieval: New Tools for New Needs.  
Integration of Knowledge Across Boundaries 149
- Hur-Li LEE and Allyson CARLYLE  
Academic Library Gateways to Online Information: a Taxonomy of  
Organizational Structures 158
- 5. Models and Methods for Knowledge Organization and Retrieval**
- Gerhard J.A. RIESTHUIS and Maja ZUMER  
The Functional Requirements for Bibliographic Records and Knowledge  
Organization 165
- Rochelle KEDAR and Snumith SHOHAM  
The Subject Cataloging Of Monographs With The Use Of a Thesaurus  
Ana PEREZ LOPEZ, Mercedes DE LA MONEDA and  
Ángel MOROS RAMÍREZ  
Application of the Cantor set Theory in Making Decision about the  
Collection Development 173
- Hemalata IYER and Jeanne M. KEEFE  
The WordNet as an Auxiliary Resource To Search Visual Image  
Database In Architecture 181
- 186

- Douglas TUDHOPE, Ceri BINDING, Dorothee BLOCKS and Daniel  
CUNLIFFE  
Representation and Retrieval in Faceted Systems 191
- 6. Organization of Integrated Knowledge in the Electronic  
Environment. The Internet**
- José Antonio SALVADOR OLIVÁN, José María ANGÓS ULLATE  
and María Jesús FERNÁNDEZ RUIZ  
Organization of the Information about Health Resources on the Internet 198
- Eduardo PEIS, Antonio RUIZ, Francisco J. MUÑOZ-FERNÁNDEZ and  
Francisco de ALBA QUIÑONES  
Practical Method to Code Archive Findings Aids in Internet  
Marthinus S. VAN DER WALT  
An Integrated Model For The Organization Of Electronic  
Information/Knowledge in Small, Medium and Micro Enterprises  
(Sime's) in South Africa 205
- Ricardo EITO BRUN  
Software Development and Reuse as Knowledge Management Practice  
Roberto POLI  
Framing Information 211
- 218
- 225
- 7. Models and Methods for Knowledge Organization and  
Conceptual Relationships**
- Terence R. SMITH, Marcia Lei ZENG, and ADEPT Knowledge  
Organization Team  
Structured Models of Scientific Concepts for Organizing, Accessing, and  
Using Learning Materials 232
- M. OUSSALAH, F. GIRET and T. KHAMMACI  
A kr Multi-hierarchies/Multi-Views Model for the Development of  
Complex Systems 240
- Jonathan FURNER  
A Unifying Model of Document Relatedness for Hybrid Search Engines 245
- José Manuel BARRUECO and Vicente Julián INGLADA  
Reference Linking in Economics: The Citec Project  
Allyson CARLYLE and Lisa M. FUSCO  
Equivalence in Tillett's Bibliographic Relationships Taxonomy: a  
Revision 251
- 258
- José Antonio FRÍAS and Ana Belén RÍOS HILARIO  
Visibility and Invisibility of the Kindship Relationships in Bibliographic  
Families of the Library Catalogue 264

- 8. Integration of Knowledge in the Internet. Representing Knowledge in Web Sites**
- Houssein ASSADI and Thomas BEAUVISAGE  
A Comparative Study of Six French-Speaking Web Directories  
Barbara H. KWAŚNIK  
Commercial Web Sites and The Use of Classification Schemes:  
The Case of Amazon.Com 271
- Jorge SERRANO COBOS and Ana M<sup>a</sup> QUINTERO ORTA  
Design, Development and Management of an Information Recovery  
System for an Internet Website: from Documentary Theory to Practice  
José Luis HERRERA MORILLAS and M<sup>a</sup> del Rosario FERNÁNDEZ  
FALERO 279
- Information and Resources About Bibliographic Heritage on The Web  
Sites of the Spanish Universities 286
- J.F. ALDANA, A.C. GÓMEZ, N. MORENO, A. J. NEBRO, M.M.  
ROLDÁN  
Metadata Functionality for Semantic Web Integration  
Uta PRISS  
Alternatives to the "Semantic Web": Multi-Strategy Knowledge  
Representation 291
- 9. Models and Methods for Knowledge Integration in Information Systems**
- Rebecca GREEN, Carol A. BEAN and Michèle HUDON  
Universality And Basic Level Concepts  
Grant CAMPBELL 298
- Chronotope And Classification: How Space-Time Configurations Affect  
the Gathering of Industrial Statistical Data  
Marianne LYKKE NIELSEN and Anna GJERLUF ESLAU  
Corporate Thesauri - How to Ensure Integration of Knowledge and  
Reflections of Diversity  
Nancy WILLIAMSON 305
- Knowledge Integration and Classification Schemes  
M.V. HURTADO, L. GARCÍA and J.PARETS  
Semantic Views over Heterogeneous and Distributed Data Repositories:  
Integration of Information System Based on Ontologies  
Fernando ELCHIRIGOTTY and Cheryl KNOTT MALONE  
Representing the Global Economy: the North American Industry  
Classification System 311
- 10. Applications of Artificial Intelligence Techniques to Information Retrieval (Part I)**
- Christopher S.G. KHOO, Karen NG and Shiyuan OU  
An Exploratory Study of Human Clustering Of Web Pages 318

- Stéphane CHAUDIRON, Majid IHADJADENE and François ROLE  
Authorial Index Browsing in an Xml Digital Library 358
- Xavier POLANCO  
Clusters, Graphs, and Networks for Analyzing Internet-Web-Supported  
Communication within a Virtual Community 364
- E. HERRERA-VIEDMA, O. CORDÓN, J.C. HERRERA, M. LUQUE  
An IRS Based on Multi-Granular Linguistic Information  
Pedro CUESTA, Alma M. GÓMEZ and Francisco J. RODRÍGUEZ  
Using Agents for Information Retrieval 372
- 11. Integration of Knowledge in Multicultural Domain-Oriented and General Systems. (Part I)**
- Antonio GARCÍA JIMÉNEZ, Alberto DÍAZ ESTEBAN and Pablo  
GERVÁS  
Knowledge Organization in a Multilingual System for the  
Personalization of Digital News Services: How to Integrate Knowledge  
María J. LÓPEZ-HUERTAS and Mario BARITÉ  
Knowledge Representation and Organization of Gender Studies on the  
Internet: Towards Integration  
Victoria FRANCU 386
- Language-Independent Structures and Multilingual Information Access  
Annelise Mark PEJTERSEN and Hanne ALBRECHTSEN  
Models for Collaborative Integration of Knowledge 393
- 12. Applications of Artificial Intelligence Techniques to Information Retrieval (Part II)**
- C. LOPEZ-PUJALTE, V.P. GUERRERO, F. de MOYA-ANEÓN  
Evaluation of the Application of Genetic Algorithms to Relevance  
Feedback 404
- O. CORDÓN, E. HERRERA-VIEDMA, M. LUQUE, F. de MOYA,  
ANEÓN and C. ZARCO  
An Inductive Query by Example Technique for Extended Boolean  
Queries Based on Simulated Annealing-Programming  
Victor HERRERO-SOLANA and F. de MOYA-ANEÓN  
Graphical Table of Contents (GTOC) for Library Collections: the  
Application of UDC Codes for the Subject Maps 422
- Luis M. CAMPOS, Juan M. FERNÁNDEZ-LUNA and Juan HUETE  
Managing Documents with Bayesian Belief Networks: A Brief Survey  
of Applications and Models 429
- 437  
443

### 13. Epistemological Approaches to Classification Principles, Design and Construction

- Birger HJØRLAND  
The Methodology Of Constructing Classification Schemes: A discussion of the State-of-the-Art 450
- Hope OLSON, Juliet NIELSEN and Shona R. DIPP  
Encyclopaedist Rivalry, Classificatory Commonality, Illusory Universality 457
- Jian QIN  
Evolving Paradigms of Knowledge Representation and Organization: A Comparative Study of Classification, XML/DTD and Ontology 465
- Jens-Erik MAJ  
Is Classification Theory Possible? Rethinking Classification Research 472
- I.C. MCILWAINE  
Where Have All The Flowers Gone? An Investigation Into The Fate of Some Special Classification Schemes 479

### 14. Professional Ethics, Users and Information Structures. Evaluation of Systems

- J. Carlos FERNÁNDEZ-MOLINA and J. Augusto C. GUIMARAES  
Ethical Aspects of Knowledge Organization and Representation in the Digital Environment: Their Articulation in Professional Codes of Ethics 487
- Ali Asghar SHIRI, Crawford REVIE and Gobinda CHOWDHURY  
Assessing the Impact of User Interaction with Thesaural Knowledge Structures: A Quantitative Analysis Framework 493
- Carmen CARO CASTRO and Crispulo TRAVIESO RODRÍGUEZ  
Ariadne's Thread: Knowledge Structures for Browsing in OPAC's 500
- Linda BANWELL  
Developing and Evaluation Framework For a Supranational Digital Library 509
- Antonio L. GARCÍA GUTIÉRREZ  
Knowledge Organization From a "Culture of the Border": Towards a Transcultural Ethics of Mediation 516
- Christopher KING, David H. MARWICK and M. Howard WILLIAMS  
The Importance of Context in Resolving of Conflicts when Sharing User Profiles 523

### 15. Integration of Knowledge in Multicultural Domain-Oriented and General.(Part II)

- Richard P. SMIRAGLIA  
Crossing Cultural Boundaries: Perspectives on the Popularity of Works A. NEELAMEGHAN and Hemalata IYER 530
- Some Patterns of Information Presentation, Organization and Indexing for Communication Across Cultures and Faiths 539

- María Odaisa ESPINHEIRO DE OLIVEIRA  
Knowledge Representation from Amazonian Narratives 546
- Evelyn Goyanes Dill ORRICO  
Metaphorical Representations of the Thematic Identity of Social Groups in the Assistance of Information Retrieval 552

### 16. Applications of Artificial Intelligence Techniques to Information Retrieval (Part III)

- F. MARTÍNEZ, M.T. MARTÍN, V. M. RIVAS, M.C. DÍAZ and L.A. UREÑA  
Using Neural Networks for Multitword Recognition in IR 559
- E. PEIS, E. HERRERA-VIEDMA, J.C. HERRERA  
On the Evaluation of XML Documents Using Fuzzy Linguistic Techniques 565
- V.P. GUERRERO, C. LÓPEZ-PUJALTE, C. FABA, M.J. REYES, F. ZAPICO and F. de MOYA-ANEÓN  
Artificial Neural Networks Applied to Information Retrieval 572
- I. BLANCO, M.J. MARTÍN-BAUTISTA, D. SÁNCHEZ, A. VILA  
Fuzzy Logic for Measuring Information Retrieval Effectiveness 578

### Subject Index

### List of Contributors



F. Martínez, M.T. Martín, V. M. Rivas, M.C. Díaz and L.A. Ureña  
Department of Computer Science, University of Jaén, Spain

## Using Neural Networks for Multiword Recognition in IR

**Abstract:** In this paper, a supervised neural network has been used to classify pairs of terms as being multiwords or non-multiwords. Classification is based on the values yielded by different estimators, currently available in literature, used as inputs for the neural network. Lists of multiwords and non-multiwords have been built to train the net. Afterward, many other pairs of terms have been classified using the trained net. Results obtained in this classification have been used to perform information retrieval tasks. Experiments show that detecting multiwords results in better performance of the IR methods.

### 1. Introduction

In this paper we present a new approach to solve the task of effectively detecting multiwords. A multiword is a succession of words whose sense taken as a whole differs from the sum of the senses of its single words. Thus, a multiword can be considered in fact as a new concept.

There exists a second kind of multiword composed of a set of words that complement their senses. Nevertheless, considering this kind of successions as multiwords does not provide information useful for information retrieval tasks. For this reason, this kind of multiwords have not been considered in the present work.

Multiword detection can be successfully used in many different tasks. Information Retrieval (IR) methods, for instance, use the word as the basic unit of information; thus, detecting multiwords in corpus and queries make IR systems get better results. Cross Language Information Retrieval can be also improved by detecting multiwords, given that translating word by word results in lost of information. Natural Language Processing can also be helped when multiwords are correctly detected, because it makes the text easier understanding.

The new approach presented in this paper uses neural nets to discriminate pairs of terms that really are multiwords from those that are not. We propose a well-known supervised neural network: Kohonen's Learning Vector Quantization (LVQ), widely used for classification tasks (Kohonen, 1995, 1992). Inputs for the nets are the values yielded by estimators used in literature to perform this same task, and the output the nets provide is a class determining if that values corresponds to a multiword or a non-multiword. Nets learning is performed by training them with the values yielded by the cited estimators when they are applied to pairs of terms known to be either multiwords or non-multiwords. In order to test this new method, the network has been used to classify new pairs of terms; then, the information obtained has been used to perform some IR tasks. Experiments show that results obtained in these task are better than those obtained using only the estimators.

The rest of the paper is organized as follows: Section 2 gives a little introduction to the state of the art, briefly showing some of the currently available methods used to detect multiwords. These methods include the different estimators

that will be lately used in our method. Section 3 describes a new estimator developed for this work as well as the neural network that has been used. Section 4 shows the experiments carried out and the results obtained. Finally, Section 5 outlines some conclusions, and also future research lines.

## 2. A new approach

Multword recognition has been explored by many researchers as a way to improve traditional Text Retrieval, in general with a moderate degree of success. However, David Hull and Gregory Grefenstette (Hull, 1996) show that multword detection and correct translation largely improve the precision in a CLIR system.

Usually methods for automatizing terminological procedures have traditionally been statistical (Hull, 1996; Ballesteros, 1998), and based on the occurrence of each particular pair of words in the text of work or corpus. Other works (Adriani, 1999) obtain the degree of similarity between terms using the co-occurrence factor, and the standard  $f^2$  term weighting formula. Recently, hybrid approaches incorporating linguistic information have been developed: Diana Maynard and Sophia Ananiadou (Maynard, 2000) make use of different types of contextual information: syntactic, semantic, terminological and statistical. Nevertheless, managing different types of information must be done by integrating them in any given way. The most straightforward way is by using a linear function, although this does not mean it is the best way this problem can be faced.

For any of the features (syntactical, semantical, terminological and statistical) to be integrated to perform multword detection, there are some well-known estimators. This paper introduces a neural network based approach that integrates terminological and statistical estimators. Multword detection is then thought as a categorization problem where only two categories have to be managed: multword and non-multword. Consequently, classifying a pair of terms turns into a two step process: firstly, obtain the values yielded by the different estimators; secondly, use those values as inputs for the neural network, and obtain the class to which the pair of terms belongs. More precisely, the estimators that have been used in this work are the following:

1. *Pearson's  $\chi^2$* . A variant of the  $\chi^2$  statistic (Hull, 1996)
2. Measure the importance of co-occurrence of the elements in a set by the *em* metric (Ballesteros, 1998)
3. *Dice similarity coefficient* obtain the degree of similarity or association-relation between terms using a term association measure and the t.f.idf weighting formula (Adriani, 1999).
4. The *mutual information ratio*, or association ratio,  $\mu$  (Johansson, 1996).
5. Finally, a new estimator, a variant of Dice similarity coefficient based on the Simpson index, has been developed.

### 2.1. A new estimator: Simpson Similarity coefficient

Roughly, Dice index is based on the association between two terms by calculating the coefficient of the intersection of two sets and their union. Usually, this approach is convenient to estimate the correlation between words, but not always. "Bill Clinton", for instance, is a multword, but "Bill" is a very common word, so the term frequency is very high, and "Bill" set is huge. In the other hand,

"Clinton" is not too frequent, so "Clinton" set is small. Thus, the coefficient of the intersection and the union of both sets will be small, because "Clinton" set is small. In another way, the Simpson index estimates the association between two sets by calculating the coefficient of the intersection of two sets and the *smaller* of them, so "Bill Clinton" will reach a high value for the Simpson coefficient, and a low value for the Dice coefficient.

$$DICE: xy = 2 \frac{\sum_{i=1}^n (w'_x \cdot w'_y)}{\sum_{i=1}^n w_x^2 + \sum_{i=1}^n w_y^2} \quad SIMPSON: xy = 2 \frac{\sum_{i=1}^n (w'_x \cdot w'_y)}{\min \left( \sum_{i=1}^n w_x^2, \sum_{i=1}^n w_y^2 \right)}$$

where:

$w_x$  = the weight of term  $x$  in the document  $i$ .

$w_y$  = the weight of term  $y$  in document  $i$ .

$w'_x$  =  $w_x$  if term  $x$  also occurs in document  $i$ , or 0 otherwise.

$w'_y$  =  $w_y$  if term  $y$  also occurs in document  $i$ , or 0 otherwise.

$n$  = the number of documents in the collection.

## 2.2. Neural Network approach: The LVQ algorithm

The LVQ algorithm is a classification method based on neural competitive learning, which allows to define a group of categories on the space of input data by a reinforced learning, either positive (prize) or negative (punishment). LVQ uses supervised learning to define class regions in the input data space. To this end a subset of similarly labeled codebook vectors is placed into each class region.

Given a sequence of input data, an initial group of reference vectors  $w_k$  (codebook) is selected. In each iteration, a input vector  $x_i$  is selected and the vectors  $W$  are updated, so that they fit  $x_i$  in a better way. The LVQ algorithm works as follows:

For each class,  $k$ , a weight vector  $w_k$  is associated. In each repetition, the algorithm selects an input vector,  $x_i$ , and compares it with every weight vector,  $w_k$ , using the euclidean distance  $\|x_i - w_k\|$ , so that the winner will be the weight vector  $w_c$  nearest to  $x_i$ , being  $c$  its index:

$$\|x_i - w_j\| = \min_k \|x_i - w_k\|$$

The classes compete between them in order to find the most similar to the input vector, so that the winner is the one with less euclidean distance regarding the input vector. Only the winner class will modify its weights using a reinforced learning algorithm, either positive or negative, depending on the classification being correct or not. Thus, if the winner class belongs to the same class of the input vector (the classification has been correct), it will increase the weights, coming slightly close to the input vector (prize). To the contrary, if the winner class is different from the input vector class (the classification has not been correct), it will decrease the weights, coming slightly far from the input vector (punishment).

Let  $x_k(t)$  be an input vector at time  $t$ , and  $w_k(t)$  represents the weight vector for the class  $k$  at time  $t$ . The following equation defines the basic learning process for the LVQ algorithm.

$$w_c(t+1) = w_c(t) + s \cdot \alpha(t) \cdot [x_i(t) - w_c(t)]$$

where  $s = 0$ , if  $k \leq c$ ;  $s = 1$ , if  $x_i(t)$  and  $w_c(t)$  belong to the same class; and  $s = 1$ , if they do not, and where  $\alpha(t)$  is the learning rate, being  $0 < \alpha(t) < 1$ , a monotonically decreasing function of time. It is recommended that  $\alpha(t)$  be rather small initially, say, smaller than 0.5, and that it decrease to a given threshold,  $u$ , very close to 0 (Kohonen, 1995).

The experiments showed in section 4 were carried out using the implementation described in LVQ\_PAK documentation (Kohonen, 1991) with default parameters. Thus, every experiment started with the same number of codebooks per class (10 for class 0 and 10 for class 1) and the learning rate being initialized to 0.3.

### 3. Experiments and results

In order to train and test the neural nets, a set of samples composed of input-output pairs had to be built, every sample corresponding to a pair of terms. In one hand, input values were obtained by applying the different estimators described in section 3. In the other, every output value consisted on a single number classifying the sample as multiword or non-multiword. In our experiment only multiwords with two relevant terms have been used, and stop words have been removed from the multiword.

Obtaining a list of multiwords was done by resorting to WordNet (Miller, 1995), a lexical database where multiwords can be found. Nevertheless, not all the pairs of terms said to be multiwords really were. For this reason, each multiword returned by WordNet was newly searched in the electronic dictionary Encarta<sup>1</sup> to remove pairs of terms that, even appearing together very frequently, were not real multiwords.

Non-multiwords list (needed to train the nets) was taken from the corpus used in CLEF 2000<sup>2</sup>. Pairs of terms were taken from this corpus and then searched in the list of multiwords previously described, checking that they did not appear in it. If they did not appear, they were once more searched in the electronic dictionary to assure they did not formed a multiword.

Once both multiwords and non-multiwords lists have been created, the above cited estimators were applied to them, obtaining the file with the samples to be used with the supervised network. This file was split to use 75% of the samples to train the neural network and the remaining 25% to validate it.

Los Angeles Time 1994 collection, borrowed from the English CLEF 2000 collection, was used to test the method. This collection is composed of 113,005 articles of the 1994 edition of Los Angeles Times, and 40 queries (Title + Description) with relevance judgments. The collection was indexed twice using Zprise software<sup>3</sup>, with Okapi (Robertson, 2000) weighting formula. First index was created without carrying out multiword detection, while second index uses multiword detection, as depicted above.

Table 1 shows average precision reached by both methods. It shows that multiword usage improves

Original query set	Query set with multiwords detection
0.375	0.410

Table 1 - Average Precision

precision scarcely. Anyway, a more detailed analysis of the results leads to conclude that multiword detection is useful for IR task. Table 2 shows the precision reached by some queries, and the detected multiwords for each one.

Query	Original AvgP.	AvgP. With Multiwords	Detected multiwords
#7	0.3969	0.4452	"world soccer"
#9	0.1022	0.2027	"war ii" "ii war" "war rwanda" "world war"
#3	0.3912	0.3220	"decisions made", "hard soft"
#32	0.4126	0.2511	"women priest", "change direction"

Table 2 - Four detailed queries.

As Table 2 shows, query #7 gained 5% of absolute precision because "world soccer" was effectively detected as a multiword. Results were even better in query #9, in which "world war", "war rwanda" "war ii" and "ii war" multiwords were correctly detected. As can be seen, precision obtained in this query by the new method is twice the precision obtained without multiword detection.

On the other hand, query #3 lost 7% of precision with multiword inclusion. Bigrams "decisions made" and "hard soft" are in fact non-multiwords, but the neural network method marked both of them as being such. Finally, query #32 lost 16% of precision: "women priest" and "change direction" again, are not multiwords.

### 4. Conclusions and future work.

This paper presents a new method to detect multiwords. This method uses the values obtained by estimators, present in literature and developed to perform this same task, as inputs for a neural net that automatically determines whether those values belong to a real multiword or simply to a pair of terms that appear together in a document.

Results show that automatic multiword detection is useful for IR. Nevertheless, the method used must get a higher accuracy, because poor detection of multiwords damages precision of the IR system. Conservative methods must be used to assess multiwords. Classifying multiwords as non-multiwords is better than recognizing too many multiwords. In other words, multiword detection must improve precision over recall.

Future lines of research include the use of new kind of neural networks, such as Radial Basis Function Nets (Broomhead, 1988; Rivas, 2001), as well as RCE (Zhorit, 2000), and also unsupervised training networks as Self-Organization Maps (Kohonen, 1995).

New estimators based on semantic information can be used to improve the results. Others applications for this method must also be investigated, especially its influence in Cross Language Information Retrieval.



## 2. Notes

1. Encarta is available at <http://www.encarta.com> [2/2/2002]. Encarta has been used because it includes proper nouns that are considered to be multiwords.
2. Cross Language Evaluation Forum (CLEF) aims at promoting research and development in CLIR. For more information, see: <http://www.clef-campaign.org>
3. ZPrise is a software developed by NIST. It is available at <http://www.itl.nist.gov/iaui/894.02/works/papers/zp2/zp2.html> [2/2/2002]

## References

- Adriani, M. and C.J. van Rijsbergen (1999). Term Similarity Based Query Expansion for Cross Language Information Retrieval. In *Proceedings of Research and Advanced Technology for Digital Libraries*, Third European Conference (ECDL'99), p. 311-322. Springer Verlag: Paris, September 1999.
- Ballesteros, L. and Croft, W.B (1998). Resolving ambiguity for cross-language retrieval. In Croft, W.B., Moffat, A., van Rijsbergen, C.J., Wilkinson, R. and Zobel, J. eds. *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY: ACM, 1998, 64--71.
- Broomhead, D.S., Lowe, D (1988). Multivariable Functional Interpolation and Adaptive Networks. In *Complex Systems*, vol. 11, pp.321-355.. 1988
- Hull, David A., Grefenstette, Gregory (1996). Experiments in Multilingual Information Retrieval. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1996
- Johansson, Christer (1996). Good Bigrams. In *Proceedings from the 16th International Conference on Computational Linguistics (COLING-96)*. Copenhagen: 592-597. 1996
- Kohonen, T., Hynninen, J., Kangas, J., Laaksonen, J., Torkkola, K. (1991). *LVQ\_PAK: The Learning Vector Quantization program package*. Helsinki University of Technology Laboratory of Computer and Information Science. Finland, 1991-1995.
- Kohonen, T., Kangas, J., Laaksonen, J., Torkkola K. (1992). LVQ\_PAK: A program package for the correct application of Learning Vector Quantization algorithms. In *Proceedings. of the International Joint Conference on Neural Networks*, Baltimore, June 1992. IEEE, I, 725-730
- Kohonen, T. (1995). *Self-Organization and Associative Memory*. 2nd Ed. Springer-Verlag, Berlin, 1995.
- Maynard, Diana and Ananiadou, Sophia (2000). TRUCKS: a model for automatic term recognition. *Journal of Natural Language Processing*, December 2000.
- Miller, G. (1995). *WORDNET: A lexical database for English*. *Communications of the ACM*, 38 (11), 1995.
- Salton, Gerard, and McGill, Michael J. (1983) *Introduction to Modern Information Retrieval*. New York: McGraw-Hill, 1983.
- Rivas, V.M., Merelo, J.J., P.Castillo, P.A. (2001). Evolving RBF Neural Networks. *Lecture Notes in Computer Science*, vol. 2064, pp.506-513. 2001
- Robertson, S. E., Walker, S. & Beaulieu, M (2000). Experimentation as a way of life: Okapi at TREC. *Information Processing & Management*, 36(1), 95-108. 2000
- Zboril, F. Zboril, F. (2000) The use of the RCE network in a Pattern Recognition. *Proceedings of MOSES 2000*, pp. 65-70. 2000