Introduction to the Thematic Issue on Language & Knowledge Engineering

This thematic issue of *Computación y Sistemas* presents a selection of papers pointing out the latest advances carried on into the field of Language & Knowledge Engineering and their applications. with topics covering natural language processing, computational linguistics, knowledge engineering, pattern recognition, artificial intelligence and other paper covering computer science in general.

Language engineering is an area of artificial intelligence and applications aiming to bridge the gap between traditional computational linguistics research and the implementation of potentially real-world applications. It looks to meet the needs of the research community working in all areas of automatic language processing, whether from a theoretical or applied perspective including some tasks such as machine translation, word sense disambiguation, reputation analysis, etc. As we will further describe, this thematic issue contains ten papers associated to the natural language engineering area, presenting specific natural language processing methods, tasks or applications.

Knowledge engineering, on the other hand, refers to all technical, scientific and social aspects involved in designing, building, maintaining and using knowledge-based systems. The aim is to support human decision-making, learning and action, with emphases the practical significance, computer development and usage of knowledge-based systems including design process, models and methods, software tools, decision-support mechanisms, user interactions, organizational issues, knowledge acquisition and representation, and system architectures.

As we will further describe, this thematic issue contains nine papers associated to the knowledge engineering area. The remaining six papers can be classified in the topic of artificial intelligence and computer science in general. In summary, this thematic issue includes twenty-five papers, representative of different tasks, techniques, and applications of language and knowledge engineering.

Thus, we start this thematic issue with eleven papers, which are devoted to the language engineering area. Their general description follows.

J. Rojas Simón et al. from Mexico in their paper "Calculating the upper bounds for Multi-Document Summarization using Genetic Algorithms" describe a Genetic Algorithm-based method for calculating the best sentence combinations of DUC01 and DUC02 datasets in Multi-Document Summarization through a Meta-document representation. Moreover, they have calculated three heuristics mentioned in several works of state-of-the-art to rank the most recent Multi-Document Summarization methods, through the calculus of upper bounds and lower bounds.

B. Priego Sánchez et al. from Mexico in their paper "Idiom Polarity Identification using Contextual Information" present experiments towards the evaluation of different polarity lexicons for the particular task of automatic identification of polarity for linguistic units known as idioms based on their contextual information. The lexicons employed were: ElhPolar dictionary, iSOL and ML-SentiCON Sentiment Spanish Lexicon, all of them containing the polarity of different words. The results reported by the authors shown that the best combination obtained results close to 57.31% when the texts were lemmatized and 48.87% when they were not lemmatized.

A. Quirin et al. from France in their paper "Analyzing polemics evolution from Twitter streams using author-based social networks" carry out experiment towards the analysis of the evolution of very short and unpredictable events, called polemics. They conducted a time-related social network analysis. Firstly, dedicated to the study of the evolution of actor interactions, using time-series built from a total of 33 graph theory metrics. Secondly, they have validated the employed techniques on a complex dataset of 284 millions of tweets, analyzing 56 days of the Volkswagen scandal.

Computación y Sistemas, Vol. 22, No. 1, 2018, pp. 5–9
ISSN 1405-5546
doi: 10.13053/CyS-22-1-2924

H. Gómez-Adorno et al. from Mexico in their paper "Stylometry-based Approach for Detecting Writing Style Changes in Literary Texts" present an approach to identify changes in the writing style of 7 authors of novels written in English. They defined 3 stages of writing for each author, each stage containing 3 novels with a maximum of 3 years between each publication. They proposed several stylometric features to represent the novels in a vector space model, and finally they employed supervised learning algorithms for determining whether or not by means of this stylometric-based representation is possible to identify to which stage of writing each novel belongs.

K. Asnani et al. from India in their paper "Extraction of Code-mixed Aspect Topics in Semantic Representation" propose a knowledge based language independent code-mixed semantic LDA (lcms-LDA) model, with an aim to improve the coherence of clusters. They have found that the proposed lcms-LDA model infers topic distributions without language barrier, based on semantics associated with words. The experimental results showed an increase in the UMass and KL divergence score indicating an improved performance in the resulting coherence and distinctiveness of aspect clusters in comparison with the state-of-the-art techniques used for aspect extraction of code-mixed data.

P. V. Veena et al. from India in their paper "Character Embedding for Language Identification in Hindi-English Code-mixed" propose a technique for identifying the language of Hindi-English code-mixed data used in three social media platforms namely, Facebook, Twitter, and WhatsApp. The classification of Hindi-English code-mixed data into Hindi, English, Named Entity, Acronym, Universal, Mixed (Hindi along with English) and Undefined tags were performed. Popular word embedding features were used for the representation of each word. Two kinds of embedding features were considered - word-based embedding features and character-based context features. The proposed method was done with the addition of context information along with the embedding features. A well-known machine learning classifier, Support Vector Machine was used to train and test the system. The work on Language Identification in code-mixed text using character-based embedding is a novel approach and showed promising results.

S. Mansouri et al. from Tunisia in their paper "A Heuristic Approach to Detect and Localize Text on Arabic News Video" attempted to solve the problem of automatic text detection in video sequences by proposing a robust detection–validation schema for text localization in Arabic news video. Candidate text regions were first detected by using a hybrid method, which combines MSER detector and edge information. Then, these regions were grouped using morphological operators. Finally, a verification process was applied to remove noisy non-text regions including specific features for Arabic text. Performance and efficacy of the proposed text detection approach have been tested by using Arabic-Text-in-Video database (AcTiV-DB).

R. Piryani et al. from India in their paper "Generating Aspect-based Extractive Opinion Summary: Drawing Inferences from Social Media Texts" presents an integrated framework to generate extractive aspect-based opinion summary from a large volume of free-form text reviews. The framework has three major components: (a) aspect identifier to determine the aspects in a given domain; (b) sentiment polarity detector for computing the sentiment polarity of opinion about an aspect; and (c) summary generator to generate opinion summary. The framework was evaluated on SemEval-2014 dataset obtaining better results than several other approaches.

L. A. Penichet et al. from Cuba in their paper "New Similarity Function for Scientific Articles Clustering Based on the Bibliographic References" implemented a new similarity function for clustering of scientific articles based on the information provided by the references of the articles. They claimed that the use of the implemented function contributes significantly to discover relevant knowledge from scientific literature.

A. Gelbukh from Mexico in the paper "Inferences for Enrichment of Collocation Databases by means of Semantic Relations" describes an interesting lexical resource – a very large database of collocations compiled manually, but with additional inference

mechanisms, which allow construction of possible collocations that never occurred in texts.

J. Lavalle et al. from Mexico in their paper "Automatic Theorem Proving for Natural Logic: A Case Study on Textual Entailment" present an Automatic Theorem Prover for Natural Logic that allows to know precisely the relationships needed in order to reach the entailment in a class of natural language expressions. This paper approaches the recognition of Textual Entailment as a Natural Language Processing task, which is in fact very important in other tasks such as Semantic Search and Text Summarization. The most interesting attractive of this paper is that it tackles the problem of explaining why the entailment is carried out, because when Natural Logic is used, it is possible to do reasoning from the syntactic part of a natural language expression, and using very little semantic information.

The next ten papers of the thematic issue are devoted to the knowledge engineering area. Their general description follows.

M. Somodevilla García et al. from Mexico in their paper "An Overview on Ontology Learning Tasks" present a general review of work related to types and tasks involving Ontology Learning for the Semantic Web. The research works reviewed consider fundamental types of Ontology Learning, schema extraction, creation and population,

M. Tovar et al. from Mexico in their paper "A metric for the evaluation of restricted domain ontologies" propose a metric for the automatic evaluation of restricted domain ontologies. The metric is defined in terms of the evaluation of different lexico-syntactic, statistical and semantic approaches. A syntactic approach employed is the use of lexical syntactic patterns, other approaches as grouping by formal concept analysis, similarity, latent semantic analysis and dependence graphs are used as well. These approaches focus on reference corpora to find evidence of the validity of concepts and semantic relationships stored in the target ontology. The proposed evaluation approach is able to provide a score obtained through the metric, which is based on the accuracy measure used for each ontology evaluated. The score is associated in some way with the ontology quality. This score is given with a certain degree of reliability, and it is obtained by comparing the results given against the evaluation of human experts and a baseline.

M. Anzures-García et al. from Mexico in their paper "A Workflow Ontology to support Knowledge Management in a Groups organizational structure" propose a workflow ontology to control an organizational. The workflow manages and controls the process, via a set of steps ordered and executed by different organization entities, whereas the ontology specifies the domain of knowledge through concepts, relations, axioms, and instances in a formal, explicit, way. A case of study, to demonstrate the knowledge management of the group's organizational structure, through workflow ontology is also shown in this paper.

E. Granillo Martínez et al. from Mexico in their paper "A neighborhood combining approach in GRASP's local search for Quadratic Assignment Problem solutions" describe a study for the search of solutions of the combinatorial optimization problem named Quadratic Assignment Problem (QAP) through the implementation of a Greedy Randomized Adaptive Procedure Search (GRASP). They compared their results with the best solutions known in literature, obtaining robust results in terms of the value of the objective function and the execution time. A comparison with the ant algorithm was also presented. The most important contribution of this paper is the use of the combination of different neighborhood structures in the GRASP improvement phase. The experiment was performed for a set of test instances available in QAPLIB.

Y. Limón et al. from Mexico in their paper "Depth-First Reasoning on Trees" propose a satisfiability algorithm for the μ-calculus extended with converse modalities and interpreted on unranked trees. In contrast with known satisfiability algorithms, the proposal given in this paper is based on a depth-first search. Authors proved the algorithm to be correct (sound and complete) and optimal. They also described an implementation. The extension of the μ-calculus with converse modalities allowed to efficiently characterize standard reasoning problems (emptiness, containment and equivalence) of XPath queries. Finally, they described several

query reasoning experiments, which shows their proposal to be competitive in practice with known implementations.

Y. Pacheco et al. from Cuba in their paper "Application of multi-criteria decision analysis to the selection of software measures" propose the application of a multi-criteria decision analysis to make documented and transparent decisions about software measures' selection. The Pareto's dominance method was utilized to narrow down the initial measures' list. The multi-attribute value theory was applied for ranking the final set of measures. As a result there was eliminated about 40% of the initial measures and the final measures' list was ranked.

A. Morfa Hernández et al. from Cuba in their paper "Integration of Visualization Techniques to Algorithms of Optimization of the Metaheuristics Ant Colony" propose a model of integration of visualization techniques in algorithms that use the metaheuristic Ant Colony (ACO) that allows the user to interact with real-time search and guide her. A software tool was implemented to solve Traveling Salesman Problem (TSP) with ACO algorithm according to the proposed model. An experimental analysis with the developed tool was performed and the results showed the efficiency of the model.

G. Molero-Castillo et al. from Mexico in their paper "Interactive system for the analysis of academic achievement at the upper-middle education in Mexico" analyze and process data from different sources through user-centered data mining, based on the fundamentals of the usability engineering and accessibility. The academic achievement, at Language and Communication and Mathematics, of students at the upper-middle education in Mexico was analyzed through a partitional clustering algorithm. A variety of academic achievements were observed, highlighting Insufficient and Elementary in the evaluated population, while Good and Excellent achievements were achieved by a reduced number of schools.

J. Pérez et al. from Mexico in their paper "A Storage Pattern-Based Heuristic Algorithm for Solving Instances of Hard28 Datasets for the Bin Packing Problem" propose a heuristic algorithm that solves some instances of the data set Hard28, applied to the problem of packaging objects in containers of one dimension. The algorithm is based on storage patterns of objects in containers. In order to detect how objects are stored in containers, the HGGA-BP algorithm was used. This algorithm has optimally solved the largest number of instances of the data set Hard28.

J. A. Hernández-Castaño et al. from Mexico in their paper "Experimental platform for Intelligent Computing" present the architecture and user interface of a novel Experimental Platform for Intelligent Computing (EPIC). Authors claimed that unlike the two most popular platforms (WEKA and KEEL), the proposed EPIC tool has a very friendly user interface, and offers some advantages with respect to existing tools for Intelligent Computing experiments. They emphasized that EPIC handles mixed and incomplete data directly, without preprocessing, and its architecture supports multi-target supervised classification and regression. Authors said that it also contains a module for two dimensional dataset visualization, which includes the visualization of the decision frontier for several supervised learning algorithms.

The following three papers can be categorized in the artificial intelligence area, in particular, in the human computer interaction topic.

J. M. González Calleros et al. from Mexico in their paper "Is Natural User Interaction Really Natural? An evaluation of gesture-based navigating techniques in Virtual Environments" analyze the use of gesture-based interaction for the navigation of virtual worlds. In their paper they presented a virtual world and contrasted the use of interactive interfaces based on gesture of hands or body, as well as interaction based on mouse and keyboard. The results of this paper indicate that the natural is not as it is even though we imitate what we do in real life.

G. Saldaña González et al. from Mexico in their paper "Recognition and Classification of Sign Language for Spanish" present a computational system for recognition and classification of letters of the sign language in Spanish, designed for helping deaf-mute people to communicate with other persons. A low- cost glove that captures the hand movements was also constructed, which it is said that contains an accelerometer for each finger allowing to detect

its position by using an acquisition data board. Sensor information is sent wirelessly to a computer having a software interface, developed in LabVIEW, in which the symbols dataset is generated. For the automatic recognition of letters they have applied a statistical treatment to the dataset obtaining accuracy greater than 96%.

R. O. Klenzi et al. from Argentina in their paper "Visualization in a Data Mining Environment from a Human Computer Interaction Perspective" analyze the visualization issue from a Human Computer Interaction setting with the aim of providing support to the design, analysis and evaluation of result visualization mechanisms used to supply information in Data Mining Environments. Three practical examples based on numerical, textual and georeferenced data are described by means of the KNIME Analytics tool. In addition, the use and importance of graphs are emphasized for a correct information interpretation.

Finally, the following two papers present approaches associated with the image processing topic.

J. Quintanilla-Domínguez et al. from Mexico in their paper "Microcalcifications detection using image processing" present an effective methodology to detect microcalcifications in digitized mammograms is presented. This methodology is based on the synergy of image processing, pattern recognition and artificial intelligence. The methodology consists in four stages: image selection, image enhancement and feature extraction based on mathematical morphology operations applying coordinate logic filters, image segmentation based on partitional clustering methods such as k-means and self organizing maps and finally a classifier such as an artificial metaplasticity multilayer perceptron. The proposed system constitutes a promising approach for the detection of Microcalcifications. The experimental results show that the proposed methodology can locate Microcalcifications in an efficient way. The best values obtained in the experimental results are: accuracy 99.93% and specificity 99.95%, These results are very competitive with those reported in the state of the art.

G. Saldaña González et al. from Mexico in their paper "Vision System for the Navigation of a Mobile Robot" present the development of an object detection system in a controlled two-dimensional space using computer vision techniques. The detected objects have a rigid geometry and are exposed to real light; therefore, the system is robust to changes in lighting and shading. In order to handle the large amount of data to be processed in real time, a MyRIO device which contains an FPGA was used. This device allowed communication with the LabVIEW software where the user interface resides. Using LabVIEW a tracking by color algorithm was implemented, in order to attend to a reactive agent, which uses an infrared sensor to detect the distance to an obstacle and perform the functions of foraging and storage. In order to improve performance, a supervisory system was implemented using a Kinect device that provides information relative to the position of the objects in the test area. This information allowed eliminating occlusion problems.

David Pinto
Darnes Vilariño
Beatriz Beltrán

BUAP, Mexico

Guest Editors

Computación y Sistemas, Vol. 22, No. 1, 2018, pp. 5–9
ISSN 1405-5546
doi: 10.13053/CyS-22-1-2924