ISSN 2007-9737

Introduction to the Thematic Issue on Language & Knowledge Engineering

This thematic issue of *Computación y Sistemas (CyS)* contains a selection of papers presenting advances into the field of Language & Knowledge Engineering and their applications. with topics covering, among others, machine learning methods for classification of unstructured data, information retrieval, ontologies, semi-supervised methods, image analysis, sentiment analysis, text mining, language engineering, named entity recognition, knowledge engineering, video analysis, graph based analysis for text analytics, text clustering, opinion mining, recommendation systems and computational linguistics.

"Language engineering" is an emergent research area of artificial intelligence and applications aiming to bridge the gap between traditional computational linguistics research and the implementation of potentially real-world applications. Language engineering attempts to satisfy the research community needs in all areas of automatic natural language processing, from a theoretical and applied perspective. As we will further describe, this thematic issue contains several papers associated to the natural language engineering area.

On the other hand, "knowledge engineering", on the other hand, refers to technical, scientific and social aspects involved in the design, construction, maintenance and use of knowledge-based systems. The aim of knowledge engineering is to support human decision making, learning and action, emphasizing practical appliance, computer development and usage of design processes, models and methods, software tools, decision-support mechanisms, user interactions, organizational issues, and acquisition and representation of knowledge. As we will further describe, this thematic issue contains several papers associated to the knowledge engineering area. The remaining papers can be classified in the topic of artificial intelligence and computer science in general.

In summary, this thematic issue includes thirty-one papers of 111 submitted (28% of acceptance rate), representative of different tasks, techniques, and applications of the Language and Knowledge Engineering (LKE) area.

In the following paragraphs, we provide an overview of the papers that made up this volume. Their general description follows.

E. Montelongo et al. from Mexico in their paper entitled "Machine Learning Models for Cancer Type Classification with Unstructured Data" present an approach to obtain information from clinical notes, based on Natural Language Processing techniques and Paragraph Vectors algorithm. A comparison and evaluation process of chosen machine learning models with varying parameters were conducted to obtain the best one. Results obtained are promising and they show the best model for classification is the MLP model with a precision 0.89 and f1-score 0.87.

S. Goyal et al. from India in their paper "Information Retrieval from Software Bug Ontology Exploiting Formal Concept Analysis" presents an ontology-based retrieval approach that visualizes and structure the data of software bug reports domain. It exploits formal concept analysis (FCA) to elicit conceptualizations from bug reports datasets and a hierarchical taxonomy is generated of extracted knowledge. A lattice diagram of concepts and relationships is constructed from concept-relationship matrix created by FCA. The proposed approach is evaluated on 21 bug reports of apache projects of the JIRA repository.

R. Gallardo et al. from Mexico in their paper "Comparison of Clustering Algorithms in Text Clustering Tasks" compare the performance and accuracy of several clustering algorithms in text clustering tasks. The clustering tasks were carried out by employing the PAN dataset and three different algorithms: Affinity Propagation, K-Means and Spectral Clustering. A comparative table with precision, recall and f-measure scores is presented.

J. Valdez et al. from Mexico in their paper entitled "Single-Stage Refinement CNN for Depth Estimation in Monocular Images" propose five different models for solving this task, ranging from

a simple convolutional network, to one with residual, convolutional, refinement and upsampling layers. They compare the presented models with the current state of the art in depth reconstruction and measure depth reconstruction error for different datasets (KITTI, NYU), obtaining improvements in both global and local error measures.

K. Jobczyk et al. from Poland in their paper entitled "A Fuzzy Multi-Agent Problem: A General Depiction and its Logic Programming-based Application" a Fuzzy Multi-Agent Problem (FMAP), as referred to Constraints Satisfaction Problems (CSP), and its PROLOG-based solutions are considered. An effectiveness of this PROLOG-based approach exploits a multi-valency-based approximation of fuzziness in programming contexts.

J. Navarro et al. from Mexico in their paper "Fast Convergence of the Two Dimensional Discrete Shearlet Transform" carried out experiment in order to illustrate the continuity of f at (0,0) by means of the convergence of the discrete Shearlet transform as the dilation parameter converges to zero. This property is studied by detecting edges in images that correspond to discontinuities.

J. Gómez et al. from Mexico in their paper entitled "OPAIEH: An Ontology-based Platform for Activity Identification of the Elderly at Home" focused on the activity recognition of the elderly who live independently at home. Authors presented OPAIEH, an ontology-based platform for activity identification of the elderly at home. The platform includes the ontological model, a sensors network, a client device, and web server to perform the recognition of different activities that the elderly do inside their home. Furthermore, the platform generates a set of graphs that shows different statistics and user behaviors.

P. Sendash et al. from India in their paper "Local Binary Ensemble based Self-training for Semi-supervised Classification of Hyperspectral Remote Sensing Images" a novel efficient self-training approach for handling the deficiency of labelled samples for semi-supervised classification of hyperspectral remote sensing images is proposed. Experimental results on two benchmark hyperspectral image datasets prove the effectiveness of the proposed method over supervised and traditional self-training based semi-supervised pixelwise classification approach in terms of different classification measures.

M. Contreras et al. from Mexico in their paper entitled "Knowledge Representation in TOEFL Expository Texts" characterize expository passages semantic structures through FOL and situation calculus with the question-answer block. The aim is to look for semantic patterns that allow the creation of intelligent tools to train students in the reading comprehension of English language.

E. Shushkevich et al. from Ireland in their paper "Offensive Language Recognition in Social Media" propose an approach for solving the problem of multiclassification within the framework of aggressive language recognition in Twitter. The model created is an ensemble of classical machine learning models included Logistic Regression, Support Vector Machines, Naive Bayes models and a combination of Logistic Regression and Naive Bayes. The obtained value of macro F1-score for one of the experiments achieved is 0.61.

O. Ramos et al. from Mexico in their paper "Proposal for NERC and the Automatic Generation of Rules on Mexican News" introduce a proposal for extracting facts from news on Mexican online newspapers through their RSS (Really Simple Syndication). The problem is addressed by using the task of automatic named entities recognition and classification (NERC), as well as the semantic relation extraction among entities, so that they can build a database of facts and rules from the obtained entities in an automatic manner. The final aim is to be able to infer new rules through the use of the knowledge databases constructed and an inference engine.

B. Pedroza et al. from Mexico in their paper entitled "Fuzzy Models for Implement of the Decision-Making Module in Networked Didactic Prototypes" present the results of the combination of fuzzy models (inference models and Fuzzy Cognitive Maps - FCM) to identify cognitive skills and types of problems that help the student reach the appropriate levels in the domain of algebra topics and the differential calculus, are presented. The objective is to implement fuzzy models in electronic devices based on tangible interfaces.

D. Muñoz et al. from Mexico in their paper "Named Entity Recognition based on a Graph Structure" propose a graph structure for storage and enrichment of named entities. It makes use of synonyms and domain-specific ontologies in the area of computing. The performance of the proposed structure is measured and compared with other NER classifiers in the experiments carried out.

M. Etcheverry et al. from Uruguay in their paper entitled "Order Embeddings for Supervised Hypernymy Detection" present a supervised approach to partially order word embeddings, through a learned order embedding, applying it in supervised hypernymy detection. They use neural network as an order embedding to map general purpose word embeddings to a partially ordered vector set. They show that this distributional approach presents interesting results in comparison to other distributional and path-based approaches.

B. Beltrán et al. from Mexico in their paper "Survey of Overlapping Clustering Algorithms" present a study of the overlapping clustering algorithms that have been developed in the last years. The algorithms included in this paper are: ADditive CLUstering, Overlapping K-means, Dynamic Overlapping Clustering based on Relevance, Overlapping Clustering based on Density and Compactness, MCLC, A tree-based incremental overlapping clustering method, INDCLUS and Hybrid K-means.

N. Hernández et al. from Mexico in their paper "Evolutionary Automatic Text Summarization using Cluster Validation Indexes" propose a comparison of the correlation of the quality of a human-made summary to the internal quality of the clustering validation index for finding the best correlation with a clustering validation index. Additionally, they propose an evolutionary method based on the best internal clustering validation index for an automatic text summarization task. Authors of this paper claim that their method maintains a high correlation with human-made summaries employed as gold standard for the task evaluation.

D. Guevara et al. from Ireland in their paper "Analysis of Automatic Annotations of Real Video Surveillance Images" present results of the analysis of automatic annotations of real video surveillance sequences. The annotations of the frames of surveillance sequences of the parking lot of a university campus were generated with the purpose of analyzing the quality of the descriptions and the relationship between the semantic content of the images and the corresponding annotation.

R. Guzmán et al. from Mexico in their paper entitled "Classification of Opinions in Cross Domains Involving Emotive Values" perform an automatic categorization of textual opinions corresponding to four products: books, DVDs, kitchens, and electronics. Both negative and positive opinions were considered for the experiment. Further categorization experiments were performed using different domains of learning with the aim of investigate wheter or not is possible to undertake classification of opinions, positive and negative, of any given domain using instances of training from a different domain.

A. Garcés et al. from Mexico in their paper "A Logical Interpretation of Silence" focus on a puzzle formerly expressed and solved in Answer Set Programming, to analyze the implications of two different interpretations of silence (Defensive and Acquiescent Silence), in terms of the Says() predicate. Several conclusions are derived from the different possibilities that opened for analysis. Additionally, they propose a general strategy for analysis of problems involving testimonies and silence.

L. Sánchez et al. from Mexico in their paper entitled "Weighted Bidirectional Graph-based Academic Curricula Model to Support the Tutorial Competence" propose a weighted bidirectional graph, which will represent the intrinsic and wired restrictions in the flow of subjects that students must follow and pass, establishing an organized context (which is regulated by the weight of each subject) within which they can be oriented and guided by the tutor. A system to monitor the students' academic progress is presented providing an interaction between the students and the tutor, where the weighted bidirectional graph is integrated.

S. Frenda et al. from Italy in their paper entitled "Do Linguistic Features Help Deep Learning? The Case of Aggressiveness in Mexican Tweets" presents an approach that combines the deep learning framework with

linguistic features for the recognition of aggressiveness in Mexican tweets. This approach has been evaluated relying on a collection of tweets released by the organizers of the shared task about aggressiveness detection in the context of the Ibereval 2018 evaluation campaign, concluding that linguistic features seem not to help the deep learning classification for this particular task.

P. Silva et al. from Ireland in their paper "An Univariable Approach for Forecasting Workload in the Maintenance Industry" discuss the problem and the challenges of forecasting of the workload in the maintenance industry. They analyze data from a company operating in the industry and present the results of several forecasting models.

J. Huetle et al. from Ireland in their paper "On Detecting Keywords for Concept Mapping in Plain Text" presents a way to obtain the key terminology based on labels that were manually obtained by an expert in the area. Four different experiments were reported in this paper with different results.

S. Ketu et al. from India in their paper entitled "Performance Analysis of Distributed Computing Frameworks for Big Data Analytics: Hadoop Vs Spark" show a comparative analysis of Hadoop MapReduce and Spark has been presented on the basis of working principle, performance, cost, ease of use, compatibility, data processing, failure tolerance, and security. Experimental analysis has been performed to observe the performance of Hadoop MapReduce and Spark for establishing their suitability under different constraints of the distributed computing environment.

V. Pratap et al. from India in their paper "An LSTM Based Time Series Forecasting Framework for Web Services Recommendation" LSTM based deep learning models were used for the prediction of these time aware QoS parameters and the results are compared with the previous approaches. The experimental results show that the Long Short-Term Memory (LSTM) based Time Series Forecasting Framework is performing better than other approaches.

J. Ramos et al. from Mexico in their paper entitled "Determining the More Adequate Web Page node for Advertising Placement" explore the underlying tree-like structure of a web page, extracting the text from each (X)HTML node and computing the semantic similarity (by employing latent semantic analysis) with respect to the advertising source text. They introduce a unique formula for the numerical calculation of the web page node relevance with the aim of using it for measuring the concordance among web page nodes and the commercial information so as for the design of dynamic ads insertion methods.

A. Vazquez et al. from México in their paper entitled "Grammatical Inference of Semantic Components in Dialogues" the construction of a model that recognizes semantic components of spontaneous dialogues about telephonic queries of schedules and prices of long distance train tickets is reported. Grammatical inference techniques were used to infer an automaton. The accuracy of the automaton recognizing sequences of semantic components is around 96%.

M. Bernábe et al. from Mexico in their paper "Algorithm for Collecting and Sorting Data from Twitter through the Use of Dictionaries in Python" introduce a tool for the classification of opinions written in natural language in Twitter, well-known social network. The main purpose is to split up, into two classes, the opinions expressed by Twitter users about the Mexican presidential polling carried out in 2018.

B. Priego from Mexico in their paper entitled "A New Lexical Resource for Evaluating Polarity in Spanish Verbal Phrases" presents a new manual annotated corpus for the news genre, made up of sentences containing each a verbal phrase. The supervised corpus has been annotated with respect to polarity and compositionality for identifying polarity in Spanish verbal phrases, so as to determine whether or not a given verbal phrase has a figurative or literal meaning. A lexicon of verbal phrase likelihoods in Spanish language is also given.

L. Colmenares et al. from Mexico in their paper "Face Classification in Adults and Minors, an Approach based on Facial Anthropometry" a geometric approach is developed by using distances and proportions existing in every face. The location of 15 fiducial points is considered through the ASM algorithm and 8 distances between fiducial points, which statistically verify the difference in facial proportions between adults and minors. The distances obtained are analyzed

ISSN 2007-9737

with the Discriminant Analysis procedure to study the correlations between distances and the corresponding age group.

Y. Aleman et al. from Mexico in their paper entitled “An Analysis of Variance Method to Detect Collocations in a Pedagogical Domain Corpus” an exploratory experiment, based on analysis of variance was carried out in order to obtain collocations in a pedagogical domain corpus. The proposed method divides the list of bigrams in quartiles and analyzes the variance on each one of them.

This issue also contains several regular papers reviewed by the journal editors.

David Pinto, Beatriz Beltrán, Andrés Vázquez
BUAP, Mexico
Guest Editors