Journal of Intelligent & Fuzzy Systems 42 (2022) 4299–4305
DOI:10.3233/JIFS-219220
IOS Press

Guest Editorial

Recent advances in language & knowledge engineering

David Pinto[a,*], Beatriz Beltrán[a] and Vivek Singh[b]
[a] *Faculty of Computer Science, Benemérita Universidad Autónoma de Puebla, San Claudio, Puebla, Mexico*
[b] *Computer Science Department, Banaras Hindu University, Varanasi, India*

Abstract. Language & Knowledge Engineering is essential for the successfully development of artificial intelligence. The technologies proposed in international forums are meant to improve all areas of our daily life whether it is related to production industries, social communities, government, education, or something else. We consider very important to reveal the recent advances Intelligent and Fuzzy Systems applied to Language & Knowledge Engineering because they are the base for the society of tomorrow. Thus, the aim of this special issue of Journal of Intelligent and Fuzzy Systems is to present a collection of papers that cover recent research results on the two wide topics: language and knowledge engineering. Even if the special issue is structured into these two general topics, we have covered specific themes such as the following ones: Natural Language Processing, Knowledge engineering, Pattern recognition, Artificial Intelligence and Language, Information Processing, Machine Learning Applied to Text Processing, Image and Text Classification, Multimodal data analysis, sentiment analysis, etc.

Keywords: Language engineering, knowledge engineering

Language engineering is an active area of artificial intelligence having the aim to bridge the gap between traditional computational linguistics research and the implementation of potentially real-world applications. The purpose of language engineers is to provide new insights which foster the development of theoretical or applied methods which can advance the research in artificial intelligence from a human language research perspective in some tasks such as machine translation, sentiment analysis, reputation analysis, etc. As we will further describe, this thematic issue contains twenty-four papers associated to the natural language engineering area, presenting specific natural language processing methods, tasks or applications.

On the other hand, knowledge engineering is related to all technical, scientific and social aspects involved in designing, building, maintaining and using knowledge-based systems. The final aim of knowledge engineering is to support human decision-making, learning and action, with emphases to 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.

The call for papers of this special issue received an overwhelming response from the community. After rigorous review only 46 papers representative of different tasks, techniques, and applications of language and knowledge engineering were selected from more than 100 papers submitted to the special issue. These papers represent the most up-to-date research work covering the aforementioned topics. We hope the

[*] Corresponding author. David Pinto, Faculty of Computer Science, Benemérita Universidad Autónoma de Puebla, 14 Sur & Av. San Claudio, 72570, Puebla, Mexico. E-mails: david.pinto@correo.buap.mx; davideduardopinto@gmail.com

reader will find this special issue informative and stimulating.

The broad two themes covered in this special issue are described as follows. First, we start this special issue with 24 papers devoted to the language engineering area. The general description of all the papers for this particular topic follows.

Agarwal et al. in their paper entitled "Query-focused Multi-Document Text Summarization using Fuzzy Inference" present a fuzzy inference system for query-focused multi-document text summarization with promising results which helps in designing fuzzy rule base for inferencing about the decision variable from a set of antecedent variables.

Crespo-Sanchez et al. in their paper entitled "A Content Spectral-based Text Representation" propose a content spectral-based text representation applicable to machine learning algorithms for text analysis which integrates the spectra from the lexical, syntactic, and semantic components of text producing an abstract image, which can also be treated by both, text and image learning algorithms. The proposal has been tested on text classification and complexity reading score prediction tasks obtaining promising results.

Tahir et al. in their paper entitled "Anbar: Collection and Analysis of a Large Scale Urdu Language Twitter Corpus" build and analyze a large scale Urdu language Twitter corpus Anbar. They collected 106 million Urdu tweets posted by 1.69 million users for one year. The constructed corpus contains 3.8 million unique tokens along with 58K hashtags and 62K URLs. Moreover, it contains 76 million (71.6%) retweets and 847K geotagged tweets. Finally, they examine Anbar using a variety of metrics like temporal frequency of tweets, vocabulary size, geo-location, user characteristics, and entities distribution.

García-Mendoza et al. in their paper entitled "An Autoencoder-based Representation for Noise Reduction in Distant Supervision of Relation Extraction" propose an Adversarial Autoencoders-based approach for obtaining a new representation that allows noise reduction in Distant Supervision. The representations obtained using Adversarial Autoencoders minimize the intra-cluster distance concerning to pre-trained embeddings and classic Autoencoders. Experiments conducted in their paper, allowed the authors to claim that in the noise-reduced datasets the macro precision values obtained over the original dataset are similar using less instances considering the same classifier.

Priego-Sanchez et al. in their paper entitled "Polarity Identification of a text given the emotion of its author" implement different methods for automatically determining the polarity of sentences and categorize them as positive, negative or neutral considering their lemmas. The results presented in their paper allow them to observe a direct relationship between the categorized emotional tone and it is positive, negative or neutral classification, which provide additional information to discovering the intention that the author had when he created the phrase.

Sanchez-Fernandez et al. in their paper entitled "Latent Semantic Analysis for tagging Activation States and Identifiability in Northwestern Mexican news outlets" study the relation between, on the one side, the measures obtained from Latent Semantic Analysis (LSA) as well as from a variant known as SPAN and, on the other, the activation and identifiability states (Informative States) of referents in noun phrases present in journalistic notes from Northwestern Mexican news outlets written in Spanish. This semi-supervised method is aimed at finding a strategy to achieve labeling of new/old information in the discourse. The experiments presented in their paper have shown good results for the binary division, detecting which sentences introduce new referents in discourse.

Balouchzahi et al. in their paper entitled "Fake News Spreaders Profiling using N-grams of Various Types and SHAP-based Feature Selection" explore char, character sequences, syllables, word n-grams as well as syntactic n-grams to train machine learning classifiers. Additionally, authors look to improve the performance of classifiers for Fake News Spreaders Profiling task by employing most frequent features and apply feature selection of these features based on SHAP values. The Voting Classifier with soft voting exhibited best performances on frequent n-grams features and features selected based on SHAP values and the proposed models resulted in an average accuracy of 0.785.

Damian et al. in their paper entitled "Fake News Detection using n-grams for PAN@CLEF Competition" address the task of detecting possible fake news spreaders accounts through word n-grams and statistical features, which allow them to select the best classifier that can accomplish the objective. They also use Explainable Artificial Intelligence methods (XIA) as feature selected method with the aim of preventing the consume of fake news in two of the most spoken languages in the world.

Kostiuk et al. in their paper entitled "Prior Latent Distribution Comparison for theRNN Variational

Autoencoder in Low-Resource Language Modeling" study the priors selection impact of continuous distributions in the Low-Resource Language Modeling task with Variational Autoencoder (VAE). The experiment presented in their paper let them to conclude that there is a statistical difference between the different priors in the encoder-decoder architecture.

Fuentes-Ramos et al. in their paper entitled "Neurodegenerative diseases categorization by applying the automatic model selection and hyperparameter optimization method" propose to employ an automatic model selection and hyperparameter optimization method that has not been addressed before for the problem of neurodegenerative diseases categorization. The results presented in their paper showed highly competitive percentages of correctly classified instances when discriminating binary and multiclass sets of neurodegenerative diseases: Parkinson's disease, Huntington's disease, and Spinocerebellar ataxias.

Zhang et al. in their paper entitled "Question Type and Answer Related Keywords Aware Question Generation" propose a framework for question type and answer related keywords aware question generation by taking experiments based on the GPT-2 model and the SQuAD dataset. They claim their framework can improve the performance measured by similarity metrics, while it also provides appropriate alternatives for controllable diversity enhancement.

Tandon et al. in their paper entitled "Multi-label Text Classification with an Ensemble Feature Space" describe a scheme of feature extraction where the training document set, and the prescribed label set are intertwined in a novel way to capture the ambiguity in a meaningful way. Experiments were conducted using Topic Modeling and Fuzzy C-means clustering which aim at measuring the underlying uncertainty using probability and membership-based measures, respectively. Several Nonparametric hypothesis tests establish the effectiveness of the features obtained through Fuzzy C-Means clustering in multi-label classification.

Ruiz et al. in their paper entitled "multi-label classification of feedbacks" classify the feedback generated by teachers in online courses to the activities sent by students according to the model of Hattie and Timperley, considering that feedback may be at the levels task, process, regulation, self and other.

Ruiz et al. in their paper entitled "Hyperparameter tuning for multi-label classification of feedbacks in online courses" propose the extension of a methodology for the multi-label classification of feedback according to the Hattie and Timperley feedback model, by incorporating a hyperparameter tuning stage. It is analyzed whether the incorporation of the hyperparameter tuning stage prior to the execution of the algorithms support vector machines, random forest, and multi-label k-nearest neighbors, allows to improve the performance metrics of multi-label classifiers that automatically locate the feedback generated by a teacher to the activities sent by students in online courses on the Blackboard platform at the task, process, regulation and praise levels proposed in the feedback model by Hattie and Timperley. Finally, the grid search strategy is used to refine the hyperparameters of each algorithm.

Bahuguna et al. in their paper entitled "A Unified Deep Neuro-Fuzzy Approach for COVID-19 Twitter Sentiment Classification" propose a unified deep neuro-fuzzy approach for COVID-19 twitter sentiment classification using Sugeno integral which succeeds over individual classifiers.

García-Gorrostieta et al. in their paper entitled "Improved Argumentative Paragraphs Detection in Academic Theses Supported with Unit Segmentation" present an exploration of lexical features used to model automatic detection of argumentative paragraphs using machine learning techniques, which combines the information in the complete paragraph with the detection of argumentative segments in order to achieve improved results for the detection of argumentative paragraphs.

Medina et al. in their paper entitled "Onto4AIR2: an ontology to manage theses from open repositories" describe Onto4AIR2, an ontology to manage theses from open repositories, this fosters unique and formal definitions of concepts from the Mexican repositories' domain in English and Spanish languages, its goal is to support the construction of machine-readable datasets that are semantically labeled for further consultations in educational organizations. The ontology instances are sample data of theses from the National Repository of Mexico, an initiative promoted by the National Council of Science and Technology. The paper suggests practical applications derived from the formalisms of the ontology and describes an assessment technique where participants are developers and potential users.

Rahman et al. in their paper entitled "Improved Neural Machine Translation for Low-Resource English-Assamese Pair" propose an approach of data augmentation-based neural machine translation, which exploits synthetic parallel data and shows significantly improved translation accuracy for English-to-Assamese and Assamese-to-English translation claiming to obtain state-of-the-art results.

Chatterjee et al. in their paper entitled "Soft Rough Set based Span for Unsupervised Keyword Extraction" propose an application of Soft Rough Set and its span for unsupervised keyword extraction. The proposed technique uses a greedy algorithm for computing spanning sets. The experimental results suggest that the extraction of keywords using the proposed scheme gives consistent results across different domains.

Ashraf et al. in their paper entitled "YouTube Based Religious Hate Speech and Extremism Detection Dataset with Machine Learning Baselines" present a methodology for the detection of religion-based hate videos on YouTube. Messages posted on YouTube videos generally express the opinions of users related to that video. The proposed methodology applies data mining techniques on extracted comments from religious videos in order to filter religion-oriented messages and detect those videos which are used for spreading hate.

Solovyev et al. in their paper entitled "Automatic generation of a large dictionary with concreteness/abstractness ratings based on a small human dictionary" present a method for automatic ranking concreteness of words and propose an approach to significantly decrease amount of expert assessment. The method has been evaluated on a large test set for English. The quality of the constructed dictionaries is comparable to the expert ones. The correlation between predicted and expert ratings is higher comparing to the state-of-the-art methods.

Vázquez-González et al. in their paper entitled "Creating a corpus of historical documents for the emotions identification" contextualize and describe the gathering and annotation of a conventual Hispanic and Novo Hispanic texts corpus for emotions identification. It is also described the manner the corpus is employed for obtaining a lexicon with the corresponding polarities and emotions tags, and how some of the documents are hand-labeled by experts for the evaluation of the Machine Learning-based emotion classification model.

Sierra et al. in their paper entitled "A Case Study in Authorship Attribution: The Mondrigo" carry out authorship attribution methods over "El Mondrigo", a controversial text created in 1968 by order of the Mexican Government to defame a student strike. In order to discover the author of the text, the authors implement methods based on textual distance, stylometry, supervised and unsupervised learning. The applied methods were consistent by pointing out a single author as the most likely one.

Martín et al. in their paper entitled "Unsupervised Authorship Attribution Using Feature Selection and Weighted Cosine Similarity" present a computational model for the unsupervised authorship attribution task, based on a traditional scheme of machine learning. By comparing different feature selection methods, an improvement was achieved with respect to the results obtained in the state of art (with the same dataset). A method to separate the tokens by types, to assign only one category to each token was used. Similarly, special characters were used as part of the punctuation marks, this to improve the result obtained when using the typed character n-grams.

The second part of this volume contains 22 papers devoted to the knowledge engineering area and their general description follows.

Bernábe et al. in their paper entitled "A Statistical evaluation of the oral vaccine S3pvac Papaya against Cysticercosis of Taenia Psiformis" evaluate the oral S3Pvac-Papaya 12 mg vaccine in a model of rabbit cysticercosis by Taenia Pisiformis. This vaccine has been applied to a group of rabbits bred in New Zealand. The descriptive and inferential statistical analysis revealed the condition of the rabbits before being vaccinated, after being infected and when they were slaughtered. A multiple means comparative analysis indicated no significant statistical difference between the 3 treatments, however an ANOVA study in combination with the box plot for S3P Wild at T2 showed a high level of data dispersion, i.e., there were rabbits with many antibodies and others with few antibodies. On the other hand, the vaccine of interest, S3Pvac-Papaya, revealed in the box diagram at T2 that the development of antibodies was high and showed little dispersion, which implies that the vaccine is efficient in the production of antibodies.

García et al. in their paper entitled "Detection of the level of attention in children with ADHD through brain waves and corporal posture" describe a methodology and the experimentation to know the level of attention of people through a test to identify colors also are shown the development and the application of a system (hardware and software) to measure the level of attention of people using two input signals: corporal posture and brain waves. The mathematical analysis to find the correlation between the corporal posture and the level of attention is shown in this paper. The results obtained indicate that the corporal posture influences on the level of attention of people directly.

Guzman-Cabrera et al. in their paper entitled “Improved approach to wave potential estimation using bivariate distributions” propose a methodology to wave potential estimation using bivariate distributions based on the use of joint probability models that allow discriminating extreme values, collected from measurements as pairs of independent points, while allowing the preservation of the essential statistics of the measurements. The outcome of the proposed methodology is an equivalent data series where large-amplitude fluctuations are suppressed and, therefore, can be used for design purposes.

Rodriguez et al. in their paper entitled “A Microlearning Path Recommendation Approach Based on Ant Colony Optimization” present the technical proposal of a novel approach based on Ant Colony Optimization (ACO) to recommend personalized microlearning paths considering the learning needs of the learner. The recommendation problem is approached as an instance of the Traveling Salesman Problem (TSP), the educational pills represent the cities, the paths are the relationships between educational pills, the cost of going from one pill to another can be estimated by their degree of difficulty as well as the performance of the learner during the individual test.

Carreón et al. in their paper entitled “A novel methodology of parametric identification for robots based on a CNN” present a novel methodology to identify dynamic parameters of a real robot with a convolutional neural network (CNN). The conventional identification methodologies are based on continuous movement signals; however, the used signals are quantified and time-discrete. The methodology presented by the authors consists of an algorithm that uses a convolutional neural network trained with data created by the dynamical model of a two degree of freedom cartesian robot. In this paper, it is proposed a processing technique to transform movement signals into image which characteristics are extracted by the convolutional network to determine the dynamic parameters.

Morveli-Espinoza et al. in their paper entitled “Handling Temporality in Human Activity Reasoning” propose an approach based on formal argumentation reasoning, specifically, Timed Argumentation Frameworks (TAF), for dealing with inconsistencies in knowledge bases with the aim of handling temporality in human activity reasoning. By considering a set of observations, a model of the world and of the human’s mind is constructed in form of hypothetical fragments of activities also called evidences.

Utsuki-Alexander et al. in their paper entitled “Towards an intelligent personal assistant for hearing impaired people” describe a device prototype that was developed by them in order to be integrated as a component of ambient intelligence (AmI) for ambient assisted living (AAL) that serves to Hearing Impaired People(HIP). The prototype detects dog barks and notifies users through both a smart mobile app and visual feedback. The prototype was tested with deaf people. Participants were satisfied with precision, signal intensity, and activation of lights. The device recognized the barking efficiently by using a machine learning model based on Support Vector Machine technique.

Gutiérrez-Soto et al. in their paper entitled “A New and Efficient Algorithm to Look for Periodic Patterns on Spatio-Temporal Databases” present a new efficient algorithm to look for periodic patterns on Spatio-Temporal Database (STDB) which is compared with Apriori, Max-Subpattern, and PPA algorithms on synthetic and real STDB. Additionally, the computational complexities for each algorithm in the worst cases are presented. Empirical results show that the algorithm proposed is more efficient than Apriori, Max-Subpattern, and PAA, but in addition, it presents a polynomial behavior.

Rangel et al. in their paper entitled “Deep Symbolic Processing of Human-Performed Musical Sequences” describe ongoing research for creating music tutors for assisting a performer during his/her practice time whenever a human tutor is not available. The current proposal uses cascading connected layers of symbolic processing as the core of a human-performed error identification and characterization module able to overcome the complexity of the studied open-ended domain.

Lopez-Rincon et al. in their paper entitled “Algorithmic Music Generation by Harmony Recombination with Genetic Algorithm” present an approach for generating new music composition by replacing the existing harmony descriptors of the original MIDI file with new harmonic features obtained by using a genetic algorithm. The performance of the proposed approach has been assessed using some computational tests, which assure goodness of entire generated music piece and show its quality and competitiveness.

Romero et al. in their paper entitled “Electro-Impedance Mammograms for Automatic Breast Cancer Screening: First Insights on Mexican

Patients" address breast cancer detection as a multi-class problem with the aim to determine the corresponding label in terms of the Breast-Imaging Reporting and Data System, the standard used by physicians for interpreting a mammogram. Different experimental settings were evaluated reaching classification rates over 0.85 in F-score.

Trevino-Sanchez et al. in their paper entitled "Hybrid Pooling with Wavelets for Convolutional Neural Networks" incorporate multi-resolution analysis (MRA) within the Convolutional neural networks (CNN) layers to reduce the feature map size without losing details in order to detect and classify objects correctly. With the aim of preventing relevant information be missed during the downsampling process, an existing pooling method is combined with MRA, keeping those details "alive" and enriching other stages of the CNN. In order to validate the model, four benchmarks' datasets were used, and the results obtained were compared with the state-of-the-art.

Ahmed et al. in their paper entitled "Ensemble-based Deep Meta learning for Medical Image Segmentation" propose a meta learning-based image segmentation model that combines the learning of the state-of-the-art model and then used it to achieve domain adoption and high accuracy. Also, they propose a pre-processing algorithm to increase the usability of the segments part and remove noise from the new test image. The proposed model was able to achieve 0.94 precision and 0.92 recall.

Ahmed et al. in their paper entitled "Deep Active Reinforcement Learning for Privacy Preserve Data Mining in 5 G networks" use deep active learning to hide sensitive operations and protect private information. In their paper, the authors combine entropy-based active learning with an attention-based approach to effectively detect sensitive patterns. The constructed models are then validated using high-dimensional transactional data with attention-based and active learning methods in a reinforcement environment. Authors claim that the proposed model can support and improve the decision boundaries by increasing the number of training instances using a pooling technique and an entropy uncertainty measure.

Herrera et al. in their paper entitled "Wavelets as activation functions in Neural Networks" wavelets are reconsidered as activation functions in neural networks and their performance is studied together with other functions available in Keras-Tensorflow. Experimental results show how the combination of these activation functions can improve the performance and supports the idea of extending the list of activation functions to wavelets which can be available in high performance platforms.

Suarez-Cansino et al. in their paper entitled "Automatic Generation of Learning Outcomes based on Long Short–Term Memory Artificial Neural Network." propose the use of a Long Short–Term Memory Artificial Neural Network (LSTM) to organize the structure and automatize the obtention of learning outcomes for a focused instructional design. Authors present encouraging results in this direction using a LSTM and employing as the training data, a small learning outcome set predefined by the user, focused on the characteristics of an educative model previously defined.

Reyes-Cocoletzi et al. in their paper entitled "Motion Estimation in Vehicular Environments based on Bayesian Dynamic Networks" show a method based on Bayesian dynamic networks to infer the paths of interest objects (IO) in vehicular environments for motion estimation. The proposed approach was evaluated using test environments considering different road layouts and multiple obstacles in real-world traffic scenarios obtaining a prediction rate of 75% for the change of direction taking into consideration the risk of collision.

Ballinas et al. in their paper entitled "Marked and Unmarked Speed Bump Detection for Autonomous Vehicles using Stereo Vision" propose a methodology for detecting both marked and unmarked speed bumps. In the case of clearly painted speed bumps, they apply a local binary patterns technique for extracting features from an image dataset. For unmarked speed bump detection, they apply stereo vision where point clouds obtained by the 3D reconstruction are converted to triangular meshes by applying Delaunay triangulation. A selection and extraction of the most relevant features is made to speed bump elevation on surfaces meshes. The authors obtained encouraging results.

Sierra et al. in their paper entitled "Classification and Enhancement of Invasive Ductal Carcinoma Samples using Convolutional Neural Networks" propose an automatic methodology for improving the performance of a convolutional neural network that classifies images containing invasive ductal carcinoma cells by highlighting cancer cells using several preprocessing methods such as histogram stretching and contrast enhancement.

Gallardo et al. in their paper entitled "Searching for Memory-Lighter Architectures for OCR-Augmented

Image Captioning" introduce two alternative versions (L-M4 C and L-CNMT) of top architectures (on the TextCaps challenge), which were mainly adapted to achieve near-State-of-The-Art performance while being memory-lighter when compared to the original architectures, this is mainly achieved by using distilled or smaller pre-trained models on the text-and-OCR embedding modules. Two of the three models presented in this work overcome the baseline (M4C-Captioner) on the evaluation and test datasets.

Lopez-Medina et al. in their paper entitled "A method for counting models on grid Boolean formulas" present an algorithm based on combinatorial operations on lists for computing the number of models on two conjunctive normal form Boolean formulas whose restricted graph is represented by a grid graph. For this class of formulas, they show that the proposal improves the asymptotic behavior of the time-complexity with respect of the current leader algorithm for counting models on two conjunctive form formulas of this kind.

Rebollar et al. in their paper entitled "Modeling a multilayered blockchain framework for digital services that governments can implement" present a multilayered blockchain framework for digital services that governments can implement which divides the information into four classes and separates it into the same number of layers. Each lawyer specializes in processing its variety of information and has unique properties that allow it to maintain and even increase the integrity of the information contained while preserving efficiency in the transactions. The proposal also enables volunteer nodes to participate on the decentralization of the information and make it more secure, verifiable, transparent, and reliable.

We would like to express our gratitude to the Editors in Chief and the publisher for giving us the opportunity to edit this special issue on Recent Advances in Language & Knowledge Engineering. We would also like to thank all the contributors who have submitted their high-quality papers. Finally, we would like to thank the IOS Press editorial staff members for their constant support throughout the preparation of the issue.

Guest Editors
David Pinto
Beatriz Beltrán
Vivek Singh