Journal of Intelligent & Fuzzy Systems 36 (2019) 4547–4552
DOI:10.3233/JIFS-179006
IOS Press

Guest Editorial

Intelligent and fuzzy systems applied to language & knowledge engineering

D. Pinto and V. Singh

Language & Knowledge Engineering is a very challenging area which is essential for the development of artificial intelligence in particular and Computer Science in general. These technologies are improving all areas of our daily life whether it is related to the education, health, production industries or something else. Thus, Recent Advancements in Intelligent and Fuzzy Systems applied to Language & Knowledge Engineering are the base for the society of tomorrow.

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 topic of language and knowledge engineering. In particular, it aims to present technical papers in some of the following areas: Natural Language Processing, Knowledge engineering, Pattern recognition, Artificial Intelligence and Language, Scholarly Information Systems, Information Retrieval, Informetrics, Information Processing, Machine Learning Applied to Text Processing, Humanoids, Social Media Analytics and Fuzzy Systems for Text.

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.

The call for papers of this special issue received an overwhelming response from the community. After rigorous review only 49 papers representative of different tasks, techniques, and applications of language and knowledge engineering were selected from more than 150 papers submitted to the special issue. These papers represent the most up-to-date research work covering the aforementioned topics. We hope the reader will find this special issue informative and stimulating.

The broad themes covered in this special issue are described as follows. First, we start this special issue with 33 papers devoted to the language engineering area. Their general description follows.

Solovyev et al. in their paper "Prediction of Reading Difficulty in Russian Academic Texts" undertook a comparative analysis of academic texts features exemplified in textbooks on Social Science and examination texts of Russian as a foreign language. Experiments for 7 classifiers and 4 methods of linear regression on Russian Readability corpus demonstrated that ranking textbooks for native speakers is a much more difficult task than ranking examination texts written (or designed) for foreign students.

ISSN 1064-1246/19/$35.00

Garcia-Gorrostieta et al. in their paper "A Corpus for Argument Analysis of Academic Writing: Argumentative Paragraph Detection" elaborated an annotation guide to identify argumentation in paragraphs. After its construction, the corpus was used to perform an exploratory analysis which aimed to identify and present the amount of argumentation in each section, as well as resulting patterns for argument identification with encouraging results.

Priego et al. in their paper "An Unsupervised Method for Automatic Validation of Verbal Phraseological Units" present an unsupervised technique for validating the existence of verbal phraseological units in raw text using the concept of internal and contextual attraction which basically considers a mathematical formula based on co-occurrence of terms inside and outside of the terms considered to be part of a verbal phraseological unit. The experiments carried out using a corpus of news stories reported a 60% of accuracy.

Reyes-Magaña et al. in their paper "A Lexical Search Model Based on Word Association Norms" introduce a lexical search model based on a type of knowledge graphs, namely word association norms in order to retrieve a target word, given the description of a concept, i.e., the query. They performed experiments over a corpus of human-definitions in order to evaluate the proposed model. The results are compared with a Boolean information retrieval (IR) model, the BM25 text-retrieval algorithm, an algorithm based on word vectors and an online onomasiological dictionary–OneLook Reverse Dictionary showed that the proposed lexical search method outperforms the other IR models employed in the comparison.

González et al. in their paper "Siamese Hierarchical Attention Networks for Extractive Summarization" present an extractive approach to document summarization based on Siamese Neural Networks. Specifically, we propose the use of Hierarchical Attention Networks to select the most relevant sentences of a text to make its summary. The experimentation carried out using the CNN/DailyMail summarization corpus shows the adequacy of the proposal with promising results.

Millán-Hernández et al. in their paper "An Evolutionary Logistic Regression Method to Identify Confused Drug Names" present an improved combined logistic regression measure based on 21 individual measures together with an evolutionary learning method for a combined logistic regression measure that allows to learn an unbalanced dataset to Identify Confused Drug Names. The proposed combined measures outperformed previous research with a statistical significance to identify pairs of confused drug names.

García-Calderón et al. in their paper "Providing Order to the Handwritten Text Line Segmentation Task: A Complexity Index" they present TLS-ICI, a Text Line Segmentation (TLS) Intrinsic Complexity Index that allows measuring the complexity of a document for the TLS task, without the necessity of a human gold standard. They argued that with the proposed complexity index it is possible to select the most appropriated method for each document of a collection, reducing the time spent in exhaustive tests and increasing the performance.

Fócil-Arias et al. in their paper "Medical Events Extraction to Analyze Clinical Records with Conditional Random Fields" propose the use of different features with a statistical modeling method called conditional random fields, in order to determine which feature selection can affect the performance of four subtasks presented in SemEval Task-12: Clinical TempEval 2016. They found that simple features can be more effective in some subtask such as event, contextual modality, and polarity detection, however word embeddings were not so helpful in those subtasks.

Brena et al. in their paper "Scalable Text Semantic Clustering around Topics" propose an alternative topic modeling paradigm based on a simpler representation of topics as overlapping clusters of semantically similar documents, that is able to take advantage of highly-scalable clustering algorithms. The experiments carried out show that the Query-based Topic Modeling framework can produce models of comparable or even superior quality than those produced by state of the art probabilistic methods.

Gupta et al. in their paper "A Quantitative and Text-based Characterization of Big Data Research" attempted to map the research work carried out in the field of Big Data through a detailed analysis of scholarly articles published on the theme during 2010-16, as indexed in Scopus. The results produce interesting inferences. Quantitative measures show that there has been a tremendous increase in number of publications related to Big Data during last few years. The paper also identifies major keywords now associated with Big Data research such as Cloud Computing, Deep Learning, Social Media and Data Analytics, which in fact helps in a thorough understanding and visualization of the Big Data research area.

Figueroa et al. in their paper "Locality-Sensitive Hashing of Permutations for Proximity Searching" present several novel hash functions for Locality Sensitive Hashing (LSH) with permutation based algorithm (PBA) in order to speed up the search process. Authors claimed that at searching, their technique allows discarding up to 50% of the database to answer the query with a candidate list obtained in constant time.

Pathak et al. in their paper "Binary Vector Transformation of Math Formula for Mathematical Information Retrieval" introduce a novel approach, called Binary Vector Transformation of Math Formula (BVTMF) useful for Mathematical Information Retrieval (MIR) systems which allows to retrieve math formulae from scientific documents. Quality of the retrieved search results and appreciable values of the evaluation measures substantiate competence of the proposed approach.

Hurtado et al. in their paper "Choosing the Right Loss Function for multi-label Emotion Classification" present a strategy to incorporate evaluation metrics in the learning process in order to increase the performance of a multi-label emotion classifier, according to the measure they are interested to favor. By using a Convolutional Neural Network trained with the proposed loss functions they reported significant improvements both for the English and the Spanish corpora.

Rodríguez et al. in their paper "Predicting emotional intensity in social networks" propose a model to predict (forecast) emotions in social networks. The model specifically predicts, for a user, the proportion of comments that will be published with a particular emotion; this proportion is defined as an emotional intensity of the user in a particular time period. Authors claim that nearly 20 models are outperformed by the proposed model (with statistically significant results) when evaluated over a dataset extracted from Twitter.

Gupta et al. in their paper "Aspect-Based Sentiment Analysis on Mobile Phone Reviews" present an integrated system that generates the opinionated aspect based graphical and extractive summaries from a large set of mobile reviews. The system has been evaluated on three mobile-reviews dataset and obtains better precision and recall than baseline approach.

Baowaly et al. in their paper "Predicting the helpfulness of game reviews: a case study on the Steam store" evaluate the helpfulness of game reviews on the online Steam store. The construct a classification model that can accurately predict the helpfulness of the reviews based on different thresholds. They analyze the importance of different features in the prediction process and develop a regression-based model that can predict the score or rating of game reviews on Steam.

Frenda et al. in their paper "Online Hate Speech against Women: Automatic Identification of Misogyny and Sexism on Twitter" present an approach that is able to detect the two sides of patriarchal behavior, misogyny and sexism, analyzing three collections of English tweets, and obtaining promising results.

Alemán et al. in their paper "Similarity metrics analysis for principal concepts detection in ontology creation" present an analysis, based on similarity metrics, was carried out in order to detect main concepts related to the superclasses in a pedagogical domain ontology. Results showed a higher precision in types of intelligences class and 5-grams representation.

Girón et al. in their paper "Rule-based expert system for detection of coffee rust warnings in colombian crops" propose an expert system based on rules, where the rules are created considering the expert knowledge of specialists and technical reports about the behavior of the disease during a crop year. Experiment results present an average accuracy of 66,67% to detect a correct warning of coffee rust levels.

Lithgow-Serrano et al. in their paper "In the pursuit of semantic similarity for literature on microbial transcriptional regulation" they use an ensemble of similarity metrics including string, distributional, and knowledge-based metric and to combine the results of such analyses in order to pursue semantic similarity for literature on microbial transcriptional regulation. They report strong correlation of the results with respect to human evaluation.

Kumar Dash et al. in their paper "Generating Image Captions through Multimodal Embedding" describe a model employed to generate novel image captions for a previously unseen im- age by utilizing a multimodal architecture by amalgamation of a Recurrent Neural Network (RNN) and a Convolutional Neural Network (CNN). Microsoft Common Objects in Context (MSCOCO) evaluation server is used for evaluation of the machine generated caption for a given image.

Majumder et al. in their paper "Measuring interpretable semantic similarity of sentences using a multi chunk aligner" propose an Interpretable Semantic Textual Similarity (iSTS) method which explains the similarities and differences between a pair of sentences. The obtained result outperforms many state–of–art aligners, which were part of SemEval 2016 iSTS task.

Srivastava et al. in their paper "Extraction of Reordering Rules for Statistical Machine Translation" present a statistical method for extraction of linguistic rules using chunk to reorder the output of the baseline statistical machine translation system for improved performance. The results are compared with those of Google translation system which has been trained on a huge corpus, obtaining 0.1 point improvement in terms of NIST score, in comparison to Google Translation.

Sengupta et al. in their paper "Word Sense Induction in Bengali Using Parallel Corpora and Distributional Semantics" report a work on sense induction in Bengali, a less-resourced language, based on distributional semantics and translation based context vectors learned from parallel corpora to improve the task performance. The proposed model, in both mono and cross-lingual settings, outperformed k- means in precision (P), recall (R) and F-scores.

Ameer et al. in their paper "Author Profiling for Age and Gender Using Combinations of Types of Features" present a Content-based approach for detection of author's traits (age group and gender) for same-genre author profiles. They use a different set of features including syntactic n-grams of part-of-speech tags, traditional n- grams of part-of-speech tags, the combination of word n-grams and combination of character n-grams, achieving the best accuracy of 0.496 and 0.734 for both traits, i.e., age group and gender respectively.

Gómez-Adorno et al. in their paper "A Convolutional Neural Network Approach for Gender and Language Variety Identification" present a convolutional neural network trained on word- and sentence-level embeddings architecture that can be successfully applied to gender and language variety identification on a relatively small corpus. Their experiments show that the deep learning approach outperforms a traditional machine learning approach on both tasks, when named entities are present in the corpus. However, when evaluating the performance of these approaches reducing all named entities to a single symbol "NE" to avoid topic-dependent features, the drop in accuracy is higher for the deep learning approach.

Álvarez-Carmona et al. in their paper "A Comparative Analysis of Distributional Term Representations for Author Profiling in Social Media" present a throughout analysis regarding the appropriateness of different distributional term representations (DTR) for the autor profiling task. They compare the performance of the DTRs against classic approaches including popular topic-based methods, achieving competitive results while providing meaningful interpretability.

Posadas-Durán et al. in their paper "Detection of Fake News in a new corpus for the Spanish Language" present a new resource to analyze and detect deceptive information that is present in a huge amount of news websites. They compiled a corpus of news in the Spanish language extracted from several websites that is annotated with two labels (real and fake) for automatic fake news detection. In their paper, the authors also present a style-based fake news detection method.

Guzman-Cabrera et al. in their paper "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 are considered for the experiments. The categorization experiments were performed using different domains of learning with promising results.

Sidorov et al. in their paper "Human interaction with shopping assistant robot in natural language" present a language-independent spoken dialog management module integrated into a human-robot interaction system. They adopt an algorithmic approach to dialog modeling, exemplyfing their approach using a mobile robot functioning as a shopping assistant.

López-Ramírez et al. in their paper "Geographical Aggregation of Microblog posts for LDA Topic Modeling" propose an aggregation strategy for geolocated Twitter posts based on a hierarchical definition of the regular activity patterns within a specific region. The results obtained show that the Geographical Aggregation performs similarly to hashtag aggregation in terms of Jensen-Shannon Divergence and outperforms other aggregation schemes in its ability to reproduce the original cluster labels.

Basak et al. in their paper "Short-Answer Grading Using Textual Entailment" present an approach for short-answer grading. They employ a large number of matching rules relying on recognizing entailment relation between dependency structures of the two answers. Comparison of the grades generated by the proposed method with given by human judges on a computer science dataset shows a quite promising maximum correlation of 0.627.

Mager et al. in their paper "Low-resource Neural Character-based Noisy Text Normalization" explore the state-of-the-art of the machine translation approach to normalize text under low-resource conditions. We also propose an auxiliary task for the sequence-to-sequence (seq2seq) neural architecture novel to the text normalization task, that improves the base seq2seq model up to 5%.

The second part of this volume contains 16 papers devoted to the knowledge engineering area and their general description follows.

Rodríguez-González et al. in their paper "Frequent Similar Pattern Mining using Non Boolean Similarity Functions" extend the similar frequent pattern mining by allowing the use of non Boolean similarity functions. Authors claim that the proposed algorithms obtain better patterns for classification than those patterns obtained by traditional frequent pattern miners, and miners using Boolean similarity functions.

Rodriguez-Torres et al. in their paper "Deterministic Oversampling Methods based on SMOTE" present SMOTE-D, a deterministic version of SMOTE, and propose new deterministic SMOTE-D-based versions of some of the most recent and successful SMOTE-based methods. Authors show that all proposed deterministic methods produce as good results as random methods, but they indicate that their proposals need to be applied just once.

Martínez-López et al. in their paper "Cellular Estimation Gaussian Algorithm for Continuous Domain" present a Cellular Gaussian Estimation Algorithm (CEGA) for solving continuous optimization problems. The experimental results showed that the present proposal reduces the number of evaluations of the fitness function in the search for optimums, maintaining its effectiveness in comparison to other algorithms of state-of-art using the same benchmark of continuous functions.

Tiwari et al. in their paper "An Intuitionistic Fuzzy-Rough Set Model and its application to Feature Selection" establish an intuitionistic fuzzy rough set model by combining intuitionistic fuzzy set and rough set, proposing a novel approach of feature selection derived from this model in order to generate an algorithm which is further applied to different benchmark data sets and compared with the existing fuzzy rough set based technique.

Pinilla-Buitrago et al. in their paper "Bag of k-Nearest Visual Words for Hieroglyph Retrieval" introduce a hieroglyph representation that takes into account the frequency of the visual words and the co-occurrence of visual word pairs. They address the problem of local descriptors similarity by replacing each local descriptor by its k-nearest visual words in the vocabulary, instead of just one visual word (the nearest). The proposed approach obtains better retrieval results than those obtained by using state of the art representations.

Martínez-Espinosa et al. in their paper "Generation of Raman images through spectral mappings" propose a practical approach to access and visualize relevant information on the spatial distribution of a given sample about its biochemical composition. They use a Raman spectroscopy technique to obtain spectral maps with specific spatial resolution. The results obtained by the authors suggest that the Raman spectroscopy imaging is a powerful tool for determining the biochemistry of organic and inorganic samples based on spectral scanning and thus determine compounds concentrations of medical interest.

Céspedes-Hernández et al. in their paper "A Methodology for Gestural Interaction Relying on User-defined Gestures Sets following a One-shot Learning Approach" propose a methodology for enabling the devel- opment of gesture-based applications, considering that accuracy and efficiency in recognition tasks must not be affected, and prioritizing the flexibility for allowing the use of gestures that are suitable for different user contexts through the exploration of user-defined gesture sets and Machine Learning techniques, and using a one-shot learning approach.

Camarillo-Abad et al. in their paper "A Basic Tactile Language to Support Leader-Follower Dancing" introduce a novel language that has been designed to guide users in leader-follower dances using tactile interaction among humans through haptic technologies. They find out that it is viable to successfully guide someone to follow dance through communication using a basic vibrotactile language.

Aparicio-Díaz et al. in their paper "Temporal Copy-Move Forgery Detection and Localization Using Block Correlation Matrix" propose a simple method to detect Copy-Move for both subregion and full-frame duplication. The method achieves high precision detecting dupli- cated regions, and the used correlation matrix shows to be versatile enough to detect frame duplication attacks as well as the specific duplicated frames.

Starostenko et al. in their paper "Real-time facial expression recognition using local appearance-based descriptors" aim to increment a recognition rate of approaches for unobtrusive face sensing and automatic interpretation of emotions. The proposed approach explores local scale invariant feature transform descriptors for extraction of face key points used for face detection, recognition and then for encoding facial deformations in terms of Ekmanś Facial Action Coding System (FACS).

Pérez-Espinosa et al. in their paper "Evaluation of quantitative and qualitative features for the acoustic analysis of domestic dogs' vocalizations" present a comparison between several acoustic characterization techniques in order to determine, qualitative and quantitative, their relevance in the classification of two aspects of the barking, which are the context in which they were generated and the identity of the dog that emitted the bark.

Herrera-Alcántara et al. in their paper "Inverse Formulas of Length Twelve Parameterized Orthogonal Wavelets" present inverse parameterizations of length 12 orthogonal wavelet filters in order to determine parameter values from filter coefficients. Authors conclude that the use of the inverse formulas accelerates the convergence and that parameterized filters provide better results as their length increases and achieve a better performance than standard filters.

Francisco-Valencia et al. in their paper A comparison between UCB and UCB-Tuned as selection policies in GGP" present a comparative analysis of two selection policies in the General Game Playing (GGP) context: Upper Confidence Bound (UCB) and Upper Confidence Bound Tuned (UCB-Tuned). The results show that UCB-Tuned is better when less than 100 simulations are used in MCTS; however, when 1000 simulations are used, both policies have similar performance.

Rodriguez et al. in their paper "Improving Data Collection in Complex Networks with Failure-Prone Agents via Local Marking" present an improvement to selected movement algorithms to collect data in complex networks in a faster way. The proposed improvement consists of local marks in nodes to avoid re-exploration combined with the previously proposed algorithms. Experiments were performed with different failures rates. Results show that there is a significant difference between the pheromone algorithm with and without local marks providing a higher robustness in data collection tasks in scenarios with a higher standard deviation in the betweenness centrality.

Krystian Jobczyk in his paper "Multi-Valued Deontic Halpern-Shoham Logic for Fuzzy Deontic-Temporal Expressions" introduces a new deontic Halpern-Shoham logic DeoHS for a representation of temporal deontic expressions, which introduce a kind of fuzziness by a gradable nature of their con- notations (obligations or permissions). The deontic- temporal expressions were derived from a unequally modified description of a deontic version of Traveling Salesman Problem.

Torres et al. in their paper "Reasoning with Preferences in Service Robots" present a a non-monotonic knowledge-base system, capable of expressing incomplete knowledge, updating defaults and exceptions dynamically, and handling multiple extensions. Authors describe the general principles underlying such protocols and their implementation through the SitLog programming language. Finally, they also show a demonstration scenario in which the robot Golem-III assists human users using such protocols and preferences stored in its non-monotonic knowledge-base service.

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 Advancements 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.

David Pinto
Faculty of Computer Science, BUAP
Puebla, Mexico

Vivek Singh
Department of Computer Science, BHU
Varanasi, India