ISSN 2007-9737

Introduction to the thematic issue on Natural Language Processing

This special issue of *Computación y Sistemas* presents a selection of papers on natural language processing and computational linguistics, along with eight regular papers.

Natural language processing is an area of artificial intelligence and its applications devoted to analysis and generation of data streams involved in human communication using language, such as English or Spanish, typically in the form of text or speech, as well as, in multimodal setting, associated facial expressions and body language. It embraces natural language processing and computational linguistics in a wider context of their real-life applications in a wide range of disciplines.

The typical tasks of natural language processing include machine translation, text classification, text summarization, information extraction, and sentiment analysis, while typical applications include opinion mining, human-computer interaction, and information retrieval, among many other tasks.

In its early days natural language processing technologies mainly followed linguistically-motivated symbolic processing path. However, in the last two decades machine-learning techniques prevail in natural language processing research and applications. In the last decade, deep-learning has revolutionized the field of machine learning, and natural language processing is on board with this new trend.

The present thematic issue includes twenty-three papers representative of different tasks, techniques, and applications of human language technologies, as well as eight regular papers on a selection of topics related with computer science.

Processing of language data begins with representing the text as some data structure. The correct data structure has crucial impact on the further analysis. The first two papers of the thematic issue are devoted to text representations, namely, to dealing with graphs as representation of a text.

Next, the final goal of text analysis (or analysis of other language data, such as speech) is representing the meaning conveyed by these data in the form of some semantic representation, in order to be able to reason logically on this contents and maintain meaningful communication in natural language. An important part of understanding the semantics of natural language, with numerous applications in many natural language processing tasks and applications, is determining semantic similarity between texts. The next three papers are devoted to semantics, natural language interfaces, and measuring semantic similarity between textual units.

In addition, identifying the structure of text is the main task of natural language processing. Traditionally, this task is performed at lexical, syntactic, and semantic levels, called morphological, syntactic, and semantic analysis, respectively. Syntactic analysis, or parsing, consists in identifying grammatical relationships between words within a sentence. It can be performed at different levels of detail: in chunking, only partial analysis is performed, thus identifying simple fragments of the whole syntactic structure of the sentence; in parsing, the complete structure is sought to be determined.

As an application area, sentiment analysis and opinion mining are very actively developed applications areas of natural language processing. They are underlying technologies for decision-support systems and business intelligence systems in industry and polities, which help businesses, political parties, and governmental bodies to better adapt their products and services to the needs of the customers. They are also underlying technologies behind recommender systems, which help the users to make informed decisions on acquiring suitable products and services or voting for

suitable political parties. Thus, these technologies contribute to building democracy in real time, in industry and political issues.

Some papers of this thematic issue illustrate how opinion mining research contributes to democracy in real time: the possibility for the citizens to influence governmental decisions at each moment, and not once in four years at the time of elections, and expressing direct opinion on a topic and not just voting for or against a party without being given an opportunity to specify the reasons for the choice made by the authors.

The social media is the primary source of information on the users' opinions. News analysis also contributes to better understanding of people's needs and opinions and constitutes an important information source for decision making, both at the side of industry (providers of products and services) and on the side of the users (consumers), that is, ordinary people.

Some papers in this issue are devoted to the analysis of social media and news analysis. Two of these papers are related with named entity recognition (NER), which seeks identifying certain (usually multiword) expressions (fragments of syntactic structure) as names of people, organizations, locations, etc.: for instance, *United States of America* is a name of certain country, as well as with the interpretation of such names entities.

As yet another application area, machine translation can be mentioned. It is one of historically first applications of natural language processing and a major motivation for its development soon after the first computers were built. The next three papers are devoted to areas related with machine translation of human languages.

Other papers exemplify such important areas of human language technologies as information extraction, text classification, text summarization, and text comprehension and automatic question answering.

Also, currently, by language analysis we usually understand text analysis. However, natural form of language is speech. Speech recognition is gaining increasing importance due to the growing amount of speech data available in the user-contributed contents in Internet. The next papers are devoted to the field of speech recognition and speech processing.

Natural language processing is often understood as using linguistic theory to help computers to deal with language data. However, the opposite relationship has proven to be useful as well: the use of computational analysis in linguistic research. The next paper, the last one in this special issue, illustrates such application of natural language processing.

Finally, this issue of the journal includes regular papers that do not belong to the thematic section on human language technologies. These papers span the topics of neural networks, computer networks, and quantum computing, and other topics.

The thematic issue includes seventeen carefully selected papers devoted to various aspects of the theory and applications of natural language processing, such as social network analysis, author profiling and plagiarism detection, sentiment analysis, recommender systems, information retrieval, text summarization, analysis of multiword expressions, parsing, language variety identification, word sense disambiguation, lexical resources, and medical applications of natural language processing.

Amal Htait, Sébastien Fournier, Patrice Bellot in their paper "Unsupervised Creation of Normalization Dictionaries for Micro-Blogs in Arabic, French and English" consider text normalization is a necessity to correct and make more sense of the micro-blogs messages,for information retrieval purposes. Unfortunately, tools and resources of text normalization are rarely shared. In this paper, an approach is presented based on an unsupervised method for text normalization using distributed representations of words, known also as "word embedding", applied on Arabic, French and English Languages.

Dwijen Rudrapal, Amitava Das in their paper "Semantic Role Labeling of English Tweets" present semantic role labeling (SRL), a task of defining the conceptual role to the arguments of predicate in a sentence. This is an important task for a wide range of tweet related applications associated with semantic information extraction. SRL is a challenging task due to the difficulties regarding general semantic roles for all predicates. It is more challenging for Social Media Text (SMT) where the nature of text is more casual. This paper presents an automatic SRL system for English tweets based on Sequential Minima Optimization (SMO) algorithm. Proposed system is evaluated through experiments and reports comparable performance with the prior state-of-the art SRL system.

Kunal Chakma, Amitava Das in their paper "A 5W1H Based Annotation Scheme for Semantic Role Labeling of English Tweets" continue with the topic of semantic Role Labeling (SRL), a well researched area of Natural Language Processing. State-of-the-art lexical resources have been developed for SRL on formal texts that involve a tedious annotation scheme and require linguistic expertise. The difficulties increase manifold when such complex annotation scheme is applied on tweets for identifying predicates and role arguments. In this paper, we present a simplified approach for annotation of English tweets for identification of predicates and corresponding semantic roles. For annotation purpose, we adopted the 5W1H (Who, What, When, Where, Why and How) concept which is widely used in journalism.

Bassem Bsir, Mounir Zrigui in their paper "Enhancing Deep Learning Gender Identification with Gated Recurrent Units Architecture in Social Text" present author profiling, inferring the authors' gender, age, native language, dialects or personality by examining his/her written text. This paper represent an extension of the recursive neural network that employs a variant of the Gated Recurrent Units (GRUs) architecture. Our study focuses on gender identification based on Arabic Twitter and Facebook texts by investigating the examined texts features. The introduced exploiting a model that applies a mixture of unsupervised and supervised techniques to learn word vectors capturing the words syntactics and semantics.

Adnen Mahmoud, Mounir Zrigui in their paper "Artificial Method for Building Monolingual Plagiarized Arabic Corpus" discuss plagiarism in textual documents, a widespread problem seen the large digital repository existing on the web. Moreover, it is difficult to make evaluation and comparison between solutions because of the lack of plagiarized resources in Arabic language publicly available. In this context, this paper describes automatic construction of a paraphrased corpus in order to deal with these issues and conduct our experiments, as follows: First, we collected a large corpus containing more than 12 million sentences from different resources. Then, we cleaned it up unnecessary data by applying a set of preprocessing techniques. After that, we used word2vec algorithm to create a vocabulary from the collected corpus.

Houssem Abdellaoui, Mounir Zrigui in their paper "Using Tweets and Emojis to Build TEAD: an Arabic Dataset for Sentiment Analysis" present a distant supervision algorithm for automatically collecting and labeling 'TEAD', a dataset for Arabic Sentiment Analysis (SA), using emojis and sentiment lexicons. The data was gathered from Twitter during the period between the 1 st of June and the 30 th of November 2017. Although the idea of using emojis to collect and label training data for SA, is not novel, getting this approach to work for Arabic dialect was very challenging. We ended up with more than 6 million tweets labeled as Positive, Negative or Neutral. We present the algorithm used to deal with mixed-content tweets (Modern Standard Arabic MSA and Dialect Arabic DA).

Tomáš Hercig, Peter Krejzl, Pavel Král in their paper "Stance and Sentiment in Czech" discuss

sentiment analysis, a wide area with great potential and many research directions. One direction is stance detection, which is somewhat similar to sentiment analysis. We supplement stance detection dataset with sentiment annotation and explore the similarities of these tasks. We show that stance detection and sentiment analysis can be mutually beneficial by using gold label for one task as features for the other task. We analysed the presence of target entities for stance detection in the dataset. We outperform the state-of-the-art results for stance detection in Czech and set new state-of-the-art results for the newly created sentiment analysis part of the extended dataset.

Nicolás Olivares, Luz María Vivanco, Alejandro Figueroa in their paper "The Big Five: Discovering Linguistic Characteristics that Typify Distinct Personality Traits across Yahoo! Answers Members" explain that in psychology, it is widely believed that there are five big factors that determine the different personality traits: Extraversion, Agreeableness, Conscientiousness and Neuroticism as well as Openness. In the last years, researchers have started to examine how these factors are manifested across several social networks like Facebook and Twitter. However, to the best of our knowledge, other kinds of social networks such as social/informational question-answering communities (e.g., Yahoo! Answers) have been left unexplored. Therefore, this work explores several predictive models to automatically recognize these factors across Yahoo! Answers members.

Daniel Villanueva, Miguel Lagares, Juan M. Gómez, Israel González in their paper "RESyS: Towards a Rule-based Recommender System based on Semantic Reasoning" say that the ability to be available and stay connected always for work or social issues has become a reality and a necessity for today's Information Society. Harnessing the potential of Semantic Technologies-based reasoning for intelligent redirection of voice calls and recommender systems has been gauged as a promising field to enhance the current voice phone calling experience. Such experience might be fostered by a disruption based on rule-based recommendation and inference leveraging current state of the art technology in smartphone apps or fixed line telecommunications standards to its full potential.

Amarnath Pathak, Partha Pakray, Alexander Gelbukh in their paper "A Formula Embedding Approach to Math Information Retrieval" claim that intricate math formulae, which majorly constitute the content of scientific documents, add to the complexity of scientific document retrieval. Although modifications in conventional indexing and search mechanisms have eased the complexity and exhibited notable performance, the formula embedding approach to scientific document retrieval sounds equally appealing and promising. Formula Embedding Module of the proposed system uses a Bit Position Information Table to transform math formulae, contained inside scientific documents, into binary formulae vectors.

Thi-Thanh Ha, Thanh-Chinh Nguyen, Kiem-Hieu Nguyen, Van-Chung Vu, Kim-Anh Nguyen in their paper "Unsupervised Sentence Embeddings for Answer Summarization in Non-factoid CQA" present a method for summarizing answers in Community Question Answering. We explore deep Auto-encoder and Long-short-term-memory Auto-encoder for sentence representation. The sentence representations are used to measure similarity in Maximal Marginal Relevance algorithm for extractive summarization. Experimental results on a benchmark dataset show that our unsupervised method achieves state-of-the-art performance while requiring no annotated data.

Zuzana Nevěřilová in her paper "Discovering Continuous Multi-word Expressions in Czech" discusses that multi-word expressions frequently cause incorrect annotations in corpora, since they often contain foreign words or syntactic

anomalies. In case of foreign material, the annotation quality depends on whether the correct language of the sequence is detected. In case of inter-lingual homographs, this problem becomes difficult. In the previous work, we created a dataset of Czech continuous multi-word expressions (MWEs). The candidates were discovered automatically from Czech web corpus considering their orthographic variability. The candidates were classified and annotated manually.

Luong Nguyen Thi, Linh Ha My, Huyen Nguyen Thi Minh, Phuong Le-Hong in their paper "Using BiLSTM in Dependency Parsing for Vietnamese" say that deep learning methods have achieved good results in dependency parsing for many natural languages. In this paper, we investigate the use of bidirectional long short-term memory network models for both transition-based and graph-based dependency parsing for the Vietnamese language. They also report our contribution in building a Vietnamese dependency treebank whose tagset conforms to the Universal Dependency schema.

Naim Terbeh, Mohsen Maraoui, Mounir Zrigui in their paper "Arabic Dialect Identification based on Probabilistic-Phonetic Modeling" explain that the identification of Arabic dialects is considered to be the first pre-processing component for any natural language processing problem. This task is useful for automatic translation, information retrieval, opinion mining and sentiment analysis. In this purpose, we propose a statistical approach based on the phonetic modeling to identify the correspondent Arabic dialect for each input acoustic signal. The main idea consists first, and for each dialect, in calculating a referenced phonetic model. Second, for every input audio signal, we calculate an appropriate phonetic model. Third, we compare this latter to all referenced Arabic dialect models.

Daniil Alexeyevsky in his paper "Word Sense Disambiguation Features for Taxonomy Extraction" claims that many NLP tasks, such as fact extraction, coreference resolution etc, rely on existing lexical taxonomies or ontologies. One of the possible approaches to create a lexical taxonomy is to extract taxonomic relations from a monolingual dictionary or encyclopedia: a semi-formalized resource designed to contain many relations of this kind. Word-sense disambiguation (WSD) is a mandatory tool for such approaches. The quality of the extracted taxonomy greatly depends on WSD results.

Luz Marina Sierra Martínez, Carlos Alberto Cobos, Juan Carlos Corrales Muñoz, Tulio Rojas Curieux, Enrique Herrera-Viedma, Diego Hernán Peluffo-Ordóñez in their paper "Building a Nasa Yuwe Language Corpus and Tagging with a Metaheuristic Approach" discuss Nasa Yuwe, the language of the Nasa indigenous community in Colombia. It is currently threatened with extinction. In this regard, a range of computer science solutions have been developed to the teaching and revitalization of the language. One of the most suitable approaches is the construction of a Part-Of-Speech Tagging (POST), which encourages the analysis and advanced processing of the language. Nevertheless, for Nasa Yuwe no tagged corpus exists, neither is there a POS Tagger and no related works have been reported.

Anayeli Paulino, Gerardo Sierra, Laura Hernández-Domínguez, Iria da Cunha, Gemma Bel-Enguix in their paper "Rhetorical Relations in the Speech of Alzheimer's Patients and Healthy Elderly Subjects: An Approach from the RST" explain that the study is aimed to extract discourse relations patterns in conversational speech of subjects with Alzheimer's Disease (AD) and adults with healthy aging processes using the Rhetorical Structure Theory (RST). By means of the RST, they analyzed semi-structured interviews of native Spanish speakers. Seven subjects were in the mild, moderate or advanced stages of AD, and 6 were cognitively intact individuals. The procedure involved the segmentation of each conversational discourse into Semantic Dialog Units (SDUs), the labeling of

their rhetorical relations and the construction of tree diagrams.

Apart from the papers devoted to the main topic of this thematic issue, which is natural language processing, the issue includes eight regular papers on diverse topics, which include medical, public security, and technological imaging applications, optimization problems, educational applications, as well as intelligent applications to computer networks and to software development technologies.

Ratishchandra Huidrom, Yambem Jina Chanu, Khumanthem Manglem Singh from India in their paper "Automated Lung Segmentation on Computed Tomography Image for the Diagnosis of Lung Cancer" describe novel medical image processing techniques used in computed tomography-based diagnosis of lung cancer. The problem the authors address is that the standard image processing techniques do not allow proper processing of the images of the juxta-pleural modules, since juxta-pleural modules look on tomographic images very similar to other tissues. The authors show that the method that they proposed does allow processing of the juxta-pleural modules in lung cancer diagnosis.

Ricardo Acevedo-Ávila, Miguel González-Mendoza, Andrés Garcia-Garcia from Mexico in their paper "A Statistical Background Modeling Algorithm for Real-Time Pixel Classification" apply artificial intelligence techniques for image classification to the practical task of detecting unattended objects in images with fixed background. The main development goal of their algorithm is a low-resource architecture suitable for real-time processing in surveillance devices. The task is important in public security scenarios, for prevention of terrorist attacks. The authors show that their algorithm can efficiently classify the pixels in low-resolution images into shadows, mid-tones, highlights and foreground pixels and detect unusual objects in the images in real time and at the standard frame rate, with very low resource consumption.

Valentín Osuna-Enciso, J. Israel Espinoza-Haro, Diego Oliva, Irán F. Hernández-Ahuactzi from Mexico in their paper "Offshore Wind Farm Layout Optimization via Differential Evolution" optimize the placement of eolic generators on a given terrain using artificial intelligence techniques. The use of eolic energy is an important part of the shift towards clean, renewable energy sources, available in most countries and proven to be efficient in different environments. The authors use differential evolution as the underlying optimization technique for optimal distribution of eolic generators. The novelty of the approach in this paper is a mixture of different configurations used for the differential evolution algorithm by combining the five standard types in the solution of one problem. The authors also show which of the five main known types of the differential evolution algorithm performs best for the considered task.

Manikandan Rajagopal, M. Balasubramanian, S. Palanivel from India in their paper "An Efficient Framework to Detect Cracks in Rail Tracks Using Neural Network Classifier" apply the effective neural network classification technique to railway management, namely, to the detection of defects in rail tracks. The technique is applied to the images obtained from the echo image display devices or from semi-conductor magnetism sensors. Then image processing techniques are applied to the obtained signals, and features are extracted, which are fed to a neural network classifier to detect images that exhibit cracks. The proposed method performs with the accuracy almost at the level of manual examination of the images.

Elizabeth Suescún Monsalve, Mauricio Toro, Raúl Mazo, David Velasquez, Paola Vallejo, Juan F. Cardona, Rafael Rincón, Vera Maria Werneck, Julio Cesar Sampaio do Prado Leite from Colombia, France, and Brazil in their extensive paper "SimulES-W: A Collaborative Game to Improve Software Engineering Teaching" present a new tool for teaching

software engineering to undergraduate students using game-based learning methodology. The tool implements a collaborative game in a highly customizable way. It is based on software cases found in real, practical projects, which makes it especially instructive and interesting for the students. Basing on extensive experiments, the authors show that their tool has positive impact on the learning process.

Maribell Sacanamboy, Freddy Bolaños, Alvaro Bernal from Colombia in their paper "Adaptive Algorithm Based on Renyi's Entropy for Task Mapping in a Hierarchical Wireless Network-on-Chip Architecture" develop an important improvement for the convergence time of the Population-Based Incremental Learning optimization algorithm, which is a popular machine-learning technique, a variation of a genetic algorithm. The improvement is based on the information theory. The authors show the effectiveness of the proposed technique by applying it to the optimization of task mapping of applications in a hierarchical wireless network-on-chip architecture. The proposed technique led to slight improvement of the convergence times without degrading the quality of the obtained solutions.

Sergio Sánchez-Reyes, Mario E. Rivero-Angeles, Noé Torres-Cruz from Mexico in their paper "Teletraffic Analysis for VoIP Services in WLAN Systems with Handoff Capabilities" study, with both analytical and simulation methods, the effects of mobility on voice-over-IP technology. The authors also propose a novel fluid model for statistical estimate of the number of active and inactive users with a given set of parameters in the voice-over-IP system. Such systems currently find increasingly important practical applications, gradually replacing more traditional, and much more costly, means of communication such as ordinary phone calls.

Maximiliano A. Mascheroni, Emanuel Irrazábal from Argentina in their paper "Continuous Testing and Solutions for Testing Problems in Continuous Delivery: A Systematic Literature Review" argue, on the basis of a detailed analysis of more than fifty published papers, that testing is an important integral part of the continuous delivery software development methodology. This methodology consists in organizing the software development cycle in such a way that the software can be delivered at any moment, with the features available by that moment. The authors discuss specific issues in applying the testing procedures to the continuous delivery software development methodology.

Alexander Gelbukh (Guest editor)

Member of the Mexican Academy of Sciences;
Head, Natural Language Processing Laboratory,
Centro de Investigación en Computación,
Instituto Politécnico Nacional, Mexico