ISSN 2007-9737

Introduction to the Special Issue on Human Language Technologies

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

Natural language processing is an area of artificial intelligence 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. 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.

While in its early days natural language processing technologies mainly followed linguistically-motivated symbolic processing path, 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.

This special issue includes twenty-three papers representative of different tasks, techniques, and applications of natural language processing, as well as three regular papers.

Any 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 text representation.

M. G. Sohrab et al. from Japan in their paper "EDGE2VEC: Edge Representations for Large-Scale Scalable Hierarchical Learning" introduce a method for embedding of edges of the edges of an arbitrary graph or network into a vector space in a statistically (semantically) meaningful way. They illustrate their method on representation of words and documents, using the well-known `word2vec` and `para2vec` embeddings. In addition, the authors present a novel variant of support vector machine classifier suitable for working with large graphs.

E. Castillo et al. from Mexico in their paper "Text Analysis Using Different Graph-Based Representations" continue the topic of graph-based techniques in natural language processing with a survey of graph-based representations useful for text representation tasks, especially representations of documents using co-occurrence graphs, which they show to be richer than traditional vector space-based representations. They propose various graph-based text representations and discuss their use for sentiment analysis, author profiling, and authorship attribution and verification, among other applications.

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 texts.

M. Duží from Czechia in her paper "Property Modifiers and Intensional Essentialism" define

intersective, subsective, and privative property modifiers in semantic representation of language data, taking special care to avoid logical paradoxes typical for other definitions, such as the conclusion that if little things are smaller than large things, then a little elephant should be smaller than a large mouse.

B. Abu Shawar from Jordan in her paper "Integrating CALL Systems with Chatbots as Conversational Partners" discusses the perspectives of using artificial intelligent conversational agents as conversational partners in the process of computer-assisted language learning for the learners to practice the foreign language being learned. She shows that learning results improve when the students use a chatbot as a conversational partner.

W. Wali et al. from Tunisia in their paper "Sentence Similarity Computation based on WordNet and VerbNet" present a novel method for computing semantic similarity between sentences, which is a key component in a great number of natural language processing tasks and applications, such as information retrieval, automatic text summarization, question answering, and plagiarism detection, among others. Their method consists in a combination of three different types of similarity measures: lexical, lexical-semantic, and syntactic.

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.

B. Indig from Hungary in his paper "Less is More, More or Less... Finding the Optimal Threshold for Lexicalization in Chunking" addresses the problem of lexicalization in chunking, which consists in considering the word itself as a part the type (label) of the syntactic chunk. In his previous joint work with I. Endrédy published in the proceedings of the CICLing 2016 conference, he has suggested that only sufficiently frequent words should be used in this process, in order to reduce the number of labels and thus speedup the chunking process (mild lexicalization). In this paper, he investigates the question of how frequent the word is to be to consider it sufficiently frequent to be used in mild lexicalization. The method he developed for taking into account the frequencies of words can also be useful for part-of-speech tagging (POS tagging) and NER.

N. Ghezaiel Hammouda and K. Haddar from Tunisia in their paper "Parsing Arabic nominal sentences with transducers to annotate corpora" develop a set of rules for parsing some types of sentences, especially nominal sentences, characteristic for Arabic language. They also present a parsing approach for Arabic language based on the use of transducers.

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 politics.

In particular, the first paper of the next group of papers paper illustrates 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 their choice.

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).

The following five papers 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.

R. Satapathy et al. from Singapore in their paper "Subjectivity Detection in Nuclear Energy Tweets" develop novel methods for classification of tweets related with the hot topic of nuclear energy into subjective (opinionated, that is, those that express opinion, sentiment, or emotions) versus objective (those that express facts independently from the author's personal opinion on them). This research provides an opportunity for the governments to make informed decisions on the controversial topic of nuclear energy development taking into account public opinions.

P. Radhakrishnan et al. from India in their paper "SNEIT: Salient Named Entity Identification in Tweets" address the problem of identification of the "object" of the opinion in social media posts: what about the opinion is given. Often the opinion is about a named entity: person, organization, place, etc., mentioned in the post. However, a post, such as a tweet, can mention several named entities or other noun phrases, of which the authors identify the main, central one. An interesting idea in this paper is using images accompanying some tweets to compile a ground truth dataset for salient named entity recognition. The authors also present their publicly available dataset for the task.

S. Banerjee et al. from India and Spain in their paper "Named Entity Recognition on Code-Mixed Cross-Script Social Media Content" address the difficulties in named entity recognition task applied to highly informal code-mixed texts, i.e., texts where words and expressions in different languages and different transliterations are intermixed, along with extensive use of informal words, ad hoc abbreviations, spelling errors and spelling variants, as well as emoticons and other means of expression emotions (such as using multiple letters and punctuation signs, as in *its soooo goooood !!!*). Such texts are characteristic of social media, social networks, and other user-contributed content in Internet. The authors analyze various machine learning-based approaches to the task, as well as introduce a new dataset of Bengali-English code-mixed dataset and domain-specific taxonomies for the NER task.

S. Ghosh et al. from USA and India in their paper "Complexity Metric for Code-Mixed Social Media Text" continue the topic of code-mixed social data, proposing a method for measuring how much code-mixing in the data affects its understandability. The proposed complexity measure features a number of improvements over previously existing code-mixed text complexity measures.

A. Balali et al. from Iran in their paper "A supervised method to predict the popularity of news articles" describe a machine-learning method for predicting the importance and thus impact of a news from a news feed. As a measure of importance, they used the number of comments that the users leave on the news page on the newsfeed website. The authors use a variety of manually engineered features. They

report results superior to those of the baseline methods.

Machine translation is yet another important application of natural language processing. 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.

S. Tripathi and V. Kansal from India in their paper "Using Linguistic Knowledge for Machine Translation Evaluation with Hindi as a Target Language" propose a new approach to evaluation of machine translation results. Their evaluation measure consists in minimizing a cost function based on linguistic knowledge. While this measure is much more difficult to implement, the authors argue that it better corresponds to human judgements on translation quality than existing measures such as BLUE in the case of linguistically rich, free word order languages such as Hindi.

K. K. Arora and S. S. Agrawal from India in their paper "Pre-Processing of English-Hindi Corpus for Statistical Machine Translation" argue for the need of curation of parallel bilingual corpora used to train machine translation systems. They experimentally show that simple curation steps, such as reducing words to lowercase (since their target language is Hindi, which does not use capital letters) or labeling named entities, simplify training of machine-translation systems and improve the results of such training.

M. W. A. Kesiman et al. from France and Indonesia in their paper "Knowledge Representation and Phonological Rules for the Automatic Transliteration of Balinese Script on Palm Leaf Manuscript" describe a rule-based system for Roman transliteration of the ancient script used on palm leaf manuscripts in Bali, Indonesia. This script is one of the most complex scripts used to write in South-East Asian languages, very difficult for transliteration. Correct transliteration of the Balinese palm leaf inscription will make this important source of historical information more easily available to researchers all over the world.

The next six papers exemplify such important applications of natural language processing as information extraction, text classification, text summarization, and text comprehension and automatic question answering (in fact, the paper by S. Banerjee et al. described above is also devoted to question answering).

S. Gopalan and S. Lalitha Devi from India in their paper "Cause and Effect Extraction from Biomedical Corpus" apply discourse analysis to automatic extraction of causal relationships in biomedical domain. Given the exponential growth of available research literature in biomedical domain far beyond the human ability of reading even a tiny part of relevant article, its automatic processing, and, specifically, automatic extraction of properties of objects of different nature, is of crucial importance for turning the huge amount of available data into useful information.

Y. Zhou et al. from China in their paper "Hybrid Attention Networks for Chinese Short Text Classification" use a combination of word- and character-level features in a deep-learning setting for classification of short Chinese texts, with attention mechanism applied to word-level and character-level features. The importance of character-level features for Chinese language analysis is emphasized by the fact that Chinese writing system does not use a space to separate characters of different words, while existing algorithms for Chinese text segmentation are imperfect.

W. Waheeb and R. Ghazali from Malaysia and Yemen in their paper "Content-based SMS Classification: Statistical Analysis for the Relationship between Features Size and Classification Performance" continue the topic of classification of short messages. They show that feature selection step is necessary for such classification: in their experiments, reducing the feature set to half gave optimal performance of

the classifier. Generally, feature selection reduces noise and prevents the model from overfitting. In addition, reducing the feature set has positive impact on memory use and classification time.

R. Verma and D. Lee from USA in their paper "Extractive Summarization: Limits, Compression, Generalized Model and Heuristics" define the concept of compressibility of a document and present a new model of the summarization task, in which they combine various types of summarization, namely, abstractive vs. extractive summarization, single document vs. multi-document summarization, and syntactic vs. semantic summarization. They also present interesting results on the limits of extractive summarization.

X. Wang et al. from Japan in their paper "Learning to Answer Questions by Understanding Using Entity-Based Memory Network" introduce a novel neural architecture for entity-based text comprehension, question answering, and other tasks that can be cast as special types of the question answering task. The architecture maintains a pool of entities mentioned in the text, along with their state, which is updated as the input text is read by the network. The question answering task is solved by retrieving from the pool the entities matching the question and interpreting their state.

J. Smailović et al. from Slovenia in their paper "Automatic Analysis of Annual Financial Reports: A Case Study" determine a range of stylistic and lexical features in financial reports of companies that predict financial state of the company (even if it is not directly indicated in the report): for example, they found that the presence of necessity or cognition words in the financial reports indicate possible financial failure of the company. These results can be used for better decision-making in financial domain.

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 paper is devoted to the field of speech recognition.

W.-H. Lai et al. from Taiwan in their paper "Post-Processing for the Mask of Computational Auditory Scene Analysis in Monaural Speech Segregation" addresses the task of separating speech fragments from noise fragments in the audio data. This task is both very important for speech analysis and quite difficult technically. The authors build a model of human auditory system, which helps them to identify features useful for the task machine-learning setting and obtain results superior than existing models.

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.

A. K. Fischer et al. from Germany in their paper "Beyond Pairwise Similarity: Quantifying and Characterizing Linguistic Similarity between Groups of Languages by MDL" present an approach to linguistic classification of languages and measuring similarity between languages and whole groups of related language using only information-theoretic concepts, with no pre-defined linguistic knowledge and with no fine-tuning of the parameters of the model. On the example of 13 Slavic languages they show that their approach accurately predicts the linguistic classification of these well-studied language group.

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

S. Valadez-Godínez et al. from Mexico in their paper "How the Accuracy and Computational Cost of Spiking Neuron Simulation are Affected by the Time Span and Firing Rate" study the effect of two important parameters of spiking neural network simulation, namely, the time span and firing rate, on the effectiveness and efficiency of the simulation. In particular, they show that, contrary to a widely known statement, and in accordance with observations of some other authors, continuous models of spiking neural networks, such as Hodgkin-Huxley model, show exhibit qualities than the Izhikevich discontinuous model.

F. Fakhfakh et al. from Tunisia and France in their paper "Proving Distributed Coloring of Forests in Dynamic Networks" propose an approach for specifying and proving distributed algorithms in a forest topology, such as dynamic computer network where devices can frequently appear and disappear and links can be frequently established and dropped. As a case study, the authors analyze a tree-coloring algorithm and propose a pattern for its development using their methodology.

M. Ávila and J. B. Elizalde-Salas from Mexico in their paper "Remedies for the Inconsistences in the Times of Execution of the Unsorted Database Search Algorithm within the Wave Approach" explain some seeming paradoxes in the behavior of the Grover's algorithm for unsorted database search in the context of quantum computing, an extremely efficient quantum search algorithm, which is one of the most important developments in the field of applications of quantum computing.

Alexander Gelbukh (Guest editor)