Editorial

It is my pleasure to introduce to the readers a special issue of the journal Computación y Sistemas devoted to computational linguistics and applications of linguistic to natural language processing. Computational linguistics is a very actively developed research area at the intersection of linguistics, applied linguistics, artificial intelligence, and computer science. Its main goal is two-fold. On the one hand, it investigates human language using very powerful modern tool, the computer, which frees the linguist or lexicographer from routine work and helps the researchers analyze statistical evidence about language that before the computer era would consume centuries of manual work. On the other hand, this area is a branch of applied linguistics aimed to the application of linguistic knowledge in order to help computers to “understand” human language, i.e., in order to develop computer programs capable of solving practical tasks that involve human language, such as automatic translation, information retrieval, or text mining.

This issue presents to the reader twenty-two papers devoted to a wide variety of topics related with computational linguistics, such as measures of semantic relatedness and identification idiomaticity of collocations, analysis of specific low-resource languages, sentiment analysis and text-based social network analysis, automatic translation and multilingualism, knowledge extraction, automatic question answering, named entity recognition, text genre classification, language learning support, and authoring aid tools.

E.-P. Soriano-Morales et al. (France) study the interplay of syntax and semantics in the context of applying linguistic knowledge in determining semantic relatedness of words. They propose a hybrid linguistic network that combines lexical and syntactic information. Their proposal addresses certain shortcomings existing in current state of the art graph-based linguistic models. Based on the distributional hypothesis, the authors leverage lexical and syntactical language information while introducing a novel method to solve the task of word sense disambiguation and induction and thus demonstrate the practicality of their proposal. Their experiments showed improved results as compared with similar existing approaches.

O. Kolesnikova (Mexico) continues the topic of word distribution and the associated semantics with a detailed survey of association measures of word co-occurrence useful for collocation detection in texts. This information is very important for semantic interpretation of collocations, to distinguish free word combinations from collocations that exhibit varying degrees of idiomaticity: idiomatic collocation exhibit high degree of cohesion between their components. The survey is organized by the underlying mathematical models of association measure used for classification of collocations.

I. Z. Batyrshin et al. (Mexico and Russia) further develop the topic of association and similarity measures, addressing general mathematical properties of such measures, which can be used in linguistic, ecological, economical, and medical applications, among other. They propose models for visualization of such measures in three-dimensional space, illustrating their method at a wide variety of association and similarity measures known in the literature. Visualization is a very important tool that helps researchers to understand better the phenomenon at hand, develop their intuition about it, and, in particular, choose the association measures suitable for a given task. The authors propose a new parametric family of measures generalizing these two families and giving the possibility to construct similarity measures occupying the intermediate position between them.

L. M. Sierra Martínez et al. (Colombia) describe a tokenizer for the Páez language, with the autonym *nasa yuwe*, “language of the people,” spoken by about 80 thousand indigenous inhabitants in Colombia, among which about 40 thousand are monolingual Páez speakers. Providing computational tools for the Páez language is very important for social and cultural adaptation of the Páez people, the second largest Colombian native community, as well as for preservation of their culture. On the other hand,

ISSN 1405-5546
doi: 10.13053/CyS-20-3-2474

computational analysis of the Páez language helps the linguistic study of its structure.

T. Hercig et al. (Czech Republic) present unsupervised methods for aspect-based sentiment analysis in Czech. Sentiment analysis is a very popular area of research, which benefits businesses with better income, benefits consumers with better quality of life via better products and services offered by the businesses, benefits the governments and political parties with better understanding of the citizen's opinions, and benefits democracy with real-time feedback of the citizens. The authors develop methods for discovery of latent semantics for aspect-based sentiment analysis, as well as present two publicly available corpora that they developed for this task.

K. Přikrylová et al. (Czech Republic) further extend the topic of sentiment analysis, also for the Czech language, with the analysis of the contribution of adjectives to the polarity—positive or negative sentiment—of sentences. They study the role of conjunctions in the effect of the polarity of adjectives, and compare the obtained results for the Czech and English languages. The authors pay special attention to exceptions from the general rules and special cases, illustrating their findings with examples extracted from a large corpus of Czech.

M. Apishev et al. (Russia) address a topic closely related with sentiment analysis, namely, the topic of social network analysis, specifically, the analysis of texts that people write in social networks. For this, they apply a recent approach to topic modeling called additive regularization of topic models. Using this method of topic modeling, they mine ethnic-related content from Russian-language blogs. They show using human evaluation that the new approach outperforms on this task the traditional approach called latent Dirichlet allocation by finding topics that are more relevant and have higher or comparable quality.

C. Mărănduc et al. (Romania) continue the topic of social network analysis with exploration of discourse structure in social network conversations. They experiment on a subset of the Dependency Treebank for Romanian that represents chat conversations. They show that application of methodologies and tools known for other languages to Romanian gives satisfactory results. Thanks to their effort, the Dependency Treebank for Romanian is augmented with types of linguistic information other than purely syntactic; in this case, discourse-level information is added.

K.-Y. Jeong and K.-S. Lee (Korea) further develop the topic of social network analysis and sentiment analysis. They analyze the texts in the Twitter social network and microblogging system to determine the type of followers of an influential user. Such follows can support the user they follow or can have negative attitude towards this user. The authors use the sentiment orientation of the texts written by Twitter users to identify whether they support the users they follow or disagree with them. They use clustering techniques with the latent Dirichlet allocation and perform sentiment analysis of such clusters.

K. Chakma and A. Das (India) present another research focused on Twitter. They address multilingual setting; moreover, code mixed setting. Code mixing is a phenomenon of using words of different languages in the same text or even in the same sentence. This is a very common phenomenon in many countries; for example, in India people frequently mix in informal communication their local language with regional language such as Bengali, national language such as Hindi, and English, especially because they often use Latin characters on their devices. Dealing with such texts faces many specific challenges in comparison with dealing with traditional, monolingual, well-formed texts.

P. Bhattacharya et al. (India) address yet another multilingual scenario: when the information retrieval query is given in one language while the document collection is in another language. This scenario is important for the users not able to formulate the query in English, but able to understand the retrieved documents written in English. The authors use automatic translation techniques to translate the query from Hindi to English. For this, they use modern neural network-based representation of distributional semantics called word embeddings, with which they greatly improve the quality of retrieval results.

L. Jakubina and P. Langlais (Canada) consider the general automatic translation task. A very important subtask in automatic translation is the compilation of a bilingual dictionary directly from the data. Typically, parallel corpora are used for this: the same texts in two languages. However, availability of such corpora is very limited. On the other hand, there exist large comparable corpora, such as Wikipedia, in which similar context, but not exactly the same text, is available in different languages. The authors analyze the feasibility of extracted translation equivalents from such corpora, which is a much more challenging task that extracting translation pairs from traditional parallel corpora.

Z. Yu et al. (China and France) also apply neural network-based technique for semantic analysis of texts, namely, for the task of information extraction for populating a knowledge base, or ontology. This task allows representing and organizing knowledge found in the vast amounts of available texts so that it can be used by computers for practical applications. In a way, this is language understanding, the final goal of automatic language processing. The authors use a state-of-the-art deep neural-network architecture to learn new entities from unstructured texts, obtaining excellent performance.

J.-X. Huang et al. (Korea) continue the topic of information extraction from unstructured texts for ontology construction. They use machine-learning techniques to extract relations from Wikipedia texts. They employ both probabilistic and semantic relatedness features, as well as other linguistic information. Surprisingly, they observed that surface information, such as words and part-of-speech tags, could outperform features based on deep syntactic information. Their approach can distinguish estimate reliability of extracted relation candidates, so that reliable candidates can be accepted for populating the knowledge base without human verification.

S. K. Dash et al. (India and Mexico) describe a methodology for constructing images from verbal descriptions of medical physiotherapy procedures. Visual representation of the documents is more understandable for the end users. It is especially important for multilingual countries, such as India, where very large groups of population have no language in common, and, in particular, do not understand English. The role of natural language processing techniques in this methodology consists in providing semantic and spatial information extracted from the texts.

N. Othman and R. Faiz (Tunisia) present a technique for the most important task in automatic question answering. Typically, given a question, a question-answering algorithm first identifies passages in the documents from the document base that are likely to contain the answer; then these passages are ranked according to the estimated probability to contain the correct answer, and finally the answer is extracted from the top-ranked passages. In this paper, the authors address the first two steps in this process using state-of-the-art machine-learning methods applied to the dependency syntactic structure of sentences, as well as various lexical, syntactic, and semantic similarity measures.

S. Dandapat and A. Way (India and Ireland) improve the accuracy of the named entity recognition task in a resource-poor language, Hindi, by involving cross-lingual information. Named entity recognition consists in identification single- or multiword expressions that are names of persons, organizations, countries, etc. Typically, special dictionaries are used for this. However, in the case of resource-poor languages, such dictionaries are not available. To solve this problem, the authors use an automatic translation system to translate Hindi sentences into English, with a special procedure for word alignment, and use existing English named entity recognizer to train a statistical classifier for Hindi.

J. C. Ross et al. (India) introduce yet another approach to the task of named entity recognition: the first approach that proves viability of the use of information collected from the web for named entity class identification. They use a search engine-based approach that acquires context from the web for an entity and performs named entity class identification using this information; further improvement is achieved with hierarchical classification. Thy experiment with texts in a specific thematic domain: discussions of Indian classical music; the possible semantic classes of the recognized named entities include, apart from the usual classes such as personal names, musical terms such as song, instrument, and music concept.

B. G. Patra et al. (India) also address musical domain, now for a sentiment analysis task: they identify mood for Hindi songs. The task has practical applications in music information retrieval. To date, there have been very few works dedicated to sentiment analysis of Indian music; existing works have been mostly based on only audio data. The authors introduce a mood taxonomy and describe a methodology for compiling a multimodal dataset that includes both audio and lyrics for Hindi songs. They show that only audio analysis is not sufficient for accurate detection of mood in Indian songs, while analysis of lyrics significantly improves the results.

A. R. Nabhan and K. Shaalan (Egypt, USA, UAE, and UK) leverage graph-based models of text for classification of text genre. Unlike traditional methods for this task, the proposed method reveal important macro-structural features embedded in text. Representing text in the form of a word graph enables analysis and identification of important topological features useful in text genre classification. The graph features employed by the authors include clustering coefficients, centralization, diameter, and average path lengths for eight text genres. The authors identify patterns that vary from one genre or sub-genre to another according to the style of the texts.

C. Mi et al. (China) introduce a model that predicts pronunciation of Chinese characters by the order of the strokes. Such model is of practical importance for learners of Chinese, which is considered a language very difficult to learn for foreigners, especially because of its complicated writing system. The authors use translation techniques to convert the order of strokes in Pinyin letter sequences, which are easier to interpret phonetically. In their practical implementation, they use strategies tolerant to errors, obtaining promising results.

I. Vargas-Campos and F. Alva-Manchego (Peru) describe an authoring tool for scientific writing in Spanish. Their tool identifies the structure of the abstract of a scientific work, such as article or thesis, and compares it with the guidelines for good scientific writing. Sentences of the abstract are classified by six rhetorical categories: background, gap, purpose, methodology, result, and conclusion. The tool warns the writers of missing components. The authors also present a manually annotated corpus of abstracts of computer science theses. Their work probably can also be used in automatic summarization of scientific texts to ensure correct and complete structure of the generated abstracts.

This special issue of the journal will be useful for researchers and students interested in computational linguistics and its applications to natural language processing.

Dr. Alexander Gelbukh

Guest Editor

Research Professor and Head,
Natural Language and Text
Processing Laboratory, CIC,
Instituto Politécnico Nacional;
Member of Mexican Academy of Sciences