Editorial

It is my pleasure to present to the readers of *Computación y Sistemas* a special issue on computational linguistics and natural language processing.

Language and verbal communication plays crucial role in human society, our everyday life, and thinking and reasoning of an individual. The main treasure of the humankind is knowledge, and this treasure is represented in the form of texts: books, electronic texts in Internet, etc. Its efficient management is a necessary condition for the functioning of our modern information society.

Natural language processing is an area of artificial intelligence devoted to the development of methods for efficient and accurate analysis, understanding, translating, and generation of human language. Its most important applications are Internet search, automatic translation, and human-computer interaction.

As in any scientific discipline, researchers in this area formulate hypotheses about the properties of human language and performance of specific algorithm and design experiments to support or refute such hypotheses. Their research results in both better understanding of the regularities and functioning of human language and in novel techniques for practical applications.

This special issue contains 15 carefully selected papers that report the most recent advances in various areas of this fascinating discipline.

Zahurul Islam and Alexander Mehler from Germany consider evaluation of readability of a text. Automatic evaluation of readability is important for authoring well-written and easy to understand documents. Using automatic spelling and grammar checkers (such as the one that is part of Microsoft Word) is common practice nowadays, so that it is increasingly uncommon to see texts with spelling or grammatical errors even written by people of low education level who don't know, or don't care, how to write correctly. It is still common, however, to see documents difficult to understand, for example, full of very long and complicates sentences and paragraphs or overall poorly structured. Automatic readability checkers is a next step towards clean and well-written documents to be authored by ordinary people. Islam and Mehler show that simple purely statistical, information-theoretic measures of readability perform for the task with quality comparable with resource-demanding and slow linguistic analysis.

Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni from Italy show how to improve dependency syntactic analysis. Syntactic analysis consists in revealing the structure of each sentence: determining its subject, predicate, object, etc.). It is a common internal task underlying many applications dealing with text and language. During last half century the most common framework for such analysis was constituency parsing. Dependency parsing is a much more informative kind of analysis; however, it is much more difficult to perform. In the recent decade it gains more and more attention of researchers in both theoretical and applied perspectives; hence the importance of its improvement. Traditionally, resource-demanding supervised learning was used to construct parsers. Dell'Orletta et al. show how to use much less expensive unsupervised learning for this goal.

Utpal Sikdar, Asif Ekbal and Sriparna Saha from India and Olga Uryupina and Massimo Poesio from Italy deal with anaphora resolution: the process of detecting what was meant by a general word, such as, for example, a pronoun: *it*—what specifically? This is important for all applications that involve some degree of text understanding, for example, for opinion mining: "I bought a new camera with my credit card and I like it"; what does she like? While such systems exist for English, building a system for each language is an expensive task. Sikdar et al. show how to adapt a system originally built for English to a very different language, Bengali in this case.

The next two papers are devoted to cost-effective development of very large dictionaries. Dictionaries are the heart of most systems dealing with language. Writing a very large dictionary for a realistic-scale language processing system is a too slow and costly task for a linguist or a team of linguists. This leads researcher to exploration of alternative mechanisms for their construction.

Manel Zarrouk, Mathieu Lafourcade, and Alain Jouber from France discuss the construction and use of large dictionaries built with collaboration effort of very many contributors. Inspired by the triumph of Wikipedia's authoring model including Wiktionary, recently the researchers successfully explore the possibilities of building very large dictionaries by means of small contributions of a huge number of non-expert users. While this yields very large and rapidly growing body of data, the lower quality of the obtained resource is its possible downside. Zarrouk et al. discuss how to deal with the varying quality and inconsistencies inevitably present in such very large crowd-sourced resources.

Ajay Dubey and Vasudeva Varma from India continue the topic of generation of very large dictionaries with little effort. Specifically, they extract a very large bilingual dictionary from existing Wikipedia data. They show how to improve the quality of the generated dictionary using the information on the internal structure of the Wikipedia articles. While they experiment with the English and Hindi language pair, their approach can be applied to other languages.

The next three papers are devoted to information extraction: the way of automatically extracting structured information (such as databases or predicate networks) from free-form texts. This is currently the nearest practical approximation to understanding the text in the way humans learn new information by reading.

Christina Feilmayr from Austria shows how to enrich and complete the inevitably incomplete results of automatic information extraction. For this she suggests a carefully selected set of mutually complementary techniques.

Gábor Berend and Richárd Farkas from Hungary deal with the task of extracting keywords or key phrases from a large amount of documents. Such keywords, which are the most important words or phrases across all documents in the set, can, for example, meaningfully describe the contents of a large document collection for the user. Berend abd Farkas show that the quality of the task of extraction of keywords from a large collection crucially depends on the quality of extraction of keywords describing each single documents.

Ludovic Jean-Louis, Michel Gagnon, and Eric Charton from Canada, in their turn, address the quality of extraction of keywords from a document. They show how general encyclopedic knowledge extracted from Wikipedia can help in determining the importance and novelty of words and phrases in the text. They also personalize the keyword sets extracted from documents basing on the preferences of a specific user.

The next three papers belong to a very active area of research related with opinion mining, sentiment analysis, and analysis of social networks—topics that recently experiences a boom of interest from business, industry, and governmental bodies. Numerous applications of such research help companies understand the customers' opinions about their products, help users to make well-informed buying decisions, and help governmental bodies and political parties realize the effect of their actions and public opinion about their programs and candidates, which leads to better income businesses, better quality of life for consumers, and better democracy for citizens.

Narendra K. Gupta from the USA presents a method for extracting from Twitter messages phrases that describe problems with products or services. Using this information a company can improve the quality of its products and services by addressing the problems people note and mention in Twitter messages. Gupta shows that for this task simple statistical machine learning methods outperform expensive approaches based on manually crafted rules.

Ali Balali, Hesham Faili, Masoud Asadpour and Mostafa Dehghani from Iran address the problem or re-constructing the thread structure in sets of user-contributed comments on company's webpages or in blogs: which messages were independent and started discussions, and which messages were replies or reactions on which other messages. This information is usually not available from the website, or is even not known to the hosting site: the users just add comments, in which they argue with, or reply to, previous posts without explicitly specifying this. However, knowing this structure is crucial for understanding the meaning of the conversation. Once more Balali et al. show that statistical machine learning methods perform well for the task.

Rachel Cotterill from the UK discusses the possibilities of revealing the structure of the relationships between people in social networks, such as identifying informal leaders and followers, influential people and their areas of influence. Traditionally only metadata of the social network messages have been used for this task. Cotterill shows how to use the full text of the messages in a social network to improve the accuracy of such analysis.

The next two papers are devoted to text mining, an area closely related to opinion mining. Text mining techniques allow to identify trends and commonalities in large collections of documents, with applications similar to those of opinion mining.

Delphine Battistelli, Thierry Charnois, Jean-Luc Minel, and Charles Teissèdre from France present a framework and a system capable of extracting events relevant for a given query from a large text corpus. Their system can re-construct a timeline of relevant events, even though they are mentioned in different documents. They apply their system to the analysis of a large body of news messages from a large news agency, and demonstrate that their results are deemed meaningful by professional journalists.

Solen Quiniou, Peggy Cellier, Thierry Charnois, and Dominique Legallois from France apply graph mining techniques to the task of detecting regularities in a large text corpus. Detecting such regularities is important for text mining and summarization as well as for linguistic analysis of large bodies of texts.

Sanja Štajner from the UK and Biljana Drndarević, and Horacio Saggion from Spain address the text simplification problem. Text simplification is a technique that automatically adapts complex texts for better understanding by children, non-native speakers, and persons with cognitive disorders. Štajner et al. experiment with Spanish texts using machine learning techniques to decide when long sentences should be split, some less significant parts of the text can be omitted, or text can be rephrased to be shorter. They identify the features of the sentences that allow for such decisions.

José A. Reyes, Azucena Montes, Juan G. González and David E. Pinto from Mexico, finally, deal with the semantic role classification task. This task is important in many higher-level text processing operations. It represents a step to semantic analysis of the text, that is, text understanding. Reyes et al. use a number of linguistic features and techniques of different nature to improve the identification of semantic roles.

We hope that this special issue will be useful for researchers working in natural language processing, human language technologies, and related areas, as well as for general public interested in modern computational linguistics and artificial intelligence.

Alexander Gelbukh

Head of the Natural Language Processing
Laboratory of the Center for Computing Research
of the Instituto Politécnico Nacional, Mexico,
Member of Mexican Academy of Sciences,
National Researcher of Mexico,
President of the Mexican Society of Artificial
Intelligence (SMIA)