Editorial

This issue of IJCLA presents papers on lexical resources, text features, named entity recognition, detection of lexical functions, information extraction, learning concept identification, opinion mining, and speech processing.

J. M. Torres-Moreno *et al*. (France, Canada, Mexico, and Germany) present a corpus of German texts for similarity detection. Similarity detection between texts is a basic task underlying many natural language processing tasks such as information retrieval, text classification, opinion mining, and text summarization, to name only a few. The authors introduce a valuable lexical resource designed to train and evaluate algorithms for measuring text similarity.

G. Sidorov (Mexico) discusses various types of syntactic n-grams. Syntactic n-grams can be used as features to characterize texts instead of individual words or usual n-grams. Individual words are insufficient for describing meaning or sentiment carried by the text: say, *large screen* may carry positive sentiment when speaking about a phone, while individual words *large* and *screen* do not carry any sentiment. The technique of syntactic n-grams permits to capture long-distance relationships between words, such as *large screen* in *the phone has a large and brightly backlit screen*.

S. Zhou et al. (Japan) present a study of linguistic features useful for the task of named entity disambiguation. Different named entities can share the same name: for example, there is a number of different cities named *New York* or *Moscow*, *St. Andrew* may refer to a person or a university, etc. This justifies the need for algorithms capable of disambiguating such expressions in specific contexts. The authors discuss linguistic features of the texts that help in measuring local coherence of the text; choosing the candidate variant that maximizes the local coherence is a good heuristic for named entity disambiguation.

O. Kolesnikova (Mexico) considers the task of detecting lexical functions in verb-noun collocations in Spanish texts. Verb-noun collocations are pairs of words such as *make a decision*, *meet resistance*, or *keep a secret*. Lexical function in such a word pair, called collocation, is a very general meaning that the verb expresses with a specific noun.

In our examples, the meaning of *make* in *make* decision is, roughly, to indicate that the agent does what the noun expresses; the meaning of *meet* in *meet resistance* is to indicate that the subject undergoes what the noun expresses; the meaning of *keep* in *keep secret* is that the agent does what is expected to do with what the noun expresses. For many text processing tasks it is important to distinguish between such general meanings. The author discusses a technique for their automatic prediction.

M. F. Abdel Hady *et al*. (Egypt) suggest an unsupervised training method for information extraction using aligned bilingual corpora. Usual techniques for information extraction, and in particular for named entity recognition, require expensive lexical resources manually annotated by trained experts. Active learning reduces the amount of training examples by requesting the annotator to annotate a specific example, the most informative one at each step. In this paper the authors show how bilingual corpora can eliminate at all the need in specific manual annotation for the information extraction task, while active learnings significantly speeds up the process.

R. Piryani *et al*. (India) consider the task of identification of the learning concepts in educational texts. The system they present helps students using three-step approach. First, the system identifies the learning needs of a specific student. Next, the system retrieves the relevant learning materials and ranks them according to the suitability for the needs of the learner. Finally, the system presents the learning material to the student and monitors the learning process. The system combines methods from a number of computer science disciplines, such as natural language processing and information retrieval, recommender systems, and educational psychology, among others.

I. Cruz *et al*. (Mexico) describe a publicly available corpus for aspect-based opinion mining that they developed and a supervised machine learning method for identifying opinion aspects in texts that uses this corpus for training. Opinion mining is a fast-growing and very popular area of text processing. An important part of opinion mining process is identifying polarity of a text: whether, say, a user review of a phone Galaxy X expresses positive or negative opinion. However, it is common that people express opinions not about the product in general but of some its aspects: say, the user likes the screen of the phone but does not like its battery. Aspect-based opinion mining operates at this granularity level. While explicit aspects correspond to words present in the document, such as in *the screen is large* (the aspect is *screen*), implicit aspects are implied by other words but not mentioned explicitly, such as in the phone is inexpensive (the aspect is *price*). The authors present a

manually annotated corpus, the first of its kind, that gives such implicit aspects along with the words that imply them (indicators).

The last two papers of the volume are devoted to speech processing: one to speech synthesis and the other to speech recognition.

L.-Q. Tran *et al.* (Vietnam) address prosodic characteristics of airport announcements in Vietnamese. Generating correct prosodic information is important for the generated speech to sound naturally and to carry the correct sentiment. Vietnamese language is a tonal language, in which the role of stress is very different from the role of stress in European languages such as English or French. The authors present a description of stress and other prosodic features in airport announcements in Vietnamese and describe a Hidden Markov Model-based text-to-speech system in this domain that aims to correctly model this information.

C. Lhioui *et al*. (Tunisia) give a detailed analysis of the task of recognition of spontaneous speech in Arabic. The structure of spontaneous speech is quite different from that of written text. While semantic analysis of written text is a common research subject in natural language processing, much fewer effort has been dedicated to understanding spontaneous speech, which can be attributed to the fact that this task is much more difficult. The authors describe in detail their method, which combines two very different approaches: syntax-based approach and stochastic semantic decoding, which improves the robustness of the system. The authors discuss in detail specific difficulties of Arabic language and Arabic spontaneous speech. They also present a corpus that they developed for the task.

This issue of IJCLA will be useful for researchers, students, software engineers, and general public interested in natural language processing and its applications.

ALEXANDER GELBUKH
EDITOR IN CHIEF

MEMBER OF MEXICAN ACADEMY OF SCIENCES,
RESEARCH PROFESSOR AND
HEAD OF THE NATURAL LANGUAGE PROCESSING LABORATORY,
CENTRO DE INVESTIGACIÓN EN COMPUTACIÓN,
INSTITUTO POLITÉCNICO NACIONAL, MEXICO
WEB: <WWW.GELBUKH.COM>