Editorial

This issue of IJCLA presents papers on three topics: semantics and topic classification; sentiment analysis and truthfulness detection; and syntax, parsing, and tagging.

The first section consists of three papers devoted to semantics and topic classification. Extracting and using semantics, that is, the meaning of the text, is the main task of natural language understanding, and is gaining increasing importance in the majority of natural language processing tasks. Modern natural language processing systems are supposed to behave according to the meaning of the text irrespective from how and in what words this meaning is expressed.

Topic classification is a particular task of semantic interpretation. It consists in automatically recognizing the main topic of the given text—such as deciding whether a newspaper article or a blog post is on sports, politics, culture, science, etc.

D. Huynh *et al*. (Australia) propose a novel distributional measure of semantic similarity between words, using the local context. In addition to the use of words that appear in typical contexts of the two given words, they detect latent topics, which gives a more accurate measure of similarity when the contexts differ in words but are similar in topics. Their method outperforms other methods based purely on vector representation of texts, and is second best after a more sophisticated method that uses multi-prototypes.

L. Wolf *et al*. (Israel) describe a vector-based representation of words, known as word embedding, for two languages at the same time. This is useful in various ways. One is data sparseness: when data for one language is insufficient, information can be borrowed from another language for which there are more texts available. Another is disambiguation: data translated in a different language contains important information on the contextual meaning of ambiguous words. The authors describe the process of building a dataset analogous to the famous word2vec dataset provided by Google, but for a language for which much smaller amount of texts is available.

D. Inkpen and **A. H. Razavi** (Canada) develop a novel method for automatic detection of topics in texts. In this context, a topic is a group of words that have some relation with each other, often clearly felt or even interpretable for humans. Representing a document as a vector of topics instead of a vector of words leads to very significant dimensionality reduction and thus speedup of machine learning algorithms. The algorithm developed by the authors offers different levels of granularity of the topics, so that the users can balance speed of processing with accuracy of the representation.

The second section presents two papers devoted to sentiment analysis and truthfulness detection. Sentiment analysis is a relatively young but very actively developed and very popular area of natural language processing. A typical sentiment analysis tasks consists in detecting whether the text expresses positive or negative opinion about something or some emotion, again positive, such as joy or surprise, or negative, such as sadness, disgust, anger, or fear. Analysis of this type has very important practical and commercial applications: in order to make buying choices, consumers need to know what other people think or feel about a specific product or service; companies need to know what the users feel about their products or services; political parties need to know what voters think and feel about a candidate or political program.

Truthfulness detection is the task of automatically deciding whether a given text, such as a speech of a politician, is a lie or truth. The importance of such a task cannot be overemphasized and does not need to be explained here. Technically, the task is difficult because of limited availability of examples of real-world false texts, that is, real lies and not literary fiction texts. The methods that can be used in such a task are akin to those used to detect emotions; actually, what such a program can detect is an "emotion of lying", similar to what a physical lie detector measures.

C. Vania *et al*. (Indonesia) suggest a method for developing sentiment lexicons for languages for which not many texts are available, as well as the use of such lexicon for classification of texts in that language into positive or negative polarity. The method for compilation of the sentiment lexicon is based on the use of seed words translated from an existing polarity lexicon, in this case for English.

V. V. Datla *et al*. (USA and The Netherlands) present a technique to predict whether a short political statement is true or false. Since automatically checking the facts expressed in the statement is unfeasible, they guess the truthfulness of the text by the way it is expressed, using

only linguistic features. They achieve up to 59% accuracy, which is quite encouraging.

The last section consists of four paper devoted to syntax, parsing, and tagging. Syntactic structure of a sentence describes how the words in the sentence are grouped together, or—in a different view on syntactic phenomena—which words of the sentence add details to the meaning of other words. For example, in the sentence "John loves Mary", the words "loves Mary" are grouped together to describe a state of John, and the whole situation is described by grouping together "John" with the expression "loves Mary", which describes his state. Or, in another view on syntax, both words "John" and "Mary" add details to the word "love", thus describing a more specific type of the situation: specifically John's love and loving specifically Mary. This latter approach is called dependency parsing.

Accordingly, automatic detection of such relationships in a sentence, a process called parsing, helps understanding its meaning and plays important role in various tasks of automatic language processing.

In particular, tagging is a process of disambiguating the possible syntactic role of a word in context, that is, determining its part of speech and related properties, when the word is used in sentence: for example, the word "deep" may refer to an adjective in some contexts, to a noun in other, and to a verb yet in other contexts. Tagging is a simpler task than parsing and is much faster. With this, tagging is a step commonly used in the processing chain of many practical natural language processing applications. It is usually the first step of parsing, too, which greatly improves the speed of parsing.

O. Lacroix *et al*. (France) consider the dependency parsing formalism, which is very popular nowadays due to its adequacy for many applications. Accurate dependency parsers are trained on manually labelled text corpora. Manual labelling is a very expensive, tedious, and error-prone process. The authors describe a technique for automatically pre-labeling the corpus, so that the human annotators are offered the most probably labels, or a set of choices ordered by their probability, much like a word processor offers the user a choice of orthographic corrections for a word. Choosing a correct label of a pre-computed set, or most often just confirming the highest-ranked variant, is much faster than assigning labels from scratch.

B. Galitsky *et al*. (USA and Russia) present a method to build syntactic structures that extend to whole paragraphs instead of only one sentence. The trees of individual sentences are connected by co-referent

nodes into a larger graph. Such representation can be used for measuring similarity between texts, which in turn is at the core of a wide range of natural processing tasks such as information retrieval, text classification, and many others. With this extended structure, the authors obtain up to 8% improvement in accuracy of information retrieval of short texts, such as blog posts. The authors provide an open-source implementation of their algorithm.

O. Krůza and **V. Kuboň** (Czech Republic) describe a lightweight method for recognition of clauses and their relationship in complex sentences, relying only on morphological information. Such recognition may in the future improve the performance of syntactic parsers when dealing with complex sentences. In addition, fast and simple clause recognition is useful in those tasks that do not need complete, and thus costly, syntactic analysis—for example, in information extraction or information retrieval.

G. Orosz *et al*. (Hungary) give an extensive discussion of the lessons learned from tagging medical texts. They concentrate the discussion on Hungarian language, an under-resourced agglutinative language. The authors show how to extend and adapt existing resources to this task. They achieve about 50% reduction in the error rate. Their conclusions and advice would probably be useful for implementing tagging methods for medical domain in other under-resourced languages, especially agglutinative languages.

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

RESEARCH PROFESSOR,
CENTRO DE INVESTIGACIÓN EN COMPUTACIÓN,
INSTITUTO POLITÉCNICO NACIONAL, MEXICO
WEB: <WWW.GELBUKH.COM>