International Journal of Computational Linguistics and Applications, Vol. 6, No. 2, 2015, pp. 7–10
ISSN 0976-0962, http://ijcla.bahripublications.com

Editorial

This issue of IJCLA presents papers on cross-document event co-reference, web mining, statistical machine translation, lexical resources, question answering, metaphor detection, Twitter analysis, opinion mining,and assessing text complexity.

I. Pilán et al. (Sweden and Germany) propose a machine-learning approach to assess readability of Swedish texts for learners of this language as second language – that is, the language proficiency level required to understand the given text. They show that the most popular existing measure of readability used for Swedish as native language, LIX, is practically useless for assessing the language proficiency level required from learners of Swedish as second language, while their method distinguishes readability levels quite reliably.

A. Cybulska and **P. Vossen** (The Netherlands) introduce a robust approach to cross-textual event co-reference resolution and test it on a corpus of news articles. The approach compensates for the fact that even if the whole document, a news article in this case, is dedicated to a description of an event, the name of the event is not repeated several times in the text, which makes it difficult to detect automatically that the text is dedicated to this event. The authors use supervised classification of so-called sentence templates in order to detect co-reference of event mentions in different documents.

E. Chandrasekharan and **S. Chakraborti** (India) address the problem of gathering biographical information on people from the huge amount of traits left in Internet by digital interaction of the users. Gathering such information from the web can be both of a great historical, humanistic, and practical value and very dangerous when used maliciously. In either scenario, it is very important to understand what information and how can be collected from the web about each of us. The authors analyze the problem and give the results of their preliminary experiments.

M. Ammar and **S. Jamoussi** (Tunisia) consider decoding, the most important part of statistical machine translation algorithms. Decoding is an NP-complete problem and thus requires good heuristics for acceptable performance. The authors introduce a decoder based on artificial immune system-based metaheuristic algorithm. Evaluation, performed on two publicly available English-French corpora, shows that their approach is promising as compared with existing state-of-the-art decoders.

V. Nastase and **C. Strapparava** (Italy) argue for the importance of working with original lexical resources as opposed to resources artificially mapped to more standard resources such as WordNet, since in the latter case much of the information contained in the resource is lost, oversimplified, or altered by forceful and unnatural mapping. They present a machine-readable version of Wiktionary, a rich, and constantly growing, source of monolingual and cross-lingual lexical and semantic information about words. Unlike other authors who attempt to impose WordNet's sense inventory on Wiktionary, Nastase and Strapparava facilitate computational use of its own intrinsic structure.

C. Mărănducand **C. A. Perez** (Romania) present a Romanian dependency treebank, compiled by a combination of manual annotation and manually checked automatic annotation. The treebank will serve as a source of rules involving syntactic and semantic information on Romanian words, to be used in the construction of rule-based and hybrid dependency parsers for the Romanian language. The treebank is available in the universal dependency treebanks format for improved interoperability.

D. Clarke (UK) describes a simple and fast semantic parser based on a tensor product kernel. Semantic parsing is a sub-task of question-answering task; however, recently it tends to be understood as an approach to question answering that involves construction of a structured query from a natural-language question. The parser proposed by the author shows state-of-the-art performance while being much simpler in implementation and using less resources than existing state-of-the-art semantic parsers.

M. Mohler et al. (USA) present a methodology for detecting metaphors, both conventionalized and novel. The methodology uses supervised learning in a cross-lingual setting, so that metaphors in one language can be detected basing on the information learnt from another language. Metaphors are detected by generalization and analogical reasoning using semantic similarity, as well as transfer learning. An impressive performance of around 90% is achieved on English, Spanish, Russian, and Persian data.

V. Pinheiro et al. (Brazil) show how technology originally developed for tokenization of Chinese texts can be used for splitting composite hashtags used in social media, such as *#fergusondecision*, intro their component elements: *#*, *ferguson*, *decision*. This task is important in many natural language-processing techniques when they are applied to the succinct language of social media such as Twitter, where users give important information in the form of hashtags, which must be formally written as single words while in fact consisting of several words, such as a concept, idea, or a named entity.

L. Noce et al. (Italy) improve the performance of detection of anomalous user-generated context by combining text analysis with image analysis when the user-generated content includes images. Automatic detection of anomalous contents is a key element in fighting fake reviews in opinion mining: with such a technique, the companies such as TripAdvisor or Amazon, whole operation crucially depends on reliable user-generated reviews and on which we all, the users, crucially depend in making important decisions, can implement manual re-checking of suspicious user-generated contributions.

I. Pilán et al. (Sweden and Germany) propose a machine-learning approach to assess readability of Swedish texts for learners of this language as second language – that is, the language proficiency level required to understand the given text. They show that the most popular existing measure of readability used for Swedish as native language, LIX, is practically useless for assessing the language proficiency level required from learners of Swedish as second language, while their method distinguishes readability levels quite reliably.

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