Editorial

This issue of IJCLA presents papers on four topics: co-reference resolution; machine translation; information extraction and biomedical applications; and natural language generation and grammar checking.

The first section consists of one paper devoted to co-reference resolution, which is a process of automatically detecting whether two different words in the text refer to the same entity in real world. The simplest example are pronouns, but other words can also participate in co-reference: for example, *Barack Obama*, *the President*, and *he* can, in a suitable context, refer to the same person. Thus linking these words together is important for text understanding, as well as for many applications ranging from information retrieval and question answering to opinion mining and machine translation.

D. Weissenbacher and **Y. Sasaki** (France and Japan) study the approach to co-reference resolution with Bayesian networks. Different factors can affect the quality of the process of co-reference resolution in a machine learning framework. The most studied ones are feature selection and the learning algorithm used; others are less studied. The authors present a comprehensive study of various factors that affect this process, and conclude that two factors have important impact on its quality: how noisy the features used for classification are, and how reliably the algorithm detects whether a given word is a reference to some another word in the text. For example, in the text *it is clear that this idea is novel* the word *it* does not refer to any other word in the text, while in the text *the idea was difficult to understand but now it is clear* the word *it* refers to *the idea*; looking for an antecedent in the first case (and thus choosing the least unsuitable one) would result in an error.

The second section presents three papers devoted to machine translation. Automatic translation technologies are quickly coming of age and become part of our everyday life. They contribute to better understanding between people of different cultures in our globalized world and help people of all nations to integrate into global community.

X. Song *et al.* (UK) show how to better evaluate the results of machine translation algorithms. The standard automatic evaluation metric nowadays is BLEU, which, despite its usefulness, has certain limitations, such as its inability to handle very short texts—which are very common in Internet and social networks, as well as rather low agreement with human judgments. The authors propose a simpler variant of this evaluation metric that is more flexible and more reliable. They show that their proposed metric has better agreement with human judgments than the standard BLEW metric currently widely used for evaluation of machine translation systems.

G. Wisniewski and **F. Yvon** (UK) suggest a much faster training method for machine translation algorithms. Slow training is a bottleneck for development of statistical machine translation systems and for experimentation with the corresponding algorithms. The authors show that recent advances in recent advances in stochastic optimization and online machine learning can lead to significant improvement in training speed with competitive quality of the resulting translation.

L. Laki et al. (Hungary) present a rule-based method for reordering of phrases in phrase-based machine translation. Reordering is the most important issue that affects quality of phrase-based machine translation when the two languages have different structure and word order. On the example of English to Hungarian translation the authors show how the system can reorder the source sentences (English) in order to make them more similar to the expected translation in the target language (Hungarian) before actual translation. For example, an English phrase *the sons of the many merchants living in the city* is transformed to, roughly speaking, *the city-in living many merchants sons-of*, which is much closer to how the phrase is going to look in Hungarian, after which only a literal translation of English words is required to complete the process.

The next section consists of four paper devoted to information extraction, especially its biomedical applications. Information extraction is a process of automatically building databases and knowledge bases by extracting structured information—such as which medicine causes which side effect—from raw unstructured texts. This process requires significant degree of understanding both structure and semantics of the text.

S. Hina *et al.* (UK and Pakistan) present a semantic tagger for medical narratives, capable of tagging complex semantic information, including paraphrases, abbreviations, and multiword concepts. Such a tagger is useful for a wide variety of applications such as question answering or statistical analysis. The tagging process suggested by the authors is based on rule patterns identified from a real world medical dataset. The proposed tagger outperforms existing methods, including both SVM-based machine learning approach and ontology-based approach.

R. Nawaz *et al.* (UK) go beyond semantics to explore discourse structure of biomedical texts. Discourse-level analysis includes identification of discourse relations between text spans and rhetorical status of sentences and clauses. It is important for identification and interpretation of meta-knowledge: knowledge about knowledge. The authors show how to detect patterns of expressions that convey meta-knowledge about events in scientific papers. They also point out differences between such patterns in the full text of scientific papers and in their abstracts.

D. Kokkinakis (Sweden) continues the topic of extraction of medical events from text. He explores the possibility of using the Frame Semantics framework for this purpose, in particular, the large FrameNet lexical resource combined with domain-specific knowledge sources. He uses a rule-based approach, though machine-learning techniques can be later incorporated in the same framework. He shows that this approach provides powerful modeling mechanism for text mining and information extraction, with high quality of achieved results.

C. Li *et al.* (Hong Kong) propose a framework for named entity detection in Internet texts. Named entities are important in information extraction since they indicate the participants of relations to be extracted. The authors use an approach that does not require training labeled examples; instead, they leverage existing resources and dictionaries for training. Via extensive experiments they show the effectiveness of their approach.

Finally, the last section consists of three papers devoted to natural language generation and grammar checking, which are important applications of natural language techniques.

Y. Hayashi et al. (Japan) show how to determine correct sentence order in a text that consists of various sentences. The problem is important in style correction, where the system can suggest the user a better ordering of the sentences to make the text more understandable. It is also important in natural language generation, where the order of the sentences is to be decided before their actual generation. Natural language generation has a number of applications, of which multi-document summarization is currently the most important one. On the example of Japanese topic-marking particles the author show how linguistic information in a rule-based approach improves the results over the more widely used probabilistic approaches.

G. Sidorov (Mexico) continues the topic of importance of linguistic information for natural language processing tasks. He explains in detail the use of a newly introduced linguistic-based feature called syntactic n-grams in the task of grammar checking of English texts written by non-native speakers. Similarly to a number of other tasks, where the usefulness of the syntactic n-grams as machine-learning features have been already demonstrated, he shows that very simple system based on this approach can show performance competitive with much more sophisticated systems, thus once more confirming that syntactic n-grams are a very useful tool for diverse language processing tasks.

L. Cinman *et al.* (Russia) address the problem of assessing text quality not in the setting of style correction for human authors but instead in the setting of automatically distinguishing human-written texts from automatically generated ones. The problems is very important in fighting spam. What is more, while probably the majority of current natural language processing systems deal with Internet texts, webpages are often full of automatically generated contents usually useless for both applications and human readers, which leads to the necessity of so-called boilerplate removal: mining for useful content in the flood of such useless texts. Even more importantly, fake automatically generated reviews hinder the applications of opinion mining. The authors achieve 85% F-measure on distinguishing between automatically generated and human-written texts, which will be extremely useful in all mentioned applications.

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

GUEST EDITOR:

EFSTATHIOS STAMATATOS
ASSISTANT PROFESSOR,
UNIVERSITY OF THE AEGEAN, GREECE
WEB: < WWW.ICSD.AEGEAN.GR/LECTURERS/STAMATATOS>