Editorial

This issue of the International Journal of Computational Linguistics and Applications is divided into four topics: quantitative linguistics, lexical resources, parsing and co-reference, and applications. Since one of the papers is twice longer than an average paper in this journal, this issue contains only nine papers and not ten as usually.

The first section of this issue, which includes only one paper, is titled “Quantitative linguistics”. Quantitative linguistics includes, among other directions, the study of statistical distributions of letters, morphs, words, sentences and their characteristics. Perhaps the most widely known law of quantitative linguistics is Zipf’s law, which relates the frequencies of words in a text with their frequency rank.

M. Perakh (USA) presents an application of serial correlation statistics to the study of meaningful texts. He shows that certain regularities of the distribution of letters are present only in meaningful texts and are not present in meaningless strings of characters. Those regularities are observed in different languages of different language families and with different writing systems. Perakh also reveals the relation between serial correlation statistics and the Zipf’s law. I believe that his research can open new perspectives in a number of long-standing research questions, from the study of the Voynich manuscript to deciphering ancient scripts to, maybe, the search for messages of extraterrestrial intelligence. Unfortunately, he did not have a chance to develop and apply this research: this prominent scientist, talented writer and outstanding fighter for democracy passed away before he could finish this paper, which is presented to the reader in the version copyedited by the Editor-in-Chief.

The next section is devoted to lexical resources. Lexical resources are crucial for development and for evaluation of computational linguistics research, providing the empirical basis for the theories and techniques created in frame of all other research directions.

V. Henrich et al. (Germany) present a method for collecting sense-annotated corpora from open Internet. Sense-annotated corpora are very important for, for example, training supervised word sense disambiguation classifiers, given that supervised techniques proved so far to be more accurate than unsupervised ones, and easier to implement and maintain. The authors show that their method is language-independent: they successfully apply their method to English and German. The two obtained corpora (English corpus annotated with the WordNet sense inventory and German corpus annotated with the GermaNet sense inventory) are freely available to download.

S. Wang (China) continues the topic of WordNet with a discussion of the perspectives of its translation and use in languages other than English, in this case Chinese. During last two decades the WordNet dictionary proved to be very successful in numerous applications, both research and practical; many existing tools and techniques rely on the WordNet structure and sense inventory. However, despite numerous attempts and long-term efforts, the problem of its translation into other languages has not been solved. Similar dictionaries do exist for a number of major languages, but they are not interoperable with WordNet-based tools; as we have seen, the authors of the previous paper used GermaNet for processing German data—even if German is the language most closely related to English. Wang describes the process of translation of English WordNet into a very different language, Chinese.

K. Kotani et al. (Japan) report the creation of first text corpus that contains material reflecting all four modalities of learners of English as foreign language: writing and speaking, in the form of essays and speech by non-native speakers of English, as well as reading and listening, in the form of student's answers to questionnaires on the texts that they read in English or stories that they listened. Such a corpus will no doubt prove very useful in identifying patterns in students' performance, errors and difficulties. The authors discuss the methodology for the selection of the material for this corpus.

R. Kumar et al. (India) present a tool for manual computer-aided annotation of words in texts with part-of-speech tags. Manually annotated corpora are the raw material for both supervised learning of rules for automatic annotation and manually detecting regularities and building corresponding theories. The tool presented by the authors permits to annotate manually all words in the text, while automatically presenting to the user the most probable variant of such annotation. The authors study the effect of such automatic hints on the accuracy and efficiency of manual annotation. The tool works with Hindi, the world's second largest language.

The next section presents papers devoted to parsing and co-reference detection. Parsing is the task of identifying the internal structure of sentences, the relations between words in the sentences. While the best studied parsing technique is constituency parsing (grouping words together, and grouping such groups together), an alternative approach, dependency parsing (subordinating words to each other: some words in the sentence add more details to other words, making their meaning more specific) gains increasing attention of the research community, especially when dealing with free word order languages. Both papers in this issue devoted to parsing consider the dependency approach.

R. Alfared and **D. Béchet** (France) address the problem of efficiency of a parser by restricting the set of the possible part-of-speech marks of the input words using a separate part of speech (POS) tagger. Given that parsing is a slow operation, the usefulness of parsers for large-scale analysis of Internet texts crucially depends on their speed. The authors show that using a POS tagger significantly increases the parser's speed, while slightly decreasing its recall: the parser misses some correct analyses. The experiments were performed on a French categorial dependency parser.

R. Goutam and **B. R. Ambati** (India) explore the effect of two bootstrapping techniques—self-training and co-training—on a dependency parser, using Hindi as case study. Self-training and co-training are simple variants of semi-supervised learning: the use of unlabeled examples to improve supervised learning techniques. The authors use for their experiments two major Hindi parsers, and compare their results with a those achieved at a competition of Hindi parsers. The authors show that in-domain self-training and co-training gives significant improvement in accuracy, while out-of-domain self- and co-training is less advantageous.

Y. Guo et al. (China) address the topic of entity linking, which can be roughly understood as co-reference. They link named entities found in the text to sources of structured knowledge, such as databases. They use rich context available for the named entity in different texts where it is mentioned to build a model of the entity, so that it can be linked to a correspondent database entry. Using two different benchmark datasets, the authors show that their approach outperforms existing state-of-the art approaches.

The last section of this issue, also consisting of one paper, is devoted to applications.

T. A. Pirinen et al. (Finland) address the problem of spell-checking, probably one of the oldest applications of natural language processing and still far from complete solution. They present a context-aware spell-checker, capable of re-ranking correction suggestions generated by a simpler spell-checker, basing on the information provided by a part of speech tagger. They also show how to adapt traditional *n*-gram models for part-of-speech tagging to morphologically rich languages, with the case study of Finnish, which is an agglutinative language with very rich morphology.

I expect that the papers published in this issue would be useful for scholars, students, and general public interested in natural language processing, applied linguistics, and language learning.

GUEST EDITOR:

YASUNARI HARADA
PROFESSOR,
WASEDA UNIVERSITY, JAPAN
DIRECTOR,
INSTITUTE FOR DIGITAL ENHANCEMENT OF COGNITIVE DEVELOPMENT
PRESIDENT,
ENGLISH LANGUAGE EDUCATION SOCIETY OF JAPAN
EX-PRESIDENT,
LOGICO-LINGUISTICS SOCIETY OF JAPAN
E-MAIL: < HARADA@WASEDA.JP >