

Editorial

This issue of IJCLA presents a selection of papers on two featured topics: sentiment analysis and information retrieval and, in addition, a paper on translation process research.

Nowadays we witness—and actively participate in—an explosion of interest to sentiment analysis, opinion mining, analysis of emotions and subjectivity in text, etc.: the vibrant human side of natural language processing technology, as opposed to the traditional study of cold and impersonalized grammar and statistics of language. We begin to discover for ourselves the huge body of real-life language, language of the “steet” of Internet, language of the blogs, tweets, and social networks, language that several years ago we would not even consider “proper” language—it’s twitting, not speaking! ;-)) Apparently it is this kind of language that, due to a huge number of native “speakers” (native twitters?) in this “thumb generation”, conveys enormous commercially important information—thus the explosion of interest to it from commercial companies.

To catch up with this trend, I selected for this issue of the journal a number of papers on this hot topic.

A. Wawer (Poland) shows how to automatically construct a dictionary of words expressing emotions, or sentiment, by analyzing a large corpus. As many other works on learning ontologies or dictionaries from corpus, this work employs automatically expanding patterns. In this particular case, the patterns make extensive use of morphology—a feature that is present in a huge number of languages but is not yet properly addressed because of its lack in English.

A. Bakliwal *et al.* (India) continue the topic of automatic construction of dictionaries of sentiment and subjectivity-related expressions. In this case, they rely on existing dictionary, WordNet, to achieve this goal. They target Hidni, a language that is native language for more than ten percent of the entire humankind but has not yet receive nearly as much attention from the computational linguistics community as it deserves. Their work (as well as the previous paper)

can also be seen as an example of how such lexica can be constructed for languages other than English.

N. N. Bora (India) suggests an algorithm for extracting the majority opinion from a large number of tweets. This task has a great number of applications. Probably one of the most interesting and important application is for a company, political party or personality, artist, etc., to know what people really think of their products, their proposals, their performances, etc.—and not what smooth-tongued helpers tell them people supposedly think.

N. Konstantinova *et al.* (UK) continues this idea with a study of which features are characteristic of indication of the customers' opinions on commercial products. Such research is indispensable for building recommender systems that could help us, as customers, users, and buyers, to choose the best products and services and to avoid those that look attractive but prove to be flawed or even fraudulent.

M. Neunerdt *et al.* (Germany) close this section with a discussion of a small piece—part of speech tagging—of a much larger problem: to achieve all those goals described above, the researchers will have to develop powerful methods of analyzing and understanding this language of “native twitters” with the same quality as we used to understand the almost extinct nowadays “correct” language. The authors conclude that we are still in the beginning of this long but fascinating journey.

The other topic on which I selected papers for this issue is, in contrast, as old as Internet itself, but is still ever more important: information retrieval, our thread of Ariadne in the exponentially exploding Web, and information extraction, our current substitute for full language understanding.

D. Melo *et al.* (Portugal) propose a novel architecture for question answering systems, based on tight integration of ontologies, natural language processing, and information retrieval techniques. The architecture that they call cooperative question answering not simply answers the user's question when the user happens to ask the right question, but actively helps the user to improve the question if it cannot be successfully answered due to, say, ambiguity. In such cases the system would conduct a dialog with the user in order to achieve mutual understanding on what their information need is.

H. Imran and **A. Sharan** (India) discuss a more traditional technique aimed to the same goal: to help the user to better formulate the query; namely, automatic query expansion: a technique that allows

the computer to rewrite the user's query so that the result would better satisfy their information need. While traditionally this is done by propagating the relevance judgments via term co-occurrence, the authors suggest going beyond co-occurrences by using more advanced term similarity measures. They call their technique lexical cohesion based query expansion and claim that this idea has a great potential for improvement of information retrieval systems.

J. Makhoul and **F. Zaraket** (Lebanon) address an important particular topic in information extraction (that is, almost language understanding): parsing and understanding temporal expressions—expressions referring to time: intervals of time, points in the past or future, age, etc. While this topic has recently enjoyed certain attention, the authors tackle it the context of a morphologically-rich language; what is more, Arabic—a language with unique and amazing Semitic inflective morphology.

R. Winnemöller (Germany) addresses another, and much wider, topic in information extraction: named entity recognition, which is a task of detecting that several words (such as *United States of America* or *Santa Claus*) form a complex name of a single entity (in this case, a country or a person). He shows that his theory of sub-symbolic semantic Text Sense Representations (whatever it means—see Section 2 of his paper if you wonder) can be successfully applied to this task, too, as it has been already applied to many other natural language processing tasks.

Finally, I decided to include in the issue a paper on a rather unusual for this community topic: how the human behavior details while solving a linguistic task can be physically recorded and studied, which can potentially shed light to the underlying cognitive mechanisms.

M. Carl (Denmark) presents Translog II, a new version of a physical device and corresponding software capable of recording and analyzing the movement of human eyes when the subject translate text from one language to another. This very valuable information, totally unavailable to the researcher in the traditional text analysis setting, carries important details on which words cause delays and difficulties, how garden path situations are dealt with by the subject, etc.—which may ultimately uncover details of how this wonderful mechanism called language, which we the natural language processing community strive so much to understand and model, has been implemented in the first place: in the human brain.

I hope that this issue of the IJCLA journal will be both interesting and useful for wide readership, including both general public and students and professionals of natural language processing. Finally, I would like to thank the Editorial Board of the journal for inviting me to guest-edit this issue—it was a hard work but I enjoyed it a lot.

GUEST EDITOR:

GRIGORI SIDOROV

MEMBER OF MEXICAN ACADEMY OF SCIENCES,
CENTER FOR COMPUTING RESEARCH,
INSTITUTO POLITÉCNICO NACIONAL,
07738, DF, MEXICO
E-MAIL: <SIDOROV@CIC.IPN.MX>