IJDAR (2009) 12:139–140
DOI 10.1007/s10032-009-0096-6

EDITORIAL

Special issue on noisy text analytics

**Daniel Lopresti · Shourya Roy · Klaus Schulz ·
L. Venkata Subramaniam**

Published online: 17 September 2009

Noise is an ever-present challenge in the field of document analysis. We face it at all levels and stages of the process. While human readers are adept at compensating for noisy inputs, machines still stumble, sometimes badly, when confronted by the consequences when noise is present. The goal of building robust systems that both tolerate noisy inputs on the one hand, and minimize the errors they pass on to downstream stages on the other, is an active area of research.

The AND series of workshops on Analytics for Noisy Unstructured Text Data were established to provide a forum for addressing noise that reflects any kind of difference between the surface form of a coded representation of text and the intended, correct, or original text. Note that this definition is intentionally broad, covering differences that arise from typographic or grammatical errors, informal use of language, and errors from the machine recognition of typeset text, handwriting, and speech. By its very nature, noisy text warrants moving beyond traditional techniques for text analytics.

This special issue includes expanded versions of six papers chosen from among those presented at the Second Workshop on Analytics for Noisy Unstructured Text Data, which was organized as part of the 31st ACM SIGIR Conference held during July 2008 in Singapore. Similar to the first AND workshop which also yielded a special issue of IJDAR (vol. 10, nos. 3–4, December 2007), AND 2008 was a successful event attended by over 40 researchers from around the world representing various academic institutions, industry, and government. Each of the invited papers selected for this issue underwent a rigorous re-review process before final acceptance for the journal.

The first two papers deal with errors introduced as a result of failures in optical character recognition and handwriting recognition, respectively. Lopresti applies a hierarchical approach to dynamic programming to classify errors across the stages of a typical text analysis pipeline: sentence boundary detection, tokenization, and part-of-speech tagging. Errors and their cascading effects are isolated and analyzed, with the ultimate goal of understanding the varying impacts of different types of errors and ways of ameliorating them. Farooq, Bhardwaj, and Govindaraju examine two methods for improving the results of unconstrained handwriting recognition. The first of these attempts to build a reduced-size lexicon based on topic identification, while the second employs topic-specific language models to drive the output of the recognizer. Both methods are tested and found to yield significant accuracy improvements on highly degraded inputs.

The next paper, by Reffle, Gotscharek, Ringlstetter, and Schulz, examines the problem of correcting "false friends," spelling mistakes that result in other words present in the lexicon. They address this challenge by building a profile for the error channel and using this to focus attention on dictionary words which are more likely to be errors requiring correction, thereby maximizing the number of proper corrections that occur and minimizing the number of false corrections. On the other hand, Acharyya, Negi, Subramaniam, and Roy take Short Message Service (SMS) data as their

D. Lopresti (✉)
Lehigh University, Bethlehem, PA, USA
e-mail: lopresti@cse.lehigh.edu

S. Roy
Xerox India Innovation Hub, Xerox Corporation, Chennai, India
e-mail: Shourya.Roy@xerox.com

L. V. Subramaniam
IBM India Research Lab, New Delhi, India

K. Schulz
University of Munich, Munich, Germany

starting point. Here, noise is introduced through aggressive use of abbreviations, intentional deletions, phonetic spelling conventions, mixed languages, genuine misspellings, etc. They demonstrate how unsupervised methods can provide effective results at a lower cost than those requiring expensive manual labeling of a training set, and illustrate their discussion with tests using a large corpus of SMS messages from a real-world call center.

The article by He, Weerkamp, Larson, and de Rijke, delves into the world of blogs, where discussions can often drift from one topic to the next, severely impacting attempts to extract and distill information from such sources. The authors define a notion of topical noise, and then develop techniques for generating a reliable coherence score for blogs in the presence of noisy input as one typically finds in the blogosphere. In the final article, Dey and Haque study the problem of opinion mining in the context of noisy text inputs. They present a semi-supervised approach to learn domain knowledge from a training set that contains both clean and noisy data. Their targeted applications, from which they wish to extract opinion expressions, include customer blogs and feedback forms.

We conclude this introduction by thanking all those who participated in the AND 2008 workshop, as well as the reviewers who provided valuable feedback both for the workshop articles and the extended versions that appear here. It is our hope that this special issue will continue to broaden awareness of the problems that arise in noisy text analytics, and thereby inspire those working in related areas to consider the challenges posed here as worthy topics of study.