IJDAR (2011) 14:111–112
DOI 10.1007/s10032-011-0162-8

EDITORIAL

Special issue on noisy text analytics

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

Published online: 3 April 2011

Noisy unstructured text data are ubiquitous in real-world communications. Text produced by processing signals intended for human interpretation, such as printed and handwritten documents, spontaneous speech, and camera-captured scene images, are prime examples. Application of Automatic Speech Recognition (ASR) systems on telephonic conversations between call center agents and customers often see 30–40% word error rates. Optical character recognition (OCR) error rates for hardcopy documents can range widely from 2–3% for clean inputs to 50% or higher depending on the quality of the page image, the complexity of the layout, and aspects of the typography. Unconstrained handwriting recognition is still considered to be largely an open problem.

Recognition errors are not the sole source of noise; natural language and its creative usage can cause problems for computational techniques. Electronic text taken directly from the Internet (emails, message boards, newsgroups, blogs, wikis, chat logs, and web pages), contact centers (customer complaints, emails, call transcriptions, message summaries), and mobile phones (text messages) is often very noisy and challenging to process. Spelling errors, abbreviations, non-standard words, false starts, repetitions, missing punctuation and case information, and pause-filling words such as "um" and "uh" in the case of spoken conversations are just a few examples of the problems that can arise.

This special issue includes expanded versions of nine papers chosen from among those presented at the Third Workshop on Analytics for Noisy Unstructured Text Data, which was organized as part of the Tenth International Conference on Document Analysis and Recognition (ICDAR) in Barcelona, Spain, in July 2009. Building on previous AND workshops, AND 2009, was a successful event attended by over 25 researchers from various academic institutions and business organizations from different parts of the world. Each of the papers selected for this issue underwent a rigorous revision and reviewing process before final acceptance for the journal.

We are pleased to present the research described here as reflecting the current state-of-the-art in noisy text analytics. The first paper by Simone Marinai addresses some of the challenges arising in text retrieval from early printed manuscripts. In such cases, typography and state of preservation can lead to character and word images which are difficult to segment. Marina applies Self-Organizing Maps to solve this problem and demonstrates the efficacy of his techniques on page images from the Gutenberg Bible.

The second paper by Kesidis, Galiotou, Gatos, and Pratikakis presents a word-spotting approach for historical documents that avoids the need for optical character recognition. They describe a set of techniques for pre-processing, word segmentation, and word matching and test their methods on early Modern Greek documents from the seventeenth and eighteenth Century.

In the third paper, Cao, Govindaraju, and Bhardwaj address the unconstrained handwritten document retrieval problem. They assume handwriting recognition will produce

D. Lopresti (✉)
Department of Computer Science and Engineering, Lehigh University, 19 Memorial Drive West, Bethlehem, PA, 18015, USA
e-mail: lopresti@cse.lehigh.edu

S. Roy
Xerox India Innovation Hub, IIT Madras Research Park, Kanagam Road Taramani, Chennai 600 113, Tamil Nadu, India
e-mail: shourya.roy@gmail.com

K. Schulz
University of Munich, 81377 Munich, Germany
e-mail: schulz@cis.uni-muenchen.de

L. V. Subramaniam
Information Quality and Discovery, IBM Research India, New Delhi, India
e-mail: lvs004@gmail.com

imperfect results and propose a term frequency estimation technique that incorporates segmentation information to improve precision and recall performance.

The next paper by Subramaniam, Prasad, and Natarajan examines named entity detection in the presence of errors. Since OCR output may be noisy in many applications of interest, they employ a complete lattice which is output from recognition engine. They then describe ways of dealing with issues that arise when using lattices: the high false alarm rate and the added computational cost. Their techniques are demonstrated on English videotext and handwritten Arabic.

Information retrieval in historical document collections is the subject of the paper by Gotscharek, Reffle, Ringlstetter, Schulz, and Neumann. The problem is challenging because of the large number of spelling variations that are present. The authors study the interaction between matching procedures and specialized lexica to determine an effective approach, conducting their experiments on collections of documents spanning several centuries embodying the evolution of the German language.

Likewise, in his paper, Reynaert shows how to handle typographical variation and spelling errors in noisy text collections using an approach based on anagram hashing. Test results are reported on digitized Dutch Parliamentary documents and the 1918 edition of the daily newspaper *Het Volk*.

The paper by Bratus, Rumshisky, Khrabrov, Magar, and Thompson describes two approaches for extracting knowledge from unstructured narrative text. One of these uses domain-specific ontologies, and the other employs Hidden Markov Models trained on a small amount of annotated data. The authors present experimental results using data from the automotive industry: lexicons and ontologies of part names, along with car repair manuals.

Giannone, Basili, Naggar, and Moschitti address a different application domain in their work: criminal investigative data based on police interrogation reports. They show how text relation mining can be accomplished in such cases using kernel-based techniques. Empirical results demonstrate that the methods are effective in the presence of non-traditional language, dialects, and jargon.

Finally, the paper by Marx and Gielissen describes efficient storage mechanisms for scanned document images. Their method reduces the size of the documents returned to users by two orders of magnitude while maintaining the same overall visual appearance.

In addition to contributed papers, we also present a working group report authored by Marinai and Karatzas. Workshops such as AND are highly interactive, and one of their most noteworthy features is the way they bring together researchers with different perspectives from around the world to focus on a common topic. This working group report captures the ideas and impressions of 12 attendees who chose as their topic "noisy text datasets." As a reflection of the live discussion that took place at AND 2009, we believe it is a valuable addition to this special issue.

We conclude this introduction by thanking all those who participated in the AND 2009 workshop, as well as the reviewers who provided valuable feedback both for the workshop papers 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.