Computers in Industry 78 (2016) 1–2

Contents lists available at ScienceDirect

Computers in Industry

journal homepage: www.elsevier.com/locate/compind

Editorial: Special issue on natural language processing and text analytics in industry

The recent years have witnessed an unprecedented growth in the volume of unstructured data expressed in natural language. This explosion in the volume of text can to a large extent be attributed to the emergence of Web 2.0 platforms and online communities, including blogs, forums, review sites, and social networks. Hidden within these novel channels is valuable information, which, if properly identified and extracted, can be employed to enhance a plethora of corporate activities. For example, product reviews written by consumers on blogs or retailer websites contain useful information about brand perception and product quality.

The scientific discipline of Natural Language Processing (or Computational Linguistics) and the applied field of Text Analytics, in which Natural Language Processing coalesces with Machine Learning and Data Mining, have also evolved over the years to keep pace with the proliferation of novel sources of text data. Natural Language Processing (NLP) and Text Analytics (TA) algorithms and techniques are increasingly being developed, adopted and deployed for addressing a wide spectrum of real-life, industrial problems. Typical examples include document classification, document clustering, topic detection and modeling, and opinion mining/sentiment analysis. However, at the same time, the successful application of such innovative technologies raises a number of pertinent research questions, with both theoretical and practical ramifications.

The main objective of this Special Issue (SI) was to collect and consolidate high-quality knowledge on state of the art research in NLP/TA methods and their applications in industrial (enterprise) activities. Approximately 70 abstracts were received following the SI call. This relatively huge number is testimonial to the growing importance of the field. They were evaluated for their pertinence based on several criteria, including a sound scientific contribution and a clear industrial application. Around 30 abstracts were selected as part of this evaluation. The respective authors were requested to submit their full manuscripts, which were reviewed by at least 2 reviewers according to the journal's guidelines. Finally, 7 articles were selected for publication (after another review round) in this SI. In addition, the SI also includes an article from the guest-editors, which discusses the trends and challenges of text analytics applications in industry. An overview of these articles is provided below.

Analyzing and evaluating the task of automatic tweet generation: Knowledge to business (Lloret and Palomar): This article presents an approach based on text summarization techniques to generate tweets. Furthermore, statistical measures are employed to estimate the informativeness and interestingness of tweets. They show that automatically generated tweets can be as informative and interesting as manually generated ones for the English language. Results are more mitigated for Spanish.

A distributional approach to open questions in market research (Evert et al.): The authors present the Klugator Engine (TKE), a system for analyzing survey responses expressed in English and German. In essence, their analysis involves determining the sentiments of responses and clustering them according to their topical similarity, which are then displayed as nodes on a semantic map. It is worth noting that this system is already incorporated in a commercial solution.

Integrating a semantic-based retrieval agent into case-based reasoning systems: A case study of an online bookstore (Chang et al.): Techniques for short-text semantic similarity (STSS) and recognizing textual entailment (RTE) are integrated in a case-based reasoning system. Experimental evaluations revealed that the proposed system outperformed existing ones. Furthermore, in a case-study, the accuracy of the proposed system in responding to users' requests (queries, questions) was also compared to that of other similar solutions found in typical e-commerce websites. The case-study showed that the proposed system was more effective and provided more accurate responses to address the information needs of users.

Turning user generated health-related content into actionable knowledge through text analytics services (Martinez et al.): A system for monitoring health-related information for pharmacovigilance is presented. The proposed system integrates several text analytics solution, such as GATE and MeaningCloud, for extracting pertinent information pertaining to adverse drug reactions from social media sources. Information extraction is mainly performed using named entity recognition and semantic relation extraction from Spanish texts. These tasks are performed by leveraging upon several domain-specific dictionaries and ontologies.

A methodology for traffic-related twitter messages interpretation (Albuquerque et al.): Tweets are a valuable source of real-time

http://dx.doi.org/10.1016/j.compind.2016.01.001

traffic information. However, an important challenge lies in the analysis and interpretation of tweets, especially for traffic monitoring which require very precise information (for e.g. exact location, number of vehicles involved in accident). The challenge is addressed in this article, which presents a system that analyzes and interprets tweets, expressed in Portuguese, by transforming them into RDF triples based on a domain-specific ontology. The application has been deployed to monitor truck fleets operated by a liquid gas and a fuel distribution company.

Text classification based filters for a domain-specific search engine (Schmidt et al.): Domain-specific search engines often on rely filters to narrow down their search results. This article discusses how text classification using Support Vector Machines (SVM) can be used to predict filters. In addition, it also shows how active learning can be applied to enhance the effectiveness of the classification task. The proposed system has been employed in a commercial German search engine for the domain of job offers and vacancies.

Natural language processing for aviation safety reports: From classification to interactive analysis (Tanguy et al.): This article presents a system for classifying aviation safety reports into incident categories. Classification is performed using SVM, and the performance is enhanced by adopting an active learning procedure. In addition to classification, the system also performs topic detection from the reports, using an approach based on Gibbs sampling. However, the results for topic detection were mitigated, with much better performance achieved in the text classification task. Furthermore, the proposed system includes an information retrieval application, enabling users to search for reports describing similar incidents. The results (i.e. reports describing similar incidents) are displayed along a temporal axis on a 2-D scatter-plot for easier visualization.

Text analytics in industry: Challenges, desiderata and trends (Ittoo et al.): The last article in this SI is a review article by the guest-editors. The current state of the art in text analytics applications industry is systematically reviewed and discussed along several dimensions, such as the application contexts, the techniques utilized and the evaluation procedure. From this review, the authors identify the challenges and constraints that real-world environments impose on text analytics applications, and subsequently, identify a set of desiderata that text analytics applications should possess for their successful deployment in industry. Finally, the article discusses future trends in text analytics and their potential application in industry, including the revival in neural-network-based methods for deep learning, word-embeddings and scene (image, video sequence) labeling.

Since 2013, **Ashwin Ittoo** is an Asst-Professor in Information Systems at the HEC Management School, University of Liége, Belgium. His research interests are minimally-supervised learning techniques for Machine Learning and Natural Language Processing and in the application of these techniques for measuring socio-economic indicators. He received his Ph.D. degree from the University of Groningen, The Netherlands in 2012 and his Bachelors and Masters from the National University of Singapore and the Nanyang Technological University, Singapore.

Le Minh Nguyen is currently an Associate Professor of School of Information Science, JAIST. He leads the lab on Machine Learning and Natural language Understanding at JAIST. He received his B.Sc. degree in information technology from Hanoi University of Science, and M.Sc. degree in information technology from Vietnam National University, Hanoi in 1998 and 2001, respectively. He received his Ph.D. degree in information science from School of information science, Japan Advanced Institute of Science and Technology (JAIST) in 2004. He was an assistant professor at School of information science, JAIST from 2008-2013. His research interests include machine learning, text summarization, machine translation, natural language processing, and information retrieval.

Antal van den Bosch (PhD 1997, Universiteit Maastricht) is professor of language and speech technology at the Centre for Language Studies at Radboud University, Nijmegen, the Netherlands. His research interests include memory-based and exemplar-based natural language modeling, text analytics applied to historical texts and social media, and proofing tools. He is a member of the Netherlands Royal Academy of Arts and Sciences and ECCAI Fellow.

Ashwin Ittoo[a,*]
Le Minh Nguyen[b]
Antal van den Bosch[c]
[a]*HEC Management School, University of Liège, Liège, Belgium*
[b]*School of Information Science, Japan Advanced Insitute of Science and Technology, Japan - Division of Data Science, Ton Duc Thang University, Ho Chi Minh City, Viet nam*
[c]*Centre for Language Studies, Faculty of Arts, Radboud University Nijmegen, Nijmegen, The Netherlands*

[*]Corresponding author
E-mail addresses: ashwin.ittoo@ulg.ac.be (A. Ittoo), nguyenml@jaist.ac.jp (L.M. Nguyen), a.vandenbosch@let.ru.nl (A. van den Bosch).

Available online 12 February 2016