Guest Editorial

Text mining and information analysis of health documents

1. Introduction

Documenting every event in healthcare, as required by law in many countries, takes a lot of healthcare professionals’ time. With electronic health (eHealth) documents being increasingly common, healthcare professionals type approximately forty per cent of the documents as free-form text and the remaining sixty per cent is either automatically-recorded or typed structured-information [1]. For example, one patient’s intensive care documents alone can contain over sixty typed pages of free-form text [2], supplemented each day by over 1,500 structured items, on average [3].

*Free-form text* as an entry type is essential to release healthcare professionals’ time from documentation for other tasks. With fully structured or centralized eHealth documents, entering information can take nearly sixty per cent of their working time, whilst with electronic records that allow free text entry at the point of care, it typically takes only a few minutes per patient [4-6].

Even more time could be released through *human/natural language technologies* (HLTs a.k.a. *computational linguistics, human/natural language processing, text analytics, text (data) mining, or text processing* (TM)) to support healthcare professionals in generating and using eHealth documents [7-9]. For example, *speech recognition* with proofing by hand has been shown to make the document turn-over time up to 5.7 times faster when compared to transcription by hand [10-13]. Moreover, structuring these free-form text documents through HLTs for automated *information extraction* (IE) is almost instantaneous; eases finding and using relevant information; makes this content available for computerized re-use; and has gradually improved to exceed ninety per cent correctness [14]. Finally, *information retrieval* (IR), that is, the core technology in search engines, is an increasingly efficient way not only to find relevant eHealth documents in healthcare information systems but also relate the findings with scientific evidence in published literature in health sciences and patients,
patients’ next-of-kin, and their caretakers’ personal experiences in the Internet blogs and discussion forums [15].

In summary, HLTs hold great promise for improving *information readability* and *accessibility* for healthcare professionals, researchers, patients and other information consumers. Thereby, they can support not only *situational awareness* (i.e., people’s perception of environmental elements in time and place, the comprehension of their meaning, and the projection of their status in the near future) [16] and *decision making* in healthcare but also *knowledge discovery* in health sciences.

However, because of the severe implications that errors may have in healthcare, this promise only holds if the quality of these HLTs has been carefully evaluated as excellent and their potential benefits outweigh the risks [17-21]. One of the most severe risks is the loss of *data subjects’ privacy*, and this is particularly evident in TM applications requiring large datasets; these data may contain personal information about patients and healthcare professionals. In order to avoid compromise of privacy, storing and using these privacy-sensitive data requires careful consideration and compliance with legal and ethical principles. Therefore, the development, progress and adoption of eHealth HLTs has been substantially hindered by the barriers of lack of access to these privacy-sensitive documents, and consequent lack of common conventions and standards for materials, methods and evaluations needed to have reproducible, scientific results [22].

### 2. Syntactic, semantic and pragmatic analysis of eHealth documents

In this special issue, we focus on syntactic, semantic and pragmatic analysis of these eHealth documents; mono- and multilingual methods, applications and resources for their automated processing and intelligent re-use; evaluation of eHealth documents and their processing methods, applications and resources; as well as experiences from practical applications of these HLTs in eHealth documentation systems. More specifically, the call for papers of this special issue targeted scientific papers, case studies and reviews presenting original research with outstanding results in, or related to, the following five topics:
• clinical documents and documentation guidelines:
  – their content and language analysis (e.g., syntax, semantics, discourse, dialogue, pragmatics and word sense disambiguation);
  – their use in personalized medicine, healthcare work-flow, clinical trials as well as science and related information systems; and
  – their collaborative exchange between provider and patient, between providers in different organizations, or between different professions and the related topics of computer-supported cooperative work and human-computer interaction;

• language generation for interactive and dialogue-based documentation and text production;

• (semi-)automated re-use of text content:
  – decision support (e.g., reasoning and meta-reasoning);
  – identification and removal of privacy-sensitive and identifying information;
  – IE;
  – information visualization;
  – IR;
  – machine translation;
  – referent tracking and management;
  – sentiment analysis;
  – text classification;
  – text clustering;
  – text summarization;
  – textual entailment and paraphrasing;
  – TM; and
  – topic analysis;

• knowledge acquisition:
  – analysis and integration of eHealth documents across languages, genres and jargons;
  – combining documents from multiple sources; and
  – combinations of text and encoded/structured information;

• annotation and evaluation:
  – annotated health resources and science of annotation;
  – dictionaries, ontologies, standards and other linguistic resources;
– evaluation metrics, methods, protocols and infrastructure; and
– evaluation results and protocols across care settings, organizations and languages.

The special issue especially welcomed papers demonstrating the collaboration of academic, industrial, and governmental partners across disciplines. It strongly encouraged papers reporting on work for minor languages that shed light on solving problems common to many languages lacking in available lexical resources for HLTs.

The special issue consists of five peer-reviewed papers. This review process has included three double-blind steps. First, fourteen submissions were reviewed independently by three program committee members. Second, after receiving revisions of the six shortlisted manuscripts, they were reviewed again by the same three members but now in relation to the original review feedback. Third, five papers were accepted, resulting in an overall acceptance percentage of 36. We thank all the authors and reviewers for their hard work!

The first paper studies the development and domain-adaptation of statistical syntactic parsers for three different eHealth domains in Finnish [23]: nurses’ daily notes from intensive care (n = 2,818 sentences), physicians’ notes related to cardiac patients (n = 1,436 sentences) and nurses’ daily notes related to cardiac patients (n = 1,561 sentences). Parsing is performed using a state-of-the-art, open-source method called Bohnet. According to the results of the paper, the use of eHealth sentences in combination with general-language sentences (n = 15,126 from Wikipedia, blogs and university newspaper articles in the Internet) to train the parser improves the labeled attachment percentage from 59–71 to 77–85.

The second paper extends and evaluates a cue-based text-analysis system for English eHealth text to assert whether a disorder is affirmed, negated or uncertain in Swedish diagnostic statements (n = 1,525) [24]. This system called pyConTextSwe includes 454 contextual cues to classify text with respect to assertion (four-class classification task), existence (binary classification task) and uncertainty (binary classification task). It achieves correctness percentages of 88, 81, 55, 63, 97, 87, 78, and 86 for the categories of definite existence,
probably existence, probable negated existence, definite negated existence, existence = yes, existence = no, uncertainty = yes, and uncertainty = no, respectively.

The third paper focuses on de-identifying privacy-sensitive, confidential eHealth documents for research purposes [25]. Its system called Anonym is based on applying a well-known probabilistic sequence-classifier called Conditional Random Field to 889 medical discharge summaries and 1,885 clinical notes in English from the USA together with 852 pathology and cytology reports in English from Australia. It considers ten categories of personal health information present in these datasets (i.e., date, patient, doctor, name, age, ID, hospital, location, institution, and phone). The resulting average correctness percentages are 93 and 98 for the US datasets and 83 for the Australian dataset.

The fourth paper addresses the linkages between patients’ personal experiences from Internet blogs and healthcare by considering the task of syndromic surveillance [26]. The developed and evaluated system first extracts common expressions that patients use to describe their medical conditions from blog posts. Then it generalizes the findings by applying clustering to identify expressions that resemble the extracted snippets and adjust the related weighting model. Finally, a terminology mapping and extension tool based on contextual cues together with the aforementioned extraction and clustering methods, is used to relate these patient-centric expressions with respective professional terms (e.g., blocked/clogged/runny/sore nose vs. rhinorrhea). The system is 96 per cent correct when evaluated using the topic of influenza-like diseases and 585,777 Tweets in English.

The fifth paper investigates machine translation of IR queries [27]. Its system is based on the state-of-the-art, open-source methods called Moses for machine translation and Lucene for IR. Its datasets include eHealth dictionaries and medicine and health-patent parallel-corpora. The experimented language pairs are Czech/German/French–English. The experimented datasets are comprised of hundreds of thousands to millions words. The system improves the average correctness in translation by 55 per cent when compared to the baseline system trained on the general-language parallel-corpora from the Internet, European
Parliament documents, European Union legislation, and newspapers, among others alone. Improvements in IR are evident only for the French–English pair.

3. Conclusions

Evaluation results of these papers demonstrate the great promise of eHealth language technologies for improving information readability and accessibility for healthcare professionals, researchers, patients and other information consumers. Thereby, these technologies can support not only healthcare professionals and patients’ situational awareness and decision making but also knowledge discovery in health sciences.

4. Background

The special issue is inspired by the First Workshop on Cross-Language Evaluation of Methods, Applications, and Resources for eHealth Document Analysis\(^1\) (CLEFeHealth2012) and the Fourth International Workshop on Health Document Text Mining and Information Analysis\(^2\) (Louhi2013) workshops. However, the call was open also for submissions not published in these workshop proceedings.

CLEFeHealth2012 was the Conference and Labs of the Evaluation Forum formerly known as Cross-Language Evaluation Forum\(^3\) (CLEF) 2012 workshop on cross-language evaluation of methods, applications, and resources for eHealth document analysis with a focus on written and spoken natural-language processing. It was organized in Rome, Italy on 17-20 September 2012.

The CLEF initiative is an activity of the Participative Research laboratory for Multimedia and Multilingual Information Systems Evaluation Network of Excellence (PROMISE), funded by the European Commission 7th Framework Programme. Its main mission is to promote


research, innovation, and development of information access systems with an emphasis on multilingual and multimodal information with various levels of structure. Since 2000, it has played a leading role internationally in stimulating investigation and research in a wide range of key areas related to information access. Each year, it organizes a wider peer-reviewed conference together with more specific workshops and evaluation labs (a.k.a. shared tasks, challenges, hackathons, or competitions where all participants’ goal is to solve the same problem, typically using the same dataset).

The first time CLEF addressed eHealth HLTs was in 2005. Namely, ImageCLEFmed was launched as an annual evaluation lab and workshop with its focus on biomedical images in papers and on the Internet [28]. In 2005–2013, it targeted language-independent techniques for annotating images with concepts; multilingual IR; and multimodal (i.e., images and text) IR. The 2014 task was to fill out a form automatically by extracting information from computed tomography scans of the liver.

In 2012, CLEF introduced CLEFeHealth as a workshop focusing on electronic clinical documents [29]. The use scenario of the workshop was people using HLTs to analyze and integrate eHealth documents across languages, genres and jargons. Its program consisted of three invited talks on collaborative datasets, resources, tools, and infrastructure by Australian and US researchers; a panel of Austrian, Finnish, Irish, Swedish, and Swiss experts in eHealth; eleven paper presentations; and three workshops for roadmap development, student mentoring, and partnership building, respectively. Eleven papers, describing original work in Australia, Austria, Finland, Norway, and Sweden, were accepted after a double-blind peer-review process. Ten of the papers focused on meeting the needs of healthcare professionals and patients in information access and ease of understanding the content via user-centered vocabularies, ontologies, abbreviation processing, language generation, search engines, and tools to support patient–healthcare professional interaction. The eleventh paper described an evaluation setting. This special issue targeted extensions of these papers.

---

4 [http://ir.ohsu.edu/image/](http://ir.ohsu.edu/image/), accessed 15 May 2014
In 2013, based on the conclusions drawn from the 2012 workshop, CLEFeHealth progressed to organizing its first evaluation lab [15]. This consisted of the following three tasks to improve patients’ understanding of their health records: identification of medical disorders and standardization of their language; mapping abbreviations and acronyms to standardized codes; and IR to address questions patients may have when reading their records. The three tasks used a subset of 300 de-identified patient reports in US English gathered from about 30,000 intensive care patients. The third task also included approximately a million documents from health and medicine sites in the Internet. The tasks attracted 34 participating teams to submit 113 systems. The evaluation lab and its workshop had approximately 180 participants in total from Australia, Austria, China, Colombia, Finland, France, Germany, Greece, Holland, India, Ireland, Japan, Norway, Portugal, Republic of Korea, Spain, Sweden, Switzerland, the UK and USA.

In 2014, the CLEFeHealth evaluation lab and workshop ran again with the following three tasks to improve patients’ understanding of their health records: visual-interactive search and exploration of electronic health data; IE from clinical text (extension of the 2013 Task 1); and user-centered health IR (extension of the 2013 Task 3).

Other CLEF evaluation labs and workshops addressing the topic of eHealth HLTs were Question Answering for Machine Reading Evaluation (QA4MRE) and Entity Recognition (CLEF-ER). The QA4MRE evaluation labs and workshop had a pilot task on machine reading on biomedical text about Alzheimer's disease in 2013 [30] and in 2014, it ran a lab on biomedical semantic indexing and question answering. The CLEF-ER2013 workshop focused on the domain of patient documents but did not have a shared task.

The Louhi workshop series began in 2008 and its fourth workshop, Louhi2013, was organized in Sydney, NSW, Australia on 11-12 February 2013 in collaboration with the

---

9 http://www.clefer.org, accessed 15 May 2014
The program consisted of three keynote talks and fourteen papers presentations by American, Australian and European researchers [31]. The papers were accepted after a double-blind peer-review process. The papers addressed a wide range of topics, including clinical terminologies; multilingual aspects; Australian perspective; classifiers; technology evaluation; pathways from linguistic engineering through machine learning to applications and products; and future directions. Extensions of these papers were targeted by this special issue.

The first Louhi workshop was co-located with the 3rd International Symposium on Semantic Mining (SMBM2008) in Finland and had its special issue in the International Journal of Medical Informatics [32]. The second workshop was a part of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT2010) in CA, the USA and had its special issue of the Journal Biomedical Semantics [33]. The third workshop was organized jointly with the 13th Conference on Artificial Intelligence in Medicine (AIME2011) in Slovenia [34]. The fifth workshop was a part of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL2014) in Sweden [35] and will have its special issue in the BMC Medical Informatics and Decision Making.

Acknowledgements

NICTA is funded by the Australian Government through the Department of Communications and the Australian Research Council through the ICT Centre of Excellence Program. NICTA is also funded and supported by the Australian Capital Territory, the New South Wales, Queensland and Victorian Governments, the Australian National University, the University of New South Wales, the University of Melbourne, the University of Queensland, the University of Sydney, Griffith University, Queensland University of Technology, Monash University and other university partners.

The members of the program committee of this special issue, chaired by the guest editor Hanna Suominen, were Alexey Tsymbal, Siemens AG, Erlangen, Germany; Antti Airola, University of Turku, Turku, Finland; David Martinez, NICTA and University of Melbourne, Melbourne, VIC, Australia; Dimitrios Kokkinakis, University of Gothenburg, Gothenburg, Sweden; Filip Ginter, University of Turku, Turku, Finland; Günter Schreier, AIT Austrian Institute of Technology, Graz, Austria; György Szarvas, Technical University of Darmstadt, Darmstadt, Germany; Henning Müller, University of Applied Sciences Western Switzerland Sierre (HES-SO), University Hospitals and University of Geneva, Switzerland; Hercules Dalianis, Stockholm University, Stockholm, Sweden; Jon Patrick, University of Sydney, Sydney, NSW, Australia; Jong C. Park, KAIST, Daejeon, South Korea; Jussi Karlgen, Gavagai, Stockholm, Sweden; Laura Slaughter, Oslo University Hospital (OUH) and Norwegian University of Science and Technology (NTNU), Trondheim, Norway; Lawrence Cavedon, NICTA and University of Melbourne, Melbourne, VIC, Australia; Leif Hanlen, NICTA and Australian National University, Australia; Martin Hassel, Stockholm University, Stockholm, Sweden; Ozlem Uzuner, University at Albany, State University of New York, Albany, NY, USA; Pierre Zweigenbaum, LIMSI-CNRS, Orsay, France; Riikka Vuokko, National Institute for Health and Welfare, Helsinki, Finland; Sanna Salanterä, University of Turku and Hospital District of Southwest Finland, Turku, Finland; Simo Vihjanen, Linfox, Turku, Finland; Sumithra Velupillai, Stockholm University, Stockholm, Sweden; Tapio Pahikkala, University of Turku, Turku, Finland; Tapio Salakoski, University of Turku, Turku, Finland; William Goossen, Results 4 Care (Netherlands), University of Minnesota (USA) and Hogeschool Windesheim (Netherlands), Amersfoort, Netherlands, and Wray Buntine, NICTA and Australian National University, Canberra, ACT, Australia. Editor-in-Chief Klaus-Peter Adlassnig from Editorial Office of Artificial Intelligence in Medicine oversaw the review process.

References


Hanna Suominen*a)-d),*

a) NICTA, National Information and Communications Technology, Australia, Machine Learning Research Group, Locked Bag 8001, Canberra ACT 2601, Australia
b) The Australian National University, College of Engineering and Computer Science, Canberra, ACT 0200, Australia
c) University of Canberra, Faculty of Health, Locked Bag 1, Canberra ACT 2601, Australia
d) University of Turku, Department of Information Technology, 20014 University of Turku, Finland

*Corresponding author. Tel.: +61 2 6267 6351;
fax: +61 2 6267 6220

E-mail address: hanna.suominen@nicta.com.au (H. Suominen)