

# Managing Free Text for Secondary Use of Health Data

## Findings from the Yearbook 2014 Section on Knowledge Representation and Management

N. Griffon<sup>1,2</sup>, J. Charlet<sup>2,3</sup>, S. J. Darmoni<sup>1,2</sup>, Section Editors for the IMIA Yearbook Section on Knowledge Representation and Management

<sup>1</sup> CISMeF, Rouen University Hospital, Normandy & TIBS, LITIS EA 4108, Institute for Research and Innovation in Biomedicine, Rouen, France

<sup>2</sup> INSERM, U1142, LIMICS, Paris, France; Sorbonne Universités, UPMC Univ Paris 06, UMR\_S 1142, LIMICS, Paris, France; Université Paris 13, Sorbonne Paris Cité, LIMICS, (UMR\_S 1142), Villetaneuse, France

<sup>3</sup> AP-HP, Dept. of Clinical Research and Development, Paris, France

### Summary

**Objective:** To summarize the best papers in the field of Knowledge Representation and Management (KRM).

**Methods:** A comprehensive review of medical informatics literature was performed to select some of the most interesting papers of KRM and natural language processing (NLP) published in 2013.

**Results:** Four articles were selected, one focuses on Electronic Health Record (EHR) interoperability for clinical pathway personalization based on structured data. The other three focus on NLP (corpus creation, de-identification, and co-reference resolution) and highlight the increase in NLP tools performances.

**Conclusion:** NLP tools are close to being seriously concurrent to humans in some annotation tasks. Their use could increase drastically the amount of data usable for meaningful use of EHR.

### Keywords

Medical informatics, knowledge representation, natural language processing, semantic web, ontology

Yearb Med Inform 2014;167-9

<http://dx.doi.org/10.15265/IY-2014-0037>

Published online August 15, 2014

## Introduction

KRM focuses on developing techniques to be used and leveraged in other medical informatics domain. This year again, the selected articles for the KRM section serve one larger purpose: re-use of health information, independent of their original purpose [1].

This year Natural Language Processing (NLP) papers were reintroduced in the Knowledge Representation and Management (KRM) section. Next year the KRM section will split in Knowledge Engineering (KE) and NLP sections, both domains of medical informatics that will certainly benefit from a larger audience.

The aim of this section is to select and present some of the best papers published in 2013 in the KRM domain, based either on their impact or their novelty approach in the knowledge representation and management field.

## About the Paper Selection

The selection of papers is the result of a comprehensive literature search: section editors pre-selected 15 papers from pubmed and web of science. Those pre-selected papers were then reviewed by a minimum of five reviewers to select the four final papers (see Table 1).

NLP tools are needed to structure the huge amount of data available in free text format in EHRs. Therefore, three of the four selected

papers focus on NLP [2–4]. They respectively center on de-identification challenges [2], gold standard creation for machine learning [3], and co-reference resolution [4]. Together, these three papers give insight in NLP advances, even if sometimes, the solution has been discussed in the past [5].

Deleger et al. [2] not only presented an efficient tool to de-identify clinical notes, they also show that such a tool performs as well as human de-identifiers, who are currently the accepted way to perform de-identification. This created a strong argument to use NLP tools even before they reach perfection. MacLean and Heer [3] proposed a method that is cheaper than expert annotation to create a corpus for NLP training: human-computing [5] with low cost workers. Similar to Deleger's work, results are not perfect but good enough for the authors to promote such a strategy. Chowdhury and Zweigenbaum [4] proposed a set of constraints to enhance co-reference resolution. Focusing on more informative training instances, the proposed approach controls creation of test instances (96% reduction of evaluated instances). The last selected paper from Wang et al. [6], makes use of structured Electronic Health Record (EHR) data to achieve interoperability with a clinical pathway knowledge base in order to adapt standardized clinical pathway to specific patients. There remains still a lot of work to improve interoperability.

A brief content summary of these selected papers can be found in the appendix of this synopsis.

## Conclusions and Outlook

A large percentage of information in EHRs is only available in free-text format. These data constitute a potential goldmine for which data extraction tools are still being designed. This year's selections focus on the enhancement of NLP tools required to make that information (re-)usable. The i2b2/VA 2011 challenge [7] provides researchers with a corpus that has been intensively studied and resulted in many publications this year. For some tasks, NLP tool performances were found to be comparable to human performance, which supports their practical use.

Other aspects of Knowledge Representation and Management are still under scrutiny. Ontology learning provided a good paper as well [8]. Interoperability remains a hot topic with an interesting paper from Tao et al [9], which promoted the need for a standardization of terminology/ontology formalization and proposed some interesting guidelines to achieve this goal. Such a standardization is necessary for any end-user to easily integrate available terminology in a terminology server without a cumbersome and error-prone interpretation of the file format and syntax. We cannot stress enough the necessity to reach interoperability as we were following these guidelines to create the HeTOP terminology server: one major difference when comparing it to BioPortal [10].

### Acknowledgement

We would like to thank Martina Hutter for her support and the reviewers for their participation in the selection process of the IMIA Yearbook.

### References

1. Safran C, Bloomrosen M, Hammond WE, Labkoff S, Markel-Fox S, Tang PC, Detmer DE, Expert Panel. Toward a national framework for the secondary use of health data: an American Medical Informatics Association White Paper. *J Am Med Informatics Assoc* 2007;14(1):1–9.
2. Lamy J-B, Séroussi B, Griffon N, Kerdelhué G, Jaulent M-C, Bouaud J. Selection of the IMIA Yearbook best papers: reducing variability by formalizing the literature search strategy. *Methods Inf Med*. Submitted 2014.
2. Deleger L, Molnar K, Savova G, Xia F, Lingren T, Li Q, et al. Large-scale evaluation of automated clinical note de-identification and its impact on

**Table 1** Best paper selection of articles for the IMIA Yearbook of Medical Informatics 2014 in the section 'Knowledge Representation and Management'. The articles are listed in alphabetical order of the first author's surname.

Section
<b>Knowledge Representation and Management</b>
<ul style="list-style-type: none"> <li>■ Chowdhury MFM, Zweigenbaum P. A controlled greedy supervised approach for co-reference resolution on clinical text. <i>J Biomed Inform</i> 2013;46(3):506–15.</li> <li>■ Deleger L, Molnar K, Savova G, Xia F, Lingren T, Li Q, Marsolo K, Jegga A, Kaiser M, Stoutenborough L, Solti I. Large-scale evaluation of automated clinical note de-identification and its impact on information extraction. <i>J Am Med Inform Assoc</i> 2013 Jan 1;20(1):84–94.</li> <li>■ MacLean DL, Heer J. Identifying medical terms in patient-authored text: a crowdsourcing-based approach. <i>J Am Med Inform Assoc</i> 2013;20(6):1120–7.</li> <li>■ Wang H-Q, Li J-S, Zhang Y-F, Suzuki M, Araki K. Creating personalised clinical pathways by semantic interoperability with electronic health records. <i>Artif Intell Med</i> 2013;58(2):81–9.</li> </ul>

- information extraction. *J Am Med Inform Assoc* 2013;20(1):84–94.
3. MacLean DL, Heer J. Identifying medical terms in patient-authored text: a crowdsourcing-based approach. *J Am Med Inform Assoc* 2013;20(6):1120–7.
4. Chowdhury MFM, Zweigenbaum P. A controlled greedy supervised approach for co-reference resolution on clinical text. *J Biomed Inform* 2013;46(3):506–15.
5. The Human Computer and the Birth of the Information Age [http://www.philsoc.org/2001Spring/2132transcript.html]
6. Wang H-Q, Li J-S, Zhang Y-F, Suzuki M, Araki K. Creating personalised clinical pathways by semantic interoperability with electronic health records. *Artif Intell Med* 2013;58(2):81–9.
7. Uzuner O, Bodnari A, Shen S, Forbush T, Pestian J, South BR. Evaluating the state of the art in coreference resolution for electronic medical records. *J Am Med Inform Assoc* 2012 Sep-Oct;19(5):786–91.
8. Liu K, Mitchell KJ, Chapman WW, Savova GK, Sioutos N, Rubin DL, et al. Formative evaluation of ontology learning methods for entity discovery by using existing ontologies as reference standards. *Methods Inf Med* 2013;52(4):308–16.
9. Tao C, Pathak J, Solbrig HR, Wei W-Q, Chute CG. Terminology representation guidelines for biomedical ontologies in the semantic web notations. *J Biomed Inform* 2013;46(1):128–38.
10. Grosjean J, Soualmia L, Bouarech K, Jonquet C, Darmoni S. Comparing BioPortal and HeTOP: towards a unique biomedical ontology portal? In 2nd Int Work Bioinforma Biomed Eng. in press; 2014.

Correspondence to:  
Prof. SJ. Darmoni, MD, PhD  
Rouen University Hospital  
Department of BioMedical Informatics  
1 rue de Gémont  
76031 Rouen Cedex, France  
Tel: +33(0)232 8888 29  
Fax: +33(0)232 8889 09  
E-mail: stefan.darmoni@chu-rouen.fr

## Appendix: Content Summaries of Selected Best Papers for the IMIA Yearbook 2013, Section Knowledge Representation and Management

Deleger L, Molnar K, Savova G, Xia F, Lingren T, Li Q, Marsolo K, Jegga A, Kaiser M, Stoutenborough L, Solti I

Large-scale evaluation of automated clinical note de-identification and its impact on information extraction

*J Am Med Inform Assoc* 2013 Jan 1;20(1):84–94

To allow breakthroughs in NLP in medicine, it is necessary to provide a corpus of text to developers. Nevertheless, as medical data are considered to be sensitive personal information, a de-identification step is required. This can be performed manually, which is the gold-standard, but comes with high costs and a small corpus. Alternatively the corpus can be created automatically.

This paper describes a three step process to improve knowledge on automatic de-identification tools: (1) the authors manually de-identified an important corpus of clinical notes to evaluate two state-of-the-art algorithms, (2) they compared algorithm performance to human performance and (3) they measured the impact of de-identification on the performance of a medication annotation algorithms. The results reveal that automatic

algorithm performed quite well for most identification information, and even better than humans for some of them. The information extraction has not been worsened by the de-identification. However, running the information extraction test only on medication names, which are very recognizable, provided only little evidence on the absence of a negative impact of de-identification on automatic annotation. According to the authors, it is time to allow research using medical data de-identified by an algorithm, which, unlike manual de-identification, scales easily to millions of medical documents.

**MacLean DL, Heer J**

**Identifying medical terms in patient-authored text: a crowdsourcing-based approach**

**J Am Med Inform Assoc 2013;20(6):1120–7**

This paper focuses on a problem one step after Deleger et al. work: clinical documents are de-identified and annotated to train machine learning algorithm matters. As for de-identification, annotation can be manual – time consuming and costly – or automatic. Resorting to crowd-sourcing is a possible way to decrease costs. However the impact of the crowd on quality annotation has to be measured.

One of the objectives of MacLean work was to annotate a corpus of lay-person sentences by *Turkers* (workers on Amazon's Mechanical Turk – <http://www.mturk.com>) and experts, the former being evaluated against the latter. A second objective was to compare machine learning algorithms trained on the turker-annotated corpus against well known automatic annotation tools.

Turkers performed their job quite well, reaching a reasonable level of consistency and agreement with expert annotators (84% F1-score). It allowed the machine

learning algorithm to outperform the four automatic annotation tools tested (78% F1-score against 47% for Open Biomedical Annotator, which outperformed the others). It seems that relying on crowd sourcing for annotation task may be an effective approach that can decrease annotation cost and allows researchers to work on a bigger corpus.

Like their human computer ancestors, Turkers involvement in crude annotation will probably decline as machine learning algorithms reach performance that rival theirs.

**Chowdhury MFM, Zweigenbaum P**

**A controlled greedy supervised approach for co-reference resolution on clinical text**

**J Biomed Inform 2013;46(3):506–15**

Co-references in a clinical note, when multiple named entities refer to the same thing or person, etc. is another disambiguation task that has to be performed for a machine to understand free-text medical note. The high un-weighted average F1 score (0.915) reached by the best available tool has to be carefully interpreted because all the evaluation metrics are flawed and difficult to interpret.

Chowdhury and Zweigenbaum suggest that controlling generation of less-informative/sub-optimal training and test instances to submit to the machine learning algorithm may enhance its performance. They propose multiple a series of linguistically and semantically motivated rules or filters to use in addition to machine learning system to achieve this goal.

The performances of their solution are slightly lower than those of the best tool available; nevertheless, results are more homogenous according to the metric used. Combining both tools may enhance co-reference resolution, possibly reaching a good level of identification for any kind of named entity.

**Wang H-Q, Li J-S, Zhang Y-F, Suzuki M, Araki K**

**Creating personalised clinical pathways by semantic interoperability with electronic health records**

**Artif Intell Med 2013;58(2):81–9**

In this paper, the authors present a way to personalize clinical pathways (CP) according to patient condition. This work is important because of the duality between the individuality of each patient and the increase in health care standardization. Authors make their semantic EHR and CP ontology/rules interoperable using semantic web technologies i.e. RDF, SPARQL, OWL, and SWRL. Reasoning using patient data is therefore possible, producing a patient-adapted CP. It allows both care standardization and personalized medicine.

From the point of view of knowledge representation, this work is witnessing the convergence on terminological repositories (e.g. ontologies) and reasoning. After 15 years of independent developments, these research axes agree on representation modalities and languages (e.g. OWL / SWRL) for coherent integrations approaches.

The authors present, as example, an acute appendicitis CP. This quite simple example illustrates how the standardized CP can be customized, either because of patient particularities with regard to inclusion criteria in the CP, or because of a particular evolution during the CP. Expanding this approach to multiple pathology CPs will require a significant amount of work.

The authors only used structured data, which are, sadly, still hard to make interoperable, but as the three other selected paper suggested, we may be close to create such an adapted clinical pathway from unstructured data.