Journal of Biomedical Informatics xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Ease of adoption of clinical natural language processing software: An evaluation of five systems

Kai Zheng [a,b,*], V.G. Vinod Vydiswaran [k], Yang Liu [b], Yue Wang [c], Amber Stubbs [d], Ozlem Uzuner [e], Anupama E. Gururaj [f], Samuel Bayer [g], John Aberdeen [g], Anna Rumshisky [h], Serguei Pakhomov [i], Hongfang Liu [j], Hua Xu [f,*]

[a] School of Public Health Department of Health Management and Policy, University of Michigan, Ann Arbor, MI, USA
[b] School of Information, University of Michigan, Ann Arbor, MI, USA
[c] Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI, USA
[d] School of Library and Information Science, Simmons College, Boston, MA, USA
[e] Department of Information Studies, University at Albany, SUNY, Albany, NY, USA
[f] The University of Texas School of Biomedical Informatics at Houston, Houston, TX, USA
[g] The MITRE Corporation, Bedford, MA, USA
[h] Department of Computer Science, University of Massachusetts, Lowell, MA, USA
[i] University of Minnesota, Minneapolis, MN, USA
[j] Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA
[k] Department of Learning Health Sciences, University of Michigan Medical School, Ann Arbor, MI, USA

ARTICLE INFO

Article history:
Received 16 February 2015
Revised 9 June 2015
Accepted 6 July 2015
Available online xxxx

Keywords:
Usability
Human–computer interaction
User-computer interface [L01.224.900.910]
Software design [L01.224.900.820]
Software validation [L01.224.900.868]
Natural language processing
[L01.224.065.580]

ABSTRACT

Objective: In recognition of potential barriers that may inhibit the widespread adoption of biomedical software, the 2014 i2b2 Challenge introduced a special track, *Track 3 – Software Usability Assessment*, in order to develop a better understanding of the adoption issues that might be associated with the state-of-the-art clinical NLP systems. This paper reports the ease of adoption assessment methods we developed for this track, and the results of evaluating five clinical NLP system submissions.

Materials and methods: A team of human evaluators performed a series of scripted adoptability test tasks with each of the participating systems. The evaluation team consisted of four "expert evaluators" with training in computer science, and eight "end user evaluators" with mixed backgrounds in medicine, nursing, pharmacy, and health informatics. We assessed how easy it is to adopt the submitted systems along the following three dimensions: *communication effectiveness* (i.e., how effective a system is in communicating its designed objectives to intended audience), *effort required to install,* and *effort required to use*. We used a formal software usability testing tool, TURF, to record the evaluators' interactions with the systems and 'think-aloud' data revealing their thought processes when installing and using the systems and when resolving unexpected issues.

Results: Overall, the ease of adoption ratings that the five systems received are unsatisfactory. Installation of some of the systems proved to be rather difficult, and some systems failed to adequately communicate their designed objectives to intended adopters. Further, the average ratings provided by the end user evaluators on *ease of use* and *ease of interpreting output* are −0.35 and −0.53, respectively, indicating that this group of users generally deemed the systems extremely difficult to work with. While the ratings provided by the expert evaluators are higher, 0.6 and 0.45, respectively, these ratings are still low indicating that they also experienced considerable struggles.

Discussion: The results of the Track 3 evaluation show that the adoptability of the five participating clinical NLP systems has a great margin for improvement. Remedy strategies suggested by the evaluators included (1) more detailed and operation system specific use instructions; (2) provision of more pertinent

* Corresponding authors at: School of Public Health & School of Information, University of Michigan, M3531 SPH II, 1415 Washington Heights, Ann Arbor, MI 48109, USA. Tel.: +1 (734) 936 6331; fax: +1 (734) 964 4338 (K. Zheng). The University of Texas School of Biomedical Informatics at Houston, 7000 Fannin St, Suite 600, Houston, TX 77030, USA. Tel.: +1 (713) 500 3924; fax: +1 (713) 500-3929 (H. Xu).

E-mail addresses: kzheng@umich.edu (K. Zheng), hua.xu@uth.tmc.edu (H. Xu).

http://dx.doi.org/10.1016/j.jbi.2015.07.008
1532-0464/

onscreen feedback for easier diagnosis of problems; (3) including screen walk-throughs in use instructions so users know what to expect and what might have gone wrong; (4) avoiding jargon and acronyms in materials intended for end users; and (5) packaging prerequisites required within software distributions so that prospective adopters of the software do not have to obtain each of the third-party components on their own.

1. Introduction

Over the past two decades, the advent of new high-throughput technologies has shifted the bottleneck in biomedical research from data production to data management and interpretation. Substantial effort has focused on developing software systems that can better manage, process, and analyze biomedical data. Moreover, biomedical software also plays a critical role in improving productivity and reproducibility of biomedical studies [1]. While some recent attention has been directed toward the challenges related to locating, re-using, and properly citing biomedical software (cf. http://softwarediscoveryindex.org/report/), another important aspect is how easy it is for prospective users and user organizations to adopt these biomedical software systems. In clinical environments, the skepticism surrounding the value and cost effectiveness of health IT had been a key factor accounting for the low adoption rate of electronic health records (EHR) in the U.S. which led to significant government interventions [2,3]. Among the deployed health IT systems, the lack of usability has further hindered their effective use and contributed to numerous unintended adverse consequences such as user frustration and distrust, disrupted workflow, decreased efficiency, and escalated risks to patient safety [4–6]. However, few studies have been conducted to formally investigate the ease of adoption of software that supports biomedical research.

Recently, large EHR databases have become an enabling resource for clinical and translational research [7,8]. One challenge of the secondary use of EHR data is that much of detailed patient information is embedded in narrative clinical documents. Therefore, natural language processing (NLP) technologies, which can extract structured information from free text, have received great attention in the medical domain. Many clinical NLP systems have now been developed and widely used to facilitate various types of EHR-based studies, such as pharmacovigilance, genomic, and pharmacogenomic research [9–13]. While the target users of clinical NLP systems are often more technologically versed, they are by no means immune to poor software adoptability and usability issues [14]. Further, the lack of adoptability could limit the use of NLP systems to a small number of experts, severely undermining their potential for widespread diffusion to broader user bases.

To develop a better understanding of why there has been a lack of adoption of medical NLP tools beyond the community that develops them, a special track, *Track 3 – Software Usability Assessment*, was introduced in the 2014 i2b2 Challenge. The goal of this track was to conduct thorough adoptability evaluations – from software discovery to software installation and use – to assess how well the participating NLP systems might be received by prospective adopters. In this paper, we report the ease of adoption assessment methods that we developed for this track, as well as the results from evaluating five NLP system submissions.

It should be noted that the objective of *Track 3 – Software Usability Assessment* of the 2014 i2b2 Challenge was not to rank the participating systems based on their ease of adoption ratings. First, these systems all serve distinctive purposes and some of them, by nature, are more complicated to adopt than others. Second, the design philosophy of these systems may vary substantially according to their intended use scenarios and method of deployment. For example, some systems may choose to only provide command-line interaction modality so they can be readily invoked from other software programs; whereas some other systems provide rich graphical user interface (GUI) interfaces intended for direct interaction with end users. Thus, the results of the Track 3 evaluation should be interpreted within its own context: a higher ease of adoption rating does not necessarily suggest that a system has superior adoptability relative to the other systems evaluated.

2. Materials and methods

2.1. Scope of evaluation and submission requirements

All current and prior i2b2 Challenge participants who had developed their systems leveraging any of the i2b2 datasets since 2006 were invited to submit their work. Participating teams were only required to provide the name of the system, the URL where its descriptions and user manuals could be found, and the URL from which its executable or source code could be downloaded.

The goal of this track was to evaluate software adoptability from end users' perspective. Therefore, we only accepted systems that had a user interface (command-line or GUI); programmable components that could not be directly operated by end users, such as classes, libraries, and controls, were not included. Further, certain NLP systems offer both an online version where users may enter text or upload input files to be processed, and a downloadable version that can be locally compiled or installed. In such cases, we always chose the downloadable version to evaluate, based on the premise that a local implementation would be the preferred method for most adopting organizations due to HIPAA concerns.

2.2. Evaluators and evaluation environment

A total of twelve evaluators assisted in the Track 3 evaluation. Each of them performed a series of scripted adoptability test tasks with each of the clinical NLP systems submitted.

The two co-chairs of the track (KZ and HX) first created a draft protocol consisting of the test tasks and an evaluation instrument for collecting evaluator feedback (detailed in the next section). Two co-authors of the paper (VV and YL) then did a test run of installing and using each system. Their experience informed the further refinement of the evaluation protocol.

Their experience also led to the recognition that installing some of the participating clinical NLP systems could be a very demanding task well beyond the capability of most average users. Therefore, only four "expert evaluators," all of whom have an undergraduate or graduate degree in computer science, were asked to perform all evaluation tasks including software installation. The remaining eight individuals represent the "end user evaluators" class in the evaluation. They were only asked to work with the systems that had been preinstalled for them.

All of these end user evaluators were graduate students enrolled in the University of Michigan's Master of Health Informatics Program (http://healthinformatics.umich.edu). Six of

Please cite this article in press as: K. Zheng et al., Ease of adoption of clinical natural language processing software: An evaluation of five systems, J Biomed Inform (2015), http://dx.doi.org/10.1016/j.jbi.2015.07.008