Dig Dis Sci (2017) 62:564–566
DOI 10.1007/s10620-016-4359-z

EDITORIAL

Identifying Drug-Induced Liver Illness (DILI) with Computerized Information Extraction: No More Dilly-Dallying

H. Shen[1,2] · **A. Monto**[1,2]

Published online: 9 December 2016

The rapid adoption of electronic medical record (EMR) has provided health-care professionals with better access to patient records while also improving the quality of medical care, reducing medical errors, and lowering medical costs. As a result, the EMR has produced a parallel growth of digitized clinical data, an important medical resource. Clinical data extracted from EMRs have helped health-care professionals support their decisions and have also aided biomedical research, clinical trial screening, adverse drug reaction monitoring, and drug–drug interaction assessment. Nevertheless, a major feature of each EMR is the inclusion of a large amount of clinical narrative text, including medical histories, social histories, laboratory studies, progress notes, discharge summaries, nursing and consultation notes, and pathology, radiology, surgery, and medical imaging reports. Such information is often presented in an unstructured format not immediately suitable for computer analysis. In order to best utilize the vast amount of medical information included in the EMR, data have to be properly extracted and encoded into a structured format suitable for predefined templates. Therefore, effective tools and techniques are required to retrieve and organize these huge volumes of clinical narrative text data in order to make this information useful for supporting medical practice, project management, research, and policy-making. Natural language processing (NLP), and more specifically information extraction (IE), is the most popular and useful technique/tools to date. IE, a subdomain of NLP, is aimed at better understanding the human process of language comprehension in order to develop tools and techniques in order to enable computer systems to manipulate natural languages and perform desired tasks [1]. One of the NLP's major tasks is the extraction of semantic information from text [2]. As a result, large amounts of text can be automatically analyzed by effective extraction tools in order to gather useful information, which can then be represented in a tabular/structured format. In development since the 1950s–1960s [3, 4], the recent literature has reported significant advances in IE, particularly in the last 30 years [5]. Nevertheless, IE has mostly been developed outside of the biomedical domain, being adapted to the biomedical field much later than for other fields.

Since narrative notes are usually written by health-care professionals for documentation and communication purposes, text can be extremely variable in style and content, presenting challenges to extraction of useful information, most notably in three respects: First, clinical narrative text is usually written in ungrammatical fragments. Second, most clinical narrative texts use shorthand orthography (i.e., abbreviations and acronyms), and many also have considerable spelling errors, especially if not spell checked. Lastly, since word meaning is context dependent, some ambiguity and uncertainty are usually present regarding what is being expressed [5]. Since utilization of clinical narrative text for searching and summarization by a computer is extremely difficult, applying NLP to clinical narrative text is an immense challenge. In order to address these difficulties, a typical IE system usually includes a preprocessing step to "clean" the narrative text by expanding abbreviations, shortcuts, and correcting spelling errors while also determining sentence boundaries, tagging parts of speech, disambiguating words, recognizing

✉ H. Shen
Hui.Shen@ucsf.edu

[1] Department of Medicine, University of California San Francisco (UCSF), San Francisco, CA, USA

[2] San Francisco Veterans Affairs Medical Center, Medicine 111-A, 4150 Clement Street, San Francisco, CA 94121, USA

phrases, recognizing named entities, parsing, and combining and extracting to templates [6].

After preprocessing, a variety of approaches have been employed to extract information from free text as to fill out predefined templates. The most common approach is pattern matching, where algorithms are created to exploit basic patterns throughout a variety of structures: text strings, part-of-speech tags, semantic pairs, and dictionary entries. Shallow and full syntactic parsing is other approaches used to analyze a sentence by identifying its constituent parts with discrete grammatical meaning. Due to the differences between general and medical language, sublanguage-driven approaches are used to formulate and exploit a sublanguage's particular constraints. The syntactic and semantic parsing approaches combine the two into one processing step. The last, most expensive and time-consuming approach is machine learning. Machine learning techniques have demonstrated remarkable results in clinical information extraction, although they require large, annotated corpora (a large and structured set of texts) for training [5]. Therefore, a typical IE in NLP is a highly evolved, complex system, often requiring the collaboration of experts from fields as diverse as computer and information science, artificial intelligence and robotics, linguistics, psychology, and medicine. Yet, in the general clinical research setting, there usually are not that many types of resources on hand. The ability to develop simpler and more practical methods which involve less programming and can be more generalized among different institutions is a major challenge yet in great demand in text searching fields.

Drug-induced liver illness (DILI) is an uncommon but important cause of liver disease. It is always challenging to diagnose and identify DILI cases in the EMR system. The article by Heidemann et al. entitled "A Text Searching Tool to Identify Patients with Idiosyncratic Drug Induced Liver Injury" in this issue of *Digestive Diseases and Sciences* uses a pattern matching approach to develop a novel text searching tool to capture idiosyncratic DILI cases from the EMR system. This direct text searching method can be implemented with relative ease because it does not require a large amount of programming or the development of new software. Hence, it is in the end the most generalizable and practical text extraction method.

With their newly developed text searching tool, the authors identified 101 true DILI cases, including 62 probable, 25 possible, 9 historical, and 5 allergy-only cases derived from 2564 potential DILI cases. After modifying search terms, the precision of the results was increased from 4 % to a astonishingly improved 64 %, with the required review time decreased from 29 to 5 h. The results from this study supported the hypothesis that the direct text searching method netted a nearly fivefold increase in the number of idiosyncratic DILI cases identified compared to a previous study that used ICD-9 codes only [7], indicating that information about DILI derived from narrative clinical notes was much more reliable than simply using the traditional identifiers of surrogate terms and ICD-9 codes. Heidemann et al. have also developed a more robust terminology used as a standard of classification in order to detect DILI. This new direct text search method also relied on aspects of manually reviewed snippets to accurately identify true DILI cases. Though the F-measures increased from 8 to 58 %, recall rates decreased from 100 to 53 % after using the modified search terms. All negative cases required tedious manual review.

The current direct text searching algorithm used in this study could be improved in several ways with preprocessing perhaps the most fundamental. The development of contextual features such as "negation" (case/not case), "certainty" (probable/or possible case), "temporality" (recent/historical), and the "event subject identification" (patient/self/others) by reducing the number of negative and unrelated terms can considerably reduce the amount of potential cases while not missing true cases. As seen before, manual review can be avoided and therefore save time as well as increasing recall rate. A state-of-the art method developed by Luo et al. uses a graph-based approach to bridge semantics and syntax to most efficiently detect and analyze contextual features [8]. The machine learning method is also an efficient way to identify DILI cases since it tags records to compare with definitions, associates drugs with symptoms, and creates annotations that bring together the cause and effect between a drug and a symptom [9]. While all of these methods will improve efficiency and accuracy of finding true DILI cases, they may also require significantly more programming work and skill.

Given that the study by Heidemann et al. has addressed a key topic, there is much upside potential to significantly influence clinical research methodology. One practical way to search text without much new programming is to develop a dictionary containing DILI features that would not only include standard terminology and classification, but also all possible lexical variations such as negation (e.g., "…did not clear virus"), temporality (e.g., "….had IFN treatment two years ago"), and event subject identification (e.g., "Her sister died from breast cancer").

Accordingly, the dictionary can be applied to encode and extract entities from narrative text. Furthermore, due to the variety of the descriptions of DILI in the narrative text owing to the diversity of clinicians and disciplines who wrote it, collaboration among multiple sites may be necessary, although shared data and limited collaboration are still the major barriers to IE for clinical text [10].

Developed and developing technology in IE is crucial as rapid digitization of medical records continues. By enabling health-care professionals to organize and wield large amounts of medical information, IE helps provide better diagnosis, treatment, and care for patients. The implications of the technologies discussed in the article will be further felt as we move ever deeper into the digital age, and reveal a growing symbiosis of technology and medicine.

References

1. Chowdhury G. Natural language processing. *Ann Rev Inf Sci Technol*. 2003;37:51–89.
2. Wikipedia. https://en.wikipedia.org/wiki/Information_extraction.
3. Weiss SM, Indurkhya N, Zhang T, Damerau FJ. *Text Ming: Predictive Methods for Analyzing Unstructured Information*. Springer Science + Business Media, Inc. 2005. p 13. ISBN 0-387-95433-3.
4. Sager N, Bross ID, Story G, Bastedo P, Marsh E, Shedd D. Automatic encoding of clinical narrative. *Comput Biol Med*. 1982;12:43–46.
5. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. *Extracting Information from Textual Documents in the Electronic Health Record: A Review of Recent Research*. IMIA Yearbook of Medical Informatics. 2008; 128.
6. Tolentino HD, Matters MD, Walop W, et al. A UMLS-based spell checker for natural language processing in vaccine safety. *BMC Med Inf Decision Making*. 2007;7:3.
7. Heidemann L, Law J, Fontana RJ. A text searching tool to identify patients with idiosyncratic drug-induced liver injury. *Dig Dis Sci*. (Epub ahead of print). doi:10.1007/s10620-015-3970-8.
8. Luo Y, Uzuner O, Szolovits P. Bridging semantics and syntax with graph algorithms—state-of-the-art of extracting biomedical relations. *Brief Bioinf*. 2016;2016:1–19.
9. Aramaki E, Miura Y, Tonoike M, et al. Extraction of adverse drug effects from clinical records. *Stud Health Technol Inform.*. 2010;160(Pt 1):739–743.
10. Chapman WW, Nadkarni PM, Hirschman L, D'Avolio LW, Savova GK, Uzuner O. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. *J Am Med Inform Assoc*. 2011;18:5.