Automated Information Extraction of Key Trial Design Elements from Clinical Trial Publications

Journal:	*AMIA 2008 Annual Symposium*
Manuscript ID:	AMIA-0508-A2008.R1
Manuscript Type:	Paper
Date Submitted by the Author:	n/a
Complete List of Authors:	de Bruijn, Berry; National Research Council, Institute for Information Technology Carini, Simona; University of California, San Francisco Kiritchenko, Svetlana; National Research Council, Institute for Information Technology Martin, Joel; National Research Council, Institute for Information Technology Sim, Ida; University of California, San Francisco
Primary Axis Classification:	II.28. Enhancing the conduct of biological/clinical research and trials

Automated Information Extraction of Key Trial Design Elements from Clinical Trial Publications

Berry de Bruijn, PhD[1], Simona Carini, MA[2], Svetlana Kiritchenko, PhD[1], Joel Martin, PhD[1], Ida Sim, MD, PhD[2]
[1]Institute for Information Technology, National Research Council, Ottawa, Ontario;
[2]University of California San Francisco, San Francisco, CA

Abstract

Clinical trials are one of the most valuable sources of scientific evidence for improving the practice of medicine. The Trial Bank project aims to improve structured access to trial findings by including formalized trial information into a knowledge base. Manually extracting trial information from published articles is costly, but automated information extraction techniques can assist. The current study highlights a single architecture to extract a wide array of information elements from full-text publications of randomized clinical trials (RCTs). This architecture combines a text classifier with a weak regular expression matcher. We tested this two-stage architecture on 88 RCT reports from 5 leading medical journals, extracting 23 elements of key trial information such as eligibility rules, sample size, intervention, and outcome names. Results prove this to be a promising avenue to help critical appraisers, systematic reviewers, and curators quickly identify key information elements in published RCT articles.

Introduction

Randomized clinical trials (RCTs) are one of the least-biased sources of evidence for the practice of medicine. Thousands of new RCT findings are published every year only as text articles. This limits the ways in which computational methods such as decision support systems can use these findings to help physicians translate the evidence into improved practice.[1] Computable RCT knowledge bases that contain the trials' key information in summarized formal fields would allow powerful access. Manually extracting the key information, however, is a laborious task, which we aim to make easier using automatic information extraction techniques.

In the Trial Bank Project, we designed RCT Bank, a computable repository of information on trial design, execution and results. RCT Bank is sufficiently detailed to support critical appraisal and meta-analysis.[2] We were able to capture a wide diversity of RCTs from different clinical domains into RCT Bank, and we showed how trial-bank reporting can be integrated with traditional journal publishing.[3]

The greatest burden associated with such integrated publishing was the need to manually extract data from the article for entry into RCT Bank. Automated information extraction could support this step of the process thereby reducing the cost of transferring RCT data into a computable repository that can be used for systematic reviews, planning of future trials and ultimately to improve the practice of medicine.

While this work focuses on extracting information from full-text journal articles, we aim to extend the same approach to other sources, such as abstracts, protocol documents, and trial registration entries.

Background

Our specific information extraction task is heterogeneous for the following three reasons. Firstly, a wide array of information elements needs to be extracted from full-text RCT journal articles, including eligibility criteria, the name of all experimental and control treatments, intervention parameters (dosage, frequency, duration, etc.), sample size, start and end date of enrolment, primary and secondary outcomes, funding information, and publication details (date, authors). Some of these elements are always present while others may be absent. Some elements can have only one value while others may have several values within the same document (e.g., various funding agencies). Some elements are short and well defined (e.g., drug route) while others could span a longer piece of text with widely varying wording (e.g., eligibility criteria).

Secondly, trials may come from a wide and unrestricted range of medical subfields, from testing pharmacological and procedural treatments to organizational and educational interventions.

Thirdly, in practice we will see a range of document standards and formatting schemas. Richness of in-document annotation may vary between publishers and ranges from detailed XML to various forms of HTML, PDF and even OCR-ed documents in ASCII.

We chose a single information extraction approach that is able to handle this diversity across the various information elements, medical subfields, and

document formats. To extract the value of a certain information element from a text, a text classifier first selects the sentence(s) in the text that is or are most likely to contain the target piece of information. After that, a regular expression matcher pulls out the snippet from the high scoring sentence(s) that contains the information. This combination of a statistical method with minimal ('weak') rules fits the diversity in the task well, since it is less likely to require extensive individual modeling for each information element, medical subfield, and document format than methods with a strong semantic and/or linguistic reliance[4].

There have been a few recent efforts to semantically annotate medical articles, including RCT reports, and clinical documents.[5-9] Most of the research has been focusing on extracting the main characteristics of a study: main condition, interventions, outcomes, and, in the context of RCTs, description of the study population. A typical approach addresses some or all of the three key problems: relevant sentence identification, named entity recognition, and information/relation extraction. Various studies have concentrated on the first step: mapping sentences onto a structured publication template (notably '*background*', '*methods*', '*results*', and '*conclusions*').[10-12] Semantically structured texts represent a richer information source for the next steps of the process. Overall, the applied techniques include state-of-the-art machine learning algorithms (Naïve Bayes, Hidden Markov Models, SVM, Conditional Random Fields)[5,6,8,10-12], manually designed or cue-word-based classification/extraction rules[5,6,8], and use of medical lexicons[5-7,9], such as UMLS, MeSH, or Semantic Groups. In addition, Paek et al. address a general task of semantic parsing of sentences and identifying the semantic roles of the words in a predicate.[13] This extra step can potentially boost the information extraction part of the typical approach. On the whole, the previous research demonstrates that machine learning and NLP techniques can successfully tackle the task of automatic information extraction in the medical domain.

Our work extends the previous research in two main directions. First, we present one general approach to automatically extract over 20 information elements with differing characteristics, while other work has focused on only one to four elements at a time. Second, we work with full-text articles, whereas the past projects use only abstracts or other short text summaries. Full–text articles present more challenges yet allow us to extract information typically not found in abstracts/summaries (e.g. funding agencies, secondary outcomes, and whether the trial was stopped early).

Methods

Data sources: In this study we used a random sample of 88 full-text articles in HTML or XML format from five top-tier medical journals: PLoS Clinical Trials, NEJM, Lancet, JAMA, and Annals of Internal Medicine. The PLoS articles were in XML conforming to the PubMed DTD; the other articles were in journal-specific HTML format. A first set of 78 articles was used for training and cross-validation, a later set of 10 was used for examining the effectiveness of the program on never seen material.

The set of information elements on which we decided to focus for the first phase of the project is based on CONSORT Plus, an extension to trial-bank publishing of the CONSORT statement.[14] The set includes: eligibility criteria, the name of experimental and control treatments, intervention parameters (dosage, frequency, duration, etc.), sample size, start and end date of enrolment, primary and secondary outcomes and relevant time points, funding information, and publication details (date, authors).

The training set of articles was randomly selected among RCTs published (mostly in 2006, a few in 2007) in the relevant journals, and it was hand-annotated to identify the information elements. Predefined tags were used to delimit the elements within the text. Cluster-randomized and cross-over trials were excluded, because they present additional challenges over simple comparison trials.

Procedure: We used a machine learning approach for information extraction. The program learns from training data, i.e., preannotated text. After that, annotation of new material is done in two steps:

(1) a text classifier determines the sentence in which the information element is present

(2) an annotator processes the sentence(s) to extract the exact information value.

Fitting this extraction engine into a general work flow gives the following system design:

- pre-processing, including sentence splitting
- for each information element
    - classification of sentences
    - application of extraction patterns
- post-processing
- presentation of results

In pre-processing, the text is split into sentences. Each sentence is annotated according to the (sub) section in which it occurred ('Abstract', 'Methods', etc.), special characters and symbols are dealt with, and, for a few concepts, occurrences are tagged as such. E.g., "17 women participated" gets tagged into "<integer>17</integer> <person>women</person>

participated". Other general concepts that are tagged include 'units', 'measurements', and 'dates'.

A text classifier is trained for each information element, using a set of 78 texts manually annotated by an expert. The text classifier is based on a support vector machine (SVM)[15] which uses the training texts to learn to identify the most promising sentences. Each sentence is represented with a bag-of-terms, and the terms are words as well as multi-word phrases (word n-grams). Output of the classification stage is a ranked list of the top five sentences.

Weak extraction rules were manually crafted for most information elements. The idea is that a simple extraction rule or pattern, which in itself is not precise enough to extract a detailed piece of information, is quite likely to be accurate when applied within the right context. For example, a *date* could mean many things if seen anywhere in an article, but seeing it in a sentence that was classified as highly relevant to the *start date of enrolment* may be enough to allow labeling it as *enrolment start date*.

Post processing: If, for a certain information element, a highly ranked sentence does not match the corresponding weak extraction pattern, it is bumped down in favour of a next highest scoring sentence that does match the pattern. Several extractions from several sentences are combined to account for redundancy of information, or to return multiple solutions for one information element.

Presentation of results: Results are currently presented in tabular form. They are organized per information element, and for those elements where an extraction pattern matched, the match is presented, linked to the sentence and the position in the text where it originates. A confidence score is also given. As it is, this presentation allows for quick scanning of the results. However, the presentation will need to be redesigned for routine use by database curators.

For a small number of information elements (including author names, doi, publication date), the information is readily available in Medline. Therefore, a separate module within the program links the article to its Medline citation, either by spotting the PMID in the article when present or by searching PubMed on title words. It then fetches the Medline record and parses out the relevant information elements.

Evaluation of the classifier itself was done using leave-one-out cross-validation. Additionally, the performance of the system on ten unseen articles was evaluated. Metrics are recall (sensitivity) and precision (positive predictive value)[16]. For the extraction phase, precision and recall of partial matching is reported, since in many cases the actual boundary of the match is only arbitrarily defined.

Results

As described, information extraction is handled by two modules: the sentence classifier and the extractor module. First, we present the evaluation results of these two modules separately, then we assess the extraction system as a whole.

The effectiveness of the classification module is measured with precision and recall of the top 5 sentences selected by the classifier as the most relevant for a given information element (Table 1, col. 2-4). For most elements, the information value is contained in a single sentence; therefore, we evaluate the capability of the classifier to find at least one sentence with the information on a particular aspect. The results demonstrate that we can frequently find (with recall of 75% and higher) a correct sentence for 80% of the information elements. At this step, we aim at high recall, as most of the invalid sentences will be eliminated at the extraction step.

For the second module, an automatic extractor, we designed a set of rules to extract snippets of required information from the sentences selected in the first step. The rules are rather general, e.g. the first occurrence of a date for the *start date of enrolment*, an integer number with a reference to people (patients, women, subjects, etc.) for the *sample size*, a sequence of words with the first letter capitalized for the *funding organization names*. The resulting set of rules is assessed on the training data (Table 1, col. 5-7). Since the exact boundaries of information bits are quite subjective, we present precision and recall of partial match: an extracted snippet is considered to be correct if it contains at least part of the piece of information selected by a human annotator.

Finally, we evaluate the extraction performance of the entire system on an unseen set of ten documents. Table 1, col. 8-10 present the precision and recall of the snippets for all information elements.

We were able to locate a correct snippet of information in 75% of the 263 cases. Eight of the 23 elements (35%) are extracted with perfect precision and recall. For another 5 elements (22%) we are able to identify all the correct answers with a reasonable precision (R=1.0, P>=0.5).

Discussion

The two-stage architecture, with a sentence classifier and a weak-pattern extractor, relies on two assumptions: first, segmentation at the sentence level is appropriate in that it combines both a large enough

context to do classification and a narrow enough context to get to the target information. Based on our experiences, this assumption holds true for all but one information element we tested: only *eligibility criteria* tends to be described in a text segment that spans multiple sentences.

The second assumption is that weak rules can be used to extract the exact snippets from candidate sentences. Applying the same weak rules across a large text (the entire article) would extract many irrelevant snippets, but not if the context is restricted enough. Our results support this assumption.

The system deals well with some particularities of the information elements. For information elements where more than one single piece of information needs to be extracted (for instance, a study that is funded by multiple organizations), the extraction algorithm does find all of these. For elements where the text contains no solution (e.g., *drug dosage* for a trial where the intervention is not a drug), the system is capable of indicating the absence with a likelihood, and can provide the closest text match in case the curator wishes to double-check.

There are only two elements that were not yet handled in a satisfactory way: *funding organization name* and *name of the experimental treatment*. For the former, performance within the development set was good but for the new set of ten unseen documents, the system was often misled by organization names from the 'author affiliations' section. While this drop in performance may be magnified due to an idiosyncrasy in the extension set, we aim for a more structural solution by looking into a better strategy for seeding the text classifier. The element *experimental treatment name* is more challenging than others because a treatment can take many forms beyond a drug or a medical procedure (e.g., education, environment factors, equipment, management interventions, diet or lifestyle changes). Some of our future work will focus on handling this component in a more sophisticated way.

The two-stage architecture can be further improved. First, some information elements may appear redundantly in the text, for instance in the abstract as well as at various points in the body of the article. Examples are the names of the experimental and control treatments. Since the same information needs to be extracted only once, the algorithm could benefit from redundancy. The current implementation does not yet fully leverage this redundancy.

Second, in some cases the multiple answers between certain elements have to be grouped with each other. For example, one trial may test several experimental treatments where each treatment has its own dosage,

Table 1: Performance of the classification and extraction modules and the entire system.
n = number of data points, P=precision, R=recall. Evaluation on *Classifier* and *Extractor* separately used 78 development articles in a cross-validation set-up. The evaluation of the *Entire System* used 10 additional and previously unseen articles.

Information Element	Classifier			Extractor			Entire System		
	n	P	R	n	P	R	n	P	R
Author name (first author only)	N/A	N/A	N/A	N/A	N/A	N/A	10	1.00	1.00
Date of publication	N/A	N/A	N/A	N/A	N/A	N/A	10	1.00	1.00
DOI (Digital Object Identifier)	N/A	N/A	N/A	N/A	N/A	N/A	10	1.00	1.00
Dose (multiple dosages possible)	46	0.53	0.91	144	0.80	0.92	21	0.90	0.90
Duration of the treatment	41	0.30	0.59	72	0.80	0.88	17	0.94	1.00
Early stopping	5	0.06	1.00	5	1.00	0.80	1	1.00	1.00
Eligibility criteria	77	0.90	0.92	N/A	N/A	N/A	37	0.69	0.54
End date of enrolment	54	0.68	1.00	63	0.98	0.98	8	0.80	1.00
Frequency of treatment	40	0.42	0.83	68	0.91	0.88	16	0.76	1.00
Funding organization name	77	0.94	0.96	177	0.84	0.94	21	0.11	0.33
Funding number (grant number)	34	0.39	0.91	78	0.98	1.00	5	0.56	1.00
Make of device	2	0.01	0.50	5	0.67	0.40	1	0.50	1.00
Manufacturer of device	7	0.08	0.86	9	0.75	0.89	2	0.17	1.00
Name of control treatment	69	0.75	0.86	158	0.91	0.50	11	1.00	0.82
Name of experimental treatment	78	0.82	0.83	343	0.89	0.28	16	0.67	0.38
Primary outcome name	77	0.87	0.90	N/A	N/A	N/A	10	1.00	1.00
Primary outcome – time point	58	0.48	0.66	106	0.78	0.84	8	0.46	0.75
Registration identifier of trial	57	0.68	0.95	64	1.00	1.00	11	1.00	1.00
Route of treatment	20	0.23	0.90	45	0.92	0.98	5	1.00	1.00
Sample size	78	0.43	0.44	95	0.81	0.80	10	0.62	0.80
Secondary outcome name	50	0.57	0.90	N/A	N/A	N/A	9	0.70	0.78
Secondary outcome - time point	21	0.20	0.76	41	0.71	0.85	14	0.91	0.71
Start date of enrolment	71	0.89	0.99	84	1.00	0.99	10	1.00	1.00
Overall	**962**	**0.51**	**0.84**	**1557**	**0.87**	**0.82**	**263**	**0.65**	**0.75**

frequency, route, and duration; also, primary/ secondary outcomes may each have their respective time points. The system will have to link the corresponding values for these elements.

The current system demonstrates a very good performance for most of the elements. Since the system is not intended to replace human annotators, but rather to assist them by proposing possible values for required information elements, we consider the achieved performance as promising.

Conclusion

In this work we present a general methodology for the problem of automatic information extraction from RCT publications. A uniform two-stage process, in which target sentence identification is followed by application of weak extraction rules, is applied to locate the trial's design, treatments, intervention parameters, and outcomes in the full text. For most of the information elements (15 out of 23), the system found the key sentence or sentences nearly every time, and the correct information was indeed correctly extracted in 75% of the 263 test cases. These preliminary results demonstrate the system's ability to assist critical appraisers, systematic reviewers, and curators in extracting essential information from RCT reports.

Future work includes a refinement of the algorithms based on the results obtained so far, the design of an interactive user interface for curators to review, and, if necessary, correct the extracted information and a more extensive system evaluation using articles from different journals. Finally, a "productivity" evaluation will need to be performed to establish the value of the extraction software (measured by time saved and extraction accuracy/completeness) in the context of a curation environment.

Acknowledgments

Valuable work on this project was done by Imad Tbahriti while he was a visiting worker at NRC-IIT.

Supported by grant LM-06780 from the National Library of Medicine.

References

1. Sim I, Gorman P, Greenes RA, Haynes RB, Kaplan B, Lehmann H, Tang PC. Clinical Decision Support Systems for the Practice of Evidence-Based Medicine. J Am Med Inform Assoc. 2001 Nov–Dec; 8(6): 527–534.
2. Sim I, Olasov B, Carini S. An ontology of randomized controlled trials for evidence-based practice: content specification and evaluation using the competency decomposition method. J Biomed Inform. 2004 Apr;37(2):108-19.
3. Sim I, Carini S, Olasov B, Jeng S. Trial bank publishing: phase I results. Medinfo. 2004;11(Pt 2):1476-80.
4. Erhardt RA, Schneider R, Blaschke C. Status of text-mining techniques applied to biomedical text. Drug Discov Today. 2006 Apr;11(7-8): 315-25.
5. Demner-Fushman D, Lin J. Knowledge extraction for clinical question answering: preliminary results. AAAI Workshop on Question Answering in Restricted Domains 2005.
6. Niu Y, Hirst G. Analysis of semantic classes in medical text for question answering. AAAI Workshop on Question Answering in Restricted Domains, 2005.
7. Borlawsky T, Friedman C, Lussier YA. Generating executable knowledge for evidence-based medicine using natural language and semantic processing. AMIA Annu Symp Proc. 2006: 56–60.
8. Xu R, Garten Y, Supekar KS, Das AK, Altman RB, Garber AM. Extracting subject demographic information from abstracts of randomized clinical trial reports. Medinfo 2007;12(Pt 1):550-4.
9. Chen ES, Hripcsak G, Xu H, Markatou M, Friedman C. Automated acquisition of disease–drug knowledge from biomedical and clinical documents: an initial study. J Am Med Inform Assoc. 2008;15:87-98
10. McKnight L, Srinivasan P. Categorization of sentence types in medical abstracts. AMIA Annu Symp Proc. 2003: 440–4.
11. Xu R, Supekar K, Huang Y, Das A, Garber A. Combining text classification and hidden markov modeling techniques for structuring randomized clinical trial abstracts. AMIA Annu Symp Proc. 2006: 824–8.
12. Chung GY, Coiera E. A study of structured clinical abstracts and the semantic classification of sentences. BioNLP 2007: 121-8.
13. Paek H, Kogan Y, Thomas P, Codish S, Krauthammer M. Shallow semantic parsing of randomized controlled trial reports. AMIA Annu Symp Proc. 2006: 604-8.
14. Complete CONSORT Plus Guideline [Internet] San Francisco (CA): UCSF Trial Bank Project c2008 [cited 2008 Jul 8] Available from: http://rctbank.ucsf.edu/consort/cplus.html
15. Joachims T: Text Categorization with Support Vector Machines; Learning with Many Relevant Features. Proc 10th Annu Eur Conf on Machine Learning. 1998: 137-42
16. Van Rijsbergen C.J.: Information Retrieval, 2nd edition. London, Buttersworth 1979.