Improving Search for Evidence-based Practice using Information Extraction

Jin Zhao[1], Min-Yen Kan[1], Paula M. Procter[2], Siti Zubaidah[3], Wai Kin Yip[3], Goh Mien Li[3]

[1]National University of Singapore, Singapore
[2]Sheffield Hallam University, Sheffield, United Kingdom
[3]National University Hospital, Singapore

Abstract

The search for applicable and valid research evidence-based practice articles is not supported well in common EBP resources, as some crucial study data, such as patient details, study design and results, are not available or presented explicitly. We propose to extract these data from research articles using a two-step supervised soft classification method. Compared to manual annotation, our approach is less labor-intensive and more flexible, hence opening up the possibility of utilizing these data to facilitate the evidence selection process in information seeking support systems.

Introduction

Despite the growing popularity of evidence-based practice (EBP) in healthcare, support for the gathering and selection of applicable and valid research articles in today's EBP resources can still be improved. Published guidelines[1] recommend that a clinical question needs to be established using PICO[2] (i.e., patient, intervention, comparison and outcome) as shown in Table 1a and 1b. These identified elements can serve as the criteria in determining whether a certain research article is applicable.

Going beyond PICO, there is also a hierarchy in the strength of evidence[2] (as shown in Table 2) for articles. This hierarchy helps to ensure the validity of the research articles, as ones of a lower grade (i.e., stronger evidence) are preferred over higher ones.

However, common EBP resources seldom provide such metadata explicitly or allow users to filter for these criteria. Although users may be able to perform keyword searches and limit their searches by gender, age and study design in PubMed, they cannot specifically target keywords that match only the text sections that discuss PICO elements or strength of evidence. As such, users must resort to reading the abstract or the full text to ascertain whether an article is indeed applicable and valid.

Name	Definition
Patient	The description of the patient. It commonly consists of five elements: sex, co-morbidity, race, age and pathology (SCORAP)
Intervention	The intervention applied.
Comparison	Another intervention examined as a comparison or control
Outcome	The outcome of the experiment

Table 1a: Definitions of PICO.

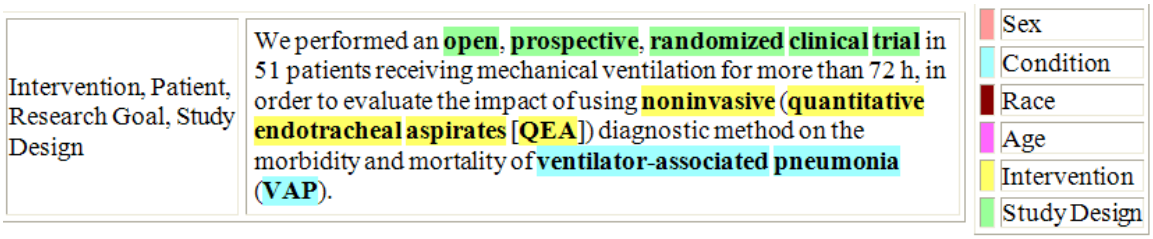
Clinical Question: For a 54-year-old woman with periodontal disease, how effective is the therapeutic use of doxcyline decrease gum bleeding and recession compared to no treatment?	
P	54-year-old (Age) woman (Sex) with periodontal disease (Pathology)
I	Doxcyline
C	No treatment
O	Decrease gum bleeding and recession

Table 1b: PICO of a sample clinical question.

Grade	Definition
I	Systematic reviews of all relevant RCTs
II	At least one properly designed RCT
III-1	Well designed pseudo-RCT
III-2	Cohort studies, case control studies, interrupted time series without control
III-3	Comparative studies with historical control, two or more single-arm studies or interrupted time series without control
IV	Case series

Table 2: Different levels of strength of evidence.

We believe the extraction of such information from articles is the key to solve this problem. With such information extracted, additional functionalities can be implemented into information systems to support the judgment process. For example, as illustrated in Figure 1, a system can display the key sentences of a research article and highlight the keywords in the sentence that reveal key information such as the intervention and study design. The users can then judge the applicability and validity of the articles immediately without the need to read them in full. Moreover, this extraction has to be automated, since manual extraction would be too labor intensive due to the large amount of research articles available.



Intervention, Patient, Research Goal, Study Design	We performed an open, prospective, randomized clinical trial in 51 patients receiving mechanical ventilation for more than 72 h, in order to evaluate the impact of using noninvasive (quantitative endotracheal aspirates [QEA]) diagnostic method on the morbidity and mortality of ventilator-associated pneumonia (VAP).	🟥 Sex 🟦 Condition 🟫 Race 🟪 Age 🟨 Intervention 🟩 Study Design

Figure 1: System display of the extraction results to assist the users in applicability and validity assessment.

Name	Definition	Example
Patient	A sentence containing information of the patients in the study.	A convenience sample of 24 critically ill, endotracheally intubated children was enrolled before initiation of suctioning and after consent had been obtained.
Result	A sentence containing information about the results of the study.	Large effect sizes were found for reducing PTSD symptom severity (d = –.72), psychological distress (d = –.73) and increasing quality of life (d = –.70).
Intervention	A sentence containing information about the procedures of interest and the ones as the comparison /control.	Children 6 to 35 months of age received 0.25 ml of intramuscular inactivated vaccine, and those 36 to 59 months of age received 0.5 ml of intramuscular inactivated vaccine. (Note: This is also a *Patient* sentence.)
Study Design	A sentence containing information about the design of the study.	A prospective international observational cohort study, with a nested comparative study performed in 349 intensive care units in 23 countries.
Research Goal	A sentence containing information about what the study aims to achieve.	The aim of this study was to investigate the balance between pro- and anti-inflammatory mediators in SA.

Table 3. Classes for key sentences.

Mining textual medical publications for different intents has been a long-standing area of interest for the community. In general, supervised machine learning and natural language processing techniques are popular and successful in medical information extraction (IE). For example, Niu and Hirst[3] answer clinical questions by categorizing sentences into three semantic classes: diseases, medications and clinical outcomes. They then analyze the relationship among them based on cue word features using the support vector machine (SVM). Demner-Fushman and Lin[4] perform extraction based on hand-crafted patterns (for elements P, I and C) and linear regression of text features (for O) on abstracts to obtain necessary information for clinical question answering. Chung and Coiera[5] classify clinical abstracts into five classes – Aim, Method, Participants, Results and Conclusion – and extract the number of patients using a Conditional Random Field (CRF) and a SVM model. Bruijin et al.[6] make use of a SVM-based sentence classifier with n-gram features and a rule-based weak-pattern extractor to identify the key trial design elements from clinical trial publications.

However, it is crucial to point out that our application of *information retrieval* (IR) differs from the above works on standalone information *extraction.* When used in an IR system, the documents gathered by the crawler often come from different sources and hence can differ greatly in formatting, layout and structure. Therefore, the extraction method should only rely on text features and make as little assumption about the document as possible. Moreover, as the amount of documents to be handled in an IR system tends to be large, it is more useful for the extraction method to focus on finding a small set of useful information instead of aiming to discover all of them. Lastly, the usability of the extracted information is critical. When fine-grained (i.e., word-level) extraction is not possible or questionable, a coarse-grained (i.e., sentence-level) extraction may still be of use in supporting the applicability and validity judgment.

Method

To address these concerns, we propose to extract information from an article in two soft classification steps: one to perform coarse-grained extraction of useful sentences that pertain to evidence search, and another to perform fine-grained extraction of word-level information from the sentences.

Step 1: Our first classifier assigns sentences into one of five classes, as listed in Table 3. The first three classes map to PICO elements: *Patient* → P, *Intervention* → I and C, and *Result* → O. In addition, we also have a fourth class, *Study Design*, which indicates the strength of evidence for the users, and a fifth class, *Research Goal*, which the users can match their own clinical questions with.

These five classes are not mutually exclusive. As shown Figure 1 and the example of *Intervention* class in Table 3, a sentence may contain more than one type of information. To implement the necessary soft clustering, for each class, we train a binary classifier to decide whether a sentence belongs to this class. A sentence is extracted when at least one classifier reports positive; and may belong to multiple classes.

Feature	Definition
Token	The N-grams (sequences of N words, where 1≤N≤3) of the sentence.
Sentence	The length of the sentence and its position in the paragraph and in the article.
Named Entity	Whether the sentence contains person name, location name and organization name. These features are extracted using the OpenNLP package[8].
MeSH	Whether the sentence contains MeSH terms and their categories among the 16 top categories of the MeSH tree.
Lexica	Whether the sentence contain a word which appears in the age/sex/race wordlists. All these wordlists are manually compiled by the first author and contain common words found in the corpus which indicate age, sex and race, respectively.

Table 4. Features for key sentence extraction.

Our classifiers are based on Maximum Entropy[7] with the features listed in Table 4. In this formulation, the relations between the features and the types of the sentences are captured in a probability distribution function. The training of the classifier is to find the probability distribution function that encodes the information given by the training data with the largest entropy (i.e., least bias).

Step 2: We further classify the words from the sentences which belong to the *Patient*, *Intervention* and *Study Design* classes. There are six classes for our word-level classification as listed in Table 5. The first four cover SCORAP elements of patients: *Sex* → S, *Condition* → CO and P, *Race* → R and *Age* →A. The last two are introduced to extract the names of the intervention and study design.

The classification process is the same as sentence classification except for the features extracted. The feature set used for this classification process can be found in Table 6. Note that, in both classification steps, we do not include any features on the layout, formatting or text structure of the document.

Name	Definition	Example
Sex	The sex of the patients.	male, female
Age	The age (group) of the patients	54-year-old, children
Race	The race of the patients	Chinese, Indian, Caucasian
Condition	The condition of the patients, usually a disease name.	COPD, asthma
Intervention	The name of the procedure applied to the patients.	intramuscular inactivated vaccine
Study Design	The name of the design of the study.	cohort study, RCT

Table 5. Classes for keywords

Feature	Definition
Token	The word itself, its stem and its part-of-speech tag.
Phrase	The position of the word in the phrase and the head noun of the phrase if it is a noun phrase. The boundary and the type of the phrases are identified using the OpenNLP package.
Named Entity	Whether the word is part of a person name, location name or organization name in the sentence. These features are extracted using the OpenNLP package.
MeSH	Whether the word is part of a MeSH term and the categories of that term among the 16 top categories of the MeSH tree.
Lexica	Whether the word appears in the age/race/sex wordlists. The wordlists used here are the same as the ones used in key sentence extraction.

Table 6. Features for keyword extraction.

Evaluation Methodology

We gathered a collection of 19,893 medical abstracts and full text articles from 17 journal websites which contain quality research materials as recommended by the nurses from the Evidence-Based Nursing Unit in National University Hospital. From this collection, 2,000 randomly selected sentences were annotated by the first author for the evaluation of sentence extraction.

Within the resulting dataset, there are 161 (8%) sentences that belong to the *Patient* class, 95 (4.8%) in *Intervention*, 449 (22.5%) in *Result*, 102 (5.1%) in

Class/Feature Group	**All**			**No Token**			**No Sentence**			**No Named Entity**			**No MeSH**			**No Lexica**		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
Patient	.70	.24	.36	**-.13**	**+.08**	**+.05**	**-.10**	-.04	-.03	+.02	-.01	0	0	+.02	+.02	-.05	-.01	-.02
Intervention	.78	.56	.65	**-.40**	**-.10**	**-.17**	**-.14**	**-.05**	**-.07**	-.04	-.03	-.04	**-.07**	**-.07**	**-.09**	-.04	-.03	-.04
Result	.90	.28	.43	**-.17**	**-.23**	**-.22**	-.02	+.02	+.01	-.01	0	0	-.01	+.02	+.01	-.02	-.01	0
Study Design	.89	.39	.54	**-.56**	**-.29**	**-.39**	**-.09**	-.04	**-.05**	-.01	**-.05**	**-.05**	-.01	+.03	+.03	0	+.01	0
Research Goal	.92	.27	.43	**-.47**	**-.08**	**-.16**	+.03	-.01	-.02	**-.05**	+.02	+.01	+.04	+.03	+.04	+.03	-.01	-.02

Table 7. Evaluation result of key sentence extraction and the performance change when omitting a feature class.

Class/Feature Group	**All**			**No Token**			**No Phrase**			**No Named Entity**			**No MeSH**			**No Lexica**		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
Sex	.98	1	.99	0	0	0	0	0	0	0	0	0	0	0	0	0	**-.12**	**-.06**
Condition	.76	.63	.69	+.03	-.03	-.01	-.03	**-.15**	**-.10**	0	**-.10**	+.01	-.01	**-.17**	**-.11**	-.02	-.01	-.01
Race	.92	.86	.89	**-.08**	0	-.03	**-.08**	0	-.03	**-.18**	0	-.03	**-.08**	0	-.03	0	**-.79**	**-.79**
Age	.85	.78	.81	-.04	**-.07**	**-.06**	+.04	**-.07**	-.03	-.02	0	-.02	0	0	-.01	-.01	-.04	-.03
Intervention	.74	.58	.65	0	-.01	-.01	**-.07**	**-.20**	**-.18**	0	+.04	+.02	+.04	**-.08**	**-.05**	-.01	+.02	.01
Study Design	.87	.73	.80	**-.15**	**-.23**	**-.21**	0	-.08	-.06	-.03	+.02	0	**-.11**	+.02	-.03	-.04	+.01	-.01

Table 8. Evaluation result of keyword extraction and the performance change when omitting a feature class.

Study Design, 70 (3.5%) in *Research Goal* and 1,333 (66.5%) others not belonging to any class.

For the evaluation of keyword extraction, 6,754 words from 667 sentences that belong to the *Patient*, *Intervention* and *Study Design* classes were separately annotated by the first author. There were 52 (0.8%) words in *Sex* class, 175 (2.6%) in *Age*, 14 (0.2%) in *Race*, 366 (5.4%) in *Condition*, 371 (5.5%) in *Intervention*, 255 (3.8%) in *Study Design* and 5,567 (82.5%) others not belonging to any class.

We trained the classifiers with the corresponding corpus and evaluated their performance using the standard IR measures: Precision, Recall and F_1-Measure defined as follows:

Precision (P) = TP / (TP+FP)
Recall (R) = TP / (TP+FN)
F_1-Measure (F) = 2 * P * R / (P + R)

where TP: true positive, FP: false positive, FN: false negative.

A 5-fold cross validation[9] is applied in all the experiment to avoid overfitting.

Results and Discussion

The resulting performance for sentence extraction is listed in Table 7's leftmost column. The results indicate that the sentence extraction is precise, especially for *Result*, *Study Design* and *Research Goal*, but there is much room for improvement on recall for all classes.

Table 7 also shows changes in performance when each group of features is removed individually. Among the features, token features are the crucial ones as removing them lead to the most significant drop in performance. In addition, sentence features also play an important role in achieving a high precision. Other than these two, the contribution of the rest of the features is mixed, i.e., benefitting some classes while harming some others.

We believe that recall is low due to the sheer variety of ways a sentence of a given class can be written. Moreover, sentence extraction is based on the type of information it contains yet such information is often too short (only a few words) compared to the total length of the sentence. This difference in length makes it difficult for the classifiers to catch all the necessary information to determine the type of the sentence. While have a low recall is not optimal, we believe this is acceptable for search engines as there is limited space for displaying each article; listing all classified sentences would not reduce the amount of text required for users to read.

As for keyword extraction, Table 8 shows the overall performance of our system and how it changes if one group of the features has been removed.

In general, our keyword extraction classifiers perform well, especially for the *Sex* and *Race* category (F_1-Measure > .9).

In terms of the contribution of different types of features, the token features contribute the most to the *Study Design* class. We believe the main reason behind is that this class is the only class that is not covered by any MeSH term category or wordlist. The phrase features also help to extract *Study Design* and

Intervention classes, as both are often expressed as noun phrases, such as “cohort study” and “hormone therapy,” where the head nouns are highly indicative of the class. Named entity features do not help much as sentences belonging to these classes do not use named entities often. Lastly, both MeSH and lexica features contribute greatly to extraction by covering the vocabulary for the classes.

We have identified two major sources of errors. As in the first sentential task, there are many different ways to express the age of the patients, such as “children,” “45-year-old” and “35 to 40 years of age.” These variations are not captured by the current set of features. Therefore, the extraction performance of the *Age* class is lower than those of *Sex* and *Race* classes that exhibit less variation. Similar problems also occur in the *Intervention* and C*ondition* classes as they can be expressed as abbreviations or in prose form instead of their canonical names. Moreover, due to the fact their vocabulary sizes are too large to be completely covered by MeSH terms, their extraction performance is the lowest among all.

Future Work

Since the current performance of keyword extraction is much better than that of key sentence extraction and the latter needs to be more informed about the small pieces of information in each sentence, we plan to implement a joint-inference framework so that key sentence extraction and keyword extraction can inform each other in one unified process. This framework should help to improve the recall of key sentence extraction.

In addition, we believe one way to improve keyword extraction is to employ a more comprehensive knowledge source, such as UMLS, to facilitate the detection of medical terms and their abbreviations. However, linguistic variation is endless and techniques for managing yet unseen variations are needed. Topic modeling techniques[10] that have begun to make headway in natural language processing and machine learning tasks may be investigated in the future.

Displaying the extracted results is just the first step in the integration of extraction results into the search process. In the future, we plan to incorporate more functionality, such as ranking search results based on how well the query matches the extracted sentences instead of the whole document, and to filter the documents based on the extracted keywords.

Conclusion

We are in the process of building an information seeking support system for evidence-based practice[11]. Key sentence and keyword extraction, which facilitates applicability and validity judgment on research articles, is one of the main features in the system. Despite the fact that our approach is simplistic, our evaluation already yields promising results. Therefore, besides improving the extraction performance with more sophisticated approaches, how to utilize the extracted results in an information seeking support system is also an important topic for further investigation.

References

1. Sackett DL, Richardson WS, Rosenberg WMC, Haynes RB. Evidence-based medicine: how to practice and teach EBM. 2nd Ed. London, Churchill-Livingstone, 2000.
2. National Health and Medical Research Council, NHMRC: A guide to the development, implementation and evaluation of clinical practice guidelines, 1999.
3. Niu Y and Hirst G. Analysis of semantic classes in medical text for question answering. In ACL 2004 Workshop on QA in Restricted Domains.
4. Demner-Fushman D, Lin J, Answering clinical questions with knowledge-based and statistical techniques. *Comput. Linguist.* 2007, 33(1):63-103.
5. Chung GY, Coiera E. A study of structured clinical abstracts and the semantic classification of sentences. In Proc. of the wksp on BioNLP 2007, 121-8.
6. Bruijn B, Carini S, Kiritchenko S, Martin J, Sim I, Automated information extraction of key trial design elements from clinical trial publications, AMIA Annu Symp Proc. 2008, 141–5.
7. Maxent: URL: http://maxent.sourceforge.net/.
8. OpenNLP: URL: http://opennlp.sourceforge.net/.
9. Wikipedia, Cross-validation. URL: http://en.wikipedia.org/wiki/Cross-validation.
10. Mark S., Tom G., Probabilistic topic models, In: Landauer T, McNamara D, Dennis S, Kintsch W (eds), Latent semantic analysis: a road to meaning. Lawrence Erlbaum Associates, 2007.
11. Zhao J, Kan MY, Procter PM, et al., eEvidence: information seeking support for evidence-based practice: an implementation case study, AMIA Annu Symp Proc. 2010.