Exploiting Classification Correlations for the Extraction of Evidence-based Practice Information

Jin Zhao, Praveen Bysani, Min-Yen Kan

National University of Singapore, Singapore

Abstract

Crucial study data in research articles, such as patient details, study design and results, need to be extracted and presented explicitly for the ease of applicability and validity judgment in evidence-based practice. To perform this extraction, we propose to use two soft classifications, one at the sentence level and the other at the word level, and exploit the correlations between them for better accuracy. Our evaluation results show that propagating the results from the first classification to second improves performance of the second and vice versa. Moreover, the two classifications may benefit each other and help improve performance through joint inference algorithms. Another key finding of our work is that irrelevant sentences in the training data need to be properly filtered out; otherwise they compromise system accuracy and make joint inference models less scalable and more expensive to train.

Introduction

The gathering and selection of applicable and valid research articles for evidence-based practice[1] (EBP) requires that healthcare practitioners locate crucial study data from those articles. For applicability assessment, PICO[2] elements (*i.e.*, patient, intervention, comparison and outcome) need to be matched with criteria in the clinical questions that healthcare practitioners have in mind. For validity assessment, the strength of evidence[2], as determined by the articles' study design, helps to establish a ranking among the articles since stronger evidence is generally preferred.

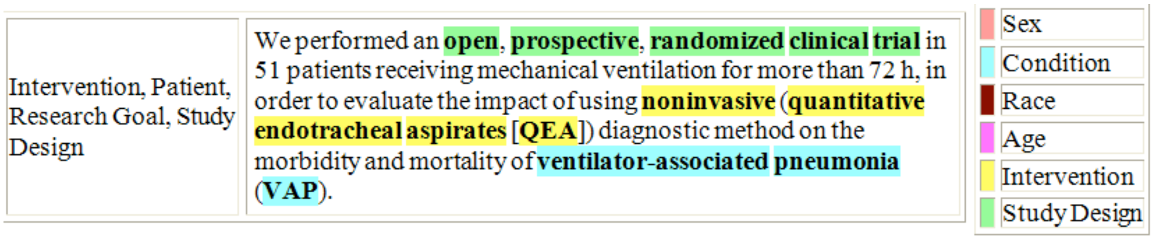
However, most of these study data are not immediately available to healthcare practitioners; common EBP resources seldom provide such metadata explicitly. Although users may be able to limit their searches by gender, age and study design in PubMed, these data are usually manually annotated and not directly associated to the relevant sections of the research articles. As such, users must resort to reading the abstract or even full text of the articles to locate these study data in the articles before proceeding to applicability and validity assessment.

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 1: Definitions of PICO elements.

An automatic extraction of such information from articles can solve this problem. For example, as illustrated in Figure 1, with such information extracted, a system can display the key sentences of a research article and highlight its keywords that express such crucial study data, such as the intervention and study design. The users can then assess the applicability and validity of the articles without the need to read the articles in full.

The identification and utilization of PICO elements and their variants have been studied extensively for various intents. Most of the previous works in this area are based on supervised learning with natural language processing techniques. For example, Demner-Fushman and Lin[3] perform sentence extraction based on hand-crafted patterns



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: Display of the extraction results to assist the users in applicability and validity assessment.

(for elements P, I and C) and linear regression of text features (for O) on abstracts as a way to obtain information for clinical question answering. Chung and Coiera[4] seek for a better understanding of the structure of clinical abstracts by classifying the sentences in them into five classes – Aim, Method, Participants, Result and Conclusion. To investigate how the accuracy of sentence classification can be improved in general, Kim et al.[5] explore the use of lexical, semantic, structural and sequential information with Conditional Random Fields, while Boudin et al.[6] test and combine multiple classifiers, such as Decision Tree, SVM and Naïve Bayes, for more robust classification.

In comparison, research on more fine-grained extraction of EBP information is less common. Existing works usually start by classifying the sentence in abstracts or articles to identify the possible locations of the EBP information and then proceed to extract the information from those locations. For example, Bruijin et al.[7] 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. Chung[8] extracts the intervention information from the method sentences of randomized controlled trials with lexical and syntactical features.

The above works either focus on sentence extraction or use sentence extraction as a basis for keyword extraction. While individually important tasks, we believe that the composition of both tasks together would be synergistic and would lessen the effort needed in relevance assessments.

- *Sentence extraction* is important because not all of the key information is fully captured by individual words. For example, research results are commonly described in prose. It is difficult to extract only a few words to represent the entire text. Extraction at the sentence level is ideal in this case. Even for information such as patient demographics that can be represented by a few words, sentence-level classification still imparts evidence that the specific key words are being used in an appropriate context.
- *Keyword extraction* is also important because recognized keywords often precisely represent the information the users need. Without keyword extraction, the practitioner needs to go through each sentence to locate the desired information. Furthermore, for the extracted sentences with multiple types of information, the recognized keywords can be highlighted differently based their classes for the ease of reading and assessment.

This work builds on our previous work[9] that tackles the same two classification problems. In contrast to the previous work, our contribution in this work is to tackle both levels of extractions simultaneously, and study how one classification task may influence the other, and investigate how to optimize both.

Method

We cast both extractions as soft-classification tasks as follows:

Key Sentence Classification: We use a five-class scheme as listed in Table 3. The first three classes map to PICO elements: Patient → P, Intervention → I/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 helps them determine whether a research article is likely to provide useful information to the clinical questions they have in mind.

Keyword Classification: We use six classes for words as listed in Table 4. The first four cover SCORAP

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 sentences.

elements of patients: Sex → S, Condition → CO/P, Race → R and Age → A. The last two are introduced to extract the names of the intervention and study design.

As discussed earlier, both classifications are equally important for the extraction of EBP information and hence a suitable technique should address both levels of classifications. In a closer inspection of our classes, we make two key observations that inform our approach.

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 4. Classes for words (tokens).

As an example, let us first examine the Patient class from the sentence classification and the Sex, Age, Race and Condition class from keyword classification. These two sets of classification tasks are correlated: if a sentence is classified as a Patient sentence, the words in this sentence are more likely to represent Sex, Age, Race and Condition information of the patients. Likewise, if some words in a sentence have been categorized into one of the Sex, Age, Race, Condition classes, this sentence contains patient information and hence should be classified as a Patient sentence. Similar correlations can be identified between the Study Design / Intervention sentence class and the corresponding keyword classes.

A straightforward approach for utilizing such correlations is to perform the classifications in sequence so that the results from the earlier classification can be incorporated into the later one. This gives rise to the two pipelined models we propose, as shown in Figure 2. In the Sentence-First model (upper right of Figure 2), the sentence classification is performed first and the resulting sentence class labels are added to the feature vectors of the word classification as additional evidence. In the Word-First model (lower left of Figure 2), this process is reversed: label results from the first word classification task are added as features for the second sentence classification task.

While these two models are able to incorporate information from earlier classification result into the later task, there is no way for the earlier classification task to benefit from the results of the later. Consequently, classification performance can improve on one level but not both. To overcome this problem, we investigate a third, Joint model (lower right of Figure 2). In this model, the two levels of classifications are mutually informed of each others' results via joint inference. Sentence labels now may influence the prediction of keyword labels and vice versa. While often advantageous in accuracy, we note that joint models significantly increase the model complexity of the classifier and impact the computational overhead in training as well as in applying the resulting models. Finally, as a baseline for comparison, we have implemented an Independent model (upper left of Figure 2), in which the two classifications are done independently.

Our second observation on sentence-level classes is that the classes are not mutually exclusive. As shown in Figure 1 and the example of Intervention class in Table 3, a sentence may contain more than one type of information. To achieve such necessary soft clustering, for each class, we train a binary classifier to decide whether a sentence (or keyword) belongs to this class. A sentence (or keyword) is extracted when at least one classifier reports positive; and may belong to multiple classes when two or more classifiers report positive on the sentence (*e.g.*, if the Patient and Intervention sentence classifiers both report positive).

We implement these models using Conditional Random Field (CRF)[10]. CRFs are a class of statistical modeling method whose strength is in modeling the context of the instances being learned / classified (*e.g.*, neighboring instances). Therefore, it is commonly applied on labeling sequential data[11] and suits the task of keyword classification since the tokens in a sentence naturally form a sequence and this sequence can serve as the context for predicting the label of any token in it. Moreover, the structure of the CRF can be arbitrarily defined such that different instances from different classification problems can be learned in the same model. This feature of CRF allows us to build the necessary joint model. We use the MALLET package[13] for the Independent, Sentence-First and Word-First models and the GRMM package[12] for the Joint model. As depicted in Figure 2, in the Independent model, the sentence classification classifies one sentence at a time based on its features, while the word classification labels a sequence of tokens in the same sentence based on their features and labels of the tokens in this sequence. In the Sentence-First model, labels predicted from sentence classification are added to the feature vectors

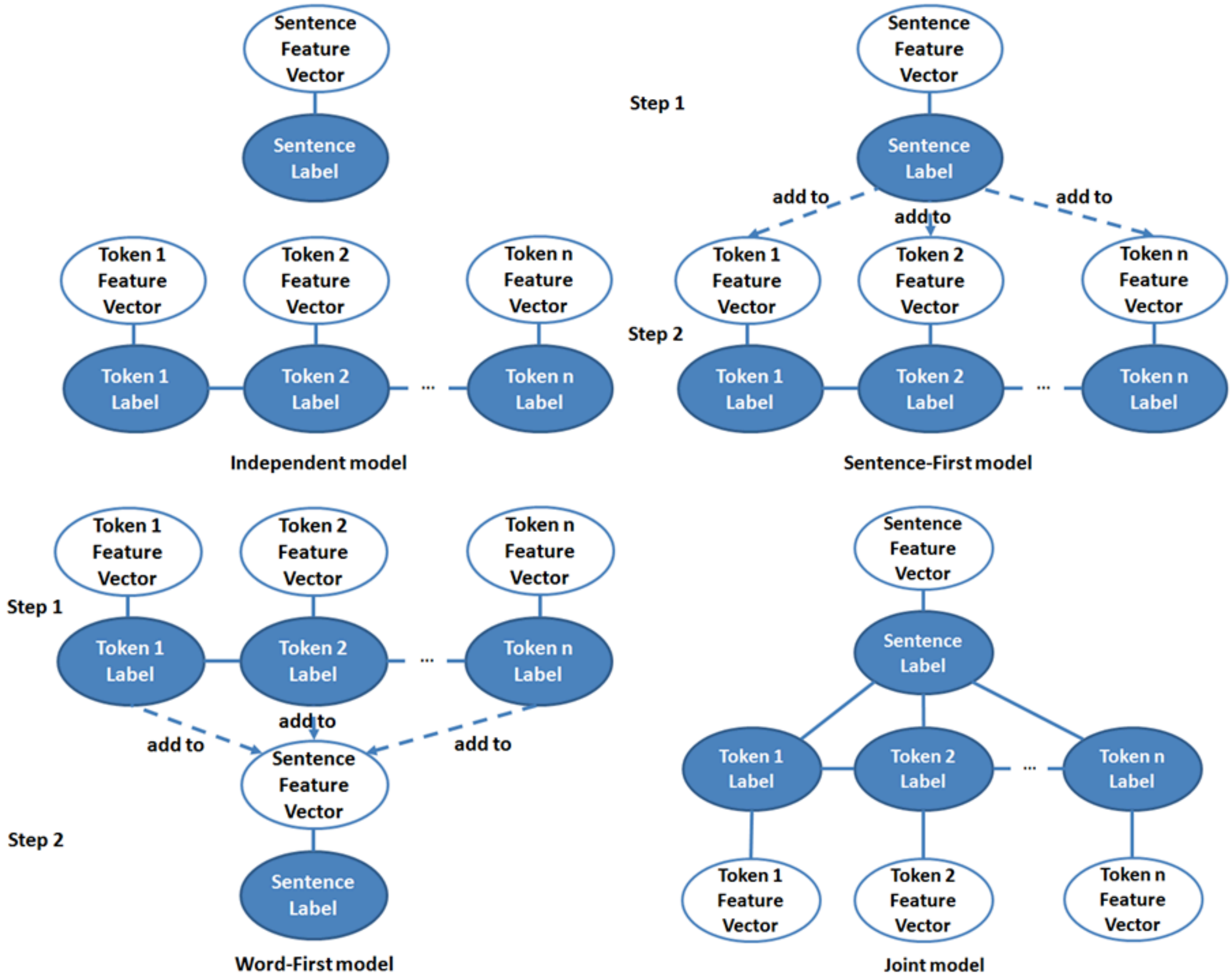


Figure 2: Four different models for key information extraction.

of the tokens as additional information for word classification, while the process proceeds in the other way round in the Word-First model. In the Joint model, the feature vectors are the same as the ones in the Independent model but both classifications are performed as one to find the most likely labels for a sentence and its tokens based on their features and labels of the others. Note that in all models, we do not use the sequence of sentences as the context for predicting the classes of a sentence since it would lead to a highly complicated graph (*i.e.*, the entire document being represented as one graph) in the Joint model.

The features we use for sentence classification and keyword classification are provided in Table 5 and 6 respectively. Both feature sets consist of generic text classification features, such as word sequences of length n (*n*-grams) and named entity information, as well as domain-specific features, such as MeSH terms and class-specific lexica.

As shown in our previous work, all the listed features contribute positively to the two classifications. For example, token features are the crucial to sentence classification as removing them can lead to significant drop in performance, while MeSH and lexica features play important roles in keyword classification by covering the necessary vocabulary for the classes. Since the main focus of this paper is to explore different ways to exploit the correlation between classifications instead of identifying new features for them, we use these two feature sets as they are.

Evaluation Methodology

We reuse the corpus from our previous work (with revised annotations) for evaluation. This corpus is a collection of 19,893 medical abstracts and full text articles from 17 journal websites that contain quality research materials as recommended by the healthcare practitioners from the Evidence-Based Nursing Unit in National University

Feature	Definition
Token	The N-grams (sequences of N words, where $1 \leq N \leq 3$) 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.
MeSH	Whether the sentence contains MeSH terms and their categories.
Lexica	Whether the sentence contain a word that appears in the age/sex/race wordlists.

Table 5. Features for key sentence extraction.

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.
Named Entity	Whether the word is part of a person name, location name or organization name in the sentence.
MeSH	Whether the word is part of a MeSH term and the categories.
Lexica	Whether the word appears in the age/race/sex wordlists.

Table 6. Features for keyword extraction.

Hospital. From this collection, 2,000 randomly selected sentences were annotated for the evaluation of sentence extraction.

Within the resulting dataset, there are 220 (11%) sentences that belong to the Patient class, 174 (8.8%) in Intervention, 448 (22.5%) in Result, 119 (6%) in Study Design, 71 (3.6%) in Research Goal and 1,329 (66.4%) others not belonging to any class.

For the evaluation of keyword extraction, 12,339 tokens (including words and punctuation) from 360 sentences that belong to the Patient, Intervention and Study Design classes were annotated. There are 72 (1.1%) words in Sex class, 177 (1.4%) in Age, 19 (0.2%) in Race, 531 (4.3%) in Condition, 607 (4.9%) in Intervention, 284 (2.3%) in Study Design and 10,651 others not belonging to any class.

Considering the fact that the Joint model is computationally expensive, we decide to first train the models only on the sentences containing at least one type of key information to obtain a preliminary sense on how the models perform. This is referred to as the reduced dataset. Later, we move on to evaluating them on the full dataset, in which a large amount of irrelevant sentences are present as noise, and examine the negative impact of such noise on the classification performance. Lastly, we apply data filtering as a preprocessing step to alleviate this negative impact of irrelevant sentences and present the final results.

We evaluate the performance of classifiers using the standard information retrieval of measures of precision, recall and F_1-Measure, which are 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[14] is applied in all experiments to avoid overfitting.

Results and Discussion

Reduced Dataset without Data Filtering. The evaluation results for key information extraction on the reduced dataset using the four models are listed in Table 7. This dataset represents an artificial case where we know *a priori* that a sentence does contain key information and the sentence classification stage is only used to determine which of the five classes it belongs to. The general classification performance, as shown in the results of Independent model, indicates that the extractions are precise for all sentences classes ($P > 0.8$) and some of keyword classes, such as Sex, Race and Study Design. However, there is still much room for improvement on recall for most classes. For sentence classification, the high precision suggests that a small portion of the sentences from each class is easily recognized, perhaps because it is written in a conventionalized style. In contrast, the low recall signals that the majority of the sentences – especially those in Intervention, Study Design and Research Goal classes – is still hard to detect, possibly due to two causes: 1) the variety of linguistic expressions, and 2) that crucial information that determines the sentence's class may be too short (1-2 words) when compared to the whole sentence's length (which could be on the order of 50 or more words). For keyword classification, the problem of linguistic variation also plagues the recall for some of the classes. For example, "children," "45-year-old" and "35 to 40 years of age" are all valid ways of expressing age information of the patients. Moreover, when the vocabulary size of a class is too large to be effectively covered by existing medical dictionaries (*e.g.*, Condition and Intervention), the extraction of

Class/Model	**Independent**			**Sentence-First**			**Word-First**			**Joint**		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
Sentence Classification												
Patient	.81	.75	**.78**	Same as Independent			**.84**	.72	**.78**	.64	**.90**	.75
Intervention	**.82**	.47	.60				.73	.55	**.63**	.62	**.59**	.61
Result	.90	.95	.92				.89	**.96**	**.93**	**.91**	.91	.91
Study Design	**.97**	.59	.73				.93	.70	**.79**	.83	**.76**	**.79**
Research Goal	**.95**	.58	.72				**.95**	.58	.72	.86	**.67**	**.76**
Keyword Classification												
Sex	.89	**.92**	**.90**	**.90**	.86	.88	Same as Independent			.88	.71	.79
Condition	.45	.31	.36	**.60**	**.41**	**.49**				.59	.36	.45
Race	.80	.42	.55	.82	**.47**	**.60**				**1**	.11	.19
Age	.71	**.57**	**.63**	.73	.49	.59				**.76**	.45	.57
Intervention	.59	.34	.43	**.77**	.35	**.48**				.57	**.39**	.47
Study Design	.85	.73	.78	.90	.62	.74				**.91**	**.75**	**.82**

Table 7. Evaluation results on the *reduced* dataset using the four models. Bolded figures indicate best performance among the models, individually for precision (P), recall (R) and F_1-measure (F).

keywords in this class is also greatly compromised. While having a low recall is not optimal, we believe this is acceptable for the purpose of workload reduction in reading. In our opinion, it is better to present to the users only a few sentences/words that are very likely to capture the key information than to list down many possible sentences/words which incurs significant effort and time for the users to review.

As for the other models, the Sentence-First model readily improves classification performance on the challenging keyword classes such as Condition and Intervention. However, it also harms the extraction of some of the other keyword classes. Based on our error analysis, we discover that when a sentence belongs to multiple classes, which happens in 272 (~40%) of the 671 sentences in the reduced dataset, multiple sentence labels are added into the feature vectors of its keywords, leading to an inaccurate learning keyword model. In contrast, keyword labels are strong indicators for the sentence labels. Therefore, the Word-First model does help to improve the sentence classification on most classes.

As for the Joint model, it is comparable to the rest of the models when the relationship between the sentence and keyword is simple. For example, we observed that it performed well for the two Study Design classes, largely because the Study Design sentences are only concerned with Study Design keywords and vice versa. In comparison, it is less effective for the Patient sentence class and the four related keyword classes since the correlations among these five classes are more complex. Nevertheless, it is the only model that can enhance sentence classification and keyword classification at the same time. This nature eliminates the need to choose between the sequence of classification tasks and thereby hindering the classification accuracy of the latter task. However, despite all these advantages, its scalability is still a major drawback. Even on the reduced dataset, it requires much more time and memory to train.

Full Dataset without Data Filtering. While the results on the reduced dataset are informative, they do not represent the complete picture, since both classifications need to be done on all sentences, not just the ones that contain the key information. Classification on the full dataset constitutes a real world trial, as entire abstracts are provided. Table 8 shows the performance of the models when the full dataset is substituted for the reduced dataset.

The 1,329 sentences that are added can be considered as noise since none of them contain key information. The presence of such noise adds to the challenge for both classifications and leads to lowered performance for all models in general. Among all the results, only the precision for the Intervention, Study Design and Research goal are maintained at the same level, indicating that the some of the sentences in these classes are still easily distinguishable even in the full dataset.

Due to this lowered performance, in the Word-First model, the keyword classification results now may harm the sentence classification instead of helping. This is due to the many occurrences of keywords outside of key sentences. The most adversely affected sentence class is the Patient class, because it is related to most (four) classes of keywords, many of which can no longer be reliably classified. In contrast, sentence classification now also functions as a filter for the sentences that do not contain any information instead of just distinguishing the sentences of one

Class/Model	Independent			Sentence-First			Word-First			Joint		
	P	R	F	P	R	F	P	R	F	P	R	F
Sentence Classification												
Patient	**.75** (-.06)	.52 (-.23)	**.61** (-.17)	Same as Independent			.67 (-.17)	.37 (-.35)	.48 (-.30)	.52 (-.12)	**.71** (-.19)	.60 (-.15)
Intervention	**.82** (0)	.34 (-.13)	.48 (-.12)				.58 (-.15)	.38 (-.17)	.46 (-.17)	.58 (-.04)	**.50** (-.09)	**.54** (-.07)
Result	**.78** (-.12)	**.63** (-.32)	**.70** (-.22)				**.78** (-.11)	.60 (-.36)	.68 (-.25)	.77 (-.14)	.58 (-.33)	.66 (-.25)
Study Design	**.97** (0)	.51 (-.08)	.67 (-.06)				.91 (-.02)	.65 (-.05)	.76 (-.03)	.84 (+.01)	**.71** (-.05)	**.78** (-.01)
Research Goal	**.97** (+.02)	.45 (-.13)	.62 (-.10)				**.97** (+.02)	.42 (-.16)	.59 (-.13)	.79 (-.07)	**.63** (-.04)	**.70** (-.06)
Keyword Classification												
Sex	.63 (-.26)	.63 (-.29)	.63 (-.27)	**.74** (-.16)	**.76** (-.10)	**.76** (-.12)	Same as Independent			.68 (-.20)	.60 (-.11)	.64 (-.15)
Condition	.20 (-.25)	.11 (-.20)	.14 (-.22)	**.53** (-.07)	**.34** (-.07)	**.42** (-.07)				.49 (-.10)	**.34** (-.02)	.40 (-.05)
Race	.62 (-.18)	**.42** (0)	**.50** (-.05)	.83 (+.01)	.26 (-.21)	.40 (-.20)				**1** (0)	.05 (-.06)	.10 (-.09)
Age	.56 (-.15)	**.44** (-.13)	.49 (-.14)	**.66** (-.07)	.42 (-.07)	**.52** (-.07)				.62 (-.14)	.36 (-.09)	.46 (-.11)
Intervention	.46 (-.13)	.25 (-.09)	.32 (-.11)	**.74** (-.03)	.26 (-.09)	.39 (-.09)				.49 (-.08)	**.36** (-.03)	**.42** (-.05)
Study Design	.81 (-.04)	.64 (-.09)	.71 (-.07)	**.93** (+.03)	.59 (-.03)	.72 (-.02)				.86 (-.05)	**.71** (-.04)	**.78** (-.04)

Table 8. Evaluation results on the *full* dataset using the four models. Bolded figures indicate best performance among the models, individually for precision (P), recall (R) and F_1-measure (F). The numbers in the brackets show the relative performance compared to the results on the *reduced* dataset.

class from others. With this classification acting as a filter, it is less likely for the keywords in the newly-added sentences to be misclassified as representing key information. As a result, keyword classification may still benefit from sentence classification despite the drop of performance in the latter.

The Joint model also suffers a drop in performance. However, the worst consequence on the shift to the larger dataset is on the resources required for training. For this experiment, it takes up to one day and more than 10GB of memory to train, while the rest of the models can still be trained with about the same level of computational resources as before (*i.e.*, within minutes).

Full Dataset with Data Filtering. To reduce the noise due to the additional irrelevant sentences and alleviate the scalability problem for the Joint model, we also investigated the use of filtering as a preprocessing step.

The idea of data filtering is to build an additional classifier to filter out the sentences that do not contain any key information. When an unseen sentence is given, this filtering classifier is first applied to determine whether the sentence contains any key information. If it does, the sentence and the words in it will be further classified into the sentence and keyword classes; otherwise this sentence and its words are discarded, deemed as not belonging to any of the sentence or keyword classes. This scenario is similar to the reduced dataset case, but incurs errors from the filtering. In this way, the level of noise is minimized and the cost of training the Joint model is alleviated.

For consistency, we implemented the filtering classifier as a one-against-all classifier using the feature set for sentence classification. All the sentences not belonging to any of the sentence classes are considered as positive examples and the rest are negative examples.

The results after applying data filtering, as shown in Table 9, are generally favorable. Improvements can be observed in both sentence and keyword classification for most classes. Moreover, the improved keyword classification is able to benefit sentence classification once again, as indicated by the performance of the Word-First model and the Joint model. Last but not least, with data filtering, the Joint model only needs to be trained those that the filtering classifier deemed positive, which is approximately the same magnitude as the reduced dataset. Therefore, the computational resources required for this model remains manageable and unaffected by the size of the full dataset.

Class/Model	Independent			Sentence-First			Word-First			Joint		
	P	R	F	P	R	F	P	R	F	P	R	F
Sentence Classification												
Patient	.64 (-.11)	.64 (+.12)	**.64** (+.03)	Same as Independent			**.67** (0)	.60 (+.23)	.63 (+.15)	.49 (-.03)	**.76** (+.05)	.60 (0)
Intervention	**.70** (-.12)	.41 (+.07)	.51 (+.03)				.63 (+.05)	.48 (+.10)	.54 (+.08)	.53 (-.05)	**.54** (+.04)	**.55** (+.01)
Result	.68 (-.10)	**.69** (+.06)	.68 (-.02)				.66 (-.12)	**.69** (+.09)	.68 (0)	**.72** (-.05)	.65 (+.07)	**.69** (+.03)
Study Design	**.92** (-.05)	.55 (+.04)	.68 (+.01)				.90 (-.01)	.68 (+.03)	**.78** (+.02)	.74 (-.10)	**.71** (0)	.73 (-.05)
Research Goal	**.94** (-.03)	.47 (+.02)	.62 (0)				**.94** (-.03)	.47 (+.05)	.62 (+.03)	.93 (+.14)	**.56** (-.07)	**.70** (0)
Keyword Classification												
Sex	.67 (+.04)	.68 (+.05)	.68 (+.05)	**.68** (-.06)	**.69** (-.07)	**.69** (-.07)	Same as Independent			**.68** (0)	.61 (+.01)	.64 (0)
Condition	.38 (+.18)	.26 (+.15)	.31 (+.17)	.46 (-.07)	**.35** (+.01)	.40 (-.02)				**.52** (+.03)	.33 (-.01)	**.41** (+.01)
Race	.72 (+.10)	.41 (-.01)	.53 (+.03)	.89 (+.06)	**.42** (+.16)	**.57** (+.17)				**1** (0)	.11 (+.06)	.19 (+.09)
Age	.60 (+.04)	.53 (+.09)	.57 (+.08)	**.68** (0)	.49 (+.07)	.57 (+.05)				**.68** (+.06)	**.54** (+.18)	**.60** (+.14)
Intervention	.49 (+.03)	.32 (+.07)	.39 (+.07)	**.65** (-.09)	.33 (+.07)	**.44** (+.05)				.51 (+.02)	**.39** (+.03)	**.44** (+.02)
Study Design	.76 (-.05)	.71 (+.07)	.73 (+.02)	**.89** (-.04)	.62 (+.03)	.73 (+.01)				.85 (-.01)	**.72** (+.01)	**.78** (0)

Table 9. Evaluation results on the *full* dataset using the four models, with *data filtering*. Bolded figures indicate best performance among the models, individually for precision (P), recall (R) and F$_1$-measure (F). The numbers in the brackets show the relative performance compared to the results on the *full* dataset, *without data filtering*.

Although the resulting performance is not as good as those on the reduced dataset, data filtering is easy to implement, meets both of our goals, and makes our method immediately applicable to current medical abstracts.

Aside from data filtering, we have also experimented with the idea of feature selection as one way to address the scalability issue of the Joint model. Ideally, by selecting a good subset of the relevant features, both the noise from irrelevant features and the dimensionality of the feature space are reduced. As such, the resulting models are often more accurate and take fewer resources to train. However, after exploring a few common feature selection techniques with different selection sizes, we observed that although the training cost can be reduced in this way, the model performance underperforms our models reported here with data filtering. Therefore, we omit the results from these experiments in this paper.

Future Work

Propagating the results from one classification to another may do more harm than good if the former cannot be done reliably. This is often referred to as cascading error. As shown in our evaluation, keyword classification may negatively influence the accuracy of sentence classification when its accuracy is compromised due to noise. Therefore, the first direction of our future work is to improve the individual classification tasks before combination. To this end, we plan to employ more comprehensive knowledge sources, such as UMLS, as well as more sophisticated statistical models – such as topic models[15] – to manage the endless possible variations of medical terms and sentence structure.

Moreover, although the Joint model provides a natural way to propagate information between the two levels of classifications, it requires much more computational resources than the other models to train and hence does not scale well to larger training dataset. Besides reducing the training data size by filtering out the irrelevant sentences, we plan to explore other joint inference algorithms to find a suitable alternative that is less computationally expensive. Approximate inference algorithms can also be considered as long as reasonable accuracy can be achieved.

Conclusion

The extraction of key information plays an important role in facilitating the applicability and validity judgment of research articles in EBP. Such key information can be most properly expressed at the word-level or at the sentence-level. As such, the extraction task needs to be performed using two separate yet correlated classifications, one for each level. In our work, we have exploited these correlations by propagating information between the two classifications through pipelined or joint models. The pipelined models are simple and inexpensive but only work in one direction. In contrast, the joint model works addresses this unidirectional shortcoming, but is costly to train, especially when large amount of irrelevant sentences are present in the training data as noise. Adding a pre-filtering classifier to remove extraneous sentences before key information classification begins yields the best compromise on performance in terms of real-world applicability, classification performance and scalability.

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. Demner-Fushman D, Lin J, Answering clinical questions with knowledge-based and statistical techniques. Computational Linguistics 2007, 33(1):63-103, 2007.
4. Chung G, Coiera E. A study of structured clinical abstracts and the semantic classification of sentences. In Proc. of the workshop on BioNLP 2007, 121-8, 2007.
5. Kim SN, Martinez D, Cavedon L, Yencken L, Automatic classification of sentences to support Evidence Based Medicine, BMC Bioinformatics 12(Suppl2):S5, 2011.
6. Boudin F, Nie JY, Bartlett JC, Grad R, Pluye P, Dawes M, Combining classifiers for robust PICO element detection. BMC Medical Informatics and Decision Making 2010, 10:29, 2010.
7. Bruijn B, Carini S, Kiritchenko S, Martin J, Sim I, Automated information extraction of key trial design elements from clinical trial publications, AMIA Annual Symposium 2008, 141–5, 2008.
8. Chung G, Sentence retrieval for abstracts of randomized controlled trials. In: BMC Medical Informatics and Decision Making, vol. 9, p.10, 2009.
9. Zhao J, Kan MY, Proctor PM, et al., Improving search for evidence-based practice using information extraction, AMIA Annual Symposium 2010, 2010.
10. Sutton C, McCallum A, Introduction to conditional random fields for relational learning. In Introduction to statistical relational learning, MIT Press, 93-128, 2007.
11. McDonald R and Pereira F, Identifying gene and protein mentions in text using Conditional Random Fields, BioCreative, 2004.
12. Sutton C, GRMM: GRaphical Models in Mallet. http://mallet.cs.umass.edu/grmm/, 2006.
13. McCallum A: MALLET: A machine learning for language toolkit. http://mallet.cs.umass.edu/, 2002.
14. Wikipedia, Cross-validation. http://en.wikipedia.org/wiki/Cross-validation.
15. 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.