An Empirical Method of Automatic Pattern Extraction for Clinical Text Classification

Musarrat Hussain, Jamil Hussain, Taqdir Ali, and Sungyoung Lee*

Abstract— Clinical text classification is an indispensable and extensively studied problem in medical text processing. Existing research primarily employs machine learning and pattern based approaches to address the stated problem. In general, pattern based approaches perform better than other methods. However, these approaches commonly require human intervention for pattern identification, which diminish their benefits and restrain their applications. In this study, we present a novel pattern extraction algorithm, which identifies and extracts patterns from clinical textual resources, automatically. The algorithm identifies the candidate concepts in the clinical text, finds the context of the concepts by discovering their context windows, and finally transforms each context window to a pattern. We evaluate our proposed algorithm on Hypertension, Rhinosinusitis, and Asthma guidelines. 70% of the hypertension guideline was used for pattern extraction while the remaining 30% and the other two guidelines were used for evaluations. The algorithm extracts 21 patterns that classify Hypertension, Rhinosinusitis, and Asthma guidelines sentences to the recommendation and non-recommendation sentences with 84.53%, 80.03%, and 84.62% accuracy, respectively. The initial results reveal the benefits and applicability of the algorithm for clinical text classification.

I. Introduction

Text classification is an essential prerequisite of most of the text mining applications [1]. The classification is more critical and challenging task in medical domain comparatively, because of noisy information, complex vocabularies, sparsity, medical measures, misspelling, abbreviations, and poor grammatical structure of the sentences. Many researchers applied machine learning and pattern based approaches to acquire promising information from unstructured documents [2]. Pattern based approaches perform better than machine learning methods in clinical text classification [3]. However, the performance of pattern based approaches depends on the set of patterns used for classification. Accurate and up to date patterns can provide precise classification, albeit with extensive manual intervention by a human expert with ample technical skills[7], [11]. Therefore, an algorithm, that can automate the pattern identification and extraction process, while maintaining the accuracy and minimizing human involvement, is necessary.

Plenty of clinical text is available online in various categories including blogs, articles, and guidelines. One of the most credible and standard forms of clinical text is the Clinical Practice Guidelines (CPGs). CPGs are “statements that include recommendations intended to optimize patient care that are informed by a systematic review of evidence and an assessment of the benefits and harms of alternative care options” [4]. It contains valuable information and detailed process flows to guide healthcare providers for providing standard healthcare services [15]. The CPGs sentences are categorized into two categories: recommendation sentences (RS) and non-recommendation sentences (NRS). RS describe the causes, consequences, and actions required in a particular patient scenario. It is one of the valuable sources of medical knowledge. While NRS represent background information and thought of the authors. The primary goal of this study is to classify RS and NRS sentences of CPGs precisely.

In this study, we devise a novel algorithm that identifies and extracts patterns for recognizing RS sentences in CPGs, automatically. The algorithm determines the initial candidate concepts based on the popularity of a concept in a CPG. This is augmented with the concept window corresponding to each concept. Context window is the combination of the candidate concepts with its neighbor concepts. Ineffective context windows with low occurrences are filtered out and the remaining context windows are transformed into regular expressions as patterns. These patterns are used to categorize CPG sentences into RS and NRS.

The proposed algorithm is beneficial to healthcare applications dealing with CPG processing. It extracts scenario specific information, which improves the healthcare services during the patient care. Additionally, the algorithm is beneficial for the preprocessing steps of CPG transformation into computer interpretable form and information extraction applications.

II. Related Work

Patterns mining has been widely explored for various purposes including information extraction, text classification, and named entity recognition [13], [12]. However, most of these approaches typically generate the patterns in the form of regular expression, manually, by software developers in collaboration with domain experts [5]. There are very few studies that automate the process of pattern extraction from example text without human involvements.

M.Hussain, J.Hussain, T.Ali, and S.Y.Lee {musarrat.hussain, jamil, taqdir.ali, sylee}@oslab.khu.ac.kr

Department of Computer Science and Engineering, Kyung Hee University, Seocheon dong, Giheung-gu, Yongin-si, Gyeonggi-do, Republic of Korea, 446-701

* Corresponding Author

978-1-7281-1990-8/20/$31.00 ©2020 IEEE

Bui et al. [5] devised a regular expression discovery (RED) algorithm to automate the generation and applications of patterns in text classification. The RED extracts patterns from an annotated set of text snippets and those are used to classify unannotated snippets. RED generates patterns in four steps including snippets alignment, key generation, pattern builder, and filtering. In the alignment step, RED matches each text snippet with all other snippets of the same class and identify the overlapping terms. The key generation creates key terms from the overlapped text snippets. Pattern Builder transforms the generated key to regular expression format. While in filtering, RED filter out the generated patterns based on the negative text nipped of the target class. The filtered patterns are considered as final RED generated patterns that can identify the targeted class text snippets in the unannotated text.

Murtaugh et al. [6] developed a supervised learning algorithm, called Regular Expression Discovery Extractor (REDEx) to detect bodyweight related measures including weight, height, and BMI. The REDEx algorithm generates patterns from annotated text, and the extracted patterns are used to get the numerical values of interest. The authors extract the patterns in four steps. Initially, each annotated snippet is split into labeled segment (LS), before the labeled segment (BLS), and after labeled segment (ALS). Then all white spaces, digits, and punctuations are converted to theirs generalized patterns. Followed by triple generation containing generalized BLS-LS-ALS obtained by trimming the front and back of the BLS and ALS, respectively, till the first false match is detected. Finally, each triple is converted to a pattern after duplicate removal. The REDEx algorithm effectively identifies bodyweight related measures and is simple to understand. However, it requires extensive annotation of the content before training, whereby, each snippet contains LS, LBS, and ALS.

III. Materials and Methods

The proposed methodology extracts patterns from CPG in a sequence of steps as depicted in Figure 1. The guideline for pattern extraction is preprocessed and align by five steps process of Guideline Preprocessing according to the required format. While five steps process of Pattern Extraction identifies and extracts the required patterns. The extracted patterns are used in classifying the guidelines sentences to RS and NRS.

The preprocessing steps prepare the documents and transform them into tokens, which are the output of Guideline Preprocessing. Initially, the Document Reader reads the document that needs to be processed for pattern extraction. Sentence Extractor splits the document into sentences, Tokenizer generates tokens of the content using "Non Letter" split technique, Transform Case transforms each token to lower case, and finally, we remove the stop words by Filter Stop words. The extracted tokens are considered for further processing to extract patterns.

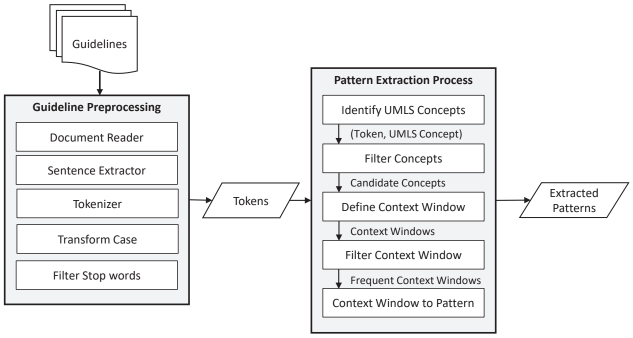


Fig. 1. Process flow of the proposed pattern extraction algorithm

The Pattern Extraction component scrutinizes the tokens to identify and extract the hidden patters. This component take tokens as input, process it and identify the patterns in the tokens that can classify the guideline sentences to RS and NRS. It achieves this goal by performing five sub steps. Initially, it identifies each token's semantic category/concept by utilizing Unified Medical Language System (UMLS) dictionary, which is one of the largest biomedical repository developed and maintained by US National Library of Medicine [10]. We represent each token in ('*token*', '$UMLSConcept$') format, where *token* represents guideline token while $UMLSConcept$ represents token's Semantic type of UMLS. The Filter Concepts component count the occurrences of each concept in the guideline and find the list of initial candidate concepts $C = [c_1, c_2, ..., c_n]$. The candidate concepts are defined in Definition 1.

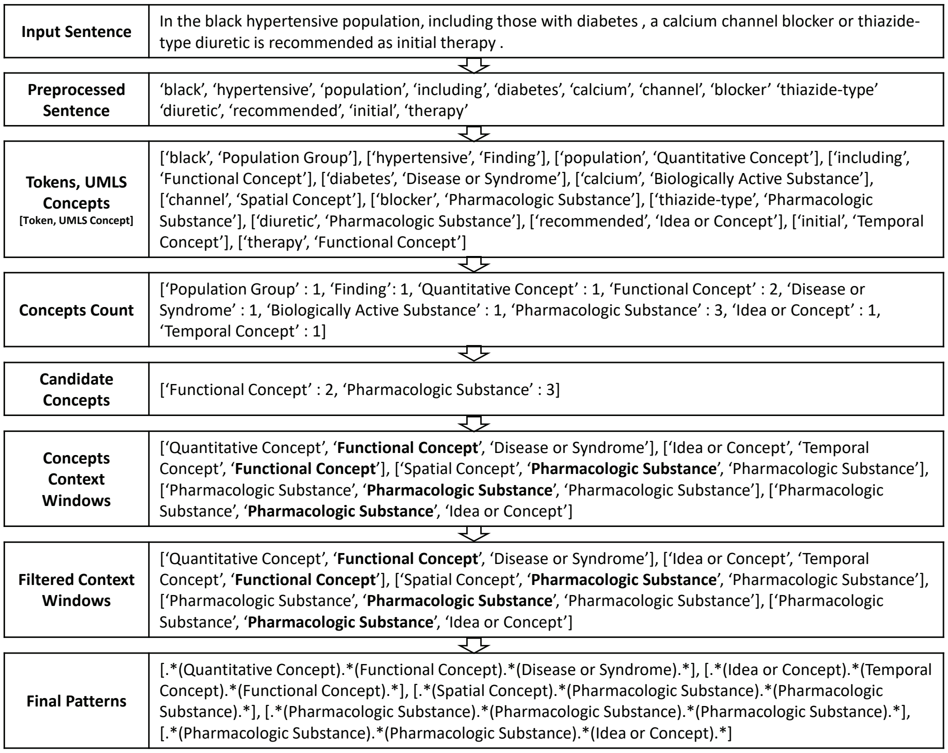
Definition 1: A concept c_i where $i = 1, 2, ..., n$ is considered as candidate concept if it is used more than a defined threshold CT value in a CPG i.e when $count(c_i) >= CT$.

The concepts in a guideline also depends on its context and neighbor concepts. Therefore, we generate a context window for each candidate concept. The context window is defined in Definition 2.

Definition 2: The context window cw of a candidate concept c_i is $cw_i = [c_{i-n}, ..., c_{i-2}, c_{i-1}, c_i, c_{i+1}, c_{i+2}, ..., c_{i+n}]$ for $i = 1, 2, ..., n$, where $c_{i-n}, ..., c_{i-2}, c_{i-1}$ represents the preceding concepts while $c_{i+1}, c_{i+2}, ..., c_{i+n}$ represents the succeeding concepts of a candidate concept c_i.

We filter out ineffective context window based on its occurrence by applying context window threshold CWT i.e $count(cw_i) < CWT$ to restrict the number of patterns and filter out ineffective patterns. The remaining context windows are transform to regular expressions as patterns $P = [p_1, p_2, ..., p_n]$, where p_i for $i = 1, 2, ..., n$ represents an extracted pattern i.

The detailed example of the proposed pattern extraction methodology is shown in Figure 2. In the example,



Input Sentence	In the black hypertensive population, including those with diabetes , a calcium channel blocker or thiazide-type diuretic is recommended as initial therapy .
Preprocessed Sentence	'black', 'hypertensive', 'population', 'including', 'diabetes', 'calcium', 'channel', 'blocker' 'thiazide-type' 'diuretic', 'recommended', 'initial', 'therapy'
Tokens, UMLS Concepts **[Token, UMLS Concept]**	['black', 'Population Group'], ['hypertensive', 'Finding'], ['population', 'Quantitative Concept'], ['including', 'Functional Concept'], ['diabetes', 'Disease or Syndrome'], ['calcium', 'Biologically Active Substance'], ['channel', 'Spatial Concept'], ['blocker', 'Pharmacologic Substance'], ['thiazide-type', 'Pharmacologic Substance'], ['diuretic', 'Pharmacologic Substance'], ['recommended', 'Idea or Concept'], ['initial', 'Temporal Concept'], ['therapy', 'Functional Concept']
Concepts Count	['Population Group' : 1, 'Finding': 1, 'Quantitative Concept' : 1, 'Functional Concept' : 2, 'Disease or Syndrome' : 1, 'Biologically Active Substance' : 1, 'Pharmacologic Substance' : 3, 'Idea or Concept' : 1, 'Temporal Concept' : 1]
Candidate Concepts	['Functional Concept' : 2, 'Pharmacologic Substance' : 3]
Concepts Context Windows	['Quantitative Concept', '**Functional Concept**', 'Disease or Syndrome'], ['Idea or Concept', 'Temporal Concept', '**Functional Concept**'], ['Spatial Concept', '**Pharmacologic Substance**', 'Pharmacologic Substance'], ['Pharmacologic Substance', '**Pharmacologic Substance**', 'Pharmacologic Substance'], ['Pharmacologic Substance', '**Pharmacologic Substance**', 'Idea or Concept']
Filtered Context Windows	['Quantitative Concept', '**Functional Concept**', 'Disease or Syndrome'], ['Idea or Concept', 'Temporal Concept', '**Functional Concept**'], ['Spatial Concept', '**Pharmacologic Substance**', 'Pharmacologic Substance'], ['Pharmacologic Substance', '**Pharmacologic Substance**', 'Pharmacologic Substance'], ['Pharmacologic Substance', '**Pharmacologic Substance**', 'Idea or Concept']
Final Patterns	[.*(Quantitative Concept).*(Functional Concept).*(Disease or Syndrome).*], [.*(Idea or Concept).*(Temporal Concept).*(Functional Concept).*], [.*(Spatial Concept).*(Pharmacologic Substance).*(Pharmacologic Substance).*], [.*(Pharmacologic Substance).*(Pharmacologic Substance).*(Pharmacologic Substance).*], [.*(Pharmacologic Substance).*(Pharmacologic Substance).*(Idea or Concept).*]

Fig. 2. Pattern extraction example

we have used concept threshold CT as two; therefore, all the concepts that appear less than twice is filtered out. The remaining two concepts having occurrences count greater than or equal to two is the initial candidate concepts. The context window size is selected as three in the example; therefore, we considered one preceding and following concepts of each candidate concept. One is selected as the context window threshold; thus, no any context window has been eliminated. Finally, all the context windows are represented in regular expression form as final patterns.

TABLE I
Details of dataset.

Guideline	RS	NRS	Total Sentences
Hypertension	78	200	278
Rhinosinusitis	151	610	761
Asthma	53	116	169

IV. Results

We examined the proposed algorithm on Hypertension [8], Rhinosinusitis [16] and chapter four of Asthma [9] guidelines. The details of the guidelines are listed in Table I. We splitted hypertension guideline into 70% and 30% for pattern extraction and evaluation process, respectively. The pattern extracted were also evaluated on the other two guidelines to check the generalization of the extracted patterns. The performance of the algorithm is mainly affected by three parameters including concepts threshold (CT), context window size (CWS), and context window threshold (CWT). We performed multiple experiments with different parameter settings to find the most appropriate values for parameters. However, in all experiment CWT value three perform the best therefore, we set $CWT = 3$. Some of the experimental results are listed in Table II. The Settings column represents the parameter values used in the experiment in the CT-CWT format. The identified patterns represent the total number of patterns extracted with the given parameter settings from a guideline. True Positive (TP) and True Negative (TN) represent the total number of RS and NRS sentences identified by patterns in a guideline, respectively. While the Accuracy column represents the accuracy achieved in percentage by the patterns.

The experiment was performed with diverse parameter settings starting CT values from 1 to 30, while CWT values from 1 to 10 to identify the appropriate values. All possible combination of the parameter values have been evaluated. However, in maximum cases, when the parameter values are higher, the number of patterns extracted was zero and achieved less than 50% accuracy for RS classification. Therefore, we showed some of the results of the instances having accuracy greater than 50% in Table 2. The parameters values (CT-CWT) mainly affect the number of patterns as shown in Table 2. The lower of the CT values, the higher is the number of patterns, which decreases the accuracy and increases the

TABLE II
Experimental Results.

Settings	Identified Patterns	Hypertension			Rhinosinusitis			Asthma		
CT-CWT		TP	TN	Accuracy	TP	TN	Accuracy	TP	TN	Accuracy
1-1	83	69	74	51.44%	138	248	50.72%	44	16	35.50%
3-1	79	69	81	53.96%	136	256	51.51%	44	25	40.83%
5-1	71	68	96	58.99%	135	277	54.14%	41	42	49.11%
1-2	34	57	137	69.78%	111	385	65.18%	31	89	71.01%
3-2	30	55	141	70.50%	109	391	65.70%	31	92	72.78%
5-2	29	55	141	70.50%	109	394	66.10%	31	92	72.78%
30-4	21	66	169	84.53%	128	481	80.03%	42	101	84.62%

amount of false negative FN value. The effect of patterns on accuracy can be seen in as Figure 3.

The 30-4 combination of the parameter values achieve better results compare to others. While the initial results presented here provided us with the direction towards achieving automatic pattern extraction from Clinical Practice Guidelines.

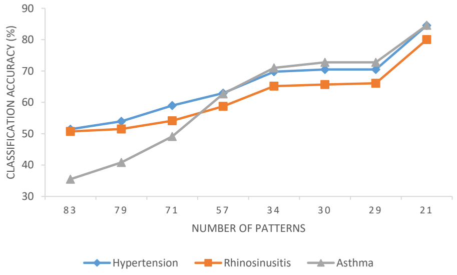


Fig. 3. Classification Results

V. Conclusion

Text classification is one of the essential steps in healthcare applications and pattern based approaches are prominent for clinical text classification. In this study, we proposed an automatic pattern identification and extraction approach that limits human experts' involvement to pattern verification in the process. We assess the proposed approach on Hypertension, Rhinosinusitis, and Asthma guidelines for RS and NRS identification and achieved an accuracy of 84.53%, 80.03%, and 84.62% respectively. The methodology can be used in the preprocessing step for applications utilizing clinical practice guidelines.

ACKNOWLEDGMENT

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2017-0-01629) supervised by the IITP(Institute for Information & communications Technology Promotion), also supported by IITP grant funded by the MSIT (No.2017-0-00655), NRF-2016K1A3A7A03951968, and NRF-2019R1A2C2090504.

References

[1] Bui, Duy Duc An, Guilherme Del Fiol, and Siddhartha Jonnalagadda. ”PDF text classification to leverage information extraction from publication reports.” Journal of biomedical informatics 61 (2016): 141-148.

[2] Mujtaba, Ghulam, et al. ”Clinical text classification research trends: Systematic literature review and open issues.” Expert Systems with Applications 116 (2019): 494-520.

[3] Yao, Liang, Chengsheng Mao, and Yuan Luo. ”Clinical text classification with rule-based features and knowledge-guided convolutional neural networks.” BMC medical informatics and decision making 19.3 (2019): 71.

[4] Pyon, Eunice Y. ”Primer on clinical practice guidelines.” Journal of pharmacy practice 26.2 (2013): 103-111.

[5] Bui, Duy Duc An, and Qing Zeng-Treitler. ”Learning regular expressions for clinical text classification.” Journal of the American Medical Informatics Association 21.5 (2014): 850-857.

[6] Murtaugh, Maureen A., et al. ”Regular expression-based learning to extract bodyweight values from clinical notes.” Journal of biomedical informatics 54 (2015): 186-190.

[7] Bartoli, Alberto, et al. ”Inference of regular expressions for text extraction from examples.” IEEE Transactions on Knowledge and Data Engineering 28.5 (2016): 1217-1230.

[8] James, Paul A., et al. ”2014 evidence-based guideline for the management of high blood pressure in adults: report from the panel members appointed to the Eighth Joint National Committee (JNC 8).” Jama 311.5 (2014): 507-520.

[9] Scottish, Intercollegiate Guidelines Network. ”British guideline on the management of asthma.” Thorax 58 (2003): i1.

[10] Bodenreider, Olivier. ”The unified medical language system (UMLS): integrating biomedical terminology.” Nucleic acids research 32.suppl-1 (2004): D267-D270.

[11] Li, Yunyao, et al. ”Regular expression learning for information extraction.” Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2008.

[12] Garofalakis, Minos, et al. ”XTRACT: a system for extracting document type descriptors from XML documents.” ACM SIGMOD Record. Vol. 29. No. 2. ACM, 2000.

[13] Galassi, Ugo, and Attilio Giordana. ”Learning regular expressions from noisy sequences.” International Symposium on Abstraction, Reformulation, and Approximation. Springer, Berlin, Heidelberg, 2005.

[14] Caragea, Cornelia, et al. ”Citation-enhanced keyphrase extraction from research papers: A supervised approach.” Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014.

[15] Wenzina, Reinhardt, and Katharina Kaiser. ”Identifying condition-action sentences using a heuristic-based information extraction method.” Process Support and Knowledge Representation in Health Care. Springer, Cham, 2013. 26-38.

[16] Chow, Anthony W., et al. ”IDSA clinical practice guideline for acute bacterial rhinosinusitis in children and adults.” Clinical Infectious Diseases 54.8 (2012): e72-e112.