Towards Evidence-based Precision Medicine: Extracting Population Information from Biomedical Text using Binary Classifiers and Syntactic Patterns

Kalpana Raja, PhD[1]; Naman Dasot, B.Tech.[2]; Pawan Goyal, PhD[2]; Siddhartha R Jonnalagadda, PhD[1]

[1]Division of Health and Biomedical Informatics, Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL

[2]Department of Computer Science and Engineering, Indian Institute of Technology, Kharagpur, West Bengal, India

Abstract

Precision Medicine is an emerging approach for prevention and treatment of disease that considers individual variability in genes, environment, and lifestyle for each person. The dissemination of individualized evidence by automatically identifying population information in literature is a key for evidence-based precision medicine at the point-of-care. We propose a hybrid approach using natural language processing techniques to automatically extract the population information from biomedical literature. Our approach first implements a binary classifier to classify sentences with or without population information. A rule-based system based on syntactic-tree regular expressions is then applied to sentences containing population information to extract the population named entities. The proposed two-stage approach achieved an F-score of 0.81 using a MaxEnt classifier and the rule-based system, and an F-score of 0.87 using a Naïve-Bayes classifier and the rule-based system, and performed relatively well compared to many existing systems. The system and evaluation dataset is being released as open source.

Introduction

The goal of precision medicine is to develop prevention and treatment strategies for individual variability based on individual patient's unique biological characteristics (e.g. inherited variation to drug response) and disease processes (e.g. tumor genomic characteristics). The approach extends beyond personalized medicine, evidence-based medicine and genome medicine. Precision medicine aims to bridge individual patient characterization and phenotype with evidence-based medicine. Recent years have witnessed the development of biological databases from human genome projects, characterization of patients using system biology approaches (e.g. proteomics, metabolomics, genomics, and diverse cellular assays), phenotype, and computational methods to enhance health and wellness of each person rather than just treating the disease.[1] The approaches for characterizing the patients incorporate knowledge derived from proteomics, genomics, metabolomics, and even social and mobile health.[1,2] Evidence-based medicine integrates the best evidence from well-designed research with clinical expertise and patient values. The four components of precision medicine have been defined to be predictive, preventive, personalized, and participatory medicine.[3,4] Khoury et al.[5] proposed the integration of "fifth P" – the *population* perspective that describes the balance between *individual and population interventions* for improving health and the evaluation of their comparative effectiveness. A population perspective implements the concept of population screening to preventive medicine, and use of evidence-based practice to personalized medicine. It was argued that the application of *population science into precision medicine* is the key for deciding the most appropriate treatment for every individual patient.[5]

The short-term goal of precision medicine is to come closer to curing cancers and diabetes, and the long-term goal is to provide access to personalized knowledge for all diseases.[1] The primary source of knowledge is biomedical literature, which is growing at an exponential rate. Natural language processing (NLP) techniques are being used to automatically extract information from biomedical literature. Various studies have explored the extraction of the number of participants, their age, sex, ethnicity, country, comorbidities, spectrum of presenting

symptoms, current treatments, etc. While most studies only highlighted the sentences containing the population data elements, six studies[6-11] extracted data elements as opposed to only highlighting the sentence containing the data element. For example, Kelly and Yang[6] extracted age of participants, duration of study, ethnicity of participants, gender of subjects, health status of participants, and number of participants on a dataset of 386 abstracts. Unfortunately, each of these studies used a different corpus of reports, which makes direct comparisons impossible. None of the studies have made their systems available as open-source, except RIDeM tool,[12] which is available as a web service. In the current study, we developed an automated approach for extracting population named entity from biomedical literature. Our hybrid approach integrates a binary classifier for preprocessing and a rule-based system for extracting the population named entity. The binary classifier classifies input sentences into those with population information using MaxEnt classifier and Naïve-Bayes classifier. The rule-based system uses a set of syntactic patterns to identify and extract the population named entity. This has the potential to provide personalized evidence updates to clinicians and patients based on their individual characteristics. The evaluation dataset and the code are available as open source to enable implementation in wider precision medicine applications.[13] To our knowledge, no evaluation dataset is publically available for testing and training population extraction approaches.

Methods

For the scope of this paper, population refers to the cohort of patients with shared characteristics such as age, gender, treatments and diseases. The algorithm for extracting population named entity operates at sentence level. Preprocessing of input sentences is carried out to first identify the sentences with population information. Two types of binary classifiers are used in the current study: MaxEnt Classifier[14] and Naïve-Bayes Classifier.[15] We use MALLET (Machine Learning for LanguagE Toolkit)[16] for sentence classification. The hybrid approach of classifiers with the population extraction algorithm is described in detail below.

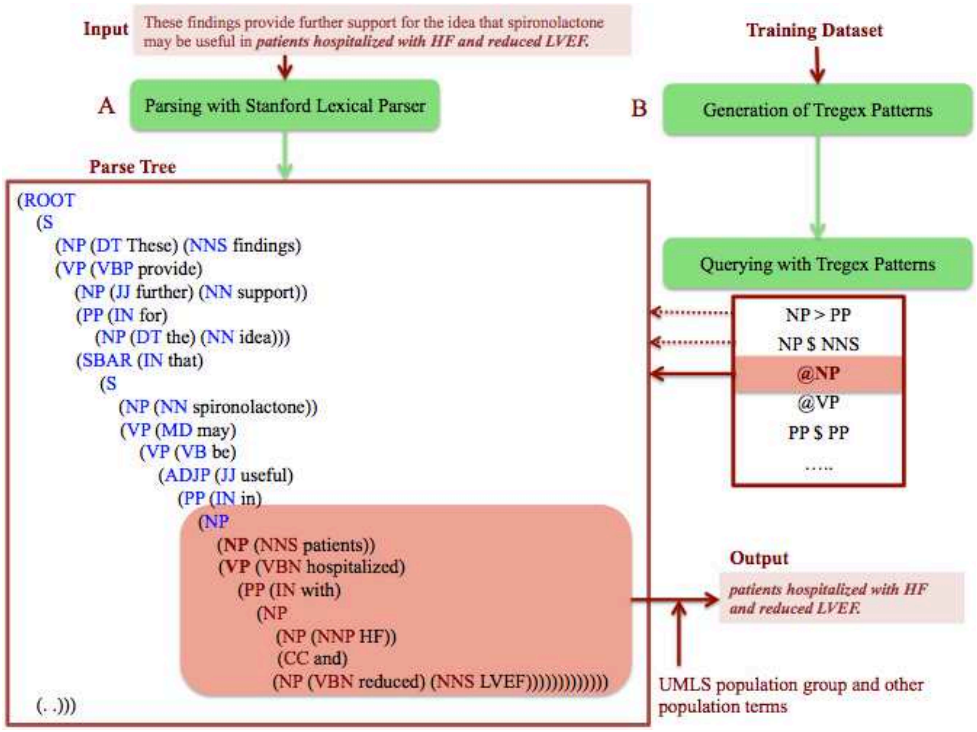

Figure 1. Extraction of population named entity

Population Extraction Algorithm

Syntax Pattern Generation: Our approach uses two NLP parsers, Stanford lexical parser[22] to parse an input sentence and Stanford Tregex (Tree regular expressions) parser, simply called as Tregex[23] to query the parse tree structure for extracting the population information. Our approach for generating syntax patterns is semi-automated and uses the Tregex-specific syntax for matching the nodes in a parse tree. In a previous study, we generated Tregex patterns for extracting protein-protein interaction

information from biomedical literature,[24] where we achieved an F-score of 66.05% to outperform most of the existing systems. In the current study, we use Tregex patterns for extracting population named entity.

The constituency parse trees for a set of sentences (used for developing Tregex patterns) are generated with Stanford lexical parser. A parse tree details the grammatical components such as noun phrase (NP) and verb phrase (VP) as shown in Figure 1A. It may contain more than one noun phrase and sometimes a noun phrase is nested within another noun phrase or verb phrase. For each parse tree, we identified the sub-tree encompassing the population named entity. We manually developed Tregex patterns that best explain the nodes of sub-tree representing a population named entity (Figure 1). Tregex patterns are similar to regular expressions and are easy to use. The patterns are then incorporated into the population extraction algorithm for automatic extraction of population named entity. The various Tregex symbols applied for identifying the relationship between the nodes are listed in Table 1. Tregex patterns developed for extracting population named entities are listed in Table 2.

Syntax Pattern Application: The input sentence is parsed with Stanford lexical parser using the probabilistic context free grammar (PCFG) model.[22] The generated parse tree is queried using Tregex patterns to identify and extract the population named entities. Not all the named entities that match the Tregex patterns are population named entities. We used UMLS to obtain a set of 130 population-related concepts belonging to "patient or disabled group" semantic type. An additional set of 22 terms related to population was manually identified from MEDLINE citations (Supplementary data 1). These concepts and terms are used to filter the population named entities (Figure 2). For named entities extracted with patterns NP > PP, PP $ NP, PP $ PP and @VP, the algorithm trims the sub-tree from the noun phrase matching population-related concepts or terms.

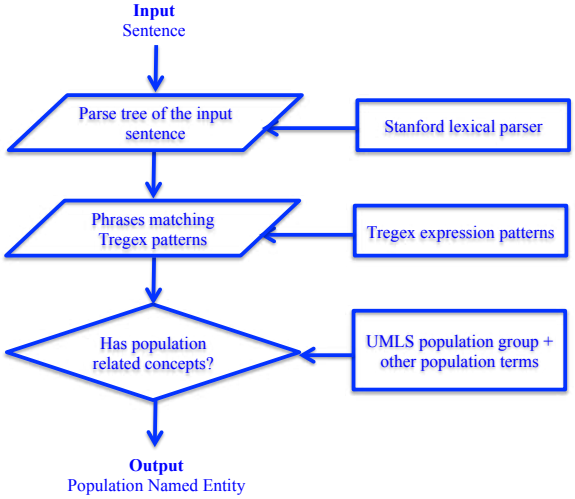

Figure 2. Rule-based approach for population

Rarely, more than one pattern is applicable for the same population named entity in a sentence and in such cases, the first matching pattern based on a predetermined order of precedence is considered. The named entity identified using the above patterns but matching selected stop phrases namely 'patient education', 'patient survival', 'patient preference', 'patient factors', 'patient characteristics', 'patient confidentiality', 'patient permission', 'patient status', 'patient selection', 'patient level data' and 'patient refusal' are filtered out.

Table 1. List of Tregex symbols for pattern generation

Tregex Symbol	Description
Node A << Node B	Node A dominates Node B
Node A >> Node B	Node A is dominated by Node B (Node B << Node A)
Node A < Node B	Node A immediately dominates Node B
Node A > Node B	Node A is immediately dominated by Node B (Node B < Node A)
Node A $ Node B	A and B are sisters i.e. at same level in the parse tree (but are not equal)
@Node	Selects the entire phrase (noun or verb) mentioned i.e. @NP

Table 2. Tregex patterns for population named entity extraction (to be used in conjunction with "population-related concepts")

Pattern	Output Phrase	Example Sentence with Output Underlined

NP > PP	Noun phrase succeeding prepositional phrase	Aldosterone blockade has been shown to be effective in reducing total mortality as well as *hospitalization for heart failure in patients with systolic left ventricular dysfunction (SLVD) due to chronic heart failure* and in patients with SLVD post acute myocardial infarction. (PMID: 15134801)
PP $ NP	Prepositional phrase and noun phrase are sisters	Rosuvastatin did not reduce mortality *compared to placebo in patients with heart failure and left ventricular systolic dysfunction due to ischaemic heart disease* in the CORONA study. (PMID: 18179987)
NP $ NP	Two noun phrases as sisters	Implantation of CRT-D rather than an implantable cardioverter defibrillator in *patients with mild heart failure and QRS >/=130 ms* reduced the risk of hospitalization for heart failure in MADIT-CRT; (PMID: 19926603)
NP $ NNS	Noun phrase and noun are sisters	Diuretics are indicated for *symptomatic patients* as needed for volume overload. (PMID: 18441861)
@NP	Noun phrase	So far, nebivolol is the only beta-blocker to have been shown effective in *elderly heart failure patients*, regardless of their left ventricular ejection fraction. (PMID: 20307222)
PP , NP	Prepositional phrase immediately follows noun phrase	It is suggested that beta-receptor blockade should be added to conventional treatment with digitalis and diuretics in *all patients with severe myocardial failure caused by congestive cardiomyopathy*. (PMID: 6107090)
PP $ PP	Two prepositional phrases as sisters	This article reviews the physiological changes that occur in the elderly and the treatment approach that can be taken *in elderly patients with heart failure*. (PMID: 9205849)
NP , PP	Noun phrase immediately follows prepositional phrase	It is suggested that potassium depletion is not a major problem *in patients with heart-failure treated with diuretics*. (PMID: 62899)
NP $ PP	Noun phrase and prepositional phrase are sisters	Piretanide, a diuretic that acts on the loop of Henle, was used to treat *patients with cardiac failure*. (PMID: 6990212)
@VP	Verb phrase	Isosorbide dinitrate and hydralazine hydrochloride should be *tried in patients who cannot tolerate ACE inhibitors or who have refractory symptoms*. (PMID: 7933398)

Rule-based approach with binary classifier

Extraction of population by the rule-based system alone may not be sufficient for sentences with complex tree structures, thereby producing a number of false negatives. Our initial study included only five Tregex patterns: NP > PP, PP $ NP, NP $ NP, NP $ NN and @NP (Table 2). Attempting to decrease the number of false negatives by developing new patterns for meeting specific requirements extracted many incorrect population phrases, thereby producing a number of false positives. Therefore, for the population extraction algorithm to be more reliable and accurate, and to accommodate the syntactic pattern matching without increasing false positives, we designed a binary classifier for pre-processing. The binary classifier determines whether a sentence contains a population named entity or not. If the sentence contains the population named entity, then it is sent to the rule-based system described above. Otherwise, the sentence is rejected. This additional layer helped us to eliminate many sentences where the probability of recognizing a population named entity is very low. Thus, we were able to concentrate on extracting the population named entity from the sentences that have a higher probability of including it. This allowed us to increase the number of Tregex patterns in our rule-based system to ten from five (Table 2). We have used two different types of binary classifiers, namely MaxEnt[14] and Naive-Bayes[15] classifiers. The system architecture of binary classifier with rule-based system is shown in Figure 3.

Figure 3. System Architecture

MaxEnt Classifier with Rule-based approach: The dataset collected for validating the population extraction algorithm was split into training dataset (80%) and test dataset (20%). Sentences from training dataset are converted to list of instances where each instance is a feature vector. We applied a set of basic filters prior to learning on the features: removal of non-ASCII or Unicode characters, conversion to lower case, removal of stop words, and lemmatization. Our feature-set includes all the unigrams and bigrams in the sentences.

Naïve-Bayes Classifier with Rule-based approach: As an alternative, we also implemented a Naïve-Bayes classifier. The feature set for training the classifier includes 50 terms (unigram and bigram) with the maximum information gain identified from the training dataset. The frequency of each feature in the training dataset is calculated and used for estimating the probability.

Evaluation approach and Dataset

We performed an experiment to extract population named entities related to congestive heart failure (CHF) and atrial fibrillation (AFib). We selected the diseases based on statistics from Centers for Disease Control and Prevention (CDC), which states heart disease as the leading cause of death in the United States.[25] The evaluation dataset for developing the population named entity extraction algorithms consists of 714 sentences from MEDLINE citations that are retrieved from SemMedDB (Table 3). Since the goal of our overall research is to apply these algorithms for precision medicine applications in cardiovascular treatment and diagnosis we focused on sentences related to diagnosis and treatment of CHF and AFib.

Table 3. Evaluation Dataset

Dataset	Citations with population	Citations without population
Diagnosis for CHF	80	120
Treatment for CHF	98	102
Diagnosis for AFib	140	60
Treatment for AFib	56	58

Citation extraction from MEDLINE

Our overall strategy aims at retrieving population information from high quality clinical journals (Figure 4A). Two Boolean queries were built to retrieve articles on systemic reviews (SR) and randomized control trials (RCTs) from MEDLINE (Figure 4B) (Supplementary data 2).

Sentence extraction from SemMedDB

A set of MEDLINE abstracts for a given clinical condition (e.g., treatments and diagnosis for CHF and AFib) is retrieved from SemMedDB for these citations.[16] Our information retrieval approach makes use of a list of UMLS concept identifiers (CUIs) (41 CUIs for CHF and 25 CUIs for AFib) (Supplementary data 3) to query SemMedDB for retrieving the sentences. For example, CUIs such as 'C0018802', 'C0264719', 'C0264722', 'C2039715' and 'C2183328' are related to CHF. Each unit of information retrieved for a condition consists of PMID, sentence and predication as in Example 1.

The evaluation dataset was divided into training dataset (80% of the dataset) and test dataset (20% of the dataset). A 5-fold cross validation is run on the training dataset for the binary classifiers. Then the test dataset is

used for evaluating the performance of each of these classifiers with the rule-based system: MaxEnt classifier and the rule-based system, and Naïve-Bayes classifier and the rule-based system.

Example 1 – PMID: 2539290
Sentence: Enalapril provides significant haemodynamic, symptomatic and clinical improvement when added to maintenance therapy with digitalis and diuretics in patients with congestive heart failure [NYHA (New York Heart Association) classes II to IV].
Predication: Diuretics-TREATS-Congestive Heart Failure

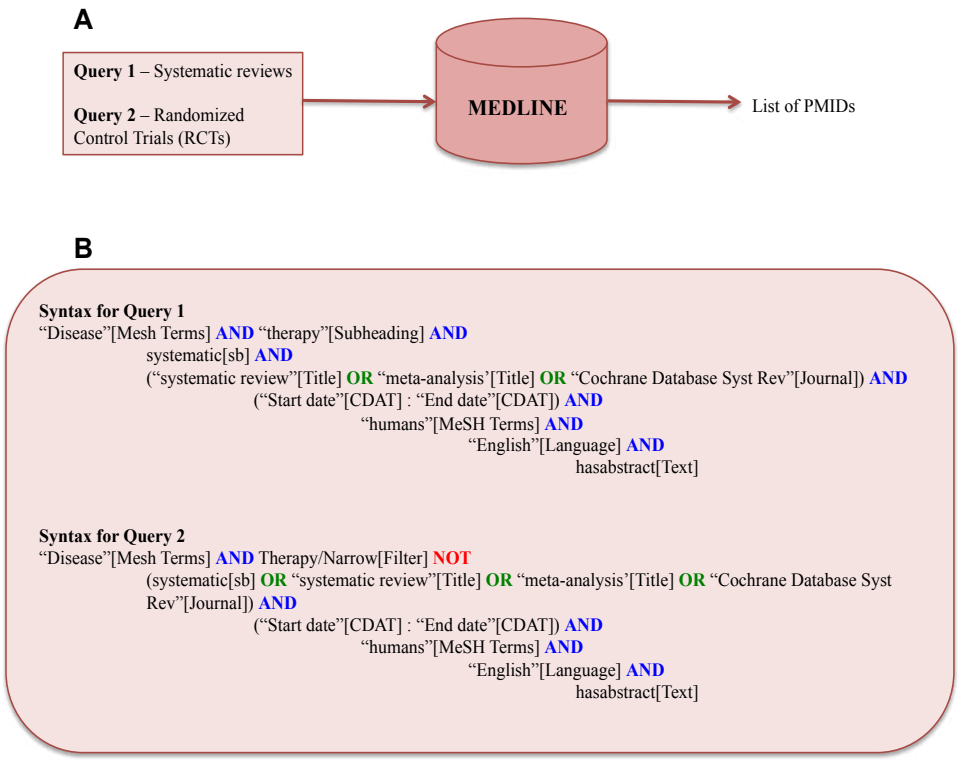

Figure 4. Citation retrieval from MEDLINE for constructing Evaluation Dataset

Results

Table 4 shows the performance of the system on the evaluation dataset consisting of sentences related to the diagnosis and treatment for CHF and AFib. The standard metrics of precision, recall and F-score were used for evaluating the system performance. The rule-based system alone achieved F-score of 0.64. The two-stage approach to classify the input sentence as having or not having population information, and to extract population named entity achieved F-score of 0.81 with MaxEnt classifier and rule-based system, and F-score of 0.87 with Naïve-Bayes classifier and rule-based system.

Table 4. System performance

System	Precision	Recall	F-score
Rule-based system	0.67	0.62	0.64
MaxEnt classifier + Rule-based system	0.87	0.76	0.81
Naïve-Bayes classifier + Rule-based system	0.90	0.83	0.87

Preprocessing with MaxEnt or Naïve-Bayes classifiers filtered sentences with potential population information. This improved the performance of MaxEnt classifier and the rule-based system by 17% (0.81-0.64), and Naïve-Bayes classifier and the rule-based system by 23% (0.87-0.64). Table 5 shows the performance of binary classifiers i.e. MaxEnt classifier and Naïve-Bayes classifier on classifying sentences with or without population information.

Table 5. Performance of binary classifiers

System	Precision	Recall	F-score
MaxEnt classifier	0.87	0.82	0.84
Naïve-Bayes classifier	0.89	0.91	0.90

Discussion

Existing systems for population extraction

Table 6 lists the performance of our system and other systems available for similar task. This gives an idea about the techniques and dataset used by our system and other systems for extracting the population named entity from a given sentence. None of these systems are available as open source, except the RIDeM tool.[12]

Table 6. Approach and dataset used by various systems

System	Dataset	Sentence Classification		Population Extraction		
		Model	F-score	Model	F-score	Remarks
Xu et al.[26]	Abstracts Only from PubMed	HMM + NLP Techniques	92%	Classification + parse Tree (Stanford)	0.51	-
Zhu et al.[27]	-	-		Partially Matched Using Metamap	0.84	-
				Partially Matched Using NLP- based method	0.83	
				Exact Matched Using Metamap	0.42	
				Exact Matched Using NLP- based method	0.64	
Demner Fushman and Lin[8]	MedLine Abstracts	-		Baseline	0.53	Returns a set of results
				Extractor	0.80	
Zhao et al.[28]	Reduced Dataset	Mallet CRF		Independent	0.78	4 different Methods Used
				Sentence-First	0.78	
				Word-First	0.78	
				Joint	0.75	
	Full Dataset			Independent	0.64	
				Sentence-First	0.64	
				Word-First	0.63	
				Joint	0.60	
Kelly[6]	Abstracts from PubMed	-		Partial match	0.877	Dependency Parse
				Exact match	0.601	Regular Expressions
RIDeM Tool[12]	Tested on Our Dataset	-		Original	0.63	Upper bound on precision
				With Add-ons	0.766	
Our Current System	Tested on Our Dataset	-		Rule-based	0.64	-
		MaxEnt	0.844	MaxEnt + Rule-based	0.81	

		Naive- Bayes	0.9	Naive-Bayes + Rule based	0.87	

The accuracies are not comparable since the datasets are different. We were able to compare our tool with the RIDeM tool developed by Demner Fushman and Thoma on our dataset. The F-score of RIDeM tool[12] on our dataset (76.6%) is comparable to its accuracy in their own dataset (80.0%). This to some extent supports the validity of our dataset. The F-score of our best approach (87.7%) is better than other approaches. However, all these approaches have to be evaluated on our dataset for drawing conclusions.

Error Analysis

Based on analysis of 20 randomly selected sentences, the following are the major reasons for errors in population named entity recognition apart from classification errors:

Parse tree can be too complex to extract the Population phrase with Tregex patterns, i.e. the patient and disease terms are in different sub-trees (65% of errors). For example, in sentence "The left ventricular partitioning device appears to be relatively safe and potentially effective in the treatment for patients with heart failure and a prior anterior myocardial infarction" (PMID: 22607859), our system extracts "patients with heart failure" as population.

There are some cases in which the Stanford Tregex parser returns an incorrect parse tree of the sentence. This results in retrieving a wrong parse tree or assigning wrong tags to nodes. Thus, Tregex patterns are not able to extract the population phrase successfully (15% of errors). For example, in sentence "Cardiac resynchronization therapy produces both short-term hemodynamic and long-term symptomatic/mortality benefits in symptomatic heart failure patients with a QRS duration >120 ms" (PMID: 19822812), our system extracts "in symptomatic heart failure patients with a QRS duration >120 ms" as population.

Extraction of population phrase can be achieved by introducing new Tregex patterns. However, including additional patterns extracted incorrect population information. This resulted in the increase of false positives to a large extent, hence lowering the precision of the system (20% of errors). Therefore, we decided to limit our use of Tregex patterns to ten (Table 2).

We believe sentence simplification, which has been shown to both improve the accuracy of parsers[31] and also of information extraction,[32] might improve the accuracy of population extraction.

Applying the system for evidence summarization

We have independently studied the efficacy of a preliminary population extraction system that uses only the rule-based component in summarizing individualized evidence for clinicians; the goal was to automatically generate clinically useful sentences that provide a specific recommendation for an intervention (e.g. medication treatment) employed with a specific patient population.[29] We found that such an approach is entirely feasible and it is possible to classify such clinically actionable sentences both from MEDLINE abstracts and also from clinical knowledge systems such as UpToDate.[30]

The gold standard used for testing the population extraction algorithm for evidence summarization is different from the one presented here. It consists of 4,499 sentences from UpToDate documents on the treatment of six chronic conditions namely coronary artery disease, hypertension, depression, heart failure, diabetes mellitus, and prostate cancer. The system achieved 90.5% precision, 96.7% recall and 93.6% F-score when tested on the gold standard generated from UpToDate document. The gold standard is available for sharing upon obtaining permission from UpToDate.

Conclusions and Future Work

Our work aimed to extract population information pertaining to evidence for supporting the retrieval of citations and ultimately evidence-based precision medicine. We used three different methods: rule-based system, MaxEnt classifier with rule-based system, and Naïve-Bayes classifier with rule-based system. In all the three methods, we used a rule-based system to extract the population named entities. F-score of the best classifier is 90% and that of whole system is 87%. We are optimistic about the use of our system to advance precision medicine, especially in being able to deliver individualized evidence summaries at the point-of-care.

Acknowledgments

This work was made possible by funding from the National Library of Medicine (R00LM011389). The authors acknowledge Dr. Guilherme Del Fiol for his valuable ideas with extracting population information from UpToDate (supported by R01LM011416) and Dr. Andrew J Sauer who is the cardiologist that helped us while formulating the problem.

References

1. Collins F, Varmus H. A new initiative on precision medicine. New Engl J Med 2015;372:793-5.
2. Robinson P. Deep phenotyping for precision medicine. Hum Mutat 2012;33:777-80.
3. Weston AD, Hood L. Systems biology, proteomics, and the future of health care: Toward Predictive, Preventative, and Personalized Medicine. J Proteome Res 2004;3:179-96.
4. Hood L, Friend SH. Predictive, personalized, preventive, participatory (P4) cancer medicine. Nature reviews Clinical oncology 2011;8:184-7.
5. Khoury MJ, Gwinn ML, Glasgow RE, Kramer BS. A population approach to precision medicine. American journal of preventive medicine 2012;42:639-45.
6. Kelly C, Yang H. A system for extracting study design parameters from nutritional genomics abstracts. J Integr Bioinform 2013;10:222.
7. Hansen M, Rasmussen N, Chung G. A method of extracting the number of trial participants from abstracts describing randomized controlled trials. J Telemed Telecare 2008;14:354-8.
8. Demner-Fushman D, Lin J. Answering clinical questions with knowledge-based and statistical techniques. Computational Linguistics 2007;33:63-103.
9. Lin S, Ng J-P, Pradhan S, Shah J, Pietrobon R, Kan M-Y. Extracting formulaic and free text clinical research articles metadata using conditional random fields. Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents; 2010: Association for Computational Linguistics. p. 90-5.
10. Xu R, Garten Y, Supekar KS, Das AK, Altman RB, Garber AM. Extracting subject demographic information from abstracts of randomized clinical trial reports. Studies in health technology and informatics 2007;129.
11. Zhu H, Ni Y, Cai P, Qiu Z, Cao F. Automatic extracting of patient-related attributes: disease, age, gender and race. Studies in health technology and informatics 2011;180:589-93.
12. Demner-Fushman D, Thoma G. Repository for Informed Decision Making (RIDeM): PICO extraction interface. (Accessed 24 Sep 2015), at http://ceb.nlm.nih.gov/ridem/ridem?pico=UserQuery.
13. https://github.com/sidkgp/PopulationExtractor
14. Nigam K, Lafferty J, McCallum A. Using maximum entropy for text classification IJCAI Workshop on Machine Learning for Information Filtering 1999;61-67.
15. Zhang H. The optimality of naive Bayes. Proceedings of FLAIRS Conference 2004;1:3.
16. McCallum AK. MALLET: A Machine Learning for Language Toolkit. 2002.
17. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research 2004;32:D267-70.
18. McCray AT. Representing biomedical knowledge in the UMLS semantic network. High performance medical libraries: Meckler Corporation; 1993;45-55.

19. Rindflesch T, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. Journal of Biomedical Informatics 2003;36:462-77.
20. Rindflesch TC, Kilicoglu H, Fiszman M, Rosemblat G, Shin D. Semantic MEDLINE: An advanced information management application for biomedicine. Inf Serv Use 2011;31:15-21.
21. Kilicoglu H, Fiszman M, Rodriguez A, Shin D, Ripple AM, Rindflesch TC. Semantic MEDLINE: a web application for managing the results of PubMed searches. Symposium on Semantic Mining in Biomedicine; 2008. p. 69-76.
22. Klein D, Manning CD. Accurate unlexicalized parsing. Association for Computational Linguistics 2003:423-30.
23. Levy R, Andrew G. Tregex and Tsurgeon: tools for querying and manipulating tree data structures. Proceedings of the fifth international conference on Language Resources and Evaluation; 2006. p. 165-70.
24. Raja K, Subramani S, Natarajan J. PPInterFinder—a mining tool for extracting causal relations on human proteins from literature. Database: The Journal of Biological Databases and Curation 2013; bas052.
25. Go AS, Mozaffarian D, Roger VL, et al. Heart disease and stroke statistics-2013 update: a report from the American Heart Association. Circulation 2013;127:e6-e245.
26. Xu R, Garten Y, Supekar KS, Das AK, Altman RB, Garber AM. Extracting subject demographic information from abstracts of randomized clinical trial reports. Studies in health technology and informatics 2007;129:550-4.
27. Zhu H, Ni Y, Cai P, Qiu Z, Cao F. Automatic extracting of patient-related attributes: disease, age, gender and race. Stud Health Technol Inform 2012;180:589-93.
28. Zhao J, Bysani P, Kan M-Y. Exploiting Classification Correlations for the extraction of Evidence-based Practice Information. AMIA Annual Symposium Proceedings 2012;2012:1070-8.
29. Morid M, Jonnalagadda S, Fiszman M, Raja K, Del Fiol G. Classification of clinically useful sentences in MEDLINE. AMIA Annual Symposium Proceedings / AMIA Symposium AMIA Symposium 2015;In Press.
30. Hoogendam A, Stalenhoef AFH, de Vries Robbé PF, Overbeke AJPM. Answers to questions posed during daily patient care are more likely to be answered by UpToDate than PubMed. Journal of Medical Internet Research 2008;10.
31. Jonnalagadda S, Tari L, Hakenberg J, Baral C, Gonzalez G. Towards effective sentence simplification for automatic processing of biomedical text. North American Chapter of the Association for Computational Linguistics - Human Language Technologies 2009.
32. Jonnalagadda S, Gonzalez G. BioSimplify: an open source sentence simplification engine to improve recall in automatic biomedical information extraction. AMIA Annual Symposium Proceedings / AMIA Symposium AMIA Symposium 2010;2010:351-5.