Quality of Life through Quality of Information
J. Mantas et al. (Eds.)
IOS Press, 2012

doi:10.3233/978-1-61499-101-4-589

Automatic Extracting of Patient-related Attributes: Disease, Age, Gender and Race

Huijia ZHU[a,1], Yuan NI, Peng CAI[a], Zhaoming QIU[a] and Feng CAO[a]
[a]*IBM China Research Lab*

Abstract. In the Evidence-based Medicine (EBM), PICO format is designed to easily and correctly search for the best available evidence. As the main element of PICO, the Patient/Problem (P) represents the attributes of patient in the clinical question and studies. In order to better understand the clinical problems, patient attribute identification is crucial and indispensable. Due to the richness of the human nature language, many issues like various term representations, grammar structures and abbreviations present challenges for automatically extracting the patient-related attributes from the unstructured data. In this paper, we employed the nature language processing (NLP) technologies to deeply analyze the linguistic characteristics of the attributes. Based on the NLP analysis results, we built the rule sets for different attributes and applied the rule-based approach to extract the patient-related attributes.

Keywords. Patient/Population/Problem, PICO, information extraction, Evidence based medicine (EBM), Natural language processing (NLP)

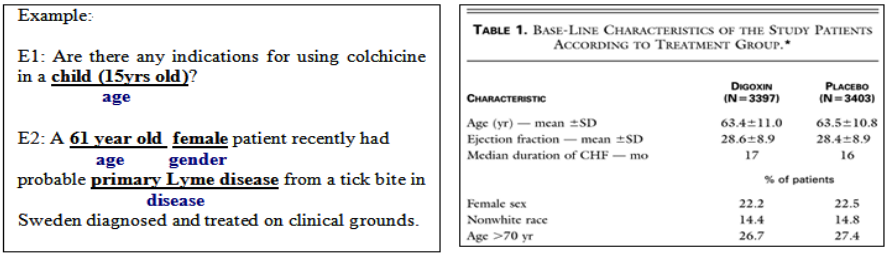
Introduction

In the clinical practice, Evidence-based Medicine (EBM) [1] aims to support the physicians to solve the clinical problems with the best available evidence. The first and most critical part of the EBM process is to ask the right question. A popular question representation involves Patient, Intervention, Comparison, and Outcome (PICO) [2, 3]. The PICO model has become the standard for stating a searchable question. In addition, critical appraisal and brief review of the PICO in the searched medical literatures is also an integral part of EBM. The task of PICO identification in free text is actually the nature language processing problem. Each element contains different attributes and need deeply linguistic analysis. However, considering the article space constrains, this paper focus on analyzing the Patient/Problem (P) which is the major body of the clinical questions and studies. NLP technologies are employed to deeply analyze the linguistic characteristics of the patient-related attributes.

For the physicians, describing the patient conditions is the first step in constructing questions. For the investigators, the characters of population in the clinical trials are important factors in clinical appraisal. In the evidence searching process, the patient similarity facilitates access to relevant evidence in medical literature. In the evidence appraisal process, the population settings reflect the evidence quality and safety. The patient can be described with the most important attributes including the findings,

[1] Corresponding Author: IBM Research – China, Building 10, NO. 399 Ke Yuan Road, Pudong New District, Shanghai, PRC, 201203. Email : zhuhuij@cn.ibm.com

disease, or co-existing conditions and the gender, age or race of a patient which might be relevant to the diagnosis or treatment of a disease. The findings and co-existing conditions can be all referred to the disease attribute. Totally, the disease, age, gender and race can describe the characteristics of patient as in Figure 1. Through using the Part-of-speech (POS) tagging and syntactic parsing technologies to deeply analyze the linguistic units of patient attributes, we applied the linguistic and rule-based approach to automatically extract the attributes.



Example:

E1: Are there any indications for using colchicine in a **child (15yrs old)**?
age

E2: A **61 year old** **female** patient recently had
age **gender**
probable **primary Lyme disease** from a tick bite in
disease
Sweden diagnosed and treated on clinical grounds.

TABLE 1. BASE-LINE CHARACTERISTICS OF THE STUDY PATIENTS ACCORDING TO TREATMENT GROUP.*

CHARACTERISTIC	DIGOXIN (N=3397)	PLACEBO (N=3403)
Age (yr) — mean ±SD	63.4±11.0	63.5±10.8
Ejection fraction — mean ±SD	28.6±8.9	28.4±8.9
Median duration of CHF — mo	17	16
	% of patients	
Female sex	22.2	22.5
Nonwhite race	14.4	14.8
Age >70 yr	26.7	27.4

Figure 1. Examples of Patient Attributes.

1. Methods

Due to the richness of the human nature language, many issues like various term representations, grammar structures and abbreviations present challenges for automatically extracting the patient-related attributes from the unstructured data. In this section, we conduct the NLP technologies to analyze the linguistic characteristics of patient attributes and build corresponding rule sets to extract these attributes.

1.1. Disease

Given a clinical free text in natural language, firstly we use a medical concept annotator such as MetaMap [4] to identify the medical concepts. The annotated text with concept types "Finding" and "Disorder" are mapped to the disease seeds (denoted as FD). The statistical result shows the initial annotated seeds have a higher partially matched precision and recall than the exactly matched result. This phenomenon denotes that boundary identification is one challenge in the disease annotation task. Normally, the linguistic structure of disease is the NP chunk. We employ a syntactic parser such as Stanford Parser [5] to improve the NP boundary identification for disease, wherein NP consists of the JJ, NN and proper noun POS types. Besides the internal structure of disease, the external syntactic relations such as coordinating (CO) relation and abbreviation (Abbr.) capitals are applied to improve the missing annotation.

In the figure 2, we give an example how we conduct the syntactic parser in disease detection. Given a clinical sentence "Temporarily relieves cough due to minor throat and bronchial irritation may occur with a cold.", disease related medical concepts are annotated with tag "FD" first. As there are lots of unknown words such as disease, intervention terms in the medical domain, the general syntactic parser may not easily understand these domain specific terms. In order to correctly parse the syntactic tree structure for medical text, we convert the letters in the initial annotated terms to the capitals normally recognized as noun words such as "FD_COUGH". Then, the syntactic parser is used to analyze the syntactic relations (e.g. NP chunk and CO) in the

text. The NP chunk containing the initial disease seed FD is recognized as a new disease term such as "bronchial IRRITATION". The CO relation is used to expand the NP chunk in adjacent disease term as the new disease term such as "minor throat and bronchial IRRITATION".

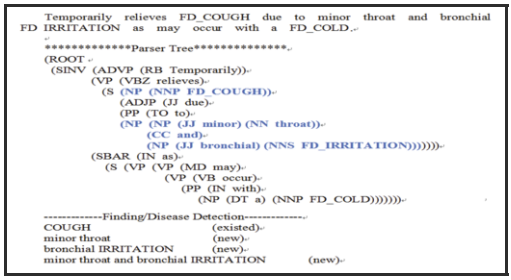


Figure 2. Example of Syntactic Parser in Disease Detection

Furthermore, many disease prop nouns are written in abbreviation with the first letters of the words such as "attention deficit/hyperactivity disorder (ADHD)". We use the capital letters in the parentheses to match the front words with their first letters and recognize the matched terms as the new ones. Iteratively using the expanded disease term list as the disease seeds can further improve the disease annotation result.

1.2. Age

The patient refers to the population group to which the physician wants to apply the information. Age is one crucial dimension for dividing the population group. The age representations have three pattern types: (1) Age number, e.g. *15yrs old*; (2) Age range, e.g. *aged 50-59 years*; (3) Age term, e.g. child. For each age pattern, we applied the following rule sets: (1) for the age number identification, the pattern [Clue Word] (<Number Pattern> [Quantifier]) [Clue Word] is applied. Number Pattern can be described by the regular expression of the digits and decimal point, i.e. *[0-9] +. [0-9]**, or the number words, i.e. one, two…; Quantifier is the time related word, i.e. *year(s), yr(s), month(s)…;* Clue Word is the hint term for age identification, i.e. *old, age(d), age group of, of age…*, which should have at least one emergence around the Number Pattern and the distance should no more than two words; (2) for the age range identification , the range symbol or word, i.e. *>, <, at, under, over, …*, around the age number is combined together as the age range. (3) 35 age related terms are manually collected, i.e. *child, elder man, adolescence, teenager, menopause, middle-aged, etc.*

Table 1. Definition of age category mapping

Age Number/Range (Years)		Age Category		Age Terms
0-0.1	→	newborn	←	neonate…
0.1-0.6	→	babe	←	little baby…
0.6-3	→	cheeper	←	toddler, …
3-7	→	children	←	child, …
7-15	→	teenager	←	teenager,…
15-36	→	youth	←	youth,…
36-61	→	middle age	←	strong man,…
61-90	→	old age	←	old man,…
90-150	→	longevity	←	longevity,…

All above identified ages are too specific with the population, which will make trouble for finding any evidence for that specific condition. Therefore, we further map

them to general age categories as in Table 1. Through categorizing the population into general age groups, more applicable evidences can be discovered.

1.3. Gender and Race

Gender and race are two major attributes of person and also the crucial dimensions of clinical population division. Gender divides the population into two groups: female and male. Besides using the common gender words such as men, women, girls and boys to identify the population gender type, we extract the gender-related concept words from MetaMap to infer the patient gender type. The concept words could be the gender-related findings or disorders, e.g. hot flushes and postmenopausal diabetic can infer the population type is female. The common racial divisions are including the white race and nonwhite (black and yellow) race. In order not to make the population too specific, we only recognize the two words "white" and "nonwhite" as two race types.

1.4. Attributes in Clinical Literature Document

Given the clinical evidence such as clinical literatures, the patient-related attributes in the article can be firstly annotated using above mentioned methods. Considering the well structured clinical trial articles, we further make use of the linguistic and structure characteristics of the literature to extract the focused attributes of document.

For the focused disease of article, a set of rules are created for candidate retrieval. Different weights are assigned to different rules. The candidate with the highest score will be the focused target. Four rules are designed for Disease: (1) if the title or introduction of the abstract section contains one of the patterns "in/for [FD] patient" or "patient with [FD]", then the [FD] part contains the focused disease candidates; (2) if there exists pre-annotated diseases in the keywords section, then the keywords contains the focused disease candidates; (3) if the title or introduction of the abstract section contains the pre-annotated disease candidates, then they are considered as the focused disease candidates; (4) the frequently appeared pre-annotated disease candidates in the abstract are also considered as the focused disease candidates. The candidates generated by different rules are given different weights which is rule (1)>rule (2)>rule (3)>rule (4). The same candidate will be merged and the final ranked candidate list will be generated. The top candidate is considered as the Focused Disease for the article.

In the literature articles, the population information, especially the age, gender and race, is always illustrated in the patient-related tables. Therefore, through finding the attribute clue word in the head and first column of table, we extract the value of patient attribute from the corresponding unit in the table. The tables containing the population information are denoted as the Patient-related Table which can be used to demonstrate the base information of patients in the literature article.

2. Results

To evaluate the performance of our extraction, we manually annotated 50 questions and 50 articles. We randomly selected 50 questions from the Trip Answers website [6] and 50 RCT articles from the PubMed website [7]. Due to the space limitation, here we only compare the disease result of our NLP-based linguistic method with the MetaMap annotator to demonstrate our performance. In our evaluation, recognized diseases with

correct boundaries are considered as the exactly matched result, and the ones having boundary overlap with the correct annotations but not exactly matched are considered as the partially matched result. We use the standard Precision (P), Recall (R), and F-measure (F = 2PR /P +R) to measure the performance. In the 50 questions, there are totally 58 disease terms and the MetaMap annotator can correctly annotate 25 exactly matched ones and 50 partially matched. The NLP-based linguistic method improves the precision and recall of result for both the exactly and partially matched terms as in Table 2. Furthermore, for the document-level focused disease extraction (described in Section 1.4) in 50 literature articles, the accuracy performance of the NLP-based linguistic method is 78% which also performed better than the MetaMap annotator.

Table 2. Comparison of patient disease extraction results on MetaMap method and NLP-based linguistic method in the 50 questions.

Method	Patient Disease Extraction Result					
	Exactly Matched			Partially Matched		
	Precision	Recall	F-Score	Precision	Recall	F-Score
MetaMap	41%	43%	42%	82%	86%	84%
NLP-based linguistic method	65%	62%	64%	87%	83%	85%
Δ= (NLP-based linguistic method/MetaMap) - 1	59%	44%	52%	6%	-3%	1%

3. Discussion

In this paper, we have introduced the patient-related attributes extraction methods. We applied the NLP technologies to deeply analyze the characteristics of the attributes and improve the annotation accuracy. Besides the attribute annotation, this paper proposed the age category mapping methodology and focused attribute extraction from article. All the mentioned methods can be used to make the evidence more relevant and searchable. All the rules and methods in this paper have been used in our automatic clinical Question & Answering System – CliniQA. Patient-related attributes have been applied for the CliniQA components, including question analyzer, evidence ranking and appraisal. In the question analyzer, Patient-related Information is extracted for understanding the background of the patient. In the evidence ranking, we generated the Patient-related Scorers, i.e. patient age, gender, race and disease scorers, to help rate the relevant evidence whose population can match the background of the patient in question. In the evidence appraisal, the Patient-related Table is shown to users to help the evidence appraisal. As future work, we plan to employ NLP-based linguistic technologies to handle other elements of PICO.

References

[1] Sackett DL, Straus SE, Richardson WS, Rosenberg W, Haynes RB. Evidence-Based Medicine: How to Practice and Teach EBM, second edition. Churchill Livingstone, Edinburgh, 2000.
[2] Richardson, WS, Wilson MC, Nishikawa J, Hayward RS. The well-built clinical question: A key to evidence-based decisions. American College of Physicians Journal Club 1995; 123(3):A12–A13.
[3] Demner-Fushman D, Lin J. Answering Clinical Questions with Knowledge-based and Statistical Techniques. Computational Linguistics 2007; 33(1):63–103.
[4] Aronson AR. Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program. AMIA Annu Symp 2001; 17-21
[5] http://nlp.stanford.edu/software/lex-parser.shtml.
[6] www.tripanswers.org
[7] http://www.ncbi.nlm.nih.gov/pubmed/.