



International Conference on Computational Intelligence and Data Science (ICCIDS 2018) Biomedical Text Analytics for Characterizing Climate-Sensitive Disease

Md. Aslam Parwez^a, Muhammad Abulaish^{b,*}, Jahiruddin^a

^aJamia Millia Islamia (A Central University), New Delhi-25, India

^bSouth Asian University, Chanakyapuri, New Delhi-21, India

Abstract

Large-scale influx of scientific literatures in the biomedical domain, enriched with various biomedical entities like genes, proteins, drugs, diseases, symptoms, microbes, pathogens etc. embed many useful information that remains untapped due to unstructured nature of texts. Processing these texts using NLP techniques, and extracting embedded entities and their relations can provide useful information in creating disease knowledge base, which is a key enabler for the development of effective disease surveillance systems. In this paper, we present a biomedical text analytics approach to identify disease symptoms and relations from biomedical texts for characterizing climate-sensitive diseases. Four climate-sensitive infectious diseases, including *Cholera*, *Dengue*, *Influenza*, and *Malaria* are considered for experimentation, and it is found that the proposed approach is able to identify new disease symptoms that are even not listed on standard websites like Center for Disease Control (CDC), National Health Survey (NHS), and World Health Organization (WHO). In addition, it also identifies generic relations between diseases and their symptoms. The proposed approach could be useful for the development of climate-sensitive disease surveillance and prevention systems.

© 2018 The Authors. Published by Elsevier B.V.

Peer-review under responsibility of the scientific committee of the International Conference on Computational Intelligence and Data Science (ICCIDS 2018).

Keywords: Biomedical text analytics, Disease characterization, Disease symptoms identification, Relation mining, eHealth

1. Introduction

Databases like PubMed containing scientific literatures are vast repository of useful biomedical information. Extracting meaningful information from such massive repository is very challenging and needs careful examination of facts. The evidences stated in clinical and biological text documents have attracted many researchers to extract novel and significant information components from these documents [1, 2, 3]. As a result, biomedical text mining has evolved as a field providing many areas to explore. Biomedical relation extraction like gene-gene relations, gene-protein relations, protein-protein relations, disease-symptom relations etc. are the key areas that have contributed a lot for the development of many useful biomedical text information processing systems. Though a significant amount of

* Corresponding author. Tel.: +0-000-000-0000 ; fax: +0-000-000-0000.

E-mail addresses: aslamparwez.jmi@gmail.com (Md. Aslam Parwez), abulaish@sau.ac.in (Muhammad Abulaish), jahiruddin@jmi.ac.in (Jahiruddin).

1877-0509 © 2018 The Authors. Published by Elsevier B.V.

Peer-review under responsibility of the scientific committee of the International Conference on Computational Intelligence and Data Science (ICCIDS 2018).

work has been done for characterizing genes and proteins using the information embedded within biomedical texts, there is a paucity of work in disease characterization or disease-symptom relation identification, which is a key enabler for the development of disease knowledge bases and disease surveillance systems. Since most of disease and symptom related information are extensively available in biomedical articles and web resources that are somehow unstructured or semi-structured in nature and scattered in different resources, taping a comprehensive knowledge about the diseases, symptoms, and their associations is a challenging task.

In this paper, we present a biomedical text analytics approach to identify disease symptoms and relations from biomedical texts. Starting with the process of disease-centric documents retrieval from PubMud repository, the proposed approach applies a set of rules based on the dependency relationships generated by Stanford¹ parser and MetaMap[4] to extract candidate information components (ICs) represented in the form of triplets consisting of disease, symptom, and their relation. We have considered four climate-sensitive infectious diseases, including *Cholera*, *Dengue*, *Influenza*, and *Malaria* for experimental evaluation, and it is found that the proposed approach is able to identify many new disease symptoms that are even not listed on standard websites like Center for Disease Control (CDC), National Health Survey (NHS), and World Health Organization (WHO). The newly identified symptoms and relations could be helpful for the development of comprehensive disease knowledge base and eHealth applications like the development of infectious disease surveillance and prevention systems.

The rest of the paper is organized as follows. Section 2 presents a brief review of the related works on biomedical text analytics. Section 3 presents the functional details of the proposed approach. Section 4 presents the experimental set-up and performance evaluation results. Finally, section 5 concludes the paper with future directions of work.

2. Related Work

Approaches like statistical, pattern-based, rule-based, and machine learning have been used to extract biomedical relations, and statistical measures have been exploited to extract associations between entities co-occurring in a sentence or in a document [5, 6]. In [7], Fundel et al. extracted gene-protein relations using simple rules derived from dependency parse trees. Hassan et al. [8, 9] used graph-based pattern mining techniques on dependency graphs to extract disease-symptom relations. Bunescu and Mooney [10] in their shortest path dependency kernel method tried to capture shortest path between two entities for extracting relations. Linguistic dependencies, though rich in knowledge representation, have been used by few researchers due to its complexity and needs attention of the research community. Accordingly, Seneviratne and Ranasinghe [11] used typed dependencies to learn rules for extracting relations between birds and their locations. They presented a good insight to exploit typed dependencies generated by the Stanford parser.

In [12], the authors proposed disease network and the molecular interaction of diseases manifestation along with genetic associations based on symptoms similarity between two diseases. Pletscher-Frankild et al. [13] presented dictionary-based text mining tool for distilling disease-gene associations. In another work for diseases-genes relation extraction by Bravo et al. [14], a text mining system called *BeFree* is proposed to identify associations between genes, drugs, and diseases. Sondhi et al. [15] proposed a symptom relation analyzing framework, *SympGraph*, for clinical notes to identify related symptoms from the given set of symptoms by expanding the symptom graph. Datla et al. [16] applied higher order co-occurrence technique for disease and symptoms co-occurrence relations using original Latent Semantic Analysis (LSA). Tran et al. [17] tried to map symptom concepts anatomically related to organ systems by exploiting the Unified Medical Language System (UMLS) Metathesaurus.

It can be seen that most of the entity and relation extraction processes mentioned above aim to identify pre-defined associations between entities. This paper presents a generic approach to identify climate-sensitive disease symptoms and relations from PubMed documents. However, the proposed approach can be applied to any set of biomedical text documents to characterize any diseases in terms of their symptoms and associations.

¹ <http://nlp.stanford.edu/software/lex-parser.shtml>

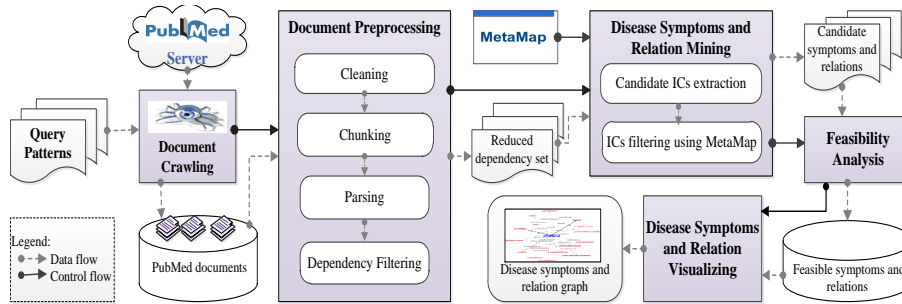


Fig. 1: Work-flow of the proposed biomedical text analytics process

3. Proposed Biomedical Text Analytics Approach

In this section, we present the functional details of the proposed biomedical text analytics approach. Figure 1 delineates various tasks performed and the flow of information during the symptoms and relation extraction processes. Further detail about each processing steps is presented in the following sub-sections.

3.1. Document Crawling

PubMed database manages a treasury of published biomedical literatures and grants access to their abstracts, including authors and their affiliations, and other meta-data. In order to retrieve query-centric PubMed abstracts, we have implemented a crawler in Java using *axis 2.1.6.2 API²* provided by the NCBI (National Center for Biotechnology Information) Entrez system. A total number of four diseases (Cholera, Dengue, Influenza, and Malaria) and their 66 symptoms listed on Center for Disease Control (CDC), National Health Survey (NHS), and World Health Organization (WHO) websites are considered to generate different query patterns. As a result, total 264 query patterns are generated and PubMed is queried using the NCBI's Entrez system API, resulting in the retrieval of total 27867 documents, including their PMIDs, titles, and abstracts.

3.2. Document Preprocessing

The crawled documents have many replicas due to presence of two or more diseases and/or symptoms in a single document, and consequently their retrieval against different queries. As PubMed documents have unique PubMed IDs, their ID's are employed to eliminate redundant documents; and as a result, only 19194 out of total 27867 documents are left in the dataset. Thereafter, each document is processed through sentence splitting and parsing to generate typed dependencies and Parts-Of-Speech (POS) tags using Stanford parser. Figure 2 depicts an exemplar dependency parse tree showing the dependency relationships and POS tags of a sample sentence produced by the Stanford parser using *DependenSee 3.7.0*.

It is found that though most of the dependency tuples are helpful in relation identification, there are some irrelevant dependencies like *det*, *dep*, etc. that need to be wiped out as they are either frivolous (*det*, *dep*) or worthless for information component extraction. A pair of exemplar sentences with POS tags, dependencies, and the reduced dependencies after removal of extraneous elements is shown in Figure 3.

Consider the dependency tuple *amod(disease-5, infectious-4)* of the sample texts outlined in Figure 3 under the caption *typed dependencies*. The governor word *disease* or the dependent word *infectious* itself seem incomplete to understand the disease, while the composite word *infectious_disease* indicates the absolute sense about the disease. Likewise, words of tuples *amod(diarrhea-9, watery-8)*, *amod(dehydration-14, severe-13)*, and *compound(cramps-18, muscle-17)* can be clubbed together to frame composite words *watery_diarrhea*, *severe_dehydration*, and *muscle_cramps*, respectively to achieve complete impression of a disease symptom or a disease concept. The concatenated

² <http://axis.apache.org/axis2/java/core/>

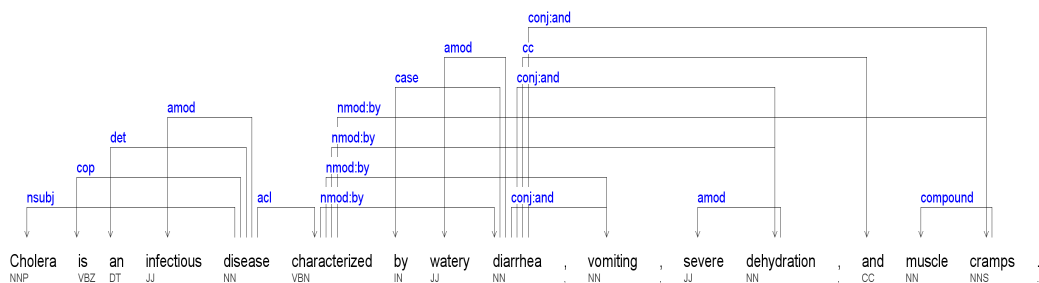


Fig. 2: An exemplar dependency parse tree generated by the Stanford parser using DependSee 3.7.0

Exemplar sentences:

Cholera is an infectious disease characterized by watery diarrhea, vomiting, severe dehydration, and muscle cramps. Cholera commonly spread by contaminated drinking water and unsafe food.

POS tagged sentences:

Cholera/NNP, is/VBZ, an/DT, infectious/JJ, disease/NN, characterized/VBN, by/IN, watery/JJ, diarrhea/NN, ,/, vomiting/NN, ,/, severe/JJ, dehydration/NN, ,/, and/CC, muscle/NN, cramps/NNS, ./.

Cholera/NNP, commonly/RB, spread/VBD, by/IN, contaminated/JJ, drinking/NN, water/NN, and/CC, unsafe/JJ, food/NN, ./.

Typed dependencies:

nsubj(disease-5, Cholera-1), cop(disease-5, is-2), det(disease-5, an-3), amod(disease-5, infectious-4), root(ROOT-0, disease-5), acl(disease-5, characterized-6), case(diarrhea-9, by-7), amod(diarrhea-9, watery-8), nmod:by(characterized-6, diarrhea-9), nmod:by(characterized-6, vomiting-11), conj:and(diarrhea-9, vomiting-11), amod(dehydration-14, severe-13), nmod:by(characterized-6, dehydration-14), conj:and(diarrhea-9, dehydration-14), cc(diarrhea-9, and-16), compound(cramps-18, muscle-17), nmod:by(characterized-6, cramps-18), conj:and(diarrhea-9, cramps-18)

nsubj(spread-3, Cholera-1), advmod(spread-3, commonly-2), root(ROOT-0, spread-3), case(water-7, by-4), amod(water-7, contaminated-5), compound(water-7, drinking-6), nmod:by(spread-3, water-7), cc(water-7, and-8), amod(food-10, unsafe-9), nmod:by(spread-3, food-10), conj:and(water-7, food-10)

Reduced dependencies:

nsubj(infectious_disease NN 5, Cholera NNP 1)
 cop(infectious_disease NN 5, is VBZ 2)
 acl(infectious_disease NN 5, characterized VBN 6)
 case(watery_diarrhea NN 9, by IN 7)
 nmod:by(characterized VBN 6, watery_diarrhea NN 9)
 nmod:by(characterized VBN 6, vomiting NN 11)
 conj:and(watery_diarrhea NN 9, vomiting NN 11)
 nmod:by(characterized VBN 6, severe_dehydration NN 14)
 conj:and(watery_diarrhea NN 9, severe_dehydration NN 14)
 cc(watery_diarrhea NN 9, and CC 16)
 nmod:by(characterized VBN 6, muscle_cramps NNS 18)
 conj:and(watery_diarrhea NN 9, muscle_cramps NNS 18)

nsubj(commonly_spread VBD 3, Cholera NNP 1)
 advmod(commonly_spread VBD 3, commonly_spread VBD 3)
 case(contaminated_drinking_water NN 7, by IN 4)
 nmod:by(commonly_spread VBD 3, contaminated_drinking_water NN 7)
 cc(contaminated_drinking_water NN 7, and CC 8)
 nmod:by(commonly_spread VBD 3, unsafe_food NN 10)
 conj:and(contaminated_drinking_water NN 7, unsafe_food NN 10)

Fig. 3: Exemplar sentences with POS tags, typed dependencies, and reduced typed dependencies

words thus obtained are substituted within the dependency relation tuples, and thereafter the *det*, *dep*, *amod*, and *compound* etc. dependency tuples are taken out to procure reduced dependency set maintaining the order of the surviving dependencies, as they stood in the initial dependencies generated by the parser. The expulsion of these unessential dependency tuples decreases the number of tuples to be dealt with for extraction of relevant information components.

3.3. Disease Symptoms and Relation Mining

Mining disease symptoms and relations involves extraction of *information components* from the reduced dependency set and identification of disease and symptom entities embedded within the triplet. Formally, an *information component* can be defined as follows:

Definition 1. (*Information Component*). An *information component* is a triplet of the form $\langle argument_1, relation, argument_2 \rangle$ where, $argument_1$ and $argument_2$ represent words/phrases consisting of entities that may be a disease or symptom concept, and $relation$ represents relational words showing relationship between $argument_1$ and $argument_2$. The relational words may have either a verb or verb with preposition or even sometimes noun with preposition.

The governor and dependent words from the reduced dependency set can be mapped to *information components* using a set of rules based on typed dependencies. A total number of 67 rules are designed to handle different sentence structures for information component extraction; out of which one rule is explained in the following paragraph.

Rule-1:

The first rule states that, if reduced dependencies have dependency relation *nsubj* with first word (say w_1) as noun and the second word (say w_2) as noun or verb, and there is a *cop* dependency relation with first word same as the first word of *nsubj* and the second word (say w_3) is a copular verb, then the identified triplet would be $\langle w_2, w_3, w_1 \rangle$, that is, second word (w_2) followed by the copular verb (w_3), which is followed by the first word (w_1). In short, it can be written as given below, where w_1, w_2, w_3 etc. represent words, NN* and VB* represent POS tags for nouns and verbs, respectively.

$$[(nsubj(w_1/NN*, w_2/NN*) \vee nsubj(w_1/NN*, w_2/VB*)) \wedge cop(w_1/NN*, w_3/VB*)] \implies \langle w_2, w_3, w_1 \rangle$$

For example, considering the first two reduced dependencies *nsubj(infectious_disease/NN/5, Cholera/NNP/1)* and *cop(infectious_disease/NN/5, is/VBZ/2)* of Figure 3 and applying *Rule-1* yields the extraction of the following information component:

$\langle cholera, is, infectious_disease \rangle$

Table 2 presents the list of extracted information components identified from the exemplar texts given in Figure 3. Once the information components are extracted, each of them is passed to *MetaMap*³, which is a tool to identify disease and symptom entities. *MetaMap* recognizes UMLS⁴ concept alluded in biomedical contents and traces their semantic types to any of the pre-defined 133 semantic categories⁵. In this study, we have considered only nine semantic categories to map entities in the extracted information components, as they portray majority of the disease and symptom concepts. Table 1 embodies these semantic categories and their descriptions. In this way, *MetaMap* helps to retain only those information components in which both left and right components contain either a disease or a symptom entity. Table 3 presents the list of retained information components after filtering the last two information components of Table 2 as their right component does not contain any disease or symptom entity.

Table 1. *MetaMap* semantic categories representing disease/symptom concepts

Semantic Category	Description
dsyn	Disease or Syndrome
neop	Neoplastic Process
anab	Anatomical Abnormality
sosy	Sign or Symptom
findg	Finding
patf	Pathologic Function
mobd	Mental or Behavioral Dysfunction
cgab	Congenital Abnormality
inpo	Injury or Poisoning

Table 2. Information components extracted from the exemplar sentences

First Entity	Relation	Second Entity
Cholera	is	infectious_disease
Cholera	characterized by	watery_diarrhea
Cholera	characterized by	vomiting
Cholera	characterized by	severe_dehydration
Cholera	characterized by	muscle_cramps
Cholera	commonly spread by	contaminated_drinking_water
Cholera	commonly spread by	unsafe_food

Table 3. Retained information components after filtering using *MetaMap* semantic categories

First Entity	Relation	Second Entity
Cholera	is	infectious_disease
Cholera	characterized by	watery_diarrhea
Cholera	characterized by	vomiting
Cholera	characterized by	severe_dehydration
Cholera	characterized by	muscle_cramps

³ <https://metamap.nlm.nih.gov/>

⁴ Unified Medical Language System

⁵ http://metamap.nlm.nih.gov/Docs/SemanticTypes_2013AA.txt

3.4. Feasibility Analysis

To identify feasible symptoms for each disease under consideration, we have applied three different ranking techniques – (i) Frequency count, (ii) Global term frequency and inverse document frequency (tf-idf^G) scoring, and (iii) Document-level (local) tf-idf^L scoring; hereafter termed as *RM1*, *RM2*, and *RM3*, respectively. We have also considered a hybrid method in which only those symptoms that are common to these three ranking methods are considered as feasible; hereafter termed as *HRM*. The first method (*RM1*) simply considers frequency count of the terms (symptoms or diseases) based on their recurrence in the list of identified information components from document collection, *D*. The recurrence count of a term *s* is represented as *freqCount(s)* and calculated using equation 1, where $|D|$ is number of documents in *D*, *freq(s, d_i)* represents the frequency of term *s* in the *i*th document *d_i* ∈ *D*.

$$freqCount(s) = \sum_{i=1}^{|D|} freq(s, d_i) \quad (1)$$

The tf-idf [18] is a robust NLP technique of weighting or ranking terms occurring recurrently in a document, but seldom in the whole document collection. This notion is used to rank extracted symptoms for a particular disease. The tf-idf computation is done in two ways. The first method (*RM2*) uses equations 2, 3, and 4 to calculate the term frequency and inverse document frequency of symptom *s_k* in the document corpus *D*, where $|D|$ is the number of documents in *D*, and $|D_{s_k}|$ serves as the number of documents containing *s_k*.

$$tf-idf^G(s_k) = tf(s_k) \times idf(s_k) \quad (2)$$

$$tf(s_k) = \sum_{i=1}^{|D|} freqCount(s_k, d_i) \quad (3)$$

$$idf(s_k) = \log\left(\frac{|D|}{|D_{s_k}|}\right) \quad (4)$$

The expression *tf-idf^G(s_k)* is global tf-idf score of symptom *s_k*, where *tf(s_k)* is the term frequency of *s_k* in document *d_i*, and *idf(s_k)* symbolizes inverse document frequency of *s_k*.

In second method of tf-idf calculation (*RM3*), tf-idf score of each term is calculated at document level and then aggregated to a single score, using equations 5, 6 and 7, where *tf-idf^L(s_k)* portrays sum of tf-idf score of the symptom *s_k* at document level and *tf-idf^{d_i}(s_k)* is the tf-idf score of *s_k* in document *d_i*. The equation 4 is adopted to determine the inverse document frequency, and the equation 7 is employed to calculate the term frequency score *tf^{d_i}(s_k)* of symptom term *s_k* in document *d_i*.

$$tf-idf^L(s_k) = \sum_{i=1}^{|D|} tf-idf^{d_i}(s_k) \quad (5)$$

$$tf-idf^{d_i}(s_k) = tf^{d_i}(s_k) \times idf(s_k) \quad (6)$$

$$tf^{d_i}(s_k) = \frac{freqCount(s_k, d_i)}{\sum_{j=1}^{|S|} freqCount(s_j, d_i)} \quad (7)$$

3.5. Disease Symptoms and Relation Visualizing

In order to visualize feasible disease symptoms and relations, we have used *Gephi 0.8.2*⁶, a leading open source tool for exploratory data analysis and visualization of graphs and networks. It facilitates to draw directed or undirected graphs by considering weight of edges as the frequency count of the edges connecting source and target nodes. This is reflected by the thickness of an edge connecting a pair of nodes. The customization of colour, size and labels of nodes and edges conveys good sense of network depiction.

⁶ <https://gephi.org>

Table 4. Performance evaluation results on different test sets of varying sizes

Test Sets	#Actual ICs	TP	FP	P	R	F1
TS1	72	37	10	78.72	51.39	62.18
TS2	147	62	20	75.61	42.18	54.54
TS3	437	164	56	74.54	37.53	49.92
TS4	757	282	97	74.40	37.25	49.69

Table 5. Common concepts among the three methods for each disease

Disease Name	Identified Symptoms	
	Symptoms already listed at CDC, NHS, or WHO websites	Newly identified symptoms
Cholera	bacterial infection, dehydration, diarrhea, diarrheal stool, intestinal infection, vomiting, watery diarrhea, watery stools	acute septicemia pneumonia, antimicrobial susceptibility, gastroenteritis, infectious disease, lesion, communicable disease, contagious disease
Dengue	bleeding complication, fever, haemorrhage, hemorrhagic shock syndrome, pain, infectious disease, mild fever, mosquito borne infection, plasma leakage, rash, shock, thrombocytopenia, vascular leakage, viral disease	arboviral disease, arthropod borne infection, encephalitis, encephalopathy, febrile illness, hepatitis, leptospirosis, neurologic complication, , severe disorder, yellow fever
Influenza	acute respiratory disease, asthma, coma, cough symptom, fever ,flu, sore throat, respiratory infection, viral infection, meningitis, encephalitis, pneumonia, respiratory symptom	encephalopathy, fatigue, febrile illness, lesions, neurologic complication, pandemic h1n1, seizures, sequelae
Malaria	anaemia, fever, low birth weight, coma, parasitemia, parasitic disease, plasmodium falciparum infection, plasmodium vivax infection, neurological abnormalities	attack, chemoprophylaxis, co-infection, congenital infection, convulsion, febrile patient, febrile illness, g6pd deficiency, hiv 1 infection, infectious disease, intermittent preventive treatment, rapid diagnostic test, renal dysfunction, seizure, sickle cell trait, splenomegaly, thrombocytopenia

4. Experimental Setup and Results

In this section, we present experimental results over a dataset of 19194 unique documents containing 173617 sentences. Out of these, we have considered only those sentences that contain at least one disease or symptom name to enhance the efficiency of the proposed text information processing method. Due to unavailability of any labelled corpora, we have manually labelled valid triplets embedded within the sentences of the dataset with the help of domain experts. Since manual labelling of whole dataset is not feasible, four different varying-size test sets *TS1*, *TS2*, *TS3*, and *TS4* comprising 100, 200, 500, and 1000 sentences, respectively are generated from the entire dataset using random sampling with replacement strategy.

For performance evaluation, we have considered standard Information Retrieval (IR) metrics *Precision*, *Recall*, and *F-measure*, which are defined using equations 8, 9, and 10, respectively in terms of *True Positive* (TP), *False Positive* (FP), and *False Negative* (FN), where *TP* is the number of positive instances identified as positive, *FP* is the number of negative instances identified as positive, and *FN* is the number of positive instances identified as negative.

$$Precision(P) = \frac{TP}{TP + FP} \quad (8)$$

$$Recall(R) = \frac{TP}{TP + FN} \quad (9)$$

$$F-measure(F1) = \frac{2 \times P \times R}{P + R} \quad (10)$$

Table 4 presents the evaluation results of the proposed information components extraction method on different test sets of varying sizes. It can be observed from this table that the *F-measure* (F1) is decreasing with increasing number of sentences in the datasets, which is mainly due to increase in false negatives (resulting in low recall) with increasing number of sentences. However, precision values remain consistent across the data sets. On analysis, we found that low recall values are mainly due to the structure and complexity of sentences composed with some implied relations, and shortcomings of existing NLP tools. Nonetheless, the uniqueness of our proposed approach lies in amalgamation of named entities and typed dependency-based rules to identify disease symptoms and their relations in unstructured text documents.

In order to identify feasible symptoms and relations, a separate list of symptoms for each disease is compiled from the information components and ordered using the ranking methods discussed in the previous section. Thereafter, common elements in top-30 of all three ranking methods are considered as the feasible symptoms. Finally, the *relation*

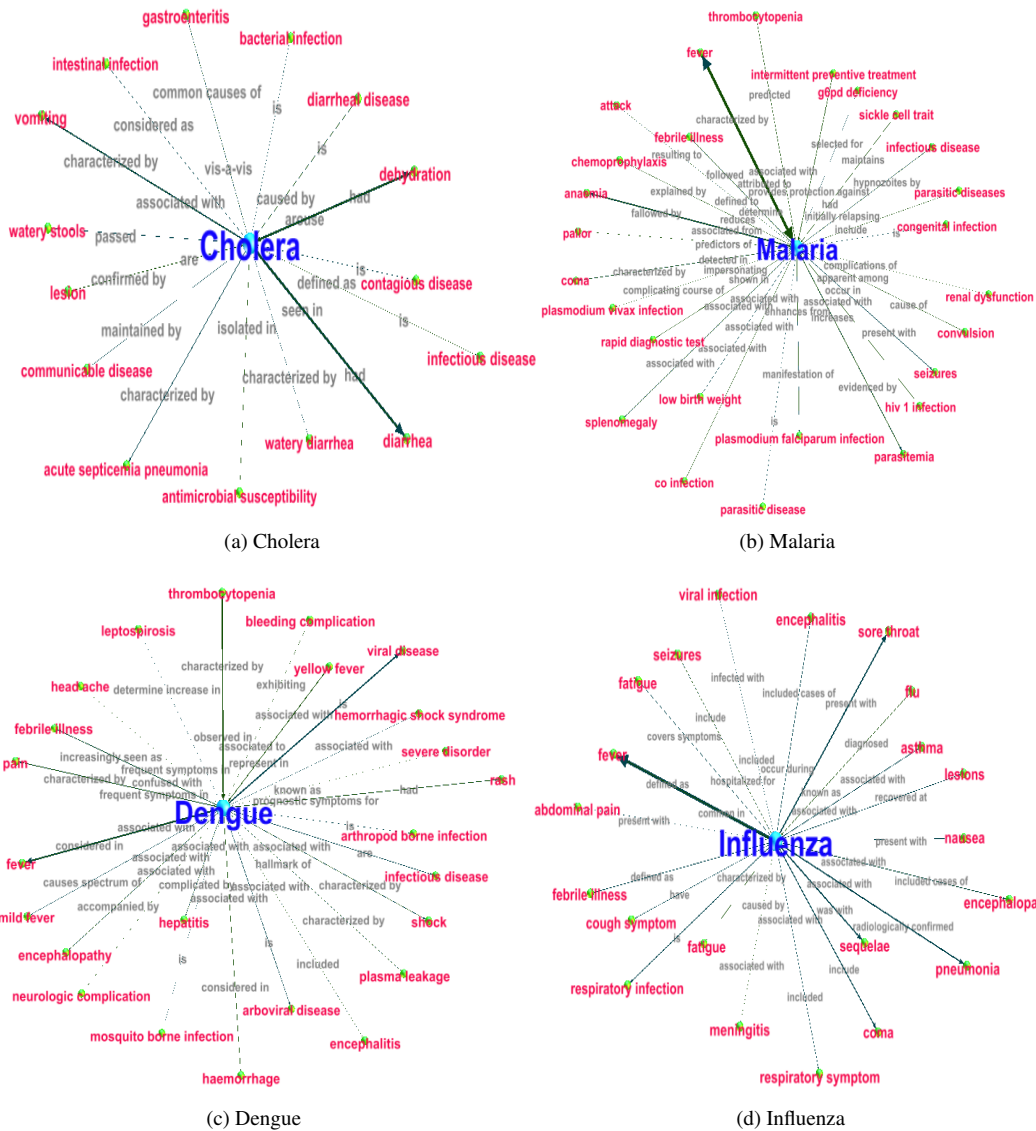


Fig. 4: Visualization of disease-symptoms relationships for Cholera, Malaria, Dengue, and Influenza using Gephi 0.8.2

Table 6. Partial list of information components containing feasible Cholera symptoms and relations

Disease/Symptom	Relation	Symptom/Disease
cholera	characterize by	dehydration
severe dehydration	associated with	cholera
cholera	characterize by	diarrhea
vomiting	associated with	cholera
cholera	considered as	acute intestinal infection
cholera	is	infectious disease
hog cholera	confirmed by	lesion
cholera	is	diarrheal disease
patient with cholera	pass	watery stool
cholera	common cause of	infantile gastroenteritis

Table 7. Partial list of information components containing feasible Dengue symptoms and relations

Disease/Symptom	Relation	Symptom/Disease
dengue	characterized by	fever
dengue fever	is	arboviral disease
shock syndrome	associated with	dengue fever
dengue fever	present with	thrombocytopenia
bleeding complication	associated with	dengue infection
dengue fever	characterized by	rash
plasma leakage	occurs in	dengue hemorrhagic fever
rash	associated with	dengue fever
encephalopathy	associated with	dengue fever
dengue	is	mosquito borne infection

Table 8. Partial list of information components containing feasible Influenza symptoms and relations

Disease/Symptom	Relation	Symptom/Disease
influenza	characterized by	high fever
high fever	suggests	influenza infection
influenza	present with	cough symptom
influenza	is	viral infection
pneumonia	associated with	influenza
influenza	included	respiratory infection
influenza	present with	sore throat
influenza	caused	lesions
influenza	associated with	asthma
bacterial meningitis	associated with	influenza

Table 9. Partial list of information components containing feasible Malaria symptoms and relations

Disease/Symptom	Relation	Symptom/Disease
malaria	characterized by	fever
malaria	associated with	anaemia
malaria	causes	anaemia
		plasmodium
malaria	complication of	falciparum infection
plasmodium	causing	malaria
vivax infection		
cerebral malaria	associated with	seizure
malaria	is	parasitic disease
cerebral malaria	characterized by	coma
hiv 1 infection	associated with	malaria
thrombocytopenia	complication in	malaria

Table 10. Detection Rate (DR) values for feasible symptoms identification using different ranking methods

Disease Name	RM1			RM2			RM3			HRM		
	TP	FP	DR	TP	FP	DR	TP	FP	DR	TP	FP	DR
Cholera	27	3	0.90	27	3	0.90	27	3	0.90	14	1	0.93
Dengue	30	0	1.00	30	0	1.00	29	1	0.97	24	0	1.00
Influenza	30	0	1.00	30	0	1.00	29	1	0.97	21	0	1.00
Malaria	26	4	0.87	26	4	0.87	26	4	0.87	23	3	0.88

constituents of all information components having a disease name and any of the feasible symptoms are considered as feasible disease-symptom relations.

As stated earlier, the proposed approach is able to identify new disease symptoms that are even not listed on standard websites like Center for Disease Control (CDC), National Health Survey (NHS), and World Health Organization (WHO). Table 5 presents the list of identified symptoms for each disease, wherein we have segregated the list into two parts – one containing the symptoms that are already listed on CDC, NHS, or WHO websites, and the other containing the newly identified symptoms.

Tables 6, 7, 8, and 9 present a partial list of information components containing feasible symptoms and relations characterizing Cholera, Dengue, Influenza, and Malaria diseases, respectively. Figure 4(a-d) presents a visualization of the identified feasible symptoms and relations for these diseases, wherein the focal sphere (blue colour) shows disease name and the outer spheres portray the symptoms. The labels assigned to edges connecting a disease and symptoms depicts the disease-symptom relations.

In order to establish the accuracy of the identified symptoms, we have used Detection Rate (DR) measure which is defined using equation 11, where *TP* is the number of identified valid symptoms related to the disease under study, and *FP* is the number of identified symptoms that are not related to the disease under study.

$$DR = \frac{TP}{TP + FP} \tag{11}$$

Table 10 presents the *DR* values for identified symptoms using different ranking methods. The last column presents the case where only those symptoms that are common to ranked lists by all three ranking methods are considered. It can be noted in Table 10 that one symptom is wrongly identified as valid symptom of Cholera by all three methods. On analysis, it is found that the identified symptom is “antimicrobial susceptibility”, which is neither a symptom nor a disease or disease type, associated with Cholera. Similarly, in case of Malaria, three terms “rapid diagnostic test”, “chemoprophylaxis”, and “intermittent preventive treatment” are wrongly identified as its symptoms. Though all these terms are related to Malaria, they are neither a symptom nor a disease or disease type. On analysis, we found that “rapid diagnostic test” is used to test the presence of malaria parasite, “chemoprophylaxis” is the administration of medications to prevent disease or infections like malaria, and “intermittent preventive treatment” is the public health intervention used to treat and prevent malaria in infants.

5. Conclusion and Future Directions of Work

In this paper, we have presented a biomedical text mining approach to identify disease symptoms and their associations by exploiting named entities and typed dependencies generated by the Stanford parser. The proposed approach is able to identify new symptoms of the diseases under study that are even not listed on standard disease-related websites despite their existence in biomedical literatures. The extracted disease symptoms and relations can be used to develop an exhaustive knowledge base in the form of ontology for biomedical text information processing. At present, we are extending our experiment to consider more diseases for the development of a climate-sensitive infectious disease surveillance and prevention system.

Acknowledgements

The authors would like to thank the University Grant Commission (UGC) of India for providing financial assistance as Senior Research Fellow (SRF) to the first author under its Maulana Azaad National Fellowship (MANF) scheme.

References

- [1] M. Abulaish, L. Dey (2007) “Biological relation extraction and query answering from medline abstracts using ontology-based text mining”, *Data and Knowledge Engineering* 61 (2) 228–262.
- [2] M. Abulaish, Jahiruddin (2013) “Web content mining for learning generic relations and their associations from textual biological data”, in: M. Elloumi, A. Y. Zomaya (Eds.), *Biological Knowledge Discovery Handbook: Preprocessing, Mining, and Postprocessing of Biological Data*, John Wiley & Sons, Inc., pp. 919–942.
- [3] L. Dey, M. Abulaish, G. Sharma, Jahiruddin (2007) “Text mining through entity-relationship based information extraction”, in: *Proceedings of the Workshop on Biomedical Applications of Web Technologies, Co-located with IEEE/WIC/ACM International Conference on Web Intelligence*, Silicon Valley, USA, Springer, pp. 177–180.
- [4] A. R. Aronson (2001) Effective mapping of biomedical text to the umls metathesaurus: the metamap program., in: *Proceedings of the AMIA Symposium*, American Medical Informatics Association, pp. 17–21.
- [5] Jahiruddin, M. Abulaish, L. Dey (2010) “A concept-driven biomedical knowledge extraction and visualization framework for conceptualization of text corpora”, *Journal of Biomedical Informatics* 43 (6) 1020–1035.
- [6] M. Abulaish, L. Dey (2009) “A relation mining and visualization framework for automated text summarization”, in: *Proceedings of the 3rd International Conference on Pattern Recognition and Machine Intelligence (PReMI)*, Delhi, India, LNCS-5909, Springer, Dec. 16-20, pp. 249–254.
- [7] K. Fundel, R. Küffner, R. Zimmer (2007) “Relex: –relation extraction using dependency parse trees”, *Bioinformatics* 23 (3) 365–371.
- [8] M. Hassan, A. Coulet, Y. Toussaint (2014) “Learning subgraph patterns from text for extracting disease–symptom relationships”, in: *Proceedings of 1st International Workshop on Interactions between Data Mining and Natural Language Processing*, Vol. 1202, pp. 81–96.
- [9] M. Hassan, O. Makkaoui, A. Coulet, Y. Toussaint (2015) “Extracting disease-symptom relationships by learning syntactic patterns from dependency graphs”, in: *Proceedings of the Workshop on Biomedical Natural Language Processing (BioNLP)*, pp. 184–194.
- [10] R. C. Bunescu, R. J. Mooney (2005) “A shortest path dependency kernel for relation extraction”, in: *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, pp. 724–731.
- [11] M. D. S. Seneviratne, D. N. Ranasinghe (2014) “Natural language dependencies for ontological relation extraction”, in: *Proceedings of International Conference on Advances in ICT for Emerging Regions (ICTer)*, IEEE, pp. 142–148.
- [12] X. Zhou, J. Menche, A.-L. Barabási, A. Sharma (2014) “Human symptoms–disease network”, *Nature communications* 5 4212.1–4212.10.
- [13] S. Pletscher-Frankild, A. Palleja, K. Tsaou, J. X. Binder, L. J. Jensen (2015) “Diseases: Text mining and data integration of disease–gene associations”, *Methods* 74 83–89.
- [14] À. Bravo, J. Piñero, N. Queralt-Rosinach, M. Rautschka, L. I. Furlong (2015) “Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research”, *BMC bioinformatics* 16 (1) 55.
- [15] P. Sondhi, J. Sun, H. Tong, C. Zhai (2012) “Symprgraph: a framework for mining clinical notes through symptom relation graphs”, in: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, pp. 1167–1175.
- [16] V. Datla, K.-I. Lin, M. Louwse (2012) “Capturing disease-symptom relations using higher-order co-occurrence algorithms”, in: *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW)*, IEEE, pp. 816–821.
- [17] L.-T. T. Tran, G. Divita, M. E. Carter, J. Judd, M. H. Samore, A. V. Gundlapalli (2015) “Exploiting the umls metathesaurus for extracting and categorizing concepts representing signs and symptoms to anatomically related organ systems”, *Journal of Biomedical Informatics* 58 19–27.
- [18] C. D. Manning, P. Raghavan, H. Schütze (2008) “Scoring, term weighting and the vector space model”, *Introduction to Information Retrieval* 100 2–4.