Final version of the accepted paper. Cite as: *Muhammad Abulaish, Md A. Parwez, and Jahiruddin, DiseaSE: A Biomedical Text Analytics System for Disease Symptom Extraction and Characterization, Journal of Biomedical Informatics, Vol. 100, 103324, pp. 1-23, 2019.*

`DiseaSE`: A Biomedical Text Analytics System for Disease Symptom Extraction and Characterization

Muhammad Abulaish, *SMIEEE*[a], Md. Aslam Parwez[b], Jahiruddin[b]

[a] *Department of Computer Science, South Asian University, Delhi, India*
[b] *Department of Computer Science, Jamia Millia Islamia, Delhi, India*

Abstract

Due to increasing volume and unstructured nature of the scientific literatures in biomedical domain, most of the information embedded within them remain untapped. This paper presents a biomedical text analytics system, `DiseaSE` (**Disea**se **S**ymptom **E**xtraction), to identify and extract disease symptoms and their associations from biomedical text documents retrieved from the `PubMed` database. It implements various NLP and information extraction techniques to convert text documents into record-size *information components* that are represented as semantic triples and processed using `TextRank` and other ranking techniques to identify feasible disease symptoms. Eight different diseases, including *dengue*, *malaria*, *cholera*, *diarrhoea*, *influenza*, *meningitis*, *leishmaniasis*, and *kala-azar* are considered for experimental evaluation of the proposed `DiseaSE` system. On analysis, we found that the `DiseaSE` system is able to identify new symptoms that are even not catalogued on standard websites such as Center for Disease Control (CDC), World Health Organization (WHO), and National Health Survey (NHS). The proposed `DiseaSE` system also aims to compile generic associations between a disease and its symptoms, and presents a graph-theoretic analysis and visualization scheme to characterize disease at different levels of granularity. The identified disease symptoms and their associations could be useful to generate a biomedical knowledgebase (e.g., a disease ontology) for the development of e-health and disease surveillance systems.

Keywords:
Biomedical text mining, Entity extraction, Disease characterization, Symptom extraction, Visualization.

1. Introduction

Biomedical literature databases like `PubMed` are treasury of scientific literatures that encapsulates an enormous amount of valuable information. Extracting relevant information from such valuable resources need strenuous effort and careful examination of the evidences. The research findings recorded in scientific literatures have drawn attention of several researchers to extract innovative and significant *information components* from biomedical texts. As a result, biomedical information retrieval and extraction has emerged as a field providing many areas to explore. Starting with biological entity extraction, mining biological relations, such as gene-gene interactions, gene-protein associations, protein-protein interactions, disease-gene associations, disease-symptom associations, etc. are the important fields that have played a vital role to develop many pragmatic and valuable biomedical text information processing systems. Despite the availability of large repository of biomedical literatures and many works in genes and proteins characterization and their associations extraction from biomedical literatures, there are limited works in disease symptom identification or disease characterization, which can facilitate to develop healthcare and disease surveillance systems. As most of the disease- and symptom-related useful information are embedded within

Email addresses: `abulaish@sau.ac.in` (Muhammad Abulaish, *SMIEEE*), `aslamparwez.jmi@gmail.com` (Md. Aslam Parwez), `jahiruddin@jmi.ac.in` (Jahiruddin)

biomedical literatures and web resources that are some way or another scattered and unorganized or semi-organized, extracting meaningful information and comprehensive knowledge associated to diseases, symptoms, and their relations still remains a challenging research problem.

This paper attempts to present the development of a biomedical text analytics system named `DiseaSE` (Disease Symptom Extraction) to extract disease symptoms and their associations from biomedical texts and characterize disease at different levels of granularity. It is a major extension of one of our previously published conference papers, [30], by considering more diseases, larger dataset, abbreviation handling mechanism, additional methods of ranking and analysis, visualization of information at different levels of granularity (disease-disease, disease-symptom, and symptom-symptom similarity), validation of the identified disease symptoms from medical doctors, and a critical discussion. Commencing with the procedure of disease-centric query-based documents retrieval from `PubMed` repository, the proposed system uses syntactic patterns of dependency relationships generated by Stanford parser[1] to distill candidate *information components* (ICs) from biomedical text documents, and represents them as sematic triples consisting of disease, symptom, and their association. `MetaMap`[2] [4], an NER annotation tool that identifies the disease symptom concepts according to their defined semantic categories, is used to retain ICs possessing valid disease and symptom. We have considered eight different diseases, including *dengue*, *malaria*, *cholera*, *diarrhoea*, *influenza*, *meningitis*, *leishmaniasis*, and *kala-azar* for experimental evaluation of the proposed system. A brief descriptions of these diseases are presented in Table 1.

Ranking identified symptoms associated with a particular disease is crucial in determining their significance with respect to a given text corpus. For ranking, four different methods relying on statistical and graph-based approaches are employed and symptoms common to all methods are regarded as the feasible symptoms. On analysis, it is found that the proposed `DiseaSE` system can identify many new symptoms of the diseases that are remained unspecified on standard websites such as Center for Disease Control (CDC), World Health Organization (WHO), and National Health Survey (NHS). The identified disease symptoms and their associations could be useful to develop a comprehensive disease knowledgebase for e-health applications, like disease surveillance, control, and prevention systems. Further, the comprehensive list including existing and newly identified symptoms is important for understanding the underlying disease and its association with other diseases. The symptom-based similarity among the diseases can also be helpful to understand their epidemiology.

In short, the key contribution of this paper can be summarized as follows:

- Development of a Disease Symptoms Extraction (`DiseaSE`) system for identifying symptoms and their associations from biomedical documents to characterize and visualize disease at different levels of granularity.
- Development of an *information component* extraction technique to identify information components from biomedical documents and represent them as semantic triples containing disease, symptoms, and their associations.
- An abbreviation processing mechanism to map biomedical abbreviations to respective disease symptom concepts, with respect to an underlying biomedical text corpus.
- A feasibility analysis approach using `TextRank` and other ranking methods to assess feasible disease symptoms and their associations from the semantic triples.

The rest of the paper is organized as follows. Section 2 presents a brief overview of the existing works on biomedical text information processing. Section 3 deals with the functional aspects of the proposed `DiseaSE` system. Section 4 presents the experimental setup and evaluation results. Critical discussion highlighting the implication of findings and recommendation for further enhancement of the proposed system is discussed in section 7. Finally, section 8 concludes the paper and directs pos-

[1] http://nlp.stanford.edu/software/lex-parser.shtml

[2] http://metamap.nlm.nih.gov/

Table 1: Disease names and their brief descriptions [source: CDC, WHO, and NHS]

S.No.	Disease name	Description
1	Dengue	• A disease having presence in tropics and subtropics and causing illness and death. • Caused by viruses transmitted by *Aedes Aegypti* and *Aedes Albopictus* mosquitoes. • Common symptoms are: (i) high fever, (ii) severe eye pain, mainly behind the eyes, (iii) severe headache, (iv) joint/muscle/bone pain, (v) mild nose or gum bleeding, (vii) rashes, (viii) low WBC count, etc.
2	Malaria	• A mosquito-borne disease caused by parasite *plasmodium* and spread by female *Anopheles* mosquitoes which predominantly bites at night between dusk and dawn. • Common symptoms are: (i) high fever, (ii) sweat and chills, (iii) headache, (iv) vomiting, (v) muscle pain, (vi) diarrhoea, etc.
3	Cholera	• A intestinal-infection disease caused by the *vibrio cholerae* bacterium. • Generally spread through contaminated food or water • Common symptoms are: (i) profuse watery diarrhea, (ii) vomiting, (iii) reduced skin elasticity, (iv) dry mucous membranes, (v) rapid heart rate and low blood pressure, (vi) repeated thirst,restlessness and muscle cramps, etc.
4	Diarrhoea	• A disease caused by different microorganisms like virus, bacteria or parasites, and a second driving reason for death in kids under five years age. • Generally spreads through intake of contaminated food or drinking water. • Common symptoms are: (i) fever, (ii) severe stomach cramps, (iii) nausea and vomiting, (iv) headache, (v) loss of appetite, etc.
5	Influenza	• A disease causing contagious respiratory illness due to influenza viruses, and also known as flu. • Common symptoms are: (i) sudden onset of high fever, (ii) sore throat, (iii) muscle and joint pain, (iv) usually dry cough, (v) headache (vi) runny nose, etc.
6	Meningitis	• A disease due to viral, bacterial, or fungal infections, causing inflammation of membranes. • Common early symptoms are: (i) fever, (ii) nausea and vomiting, (iii) severe headache, (iv) muscles/joints/limbs pain, (v) cold hands and feet, (vi) shivering pale or blotchy skin and blue lip, etc.
7	Leishmaniasis	• A parasitic disease caused by *leishmania* parasites carried by infected sand flies. • Common early symptoms are: (i) skin sores in case of *cutaneous leishmaniasis*, and (ii) affected internal organs like spleen, liver, and bone marrow in case of *visceral leishmaniasis*.
8	Kala-azar	• A *visceral leishmaniasis*, which is known as *kala-azar* in Indian subcontinent. • Can reduces RBC count (causing anemia), WBC count (causing leukopenia), and platelet count (causing thrombocytopenia). • Common symptoms are: (i) irregular bouts of fever, (ii) spleen and liver enlargement, (iii) weight loss, (iv) anaemia, etc.

sible implementation of the text information processing system in other domains.

2. Related Works

Due to complexity of linguistic structures, entity and relation extraction from text requires careful examination of sentence structure. Many approaches including pattern-based, statistical, rule-based, machine learning, and hybrid approaches have been adopted to identify biomedical entities and several statistical measures have been used to distil associations between the entities having their sentence or document-level co-occurrence [1, 2, 20, 3, 6]. Manually generated or automatically learned rules have been applied on phrase structure tree, dependency tree or dependency graphs where entities have been identified using some named Entity Recognition (NER) tools. Kernel-based [35, 11, 31] and feature-based [21] techniques for machine learning have been used by many researchers in diverse domain. Hybrid approach [9] has also been in focus by some researchers. Shallow linguistic processing [16] is very common in information extraction.

Dependencies have been used by limited researches to extract significant information embedded in biomedical texts. In [15], Fundel et al. identified gene-protein relationships by adopting three elementary rules between *effector* and *effectee* proteins applied on dependency parse tree chunks containing an *effector* and an *effectee* entities bounded by fixed set of relation terms. Hassan et al. [17, 18] applied graph-based frequent subgraph mining techniques on texts transformed as dependency graphs to extract relations between annotated disease-symptom entities. Similarly, Bunescu and Mooney [7] proposed “shortest path dependency kernel” to identify relation between two known entities *person* and *facility* by designing a kernel method that employs the shortest path between entities in an undirected dependency graph. Linguistic dependencies though abundant in representing knowledge, have been explored by few researchers, particularly for disease-symptoms extraction and their association identification considering its complexity; hence, requires further heed of the research community. In one of the study based on typed dependencies, Seneviratne and Ranasinghe [33] demonstrated the use of dependency-based rules to extract ontological relationships between birds and their locations in which they showed how to take advantage of typed dependencies generated by a natural language parser. Typed dependencies in terms of their grammatical relation (i.e. labels of typed dependencies) have been used rarely for disease symptoms and relations extraction.

Limited works aiming disease and symptoms recognition indicates great opportunity in this area. However, automatic extraction of disease symptoms and their relations requires much effort to accomplish. In a study which set out to determine disease and symptom co-occurrence relations, Datla et al. [12] applied higher order co-occurrence technique using original Latent Semantic Analysis (LSA) and customized LSA, considering negation. They found original LSA is able to apprehend disease symptom relations and customized LSA performs slightly better. Tran et al. [34] presented a method to facilitate mapping of symptom concepts anatomically associated with organ systems by utilizing the concepts from unified medical language system metathesaurus. They discussed that clinical signs and symptoms can be categorized into semantic types that include *sign or symptom*, *finding*, or *mental or behavioral dysfunction*. Similarly, mental signs and symptoms related concepts can be encapsulated in different semantic categories, such as *mental process*, *individual behavior*, or *social behavior*.

In view of the existing studies mentioned above, the majority of entity and relation extraction approaches identifies predefined associations between entities. Moreover, previous works ignore mining disease symptoms and their associations using dependency grammar comprehensively. Parwez et al. [30], however, attempted to characterize climate-sensitive disease by extracting disease symptoms and their associations with limited dataset. This paper acquaints with a generic approach to extract information components and identify disease symptoms and their associations efficiently from biomedical text documents.

3. Proposed `DiseaSE` System

In this section, we present the functional details of the proposed `DiseaSE` system for disease symptoms extraction and characterization from biomedical documents. Figure 1 depicts the architecture of the proposed system with different modules and flow of information between them during the disease symptom and association extraction process. It consists of different functioning modules – *document crawler*, *document preprocessor*, *dependency processor*, *abbreviation extractor*, *disease symptom miner*, *feasibility analyzer*, and *information visualizer*. Further details about these modules are provided in the following sub-sections.

3.1. Document Crawler

The rationale behind the development of *document crawler* is to fetch `PubMed` documents automatically based on triggered queries and store them into a local repository. `PubMed` database maintains a repository of published life science and biomedical articles and provides access to their abstracts, including authors, affiliations, and other associated information. In order to fetch query-centric `PubMed` abstracts automatically, the crawler is implemented in Java employing axis 2.1.6.2 API[3] rendered by the NCBI Entrez system. The crawler receives queries, requests `PubMed` database which uses “NCBIs Entrez search and retrieval system” [8] by calling NCBI utility server, and the server responds with `PubMed` documents based on the triggered query. The documents fetched by the crawler are stored in a

[3] http://axis.apache.org/axis2/java/core/

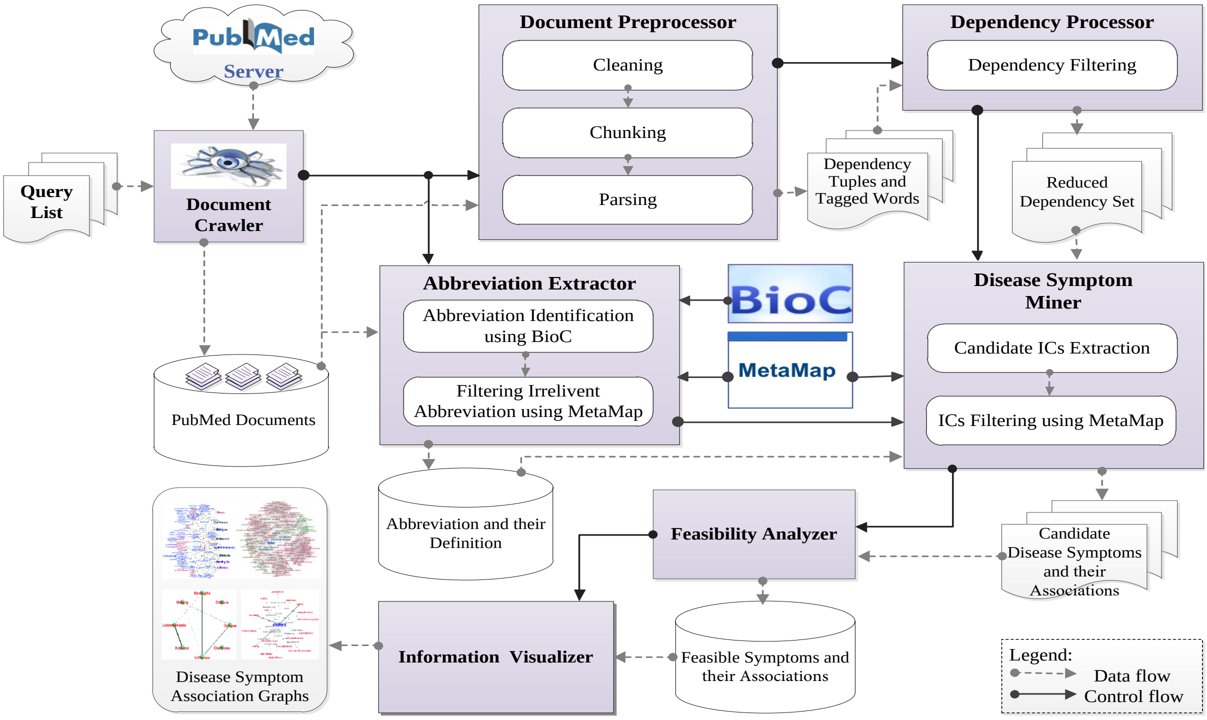


Figure 1: Architecture of the proposed DiseaSE (Disease Symptoms Extraction) system

local repository for additional processing. Eight diseases including *dengue*, *malaria*, *cholera*, *diarrhoea*, *influenza*, *meningitis*, *leishmaniasis*, and *kala-azar* reported in [22] and their 66 symptoms listed at CDC, NHS, and WHO websites are adopted to construct different query patterns. Accordingly, total 528 query patterns consisting of a disease and symptom name joined by the logical "AND" operator are produced and PubMed is queried by employing the NCBIs Entrez system API that lead to the retrieval of total 107302 documents covering their PMID, title and abstracts. PubMed processes queries by mapping terms or phrases internally using MeSH (Medical Subject Headings) translation table and provides appropriate document abstracts and other citation information. The crawler is able to retrieve documents against 317 queries, while unable to fetch any document for the remaining 211 queries due to absence of both query terms occurring together within the document abstracts.

3.2. Document Preprocessor

The *Document Preprocessor* module performs document cleaning, sentence chunking and parsing tasks to prepare documents for extraction of *information components* from which biomedical entities and their associations are to be distilled out. Due to occupancy of two or more disease-symptom query patterns within a document, the crawler retrieved many documents multiple times. To preserve unique documents for further processing, the documents PubMed IDs (a unique ID assigned by PubMed to each document) are employed to exclude multiple copies of the documents. Thereafter, each document is parsed into sentences to produce typed dependencies and Parts-Of-Speech (POS) tags by employing Stanford natural language processing parser, which is a widely used parser to generate POS tags, phrase structure tree, dependency tree, and typed dependencies. The typed dependencies[4] in the form of dependency triplets, denotes pair-wise grammatical relationships of *governor* and *dependent* words along with their respective positions within a sentence [14]. In its collapsed form, typed dependencies denote direct relationship between words of a sentence, which is helpful in extracting entities and their relations based on their syntactic patterns. In Universal Stanford Dependencies, De Marneffe et al. [13] discussed about 42 grammatical dependency relations

[4] https://nlp.stanford.edu/software/dependencies_manual.pdf

to be widely used for NLP applications. Figure 2 delineates a dependency parse tree exhibiting dependency relationships among words and POS tags of an exemplar sentence generated by the Stanford NLP parser by employing a dependency parse visualization tool *DependenSee 3.7.0.*

The Stanford parser produces POS-tagged words as well as dependencies that can be clubbed together to identify nouns, verbs, and other required tags of the *governor* and *dependent* of a dependency tuple.

3.3. Dependency Processor

This module receives the typed dependencies and POS tags of the *governor* and *dependent* words, club together their corresponding POS tags in the dependency tuple, and filters out insignificant tuples to obtain reduced dependency set from which *information components* and subsequently, disease symptoms and their associations are to be gleaned. Although most dependency relation tuples are applicable in *information component* identification, there are some irrelevant dependency relations such as *det*, *dep*, *amod*, *compound* etc. that require to be weeded out as they are either insignificant (e.g. *det*, *dep*) or become valueless (e.g. *amod*, *compound*, *nmod:poss* and *advmod*) in *information component* extraction while concatenating their governor or dependent words to form compound word or phrase. Eliminating these extraneous dependencies gives a minimal set of dependencies that can be processed efficiently to procure *information components*. The *det* relation between a noun and its determiner is a trivial dependency relation, as it does not contribute to *information component* extraction. Likewise, *dep* is an unspecified dependency labeled by the parser when it is unable to determine the exact grammatical dependency between a pair of words. Figure 3 depicts a couple of sample sentences along with their POS tags, typed dependencies, and the reduced dependencies procured after eliminating irrelevant dependency tuples. As discussed earlier, typed dependencies show grammatical relations between governor and dependent words. Sometimes, a *governor* or a *dependent* word alone does not make complete sense of a disease or a symptom concept. To grab complete sense, these words need to be concatenated to form a meaningful compound word.

Consider the dependency relation tuple *amod(disease-2, Whipple-1)* of the exemplar sentences delineated in Figure 3. It is clearly understood that the word *disease* or the word *Whipple* itself appears inadequate to perceive the meaning of the disease, while the composite word *Whipple_disease* obtained by clubbing governor and dependent words of the dependency tuple gives the complete sense of a disease. Similarly, words of the dependency tuples *amod(illness-7, systemic-6)*, *amod(diarrhea-13, chronic-12)* and *compound(loss-16, weight-15)* can be concatenated to form the composite words *systemic_illness*, *chronic_diarrhea*, and *weight_loss*, respectively. To grasp complete symptom or complete disease concepts when the symptom/disease concepts are complex, i.e., represented by more than one word, the words of dependencies like *amod*, *compound*, *nmod:poss* and *advmod* are amalgamated to form a single composite word or phrase. The dependency relation *amod* is an adjectival modifier, which modifies any noun or noun phrase, and the dependency relation *compound* represents compound nouns when a noun modifies a head noun. Similarly, *nmod:poss* and *advmod* are possessive nominal modifier and adverbial modifier, respectively that modify meaning of the head word.

In some cases, a single composite word is formed by combining two consecutive words of *amod* only or two consecutive words of *compound* only, as shown in the previous paragraph; whereas in other cases, the words of two successive *amod* dependency relations or two successive *compound* or one *amod* and one *compound* dependency relations are combined to form a composite word, keeping words arranged by their position in the sentence. In second case, the composite word consists of three consecutive words of the sentence. For example, the phrase *rare_systemic_illness* is formed by combining governor and dependent words of two consecutive *amod* dependency relations *amod(illness-7, rare-5)* and *amod(illness-7, systemic-6)*, as shown in Figure 3. The composite word so obtained is substituted as a single word in all other dependencies where any of the constituents of the composite word is present, keeping tag of the

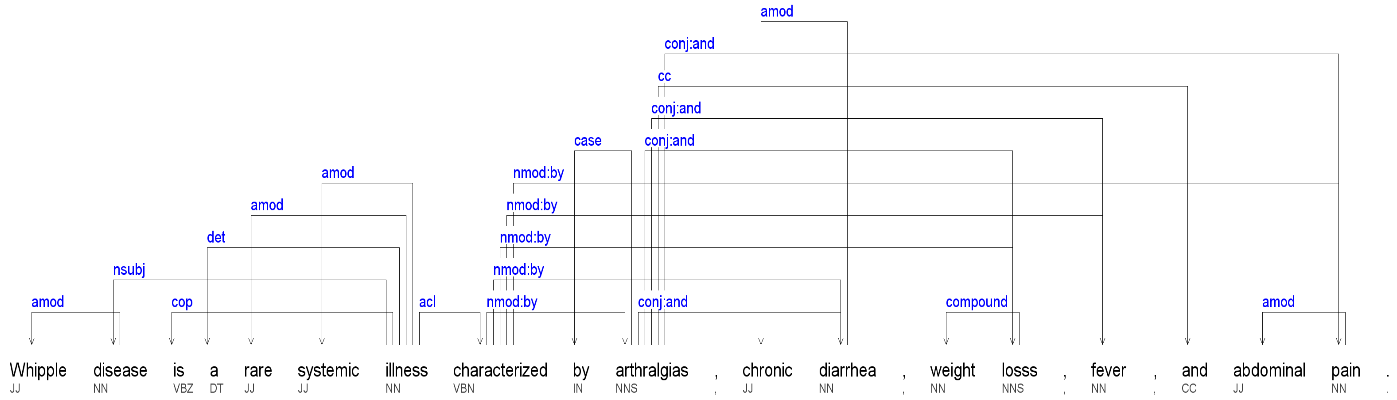


Figure 2: A sample dependency parse tree constructed by the Stanford NLP parser using the dependency visualization tool *DependenSee 3.7.0*

composite word as the tag of the head word of the dependency.

After composing single word and substituting it within the dependency tuples, the number of parser generated dependencies of a sentence is reduced by removing dependency tuples like *det*, *dep*, *amod*, *compound* etc., keeping the order of the remaining dependencies as they appear in the original dependencies produced by the parser. Removal of such extraneous dependency relation tuples reduces overall number of tuples to be processed for distillation of appropriate *information components*. The residual dependencies thus obtained from reduced dependency set with each dependency tuple consisting dependency relation, governor and dependent word (or composite word) followed by their POS tags and word positions.

3.4. Abbreviation Extractor

The motive behind abbreviation extraction is to identify the abbreviations embedded within texts that represent disease or symptom concepts, which could otherwise remain untraceable or could be wrongly classified by `MetaMap` into any semantic categories other than the categories we have considered for disease or symptom identification. As abbreviations and acronyms concerning biomedical terminologies are extensively used in biomedical literatures that may represent any disease or symptom concepts, mapping them to their intended definition introduced within the document would be helpful to identify disease symptom and their associations accurately. Many of these abbreviations are ambiguous and their actual meaning can only be judged by using their full forms or definitions.

To catalog these abbreviations, tools like *abbr* provided by *BioC*[5], and `MetaMap` are used. *BioC* is a framework for data sharing and annotations in biomedical text processing. The Java implementation of *BioC* provides an *abbr* tool in which *ExtractAbbr* class implements a simple algorithm to extract abbreviations and their corresponding definitions from biomedical texts [32, 10]. We have identified total 11637 unique abbreviations and their corresponding definitions using *BioC*. Since abbreviations and their corresponding definitions identified by *BioC* may not necessarily represent disease or symptom concepts, irrelevant abbreviations are filtered out with the help of `MetaMap`. Out of total 11637 abbreviations, `MetaMap` correctly identified 737 abbreviations as any of the nine semantic categories of our interest, while remaining are either labeled as other semantic categories or completely discarded. On close examination, it is found that the ignored or other semantic categories of abbreviation have many disease-symptom concepts hidden in their definitions. Because of these untapped or other semantic category abbreviations, the system may ignore many prospective triplets representing disease symptoms and their associations. Therefore, in order to capture such abbreviations as disease or symptom, their definitions are subjected to `MetaMap`. Thereafter, a dictionary of the untapped abbreviations is built by identifying them through their definition, resulting in 4060 entries of abbreviation-definition pairs. The abbreviations' definitions are later substituted within the extracted *information components*

[5] http://bioc.sourceforge.net/

Sample sentences:

"Whipple disease is a rare systemic illness characterized by arthralgias, chronic diarrhea, weight loss, fever, and abdominal pain. The disorder generally affects middle-aged men."

POS-tagged sentences:

Whipple/JJ, disease/NN, is/VBZ, a/DT, rare/JJ, systemic/JJ, illness/NN, characterized/VBN, by/IN, arthralgias/NNS, ,/,, chronic/JJ, diarrhea/NN, ,/,, weight/NN, loss/NN, ,/,, fever/NN, ,/,, and/CC, abdominal/JJ, pain/NN, ./.

The/DT, disorder/NN, generally/RB, affects/VBZ, middle-aged/JJ, men/NNS, ./.

Typed dependencies:

amod(disease-2, Whipple-1), nsubj(illness-7, disease-2), cop(illness-7, is-3), det(illness-7, a-4), amod(illness-7, rare-5), amod(illness-7, systemic-6), root(ROOT-0, illness-7), acl(illness-7, characterized-8), case(arthralgias-10, by-9), nmod:by(characterized-8, arthralgias-10), amod(diarrhea-13, chronic-12), nmod:by(characterized-8, diarrhea-13), conj:and(arthralgias-10, diarrhea-13), compound(loss-16, weight-15), nmod:by(characterized-8, loss-16), conj:and(arthralgias-10, loss-16), nmod:by(characterized-8, fever-18), conj:and(arthralgias-10, fever-18), cc(arthralgias-10, and-20), amod(pain-22, abdominal-21), nmod:by(characterized-8, pain-22), conj:and(arthralgias-10, pain-22)

det(disorder-2, The-1), nsubj(affects-4, disorder-2), advmod(affects-4, generally-3), root(ROOT-0, affects-4), amod(men-6, middle-aged-5), dobj(affects-4, men-6)

Reduced dependencies with embedded tags:

nsubj(rare_systemic_illness / NN 7, Whipple_disease / NN 2)
cop(rare_systemic_illness / NN 7, is / VBZ 3)
acl(rare_systemic_illness / NN 7, characterized / VBN 8)
case(arthralgias / NNS 10, by / IN 9)
nmod:by(characterized / VBN 8, arthralgias / NNS 10)
nmod:by(characterized / VBN 8, chronic_diarrhea / NN 13)
conj:and(arthralgias / NNS 10, chronic_diarrhea / NN 13)
nmod:by(characterized / VBN 8, weight_loss / NN 16)
conj:and(arthralgias / NNS 10, weight_loss / NN 16)
nmod:by(characterized / VBN 8, fever / NN 18)
conj:and(arthralgias / NNS 10, fever / NN 18)
cc(arthralgias / NNS 10, and / CC 20)
nmod:by(characterized / VBN 8, abdominal_pain / NN 22)
conj:and(arthralgias / NNS 10, abdominal_pain / NN 22)

nsubj(generally_affects / VBZ 4, disorder / NN 2)
advmod(generally_affects / VBZ 4, generally_affects / VBZ 4)
dobj(generally_affects / VBZ 4, middle-aged_men / NNS 6)

Figure 3: Sample sentences with POS tags, typed dependencies, and reduced typed dependencies

Table 2: A partial list of abbreviations overlooked by the `MetaMap`

Abbreviation	Definition (Full-form)
BBE	Bickerstaffs Brainstem Encephalitis
DF	Dengue Fever
DF/DHF	Dengue Fever/Dengue Haemorrhagic Fever
HDF	Haemorrhagic Dengue Fever
SDH	Subdural Haematoma, Subdural Hemorrhage
JSF	Japanese Spotted Fever
SEA	Spinal Epidural Abscesses
BDI	Biliary Duct Infections
WNF	West Nile Fever
ABS	Acute Brain Swelling

to correctly identify disease symptom concepts hidden within the abbreviation. A partial list of abbreviations along with their definitions overlooked by the `MetaMap` is shown in Table 2.

3.5. Disease Symptoms Miner

This module takes reduced dependency set along with the list of abbreviations and definitions as input and aims to identify meaningful *information components* (ICs), followed by the extraction of disease symptoms and their associations. Accordingly, it performs two major tasks – IC extraction and IC filtering using `MetaMap`. Further details about these tasks are presented in the following sub-sections

3.5.1. Information Component Extraction

Extraction of disease symptoms and their associations comprises the distillation of *information components* from the reduced dependency set using dependency-based syntactic patterns after analyzing a list of different English sentence structures. Formally, an *information component* is defined as follows.

Definition 1. (*information component*). An *information component* is a semantic triple of the form <$entity_i$, *relation*, $entity_j$>, where $entity_i$ and $entity_j$ are the words/phrases representing disease or symptom concepts, and the *relation* is a relational word representing the relationship between the entities. The relational word may be a verb, or verb with preposition, or even sometimes noun with preposition.

It is observed that not only *nsubj*, *nsubjpass* or *dobj* has entities and relational verbs that contribute to *information component* extraction but other typed dependencies like *acl*, *appos*, and *nmod* also have governor and dependent words contributing to *information component* extraction. At this stage, the system can extract all possible candidate ICs without imposing any named entity restriction. Table 4 presents the identified ICs from exemplar sentences of Figure 3. The ICs extraction algorithm is implemented in Java. A complete implementation details along with the codes can be found at `GitHub`[6].

Since many ICs may have abbreviated terms as a part of $entity_i$, possibly representing any disease or symptom, the abbreviations are replaced with their definitions using the compiled list of abbreviation-definition pairs. Since some ICs may contain entities not necessarily representing disease or symptom concepts, they should be restrained from further processing for efficiency purpose. In this regard, we applied ICs filtering to remove such ICs using *MetaMap*. The following sub-section provides details about the ICs filtering process.

3.5.2. IC Filtering using MetaMap

Biological entities alluded within a sentence, when extracted as ICs, need to be recognized correctly to acquire meaningful disease symptoms and their associations. Since the ultimate aim is to extract disease symptoms and their associations, `MetaMap`[7] is used to identify and annotate entities representing disease or symptom concepts. `MetaMap` identifies Unified Medical Language System (*UMLS*) concepts referred in biomedical texts and maps them into any of the pre-defined 133 semantic categories[8]. The semantic types broadly represent entities or events like physical objects, organisms, anatomical structures, cell components, nucleotides, idea or concepts, chemicals, activities, behaviors, findings, processes or phenomenon, biological functions, pathological functions, injuries or poisoning etc. In this study, we have considered only nine semantic categories representing either disease or symptom concepts, as shown in Table 3. Out of these semantic categories, *dsyn*, *neop*, and *anab* represent disease concepts, whereas *sosy*, *fndg*, *patf*, *cgab*, *mobd*, and *inpo* represent symptom concepts.

Table 3: `MetaMap` Semantic categories representing disease or symptom concepts

S.No.	Semantic Category	Description
1	dsyn	Disease or Syndrome
2	neop	Neoplastic Process
3	anab	Anatomical Abnormality
4	sosy	Sign or Symptom
5	fndg	Finding
6	patf	Pathologic Function
7	mobd	Mental or Behavioral Dysfunction
8	cgab	Congenital Abnormality
9	inpo	Injury or Poisoning

Note. Adapted from *"Biomedical Text Analytics for Characterizing Climate-Sensitive Disease"* by M. A. Parwez et al., (2018).

Table 4: *Information components* extracted from sample sentences of Figure 3

First Entity	Relation	Second Entity
Whipple disease	is	rare systemic illness
Whipple disease	characterized by	arthralgias
Whipple disease	characterized by	chronic diarrhea
Whipple disease	characterized by	weight loss
Whipple disease	characterized by	fever
Whipple disease	characterized by	abdominal pain
disorder	generally_affects	middle-aged_men

It should be noted that the $entity_i$ or $entity_j$ of an IC, when annotated by `MetaMap`, may represent spurious disease or symptom. Therefore, such ICs need to be eliminated to have genuine disease or symptom. As a result, the ICs whose $entity_i$ or $entity_j$ comprises only terms like *disease*, *symptom*, *infection*, *ill*, *complication* etc., which alone does not represent any disease or symptom, are filtered out. Table 5 shows the list of ICs retained after filtering the last *information component* of Table 4, as its left or right entity does not contain any disease or symptom concept. It can be observed from Table 5 that both taxonomic (e.g. *is*) and non-taxonomic (e.g. *characterized by*) relations exist between entities. Moreover, a particular relational word binds a disease with multiple symptoms. The constituents of retained ICs are considered as candidate symptoms and associations that are subjected for feasibility analysis as discussed in the following sub-section.

3.6. Feasibility Analyzer

This module performs feasibility analysis over the list of candidate symptoms to identify significant symptoms associated with each disease under consideration. For feasibility analysis,

[6] https://github.com/maparwez/InformationComponentExtractor

[7] https://metamap.nlm.nih.gov/

[8] http://metamap.nlm.nih.gov/Docs/SemanticTypes_2013AA.txt

Table 5: Retained ICs after filtering irrelevant ones from Table 4 using `MetaMap`

First Entity	**Relation**	**Second Entity**
Whipple disease	is	rare systemic illness
Whipple disease	characterized by	arthralgias
Whipple disease	characterized by	chronic diarrhea
Whipple disease	characterized by	weight loss
Whipple disease	characterized by	fever
Whipple disease	characterized by	abdominal pain

we have applied four different ranking approaches – i) frequency count, ii) global *tf-idf*, iii) local *tf-idf*, and iv) `TextRank`; henceforth referred to as *RM1*, *RM2*, *RM3*, and *RM4*, respectively. Additionally, we have considered a hybrid approach, henceforth referred to as *HRM*, in which symptoms mutually shared by these four ranking approaches are regarded as feasible. These ranking approaches are briefly explained in the following paragraphs.

Frequency Count: The first ranking approach (RM1) merely inspects frequency count of the symptom or disease terms based on their occurrence in the list of retained ICs for a particular disease. Given the document collection D, the frequency count of a symptom term s_i can be determined using equation 1, where $Count(s_i, d_j)$ denotes count of s in j^{th} document $d_j \in D$, and $|D|$ represents number of document in D.

$$fCount(s_i) = \sum_{j=1}^{|D|} Count(s_i, d_j) \tag{1}$$

TF-IDF-based ranking: The *tf-idf* [23] is a powerful NLP technique in information retrieval and text mining for weighting or ranking terms appearing more frequently in a document, but rarely in the entire document collection. Using this concept of ranking to determine the prominence of extracted symptoms for individual diseases, two approaches (*global tf-idf* and *local tf-idf*) have been defined based on standard document-level tf-idf. The *global tf-idf* method (RM2) applies equations 2, 3, and 4 to compute *tf-idf* score of symptom s_i in a document collection D, where $|D|$ represents number of documents in D, and $|D_{s_i}|$ represents number of documents that contain symptom s_i.

$$tf\text{-}idf_g(s_i) = tf(s_i) \times idf(s_i) \tag{2}$$

$$tf(s_i) = \sum_{j=1}^{|D|} fCount(s_i, d_j) \tag{3}$$

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

Here, $tf\text{-}idf_g(s_i)$ represents global *tf-idf* weight of symptom s_i which is basically the product of $tf(s_i)$ and $idf(s_i)$ calculated using equations 3 and 4, respectively. In these equations, $tf(s_i)$ denotes the global term frequency of s_i, and $idf(s_i)$ corresponds to the inverse document frequency of s_i.

In the *local tf-idf* ranking method (RM3), weight of each term is determined at document-level and summed to a single score by employing equations 5, 6 and 7. The local *tf-idf* socre denoted as $tf\text{-}idf_l(s_i)$ is basically the sum of *tf-idf* weights of s_i at document-level, as shown in equation 5 in which $tf\text{-}idf^{d_j}(s_i)$ represents the *tf-idf* weight of s_i in document d_j computed using equation 6. In this case, the term frequency of s_i in d_j is determined by adopting equation 7, and the inverse document frequency is calculated by employing equation 4.

$$tf\text{-}idf_l(s_i) = \sum_{j=1}^{|D|} tf\text{-}idf^{d_j}(s_i) \tag{5}$$

$$tf\text{-}idf^{d_j}(s_i) = tf^{d_j}(s_i) \times idf(s_i) \tag{6}$$

$$tf^{d_j}(s_i) = \frac{fCount(s_i, d_j)}{\sum_{k=1}^{|S|} fCount(s_k, d_j)} \tag{7}$$

TextRank: In another ranking method (RM4), we have used `TextRank` [24] algorithm, which is a graph-based approach to rank extracted symptoms of the diseases. We model the extracted symptoms as an undirected graph $G = (V, E)$, where V denotes the set of vertices representing symptoms and $E \subseteq V \times V$ denotes the set of edges representing associations between the vertices based on their *Cosine* similarity. The *Cosine* similarity between two vertices is based on the normalized term frequency vector of the symptoms in the disease-symptom triples extracted from the document collection D. Equation 8 presents the *Cosine* similarity as an weight w_{ij} of an edge between vertices s_i and s_j, where $tf_{norm}(s_i)$ and $tf_{norm}(s_j)$ represent normalized term frequency of s_i and s_j, respectively, and $|s_i|$ and $|s_j|$ represent magnitude of vectors s_i and s_j, respectively.

$$w_{ij} = CosineSim(s_i, s_j) = \frac{tf_{norm}(s_i) \cdot tf_{norm}(s_j)}{|s_i||s_j|} \tag{8}$$

Thereafter, *TextRank* algorithm is applied on graph G, which iteratively computes the weighted score of each vertex s_i using equation 9, where $WS(s_i)$ represents weighted score of s_i, w_{ji} is the weight of an edge between vertices s_j and s_i, $adj(s_i)$ denotes the set of vertices adjacent to s_i, and $d \in [0, 1]$ is the damping factor that incorporates the probability of leaping from one vertex to another vertex randomly into the computation.

$$WS(s_i) = (1 - d) + d * \sum_{s_j \in adj(s_i)} \frac{w_{ji}}{\sum_{s_k \in adj(s_j)} w_{jk}} WS(s_j) \tag{9}$$

We used damping factor $d = 0.85$ as used in [24], and iterated the score calculation equation till it converged, i.e., when the difference between the scores at two successive iterations reached to the threshold value of 0.0001, as suggested in [24]. Once the algorithm converges, the final score associated with each vertex represent its importance, and hence used to rank the vertices (symptom).

3.7. Information Visualizer

This section aims to identify associations between disease and symptoms at different levels of granularity and visualize them using different types of graph structures. For disease-symptom visualization, we have used a star-like graph structure in which disease node is placed in centre and symptom nodes are placed at periphery and connected with the disease node using labelled edges. The label of an edge connecting a disease with its symptom represents the structural or non-structural association between them. In order to derive symptom-symptom and disease-disease associations, we have generated a bipartite graph in which list of diseases constitutes one set of nodes and list of symptoms constitutes another set of nodes. Edges are drawn between the elements of these sets. Thereafter, projection is applied over the bipartite graph to generate symptoms-symptoms and disease-disease association graph.

For generating symptom-symptom association graph, symptoms are considered as nodes and edge between a node-pair is generated using Jaccard similarity [29], which is calculated as the ratio of the intersection of two sets to their union, as defined in Equation 10 in which *diseaseSet*(s_i) represents the set of diseases having symptom s_i.

Similarly, for generating disease-disease association graph, diseases are considered as nodes and edge between a disease-pair is generated using Jaccard similarity defined in Equation 11 in which *sympSet*(d_i) represents the set of symptoms associated with disease d_i.

$$JaccardSim(s_i, s_j) = \frac{|diseaseSet(s_i) \bigcap diseaseSet(s_j)|}{|diseaseSet(s_i) \bigcup diseaseSet(s_j)|} \tag{10}$$

$$JaccardSim(d_i, d_j) = \frac{|sympSet(d_i) \bigcap sympSet(d_j)|}{|sympSet(d_i) \bigcup sympSet(d_i)|} \tag{11}$$

The visualization procedure has been explained through figures and has been elaborated more clearly in section 6. All graphs are drawn using a prominent open-source data analysis, and graphs and networks visualization tool *Gephi*[9]*0.8.2*, which helps to draw graphs with different layouts.

4. Experimental Setup and Results

In this section, we present our experimental setup and results to establish the efficacy of the proposed biomedical text processing system `DiseaSE`. Starting with a brief introduction of the experimental dataset in sub-section 4.1, we present the evaluation of *information components* extraction and disease symptoms identification processes in sub-sections 4.2 and 4.3, respectively.

4.1. Dataset

For experimental purpose, we generated a dataset containing 107302 biomedical documents crawled from `PubMed` database

[9] https://gephi.org

using a set of queries based on all eight disease names and their standard symptoms, as described earlier in this paper. The crawled dataset had multiple copies of many documents because of the occurrence of two or more diseases-symptom query patterns within a document and consequently retrieved by multiple queries. To remove multiple copies and keep only unique documents in the dataset, we used documents' unique `PubMed` IDs assigned by `PubMed`. Consequently, a total 67516 out of 107302 fetched documents remained in the dataset for further processing. Thereafter, each document was parsed into sentences, and the sentences containing at least one disease or symptom entity were retained for efficiency purpose. A complete statistics of the dataset is presented in Table 6.

Table 6: Statistics of the experimental dataset

Parameters	Values
Total no. of documents fetched	107,302
No. of unique documents	67,516
Total no. of sentences	6,43,173
No. of sentences containing disease and/or symptoms	3,26,308
No. of sentences without any disease or symptom	3,16,865

4.2. Evaluation of the Information Components Extraction Process

In this section, we present the evaluation results of the *information components* (ICs) extraction process. Due to unavailability of any marked corpora, we have physically marked reasonable *information components* embodied in the sentences of the dataset by taking assistance of domain experts. As manual marking of the entire dataset is impractical, we produced three different test datasets, namely *TD1*, *TD2*, and *TD3* of varying-sizes consisting 100, 300, and 1500 sentences, respectively from the main dataset by applying the principle of random sampling with replacement. To evaluate performance of ICs extraction process, we have employed standard Information Retrieval (IR) metrics – *Precision*, *Recall*, and *F1-score* defined using equations 12, 13, and 14, respectively with respect to *True Positives* (TP), *False Positives* (FP), and *False Negatives* (FN) outcomes, where *TP* represents number of positive examples identified as positive, *FP* represents number of negative examples identified as positive, and *FN* represents number of positive examples identified as negative.

Table 7: Performance evaluation results on different varying size test datasets

Test datasets	#Actual ICs	TP	FP	P	R	F1
TD1	92	52	10	83.87	56.52	67.53
TD2	219	99	30	76.74	45.20	56.90
TD3	1194	446	153	74.46	37.35	49.75

$$Precision\ (P) = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{12}$$

$$Recall\ (R) = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{13}$$

$$F1\text{-}score\ (F1) = \frac{2 \times P \times R}{P + R} \tag{14}$$

Table 7 shows the assessment results of the proposed ICs extraction process on all three test datasets. It can be observed from this table that *F1-score* value is falling off with growing size of test datasets, which is primarily because of increment in false negatives (causing low recall) with the increasing number of sentences. Although, precision remains consistent throughout the datasets. On close examination, we found many factors contributing to low recall values, which include structure and complexity of sentence possessing some indirect relations and the limitation of existing NLP tools. Nevertheless, the uniqueness of the proposed `DiseaSE` system rests on consolidation of named entities and typed-dependency-based syntactic patterns to identify and extract disease symptoms and their associations from unstructured text documents.

4.3. Evaluation of the Disease Symptoms Identification Process

In order to evaluate the accuracy of the identified disease symptoms, a feasibility analysis is performed by compiling a list of disease-wise symptom concepts from the ICs and finding the common symptoms in the list of top-k symptoms identified by different ranking methods – *RM1*, *RM2*, *RM3*, and *RM4* that are discussed in section 3.6. In order to determine an optimal value of k, we have analyzed the percentage of common symptoms

Table 8: Percentage of common symptoms in the lists of top-k symptoms identified by *RM1*, *RM2*, *RM3* and *RM4*

Disease Name	Percentage of common symptoms to all ranking methods in top-k symptoms		
	k=10	k=20	k=30
Cholera	0.50	0.45	0.50
Dengue	0.30	0.45	0.43
Diarrhoea	0.60	0.55	0.50
Influenza	0.50	0.45	0.50
Kala-azar	0.40	0.65	1.00
Leishmaniasis	0.20	0.40	0.43
Malaria	0.40	0.40	0.43
Meningitis	0.60	0.35	0.30
Average	**0.44**	**0.46**	**0.51**

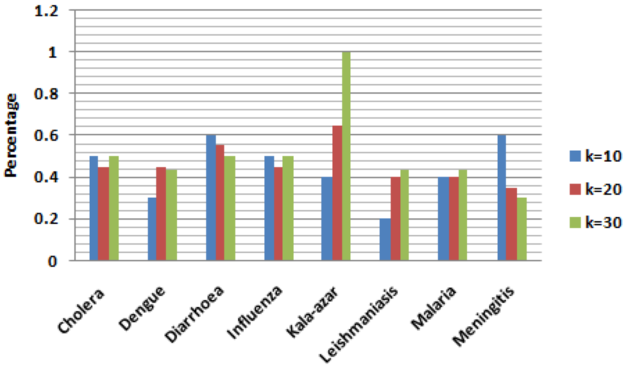


Figure 4: Percentage of common symptoms in the list of top-k symptoms identified by $RM1$, $RM2$, $RM3$, and $RM4$

in the lists of top-k symptoms identified by different ranking methods, as shown in table 8 and visualized in figure 4. It can be observed from this table that the average percentage across all disease is highest for $k = 30$. Therefore, extracted symptoms for each disease are ranked using *RM1*, *RM2*, *RM3*, and *RM4*, and the symptoms that are common to top-30 symptoms identified by each method are considered as more general and commonly identified feasible symptoms.

Table 9 presents the list of identified symptom/disease concepts for each disease, in which we have isolated the list in two parts – one consisting of the extracted symptom/disease concepts that are already recorded on websites like CDC, WHO, or NHS, and the other consisting of newly identified symptom/disease concepts. It can be observed from this table that many identified symptom concepts for almost all diseases are not listed by the standard disease related websites like CDC, WHO, and NHS.

To establish the accuracy of the identified symptom/disease concepts, we have considered the opinions of three domain experts independently who are medical professionals at two different institutions. They were given the list of disease and identified symptoms and requested to mark them as relevant or non-relevant independently. For a symptom with different opinion, majority voting was used to resolve the conflict and determine the relevance of the symptom to the respective disease. Based on experts' opinion, *Detection Rate* (DR) is defined using equation 15, where $\mathcal{S}_{\mathcal{R}}$ and $\mathcal{S}_{\mathcal{N}}$ is the number of symptoms marked as relevant and non-relevant, respectively by the experts.

$$DR = \frac{\mathcal{S}_{\mathcal{R}}}{\mathcal{S}_{\mathcal{R}} + \mathcal{S}_{\mathcal{N}}} \tag{15}$$

Table 10 presents the *DR* values for all eight diseases, and figure 5 presents its visualization using the bar-chart. It can be observed from tables 10 and 9 that three symptoms of *cholera* viz. *lesion*, *acute septicemia pneumonia*, and *contagious disease* are marked as non-relevant by the experts. After searching the Web for these terms, we found that "lesion" is an injury or abnormal damage in the tissues like skin or other organs and commonly observed in *fowl cholera* as vascular injuries; "septicemia" is a bloodstream infection and it is associated with *fowl cholera*, which is also a "contagious disease". All these terms are related to *fowl cholera*, a disease commonly found in poultry like chickens, turkeys, ducks etc., though, the experts have considered them as unrelated to *cholera*, may be due to unperceived in human. These terms are extracted by our proposed approach mainly due to the fact that we have not separated the documents whether they are related to human disease or animal disease. In case of *influenza*, concepts like "coma", "lesion", "neurological complication" and "seizures" are marked as non-relevant by all three experts, whereas "asthma" is marked as relevant by one of the experts, which may be due to the fact that sometimes asthma is a complication of *influenza*. "Haemorrhagic shock", "rare disease", "myocarditis", and "viral load" are considered as non-relevant to *kala-azar*. On analysis, we found that "haemorrhagic shock" is the condition of severe blood loss,"rare disease" is an uncommon illness whose prevalence is rare and they are mostly genetic, "myocarditis" causes inflammation and damages

Table 9: Feasible symptom/disease concepts identified by the proposed approach

Disease name	**Identified symptom/disease concepts**	
	Extracted symptom/disease concepts already listed at CDC, WHO, and NHS websites	**New symptom/disease concepts**
Cholera	bacterial infection, dehydration, watery stools, diarrhea, diarrheal stool, intestinal infection, vomiting, watery diarrhea	*acute septicemia pneumonia*, communicable disease, antimicrobial susceptibility, infectious disease, gastroenteritis, *lesion*, *contagious disease*
Dengue	bleeding complication, fever, rash, pain, hemorrhagic shock syndrome, plasma leakage, thrombocytopenia, shock, viral disease	encephalitis, encephalopathy, febrile illness, severe disorder
Diarrhoea	abdominal pain, dehydration, diarrheal illness, fever, gastroenteritis, gastrointestinal complaints, vomiting, watery diarrhea, weight loss,	clostridium difficile diarrhea, colitis, malabsorption, malnourished, pain, immune defect
Influenza	fever, flu, pneumonia, viral infection, meningitis, pandemic h1n1, asthma, encephalitis	encephalopathy, fatigue, *coma*, febrile illness, *lesions*, *neurologic complication*, *seizures*
Kala-azar	enlarged liver, fever, hepatosplenomegaly, leishmaniasis, lesions, pallor, splenomegaly, parasitic disease, parasitic infection, visceral leishmaniasis	acquired immuno deficiency syndrome, chronic disease, chronic infection, dermatosis, endemic disease, *haemorrhagic shock*, lesions skin, life threatening, mucosal lesions, *myocarditis*, pancytopenia, post kala-azar dermal leishmaniasis, *rare disease*, skin lesion, systemic infectious disease, unexplained fever, *viral load*
Leishmaniasis	spleen enlargement, fever, hepatosplenomegaly, kala-azar, lesion, prolonged fever	acquired immuno deficiency syndrome, *autoimmune disorder*, chronic disease, hiv infection, pancytopenia, ulcer, vector borne disease
Malaria	anaemia, fever, low birth weight, parasitic disease, parasitemia, plasmodium vivax infection ,plasmodium falciparum infection	chemoprophylaxis, febrile patient, renal dysfunction, seizure, splenomegaly, thrombocytopenia
Meningitis	fever, headache, hearing damage, seizure, cerebrospinal fluid leakage	febrile sepsis, sequelae, tuberculosis

Table 10: Detection Rate (DR) values to identify feasible symptoms

Disease Name	$\mathcal{S}_{\mathcal{R}}$	$\mathcal{S}_{\mathcal{N}}$	**DR**
Cholera	12	3	0.80
Dengue	13	0	1.0
Diarrhoea	15	0	1.0
Influenza	10	5	0.67
Kala-azar	23	4	0.85
Leishmaniasis	12	1	0.92
Malaria	13	0	1.0
Meningitis	9	0	1.0
Macro average			**0.905**

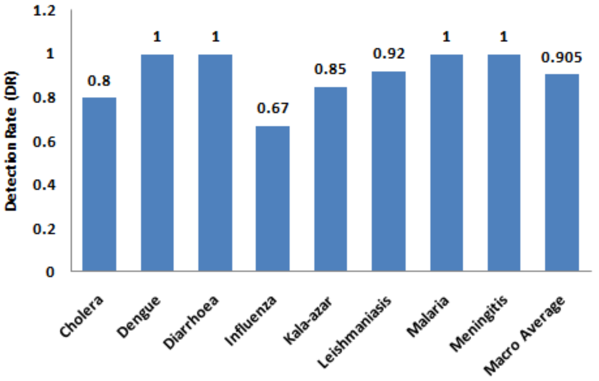


Figure 5: Visualization of Detection Rate (DR) values for feasible symptoms identification

heart muscles, and “viral load” is a type of test used to measure the amount of virus in blood, especially in the case of *HIV*. Likewise, “autoimmune disorder” is marked as non-relevant to *leishmaniasis*. Despite some limitations like extraction of indirectly related or unrelated symptoms, the high detection rate of our proposed approach suggests its applicability to identify meaningful symptoms/biological concepts in biomedical texts for disease characterization and development of enriched biomedical knowledge repository.

4.4. Cross-Validation of the Identified Symptoms with PubMed Data

In this section, we present a validation of the newly identified and experts validated relevant/non-relevant disease symptoms with respect to the underlying PubMed dataset. To this end, we considered one relevant and one non-relevant symptom for each disease (wherever available) and randomly sampled 10 sentences from PubMed dataset that contain disease-symptom pairs. For each disease-symptom pair, we analyzed the retrieved sentences manually to check whether they encode valid disease-symptom association or not. Table 11 presents the evaluation results for all disease and relevant symptom pairs. It can be observed from this table that most of the sampled sentences encode valid disease-symptom associations, except few cases. For example, consider a sampled sentence *“Cholera and enterotoxigenic Escherichia coli (ETEC) are among the most common causes of acute infantile gastroenteritis globally* [PubMed ID: 20421480]” against the *<cholera, gastroenteritis>* disease-symptom pair. This sentence explicitly mentions a valid association between *cholera* and *gas-*

Table 11: Cross-validation results of the identified relevant symptoms with PubMed data

Disease-Symptom pair	TP	FP	Precision
<cholera, gastroenteritis>	7	3	0.7
<dengue, encephalopathy>	10	0	1.0
<diarrhoea, malabsorption>	8	2	0.8
<influenza, fatigue>	10	0	1.0
<kala-azar, acquired immuno deficiency syndrome (aids)>	9	1	0.9
<leishmaniasis, pancytopenia>	9	1	0.9
<malaria, splenomegaly>	10	0	1.0
<meningitis, sequelae>	10	0	1.0

Table 12: Cross-validation results of the identified non-relevant symptoms with PubMed data

Disease-Symptom pair	TP	FP	Precision
<cholera, lesion>	9	1	0.9
<influenza, seizure>	5	5	0.5
<kala-azar, rare disease>	5	0	1.0
<leishmaniasis, autoimmune disorder>	3	2	0.6

troenteritis because *cholera* is specified as most common cause of acute infantile *gastroenteritis*. We considered such sentences as *true positives* (TP). In contrary, consider a sampled sentence *"The levels of resistance among various enteric pathogens are described, and the efficacy and safety of ciprofloxacin in treating infections such as shigellosis, cholera and Escherichia coli gastroenteritis are discussed* [PubMed ID: 9002127]". Though both disease (cholera) and symptom (gastroenteritis) terms are present in this sentence, it does not encode any valid association between them. Rather, this sentence discusses about the efficacy of the *ciprofloxacin* to treat *cholera* and *gastroenteritis*. We considered such sentences as *false positives* (FP).

We also cross-validated the symptoms that were marked by the experts as non-relevant. We followed a similar process of sampling 10 sentences for each disease and non-relevant symptom pairs. We found, only four diseases for which some of the symptoms were marked as non-relevant by the experts. Therefore, we considered four disease and non-relevant symptom pairs as shown in table 12. On analysis, we found that in some of the cases though the experts have marked the retrieved symptoms as non-relevant, the PubMed sentences encode a valid association between the disease and symptom terms. However such symptoms are reported as either a rare or as a rare complication of the specified disease or associated vaccination, convincing majority of the experts to mark them as non-relevant symptom. For example, consider the pair <*cholera*, *lesion*>. On analyzing its sampled sentences, we found that though many sentences encode a valid association between cholera and lesion, the disease *cholera* mentioned in these sentences represent an animal disease – *hog cholera* or *fowl cholera*. For example, the sentence *"Typical hog cholera lesions were observed in 2 pigs only; the other animal showed very few pathological changes* [PubMed ID: 1387299]" specifies the presence of *lesion* in case of *hog cholera*. The experts who marked *lesion* as non-relevant symptom for *cholera* may have presumed that the specified disease belongs to human cholera disease, which may not have any form of *lesion* in a person suffering from *cholera*. Hence, they have marked *lesion* as non-relevant symptom of *cholera*. Therefore, we considered such sentences as *true positives* (TP). In case of <influenza, seizure> pair, we found five sentences in which seizure is mentioned as one of the neurological complications mostly arising in the case of influenza-A and sometimes in the case of influenza-B as well. However, in remaining five sampled sentences we observed that though both terms are present but used in different context and appears unrelated. For example, consider the sentence *"Acute encephalitis, encephalopathy, and seizures are known rare neurologic sequelae of respiratory tract infection with seasonal influenza A and B virus, but the neurological complications of the pandemic 2009 swine influenza A (H1N1) virus, particularly in adults, are ill-defined* [PubMed ID: 21742505]". In this sentence, *seizure*, especially *febrile seizure*, is encoded as a neurological complication associated with influenza. However, experts marked *seizure* as non-relevant which may be due to the fact that it is an uncommon symptom and one of the rare complications in the case of *influenza*. In case of *kala-azar* and *leishmaniasis* the number of sampled sentences containing both disease and symptom terms is only five. In case of <kala-azar, rare disease>, all five sentences clearly supports the expert's decision, whereas in case of <leishmaniasis, autoimmune disorder>, only three sentences support experts' decision.

5. Comparative Analysis

To understand the efficacy of the proposed system, we compared the propsoed approach with one of the popular topic modeling approaches. To this end, we used Latent Dirichlet Allocation (LDA) [5] to induce topics using document-word frequency from the text corpus. But, we found that the topics produced using simple LDA make little sense because they were not specific to a particular disease. Additionally, there were overlapping topics, which made it difficult to assess a topic belonging to a specific disease. Therefore, to get topics specific to each disease, we used some seed words that direct the model to converge towards them. For this, we used `GuidedLDA`[10] (aka `SeededLDA` [19]) that helped to converge the topics inclined towards specific diseases. For a fair comparison, we filtered out non-relevant terms, (i.e., terms that are not related to any disease or symptom) using `MetaMap` from the list of topic terms identified by the `GuidedLDA`. The filtered result of topic modeling for all diseases is presented in table 13. It shows top-30 topic words corresponding to each topic based on the sorted probability of the words belonging to that topic. From this table it can be observed that the words corresponding to each disease are, however, related to the respective diseases but they are not specific symptoms of those diseases. The topic number 7 (i.e., eighth topic), as shown in table 13, does not correspond to the kala-azar disease that was expected, rather it appeared to be related to drugs or treatment-related topic. We observed that the words related to kala-azar are associated with the topic number 6 (leishmaniasis). This may be because kala-azar is a type of leishmaniasis called visceral leishmaniasis. Moreover, we had only 18 documents associated with kala-azar in our document corpus due to which the words associated with kala-azar may have received very less probability.

In order to compare the results of top-*k* symptoms extracted by our approach and the top-*k* topic words produced using topic modeling for each disease, we used similarity scores of the topic words and extracted symptoms with the corresponding disease. The similarity scores are calculated by employing the word embeddings taken from pre-trained PubMed word vectors[11] available in binary format.

The `PubMed` word embeddings [27] are learned from the biomedical abstracts and full-text biomedical literatures of the `PubMed` database using skip-gram model of `word2vec` [25, 26] algorithm. We used Cosine similarity to calculate the similarity score between a disease and its corresponding terms/symptoms. The overall similarity score for a disease is calculated as the normalized sum of the similarity scores of each term/symptom with the disease, and presented in table 14. It can be observed from this table that the symptoms extracted by our proposed approach have higher similarity scores with the respective disease in comparison to the terms extracted using the topic modeling approach. Further, it can be observed that the topic words of table 13 do not serve the purpose that we are intended to as we are concerned about the extraction of disease-specific symptoms for disease characterization. In addition, we are unable to capture the relational terms from the topic words that represent associations between a disease and its symptoms.

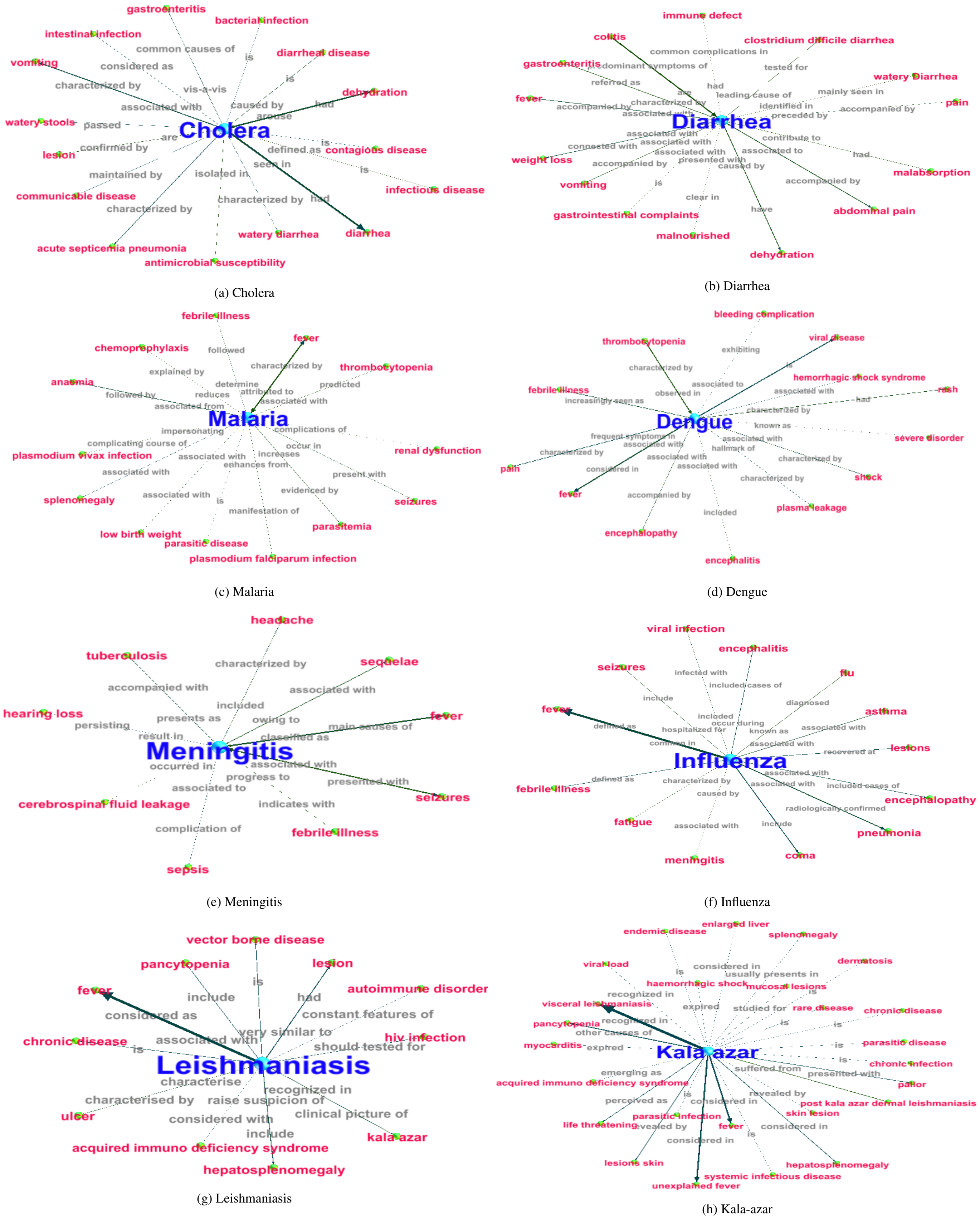
6. Disease-Symptom Association Analysis and Visualization

In this section, we present an analysis of disease-symptom associations at different levels of granularity and their visualization. Figure 6(a-h) presents a visualization of disease and related symptoms as a star like graph structure, wherein the central large blue-colored sphere demonstrates disease name, and smaller spheres at the periphery display the symptoms associated with the central node. Labels allocated to the edges linking a disease with its symptoms represent structural or non-structural disease-symptom associations.

In order to visualize disease-disease and symptom-symptom associations, we have generated a disease-symptom bipartite graph in which list of diseases constitute one set of nodes and list of symptoms constitute another set of nodes, and links are

[10] https://github.com/vi3k6i5/guidedlda

[11] http://evexdb.org/pmresources/vec-space-models/



(a) Cholera

(b) Diarrhea

(c) Malaria

(d) Dengue

(e) Meningitis

(f) Influenza

(g) Leishmaniasis

(h) Kala-azar

Figure 6: Visualization of disease-symptom associations for cholera, diarrhea, malaria, dengue, meningitis, influenza, leishmaniasis, and kala-azar using a graph visualization tool *Gephi 0.8.2*

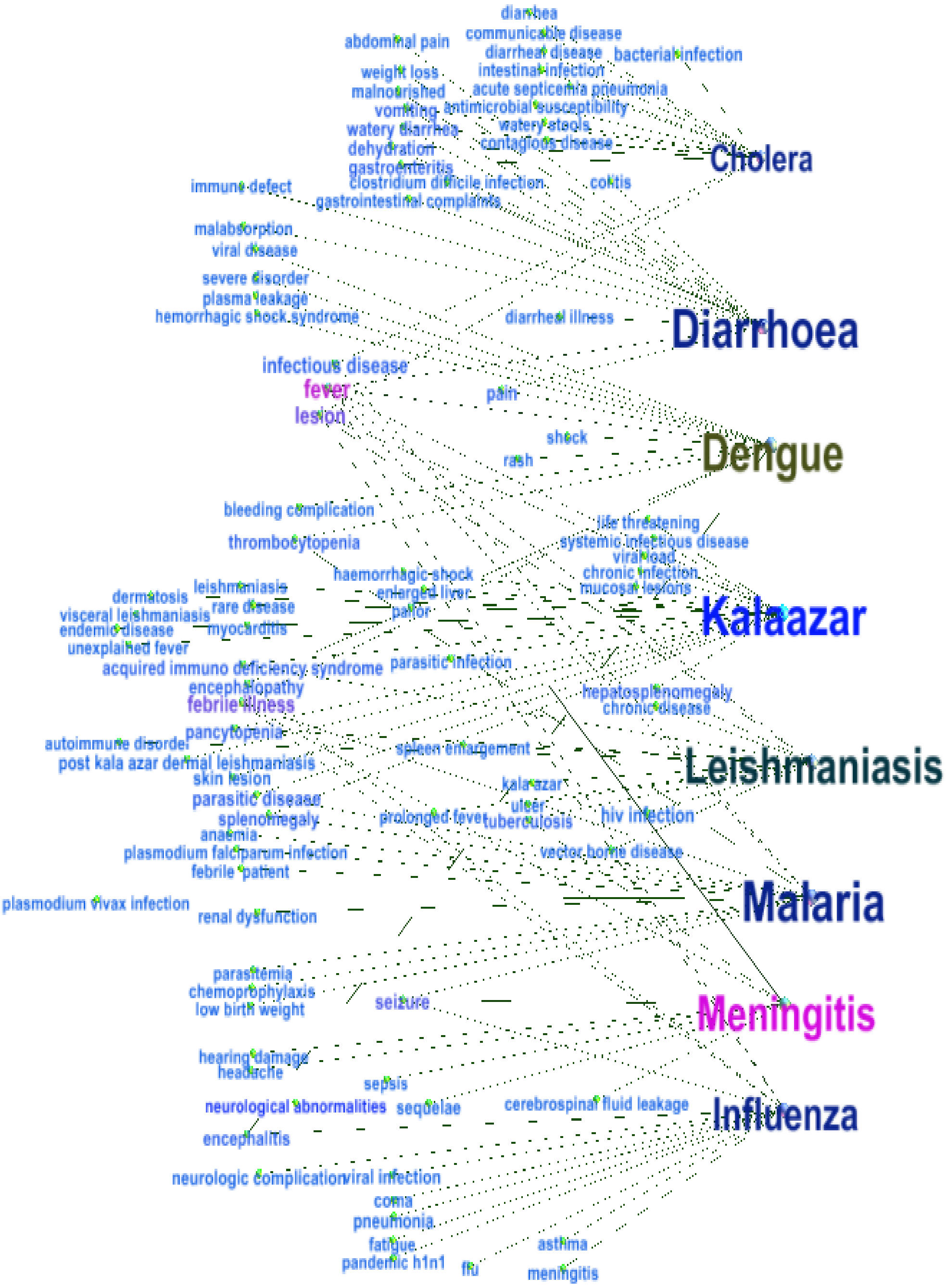
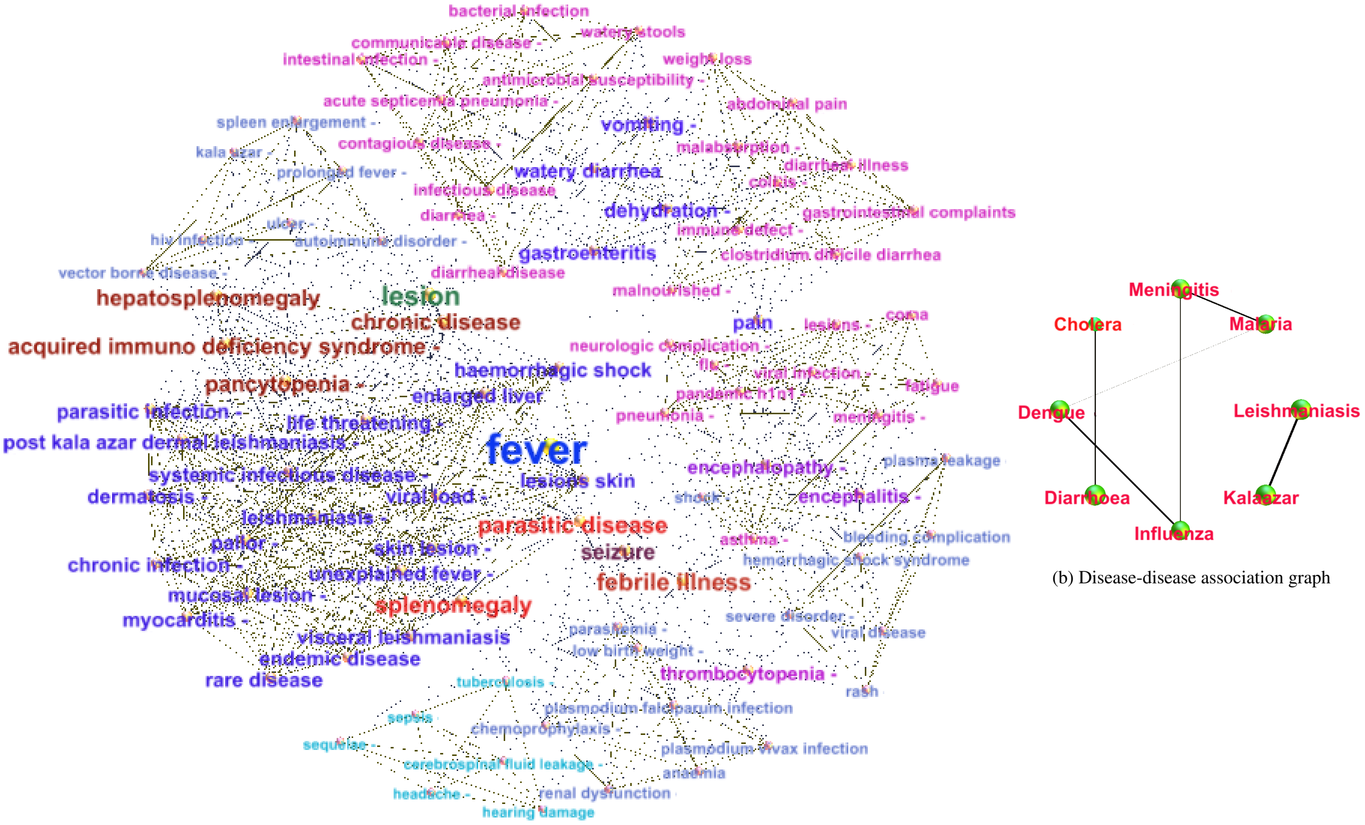


Figure 7: Visualization of disease-symptom bipartite graph using *Gephi 0.8.2*

Table 13: List of diseases (topics) and top-30 symptoms (topic terms) identified by `GuidedLDA`

S. No.	Disease name	Symptoms
1	Cholera	diarrhea diarrhoea dehydration used cholera vomiting therapy poisoning strains results infections gastroenteritis cholerae fed diseases exposure symptoms shigellosis dysentery fever diarrheal toxicity illness source toxic nausea acidosis strain enteritis conditions
2	Dengue	dengue fever infection disease pcr infections response results strains strain detected positive used pathogenesis diseases encephalitis protection wnv shock identified syndrome transmission related infectious secondary chikungunya cholera diagnosis sensitivity monoclonal
3	Malaria	malaria fever hiv infection disease diseases anaemia used anemia infections transmission diagnosis symptoms rural illness results positive febrile related malarial pregnant diarrhoea deficiency death parasitaemia diagnostic history parasitemia education infectious
4	Influenza	influenza infection fever symptoms disease infections illness pneumonia cough positive febrile identified diseases results used hospitalized infectious diagnosis pcr vaccinated emergency related flu measles gastroenteritis detected negative complications symptom pertussis
5	Diarrhea	diarrhea infection diarrhoea symptoms disease gastroenteritis positive ibs colitis strains infections pain syndrome diagnosis detected identified irritable pcr results negative constipation vomiting diagnostic therapy infectious hospitalized enteritis used strain diarrheal
6	Meningitis	meningitis fever diagnosis infection disease syndrome symptoms therapy complications headache seizures hearing infections pain diagnosed positive history vomiting nervous died tuberculosis encephalitis negative lesions sequelae hiv man shock abscess diagnostic
7	Leishmaniasis	disease diarrhea colitis symptoms diagnosis pain syndrome lesions leishmaniasis therapy diarrhoea crohns inflammation ulcerative diagnosed vomiting results infection deficiency anemia complications history bleeding response fever tumor male ibd diseases related
8	Kala-azar	cancer response therapy toxicity grade diarrhea disease nausea vomiting progression toxicities related fatigue neutropenia carcinoma results tumor pain blind rash tumors diarrhoea nsclc used evaluable thrombocytopenia headache mucositis anemia leukemia



(b) Disease-disease association graph

(a) Symptom-symptom association graph

Figure 8: Visualization of projected (a) symptom-symptom association graph, and (b) disease-disease association graph using *Gephi 0.8.2*

Table 14: Comparative analysis results of the proposed approach and `GuidedLDA`

Disease name	Normalized sum of similarity scores	
	GuidedLDA	Proposed approach
Cholera	0.15945	0.1747
Dengue	0.21218	0.2521
Diarrhoea	0.30416	0.4175
Influenza	0.23262	0.2634
Kala-azar	0.08985	0.2254
Leishmaniasis	0.13317	0.1922
Malaria	0.2657	0.3222
Meningitis	0.23995	0.3078

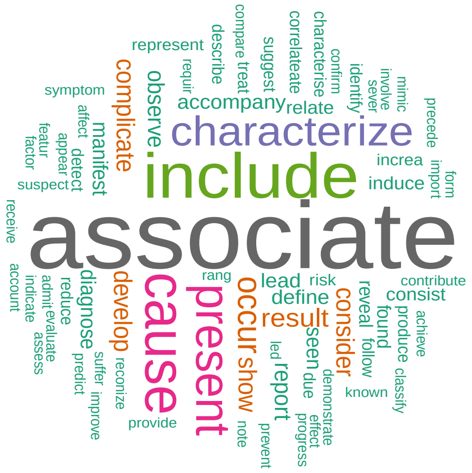


Figure 9: Word-cloud representation of top-75 relational words associating diseases and their symptoms

drawn between the elements of these sets. Figure 7 presents a visualization of the bipartite graph, which portrays individual symptoms as well as symptoms shared by the diseases. In this graph, larger nodes (larger labels) on the right side represent different diseases and the smaller nodes (smaller labels) represent identified symptom/disease concepts. It can be observed from this graph that *fever* is common to almost all diseases, except *cholera*. *Cholera* and *diarrhoea* are *intestinal infection* with common symptoms like *gastroenteritis*, *vomiting*, *watery diarrhea* and *dehydration*. *Lesion* is shared by *cholera*, *leishmaniasis*, *kala-azar*, and *meningitis*. *Kala-azar*, *leishmaniasis*, and *malaria* are *parasitic diseases*, whereas *cholera* is a *bacterial infection*. *Meningitis* may also be a *bacterial disease*. *Dengue* and *influenza* may involve *neurological complications*. *Seizure* is shared by *malaria*, *influenza*, and *meningitis*. Many other valuable insights can be observed from this bipartite graph.

In order to understand symptom-symptom associations, we have applied projection over the bipartite graph shown in figure 7. Symptoms are considered as nodes and edge between a pair of nodes is determined on the basis of Jaccard similarity discussed in sub-section 3.7. Figure 8(a) presents the projected symptom-symptom graph where edges are created if their weights are greater than 0.25. Finally, in order to study the association between different diseases, we have applied projection over the bipartite graph shown in figure 7 and the resultant graph is shown in figure 8(b). In this graph, the similarity score between a pair of disease nodes is calculated using equation 11, and an edge is created provided the similarity score between the corresponding nodes is greater than or equal to 0.25. The thickness of an edge in this graph reflects the degree of similarity between the corresponding nodes. It can be observed from this graph that the *kala-azar* and *leishmaniasis* diseases are highly similar. On analysis, we found that in fact *kala-azar* is a type of *leishmaniasis*, called *visceral leishmaniasis*, and it is a common disease in Indian subcontinent. Similarly, *meningitis* and *influenza*, and *cholera* and *diarrhoea* are somehow similar in their symptoms and their etiologies.

In order to study different types of relational words associating symptoms with a particular disease, we have compiled relation components from feasible *information components* and generated a word-cloud using the word-cloud generator package *wordcloud* in *R*, which presents a quick visualization of the words in varying font sizes and colours displaying more prominent words in bigger and bolder fonts. The size of a word represents the recurrence count of that word in the collection of feasible *information components*. Figure 9 presents a word-cloud showing the glimpse of top-75 relational words with frequency greater than or equal to 145. It clearly exhibits that the words such as *associate*, *characterize*, *cause*, *include*, etc., are most prominent and hence most important words appearing in the scientific literatures to represent disease-symptom associations.

7. Discussion

This study aims to extract disease symptoms and their associations from biomedical text documents. The experimental results demonstrate that typed dependency-based syntactic patterns play significant role in *information component* extraction. The empirical results show non-taxonomic associations between disease and symptoms, and taxonomic associations between a disease and its categories. Further, the outcome shows some heavily used relational words to represent disease-symptom associations, and some new symptoms reported in the biomedical literatures that are generally missing from standard websites.

Upon querying PubMed database, the proposed system fetched documents against only 377 queries, out of total 528 queries passed to the system. This is probably due to the absence of either of the terms of the query within the documents, limiting the number of articles to 107302. Out of these articles, many are duplicate because same disease and symptoms can be mentioned in multiple articles, and hence only 67516 articles are unique.

The *recall* value of the semantic triples extraction process is low because of many reasons. The contributors to low recall value are the limitations of the parser, limitations of the MetaMap, and sometimes very complex sentence structures. The first contributor is the dependency parser. The dependency relation *dep* is an unspecified dependency labeled by the parser when it is unable to determine exact grammatical dependency relationship between the two words of a sentence. We have ignored this dependency relation, as we are unable to detect any trend of it. For example, consider the following sentence:

"Streptococcal shock syndrome should be considered in paediatric patients with fever, vomiting, diarrhoea, abdominal pain and early shock."

and its typed dependencies generated by the parser are:

"[amod(syndrome-3, Streptococcal-1), compound(syndrome-3, shock-2), nsubjpass(considered-6, syndrome-3), aux(considered-6, should-4), auxpass(considered-6, be-5), root(ROOT-0, considered-6), case(patients-9, in-7), amod(patients-9, paediatric-8), nmod:in(considered-6, patients-9), case(pain-18, with-10), compound(pain-18, fever-11), dep(pain-18, vomiting-13), dep(pain-18, diarrhoea-15), amod(pain-18, abdominal-17), nmod:with(considered-6, pain-18), cc(pain-18, and-19), amod(shock-21, early-20), nmod:with(considered-6, shock-21), conj:and(pain-18, shock-21)]"

Due to ignorance of *dep*, the dependency tuples *dep(pain-18, vomiting-13)* and *dep(pain-18, diarrhoea-15)* are filtered out from the dependency list to produce reduced dependency set (see section 3.3). Because of this, important symptoms like *vomiting* and *diarrhoea* are absolved by the system, resulting in reduced recall value. Another contributor to low recall value of the triples extraction process is MetaMap, the tool used to annotate the disease and symptom, as it does not capture some symptoms. Most probably, they fall under different semantic categories, other than the semantic categories considered by the proposed system. However, these are captured as *information components*, but ignored by MetaMap, resulting in low recall value. The semantic category *fndg* contributed considerably to produce false positive tuples. Besides symptoms, other terms also come under this category. For example, *therapy, source, deaths, response, live, issues, old-age* etc. are categorized by the MetaMap under the symptom category, although they are actually not symptoms. Finally, presence of very complex sentences having many fragments within them also contribute to reduced recall value.

8. Conclusion and Future Work

In this paper, we have presented a disease symptom extraction (DiseaSE) system to extract disease symptoms and their associations from biomedical text documents using linguistic and semantic analyses to characterize disease at different levels of granularity. The uniqueness of the DiseaSE system lies in the amalgamation of typed dependencies and named entities using syntactic patterns to map biomedical concepts into meaningful *information components*, and identification of feasible symptoms using TextRank and other ranking approaches. In addition to the well-known disease symptoms, the proposed system is able to accurately extract and identify meaningful new disease symptoms from biomedical texts that are even not listed by the standard disease-specific websites, such as Center for Disease Control

(CDC), World Health Organization (WHO), and National Health Survey (NHS), despite their presence in biomedical literatures. The identified disease symptoms and their associations can be used to generate a comprehensive knowledgebase for the development of biomedical text information processing systems, like e-health and disease surveillance systems. Moreover, the system is generic in the sense that it can extract *information components* from text documents pertaining to any other domain that follow similar pattern of grammatical dependencies. Accordingly, the proposed approach can be tuned for open information extraction (OIE) [28], an emerging area helpful to researchers, scientists, and even an ordinary person, and for the development of an exhaustive knowledge repository in a particular domain of interest. At present, we are working to develop a disease surveillance system over `Twitter` data, using the generated disease knowledgebase.

Acknowledgements

The authors would like to express their gratitude to the University Grant Commission (UGC) of India for affording financial support as Senior Research Fellow (SRF) to the second author under its Maulana Azad National Fellowship (MANF) scheme.

References

[1] Abulaish, M., Dey, L., 2007. Biological relation extraction and query answering from medline abstracts using ontology-based text mining. Data & Knowledge Engineering 61 (2), 228–262.

[2] Abulaish, M., Dey, L., 2009. A relation mining and visualization framework for automated text summarization. In: Proceedings of the International Conference on Pattern Recognition and Machine Intelligence. Springer, pp. 249–254.

[3] Abulaish, M., Jahiruddin, 2013. Web content mining for learning generic relations and their associations from textual biological data. Biological Knowledge Discovery Handbook: Preprocessing, Mining, and Postprocessing of Biological Data, 919–942.

[4] Aronson, A. R., 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] Blei, D. M., Ng, A. Y., Jordan, M. I., 2003. Latent dirichlet allocation. Journal of machine Learning research 3 (Jan), 993–1022.

[6] Bunescu, R., Mooney, R., Ramani, A., Marcotte, E., 2006. Integrating co-occurrence statistics with information extraction for robust retrieval of protein interactions from medline. In: Proceedings of the Workshop on Linking Natural Language Processing and Biology: Towards Deeper Biological Literature Analysis. Association for Computational Linguistics, pp. 49–56.

[7] Bunescu, R. C., Mooney, R. J., 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.

[8] Canese, K., Weis, S., 2013. Pubmed: the bibliographic database.

[9] Chowdhury, M. F. M., Lavelli, A., 2012. Combining tree structures, flat features and patterns for biomedical relation extraction. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, pp. 420–429.

[10] Comeau, D. C., Doğan, R. I., Ciccarese, P., Cohen, K. B., Krallinger, M., Leitner, F., Lu, Z., Peng, Y., Rinaldi, F., Torii, M., et al., 2013. Bioc: a minimalist approach to interoperability for biomedical text processing. Database 2013, bat064.

[11] Culotta, A., Sorensen, J., 2004. Dependency tree kernels for relation extraction. In: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, pp. 423–430.

[12] Datla, V., Lin, K.-I., Louwerse, M., 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.

[13] De Marneffe, M.-C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., Manning, C. D., 2014. Universal stanford dependencies: A cross-linguistic typology. In: LREC. Vol. 14. pp. 4585–92.

[14] De Marneffe, M.-C., Manning, C. D., 2008. Stanford typed dependencies manual. Tech. rep., Stanford University.

[15] Fundel, K., Küffner, R., Zimmer, R., 2007. Relex–relation extraction using dependency parse trees. Bioinformatics 23 (3), 365–371.

[16] Giuliano, C., Lavelli, A., Romano, L., 2006. Exploiting shallow linguistic information for relation extraction from biomedical literature. In: Proceedingsof EACL. Vol. 18. pp. 401–408.

[17] Hassan, M., Coulet, A., Toussaint, Y., 2014. Learning subgraph patterns from text for extracting disease–symptom relationships. In: Proceedings of the 1st International Workshop on Interactions between Data Mining and Natural Language Processing. Vol. 1202. pp. 81–96.

[18] Hassan, M., Makkaoui, O., Coulet, A., Toussaint, Y., 2015. Extracting disease-symptom relationships by learning syntactic patterns from dependency graphs. In: Proceedings of BioNLP 15. pp. 184–194.

[19] Jagarlamudi, J., Daumé III, H., Udupa, R., 2012. Incorporating lexical priors into topic models. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. Asso-

ciation for Computational Linguistics, pp. 204–213.

[20] Jahiruddin, Abulaish, M., Dey, L., 2010. A concept-driven biomedical knowledge extraction and visualization framework for conceptualization of text corpora. Journal of Biomedical Informatics 43 (6), 1020–1035.

[21] Krallinger, M., Leitner, F., Rodriguez-Penagos, C., Valencia, A., 2008. Overview of the protein-protein interaction annotation extraction task of biocreative ii. Genome biology 9 (2), S4.1–S4.19.

[22] Kuhn, K., Campbell-Lendrum, D., Haines, A., Cox, J., Corvalán, C., Anker, M., et al., 2005. Using climate to predict infectious disease epidemics. Geneva: WHO.

[23] Manning, C. D., Raghavan, P., Schütze, H., 2008. Scoring, term weighting and the vector space model. Introduction to information retrieval 100, 2–4.

[24] Mihalcea, R., Tarau, P., 2004. Textrank: Bringing order into text. In: Proceedings of the 2004 conference on empirical methods in natural language processing.

[25] Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

[26] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., Dean, J., 2013. Distributed representations of words and phrases and their compositionality. In: Proceedings of Advances in neural information processing systems. pp. 3111–3119.

[27] Moen, S., Ananiadou, T. S. S., 2013. Distributional semantics resources for biomedical text processing. Proceedings of LBM, 39–44.

[28] Nguyen, N. T., Miwa, M., Tsuruoka, Y., Chikayama, T., Tojo, S., 2015. Wide-coverage relation extraction from medline using deep syntax. BMC bioinformatics 16 (1), 107.1–107.11.

[29] Niwattanakul, S., Singthongchai, J., Naenudorn, E., Wanapu, S., 2013. Using of jaccard coefficient for keywords similarity. In: Proceedings of the International MultiConference of Engineers and Computer Scientists. Vol. 1. pp. 380–384.

[30] Parwez, M. A., Abulaish, M., et al., 2018. Biomedical text analytics for characterizing climate-sensitive disease. Procedia Computer Science 132, 1002–1011.

[31] Reichartz, F., Korte, H., Paass, G., 2009. Dependency tree kernels for relation extraction from natural language text. In: Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, pp. 270–285.

[32] Schwartz, A. S., Hearst, M. A., 2002. A simple algorithm for identifying abbreviation definitions in biomedical text. In: Biocomputing 2003. World Scientific, pp. 451–462.

[33] Seneviratne, M., Ranasinghe, D., 2014. Natural language dependencies for ontological relation extraction. In: Proceedings of the International Conference on Advances in ICT for Emerging Regions (ICTer). IEEE, pp. 142–148.

[34] Tran, L.-T. T., Divita, G., Carter, M. E., Judd, J., Samore, M. H., Gundlapalli, A. V., 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.

[35] Zelenko, D., Aone, C., Richardella, A., 2003. Kernel methods for relation extraction. Journal of machine learning research 3 (Feb), 1083–1106.