Automatic Annotation Generation of Medical Documents

3.1 INTRODUCTION

Because of overpowered information in the medical domain, information systems endure in giving exact results. The annotation generation of medical keywords advances proficient retrieval of most fitting medical documents and related ideas.

There are just few works in the literature that deals with medical document retrieval. A large portion of the current works utilizes either Wikipedia or medical ontology for annotation generation of medical keywords. Furthermore, the existing works fail to provide an accurate result by retrieving irrelevant documents due to diversity of keywords in the user query. The annotation generation methods used in general documents are not suitable for medical documents due to the divergence of medical terms.

With a specific end goal to beat these issues, novel framework called Annotation-based Context-aware Indexing (ACI) for powerful MIR is proposed. It encourages the automatic annotation generation of medical documents. Wikipedia and medical ontology is utilized to advance the useful keywords with annotations so as to deal with the varied nature of the medical terms and user query. In this manner, semantic hole is connected by conquering the befuddle between the medical terms.

The performance change is accomplished by utilizing these annotated keywords as indexing keywords. In addition, context-aware indexing using Bernoulli Randomness Model (BRM) improves the retrieval performance by indexing documents based on lexical association between the keywords.

In this manner, indexing an annotated document gives quick and proficient information retrieval. The proposed ACI goes about as the wellspring of medical knowledge that contains a searchable database of ailment and treatment procedures. End users can utilize this source to inspect and locate the wanted medical information.
3.2 CONTRIBUTIONS

The contributions made in this work include the following:

- Improving the accuracy of the relevant medical document retrieval by addressing upon the synonymy and polysemy problems that exists in the MIR.
- Investigating how to use Wikipedia to improve MIR.

3.3 PROPOSED FRAMEWORK

The proposed medical document annotation framework ACI takes raw medical documents as input. In the medical domain, MEDLINE is the richest and most used source of medical information. Input medical documents are gathered from the PubMed\(^2\), a subset of MEDLINE.

At first, the input medical documents are pre-prepared to uproot stop words. The words that don’t contribute anything towards the semantic significance of the document are called stop words. For example, prepositions, pronouns, articles and so on. These words must not be utilized as indexing keywords. Evacuation of these words can enhance the retrieval performance.

Later, informative keywords are created from the medical documents. The informative keywords are enhanced utilizing Wikipedia and medical ontology. Both Wikipedia and medical ontology recognize related keywords for the given informative medical keyword. The last aftereffect of Wikipedia and medical ontology are amassed to acquire the list of annotated keywords for every document. Keeping in mind, the end goal to increment the efficiency of MIR, these annotated keywords are used in the indexing process.

The proposed annotation framework ACI is shown in Figure 3.1. The framework consists of three phases namely:

1. Informative keyword extraction
2. Annotation generation

\(^2\) www.ncbi.nlm.nih.gov/pubmed/
3. Context-aware document indexing and ranking. These are explained in detail in the forthcoming subsections.

**Figure 3.1:** Framework of ACI
3.3.1 INFORMATIVE KEYWORD EXTRACTION

In this, the pre-processed input document is represented as BoW.

\[ W = \{w_1, w_2, w_3, \ldots, w_n\} \]

Where,

- \( W \) indicates the collection of terms in a document,
- \( \{w_1, w_2, w_3, \ldots, w_n\} \) indicates the individual terms, and
- \( n \) indicates the total number of terms.

\[ BoW = \{w_1 (tf), w_2 (tf), w_3 (tf), \ldots, w_n (tf)\} \]

Where,

- \( tf \) indicates the term frequency.

For example, the BoW illustration of terms is presented underneath:

\[ W = \{\text{Kidney, disorder, temperature, cell, normal, infection, cardiac, infection, ophthalmic, Kidney, antibody, antibody}\} \]

\[ BoW = \{\text{Kidney (2), disorder (1), temperature (1), cell (1), normal (1), infection (2), cardiac (1), ophthalmic (1), antibody (2)}\} \]

The terms in BoW is classified as informative and non-informative keywords by using a medical glossary. All terms in BoW that occurs in the glossary are extracted as informative keywords and other terms are eliminated as non-informative keywords. BoW representation significantly reduces the time required to search. This is because it avoids the term duplication.

The term frequency information of each term is used in context-aware indexing. Consequently, these informative keywords are given to the Wikipedia and medical ontology to extract the related keywords for annotation generation. Algorithm 3.1 elucidates the procedure of informative medical keyword extraction.
Algorithm 3.1: Informative Keyword Extraction

**Input**: Pre-processed medical document

**Output**: Informative keywords

**Notations Used:**

- $n$ is the total number of input documents,
- $doc_id$ is the document id such that $doc_id = \{1,2,\ldots,n\}$,
- $BoW$ is the Bag of Words,
- $W$ is the terms in the input documents,
- $W_p$ is the unique terms in the input document,
- $Med_G$ is the medical glossary,
- $W_G$ is the terms in the medical glossary and
- $Info_Key$ is the informative keywords.

**Process:**

For $\forall$ doc_id {
    
    BoW ($W_{doc_id}$);
    
    Return List [$W_p$];

    For $\forall$ term in List[$W_p$] {
        
        Compare ($\forall[W_p] \in List[W_p] = = W_G \in Med_G$)
        
        If (True)
            
            Extract the term as informative keyword, $Info_Key$;
        
        Else
            
            Eliminate the term as non-informative word ;}

}

3.3.2 ANNOTATION GENERATION

Annotation generation adds more information relevant to documents. Informative keywords extracted from input documents are annotated by using medical ontology and Wikipedia. Therefore, both Wikipedia and medical ontology utilized as a part of the annotation generation process. Figure 3.2 portrays the strides included in the annotation generation process.
Keyword identification, Disambiguation, Keyword selection and Keyword filtration steps are performed by using Wikipedia. Whereas, Variant generation and annotation step is performed by using both Wikipedia and medial ontology.

3.3.2.1 WIKIPEDIA BASED CONCEPTUALISATION

The usage of Wikipedia is explained in this section. The acronyms are frequently used in the medical domain. As these acronyms are ambiguous, they make the indexing keyword selection process ineffective. In fact, the usage of BoW representation does not confine the semantic meaning of medical terms. Therefore, this work uses Wikipedia in the indexing keyword selection process.

The free encyclopedia collaboratively supported by huge number of volunteers is Wikipedia (Paci et al., 2010). Wikipedia is a potential source of knowledge. The tremendous growth of content in Wikipedia attracts many users.
Wikipedia’s structure and content is explored by Wikipedia Miner (Milne and Witten, 2013). It allows the developers to access information about Wikipedia page anchors and links. Wikipedia anchor-search technique allows enhanced search which retrieve a group of Wikipedia articles that refer to the same anchor (senses of the anchor). Annotation generation using Wikipedia takes informative medical keywords as an input and begins the annotation process.

In keyword identification step, Wikipedia anchor-search is performed. It identifies the keywords that can be linked with the Wikipedia pages. Wikipedia anchor-search is performed by searching the document text among all the anchors created by the Wikipedia users. Keywords identified are often ambiguous i.e. it may refer to multiple concepts. Disambiguation step is performed to disambiguate ambiguous keywords. Commonness of anchor sense, context relatedness measure and context quality are the features used for disambiguation. The concept which has got high probability in the context of the given documents is finally selected. Other unambiguous links are used to disambiguate ambiguous ones.

Let the frequency of anchor text which links to a particular sense be represented as $T_{AS}$ and the frequency of anchor text as an anchor be represented as $T_{AA}$. Commonness of anchor sense (CAS) is computed by using equation (3.1).

$$CAS = \frac{T_{AS}}{T_{AA}}$$

Context Relatedness Measure (CRM) between two senses $P$ and $Q$ is calculated by using equation (3.2).

$$CRM = \frac{\log(\max(|p|,|q|)) - \log(|p \cap q|)}{\log(|n|) - \log(\min(|p|,|q|))}$$

Where, $p$, $q$ and $n$ denote the set of links going inside and outside of Wikipedia pages $P$, $Q$ and all of the pages available in Wikipedia respectively.

Once the unambiguous keywords have been identified, other related keywords are obtained by retrieving all the Wikipedia pages it links to or those that link to it. In
keyword selection step, the anchor tags of identified unambiguous keywords and all retrieved related pages are selected as keywords. The keywords that are rarely used in Wikipedia are eliminated in keyword filtration step. Link probability is the measure used in keyword filtration process. Keywords whose link probability exceeds the threshold value are selected and others are eliminated. Thus, after keyword filtration step, final set of keywords are extracted for each input document.

Let $T_{AW}$ represents the total number of times the anchor text is used in Wikipedia. The Link Probability ($LP$) of keyword is computed by using equation (3.3).

$$LP = \frac{T_{AA}}{T_{AW}}$$ (3.3)

In variant generation and annotation step, each keyword is annotated with their variants such as synonyms, acronyms and spelling variants. Wikipedia contains redirects in order to connect each keyword with their variants.

Generation of annotated keywords using Wikipedia can improve the retrieval performance. The existing works in the literature utilize ontology to create annotations for medical related documents (Jonquet et al., 2008). However, the usage of ontology alone cannot cover all the related keywords. This work makes use of the Wikipedia to enrich the annotation process as well as to avoid the high dependency of the domain ontology which in turn leads to effective relevant retrieval of documents.

3.3.2.2 ONTOLOGY RELATED CONCEPTUALISATION

The extracted set of keywords from Wikipedia is given as input to medical ontology in the annotation process. UMLS, a medical domain-specific ontology is used in this work to enrich the annotation generation. MetaMap is a special tool which enables researchers and developers to incorporate UMLS resource into their own applications (Milian et al., 2010).

MetaMap maps the extracted set of keywords of a medical document to UMLS medical ontology in order to generate the variants. Variants are generated by using the
SPECIALIST lexicon and a database of medical related terms and their synonyms, acronyms and spelling variants for a given input medical keyword. SPECIALIST lexicon consists of syntactic, morphological and orthographic information. Annotated keywords are enhanced with the final output of the MetaMap.

Wikipedia is successfully used in the existing approaches to find all keywords related to search query (Yin et al., 2013). In order to achieve good performance improvement in the retrieval process, the proposed approach combines both Wikipedia and ontology.

The final result of annotations obtained for each keyword from both Wikipedia and medical ontology are amassed so as to get the entire domain as well as generic related keywords. The proposed ACI utilizes these annotated keywords as indexing keywords and indexes medical documents so as to enable effectual MIR.

Algorithm 3.2 describes the process involved in the annotation generation.

**Algorithm 3.2: Annotation Generation**

**Input:** Informative medical keywords

**Output:** Annotated keywords

**Notations Used:**
- Info_Key is the informative keywords,
- Iden_Key is the keywords identified using Wikipedia,
- Ambi_Key is the ambiguous keyword,
- Unambi_Key is the unambiguous keyword,
- R_P is the related page,
- Conf_th is the configurable link probability threshold value,
- Sel_Key is the keyword selected using Wikipedia,
- Fin_Key is the keyword finally obtained using Wikipedia,
- K is the list of annotated keywords,
- Acn is the acronyms,
- Syn is the synonyms and
- Spe is the spelling variant.
Process:

Wikipedia Miner ← List[Info_Key];
Wikipedia Miner (Anchor-Search);
   Return List[Iden_Key];
For ∀ Ambi_Key ∈ List[Iden_Key]
   Disambiguate. Features (Commonness of anchor sense, Relatedness measure, Context quality);
   Return Unambi_Key;
List[Iden_Key] ← List[Iden_Key] ∪ Unambi_Key
List[Unambi_Key] ← List[Iden_Key]
For ∀ Unambi_Key ∈ List[Unambi_Key]
   getAnchortag (Unambi_Key and List[R_P]);
   Return List[Sel_Key];
For ∀ Sel_Key ∈ List[Sel_Key]
   Compute Link Probability LP;
   If LP(Sel_Key) > Conf_th Then
      Select Sel_Key;
      Label Sel_Key as Fin_Key;
   Else
      Ignore the Keyword;
Wikipedia Miner. Redirects (List[Fin_Key])
For ∀ Fin_Key ∈ List[Fin_Key]
   Generate.Variant(Acn, Syn, Spe);
MetaMap. SPECIALIST Lexicon (List [Fin_Key])
   For ∀ Fin_Key ∈ List[Fin_Key]
   Generate.Variant (Acn, Syn, Spe);
Return K;
Figure 3.3 illustrates an example for better understanding of the annotation generation process in which SLE is an ambiguous term. It may occur in different senses like Sober Living Environment, Supported Leading Edge, Systemic Lupus Erythematosus. In the context of the given document, SLE refers to Systemic Lupus Erythematosus. Wikipedia has exactly mapped the term to its correct sense.

Figure 3.3: An example for annotation generation process
3.3.3 CONTEXT-AWARE DOCUMENT INDEXING AND RANKING

Indexing minimizes the time taken by the retrieval process as it provides quick access to the information. Annotated keywords are utilized by the proposed ACI for document indexing. The retrieval system has the capacity to recognize the assorted feature of the medical terms as these annotated keywords cover majority of the generic and domain related keywords of the input medical documents. The most related documents are determined by discovering the lexical link among the annotated keywords. The lexical association is computed between each pair of annotated keywords. This association is expressed as correlation co-efficient between annotated keywords. BRM can compute the lexical association between any two annotated keywords (Goyal et al., 2013). In the medical domain, context-aware indexing can locate the relevant documents speedily and aids to retrieve the accurate documents for the given user query.

3.3.3.1 COMPUTATION OF LEXICAL ASSOCIATION

In order to minimize the IR overload, the proposed framework uses the annotated medical keywords for document indexing. Identification of related documents for a particular keyword is the important task in indexing. The lexical link amongst annotated keywords helps to determine the related documents. In a document corpus, the most related keywords have high association with each other, while other keywords will have less association.

In the proposed context-aware indexing framework, the lexical relationship represents the co-occurrence of any two terms. The BRM computes the likelihood of co-occurrence of any two annotated keywords through an examination of input medical documents.

Let Num represent the total number of input medical documents. The document corpus will have $N$ number of annotated keywords and these keywords are used in document indexing. Thus, these keywords are otherwise known as indexing keywords. Let $K$ denote the set of annotated keywords. Hence, $K = \{K_1, K_2, K_3, \ldots, K_N\}$. 
Let \( K_i \) and \( K_j \) represent two annotated keywords which are used to find the co-occurrence. Let \( Num_i \) be the total number of documents in which keyword \( K_i \) occurs and \( Num_j \) be the total number of documents in which keyword \( K_j \) occurs. \( Num_i \) and \( Num_j \) is otherwise called as the document frequency of keyword \( K_i \) and \( K_j \) respectively. Let \( Prob_i \) and \( Prob_j \) be the probability of keyword \( K_i \) and \( K_j \) respectively. Let \( Prob_{ij} \) be the probability of co-occurrence of both \( K_i \) and \( K_j \).

The probability \( Prob_i \) and \( Prob_j \) of the keywords \( K_i \) and \( K_j \) can be computed by using equation (3.4).

\[
Prob_{i(j)} = \frac{Num_{i(j)}}{Num}
\]  

(3.4)

Let \( Num_{ij} \) be the number of documents in which keywords \( K_i \) and \( K_j \) co-occur. The probability of \( K_i \) and \( K_j \) keywords co-occurrence \( (Prob_{ij}) \) can be calculated using equation (3.5).

\[
Prob_{ij} = \left( \frac{Num_j}{Num_{ij}} \right) Prob_i^{Num_{ij}} (1 - Prob_i)^{Num_j - Num_{ij}}
\]  

(3.5)

3.3.3.2 INDEXING OF DOCUMENTS

The indexing weight of each keyword \( K_i \) is computed by using equation (3.4). Keywords are stored in the index table in the decreasing order of their indexing weight. The probability of co-occurrence of any two keywords can be computed by using the lexical association. Highly associated keywords are selected by using the threshold value \( (Th_{val}) \). All keyword pairs whose co-occurrence probability value exceed or equal to \( Th_{val} \) are designated as greatly associated pairs.

For example, let the annotated keywords be \( K_1, K_2, K_3, K_4, K_5 \) and \( K_6 \). The associated keyword pairs can be \( K_1K_2, K_1K_3, K_1K_4, K_1K_5 \) and \( K_1K_6 \).

Correlation coefficient value is more prominent than 0 and lesser than 1. The value is 0 when there exists no relationship between the keywords. More significant the
estimation of coefficient more flawless is the relationship. \( Th_{val} \) of 0.5 is decided for optimum results.

Let the correlation coefficient of keyword pairs are as follows:

\[
\{K_iK_2: 0.6, K_iK_3: 0.3, K_iK_4: 0.4, K_iK_5: 0.5, K_iK_6: 0.9\}
\]

Therefore, highly associated keyword pairs are:

\[
\{K_iK_2: 0.6, K_iK_5: 0.5, K_iK_6: 0.9\}
\]

The documents corresponding to a particular annotated keyword is indexed according to the frequency of the annotated keyword pair. For a particular keyword pair \( K_iK_6 \), the input medical documents containing \( K_iK_6 \) are arranged and sorted according to the frequency of occurrence. Once the process is completed for a given pair, the process is continued for the next pair \( K_iK_2 \). This iterative process will continue until all medical documents that correspond to keyword \( K_i \) have been arranged. Algorithm 3.3 describes the context-aware indexing of medical documents.

**Algorithm 3.3: Context-aware Document Indexing**

**Input:** Annotated keywords  
**Output:** Indexed documents  

**Notations Used:**

- \( K \) is the list of annotated keywords,  
- \( K_{ij} \) associated keyword pairs of \( K \) where \( j = K – i \),  
- \( Prob_i \) is the probability of occurrence of keyword \( K_i \),  
- \( Prob_j \) is the probability of occurrence of keyword \( K_j \),  
- \( Prob_{ij} \) is the probability of co-occurrence of keywords \( K_i \) and \( K_j \),  
- \( Th_{val} \) is the threshold value and  
- \( BRM \) is the Bernoulli randomness model.

**Process:**

\[
BRM \leftarrow K;
\]

For \( \forall \ (K_{ij} \in K) \) do {
   Compute \( Prob_i, Prob_j, Prob_{ij} \);
   If \( (Prob_{ij} \geq Th_{val}) \) Then
      Label \( K_i \) and \( K_j \) as highly associated keywords ;
}
Documents arrangement based on frequency of associated keywords;

3.3.3.3 DOCUMENT RANKING

Ranking takes user query as input. It aims to satisfy the user needs by giving preference to the documents which contain the keywords in the user query.

Figure 3.4 shows the storage structure of the indexed documents in which the unique keywords column represents the unique annotated keywords. Key_Pair represents the highly associated keywords and Doc_Id’s denotes the id of the document. Documents are stored based on the associated value of keyword pairs.

![Figure 3.4: Storage structure of documents](image)

Ranking fuses 3 steps such as pre-processing, searching and ranking. In pre-processing step, stop words are expelled from the user query. In searching step, keywords are extracted from the user query and compared with the unique keywords. At last, in ranking step, corresponding documents are retrieved and presented to the user as answer list.
In this way, for any user query, the proposed framework will retrieve related documents and accomplishes diversification of user’s information need in the medical domain.

3.4 RESULTS AND DISCUSSION

This section shows the result acquired in the MIR using ACI. PubMed, a subset of MEDLINE is the source of input medical documents. The document corpus consists of documents related to disease, treatment, symptoms and diagnosis. MedIndia\(^3\) is the source for creating the medical glossary. Wikipedia content is accessed by using Wikipedia Miner and UMLS is accessed by using MetaMap.

3.4.1 PERFORMANCE METRICS

3.4.1.1 PRECISION

Precision is the proportion of the total number of retrieved medical documents which are applicable to the user query to the total number of medical documents retrieved. It is computed by using equation (3.6).

\[
Precision = \frac{TP}{TP + FP}
\]  

(3.6)

Where,

\(TP\) represents True Positive and \(FP\) represents False Positive.

3.4.1.2 RECALL

Recall is the proportion of the total number of retrieved medical documents which are applicable to the user query to the total number of applicable medical documents. It is computed by using equation (3.7).

\(^3\)www.medindia.net/glossary/
Where,

$TP$ represents True Positive and $FN$ represents False Negative.

### 3.4.1.3 F-MEASURE

The harmonic mean of precision and recall value is called as F-measure. It is computed by using equation (3.8).

$$F - \text{measure} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (3.8)$$

### 3.4.2 PERFORMANCE EVALUATION

The important measures to check the performance of the IR system is the average precision, recall and F-measure values. The result analysis is presented with respect to 15 experimental queries (Hliaoutakis et al., 2009). Figure 3.5 contains the experimental queries.

<table>
<thead>
<tr>
<th>1. Menopausal woman without hormone replacement therapy</th>
</tr>
</thead>
<tbody>
<tr>
<td>2. Woman with advanced metastatic breast cancer</td>
</tr>
<tr>
<td>3. Woman with back pain</td>
</tr>
<tr>
<td>4. Patient with hypothermia</td>
</tr>
<tr>
<td>5. Male with pericardial effusion</td>
</tr>
<tr>
<td>6. Patient with fever or lymphadenopathy</td>
</tr>
<tr>
<td>7. Man with cystic fibrosis</td>
</tr>
<tr>
<td>8. Carcinoid tumors of the liver</td>
</tr>
<tr>
<td>9. Female with urinary retention</td>
</tr>
<tr>
<td>10. Stroke and systolic hypertension</td>
</tr>
<tr>
<td>11. Female with lactase deficiency</td>
</tr>
<tr>
<td>12. Female some months pregnant</td>
</tr>
<tr>
<td>13. Man with sickle cell disease</td>
</tr>
<tr>
<td>14. Adult respiratory distress syndrome</td>
</tr>
<tr>
<td>15. Young man diabetic</td>
</tr>
</tbody>
</table>

**Figure 3.5**: Experimental queries
The proposed ACI has been compared with BioDI, a medical domain-specific ontology based document indexing model (Chebil et al., 2013). In this work, the database contains 500 related documents for the given user query.

3.4.2.1 OVERALL RETRIEVAL PERFORMANCE

Figure 3.6 presents the average precision obtained as a result of the ACI in the MIR with respect to top retrieved documents at rank 100, 200, 300, 400 and 500 cut-offs respectively. When the user gives his information need in the form of query, the proposed system initially retrieves 100 documents for which the average precision value of 0.95 is obtained. Generally, if the quantity of retrieved documents is less, the retrieval system achieves the best result. On the other hand, the value of precision generally diminishes as the quantity of retrieved documents is enlarged. The proposed ACI achieves average precision value of 0.76 for 500 retrieved documents.

![Average Precision Graph](image)

**Figure 3.6:** Average precision with respect to number of retrieved documents

From Figure 3.6 it is apparent that the ACI attains better result when contrasted with the BioDI. The reason behind this is ACI utilizes both Wikipedia and a medical ontology to find the related keywords for a given input document. Generally, if retrieval is finished by using the related keywords, the quantity of relevant retrieved documents
will be diminished. The related keywords might be exact. In the event that related keywords are not accurate, then it boosts the opportunity to retrieve the irrelevant documents. In order to avoid this problem, context-aware indexing is used in the proposed ACI. This context-aware indexing expands the opportunity to retrieve the relevant documents.

Figure 3.7 demonstrates that the ACI accomplishes better average recall value when contrasted with the BioDI. This is on account of, the proposed ACI utilizes Wikipedia in addition to the medical ontology. Consequently, the retrieval system successfully recognizes more associated documents. The proposed ACI achieves average recall value of 0.19 for 100 and 0.76 for 500 retrieved documents.

![Average Recall](image)

**Figure 3.7:** Average recall with respect to number of retrieved documents

The average F-measure value is used to test the accuracy of the IR system. F-measure value depends on the attained precision and recall values. F-measure attains its best score at 1 and worst score at 0. The proposed retrieval framework attains good F-measure value because of better precision and recall values. Figure 3.8 depicts that the proposed ACI attains better average F-measure values when compared with BioDI.
CONCLUSION

In this work, a novel framework called ACI has been proposed for effective MIR. Initially, the informative medical keywords in the pre-processed input documents are extracted by using medical glossary. Later, these keywords are enriched with their variants by using both Wikipedia and medical ontology. The general related keywords are identified by using Wikipedia and domain-specific related keywords are identified by using medical ontology. The output of these two resources is aggregated and thus final set of annotated keywords are obtained for the document corpus. Finally, the documents are effectively indexed by computing lexical association between each pair of annotated keyword. The usage of Wikipedia and medical ontology for annotation generation is the special feature of this work. As it identifies various related keywords of a medical term, the proposed framework can reveal the diverse facets of the user query. The results prove that the accuracy of the ACI outperforms BioDI.

Figure 3.8: Average F-measure with respect to number of retrieved documents