Meningitis Symptoms Extraction from Published Conference Research Projects and Journals

Binyam Seyoum
binseyoum@gmail.com

Tibebe Beshah
School of Information Science, Addis Ababa University, Ethiopia
tibebe.beshah@gmail.com

Abstract

Meningitis is a potentially life-threatening infection of the meninges, the tough layer of tissue that surrounds the brain and the spinal cord. According to WHO statistics, every year bacterial meningitis epidemics affect more than 40 million people living in 21 countries of the "African meningitis belt" (from Senegal to Ethiopia). In this area over 800,000 cases were reported in the last years (1996–2010). Treating a meningitis disease is not hard if the symptoms are identified well, but identification of symptoms of meningitis diseases is a hard job when the data to be extracted is from unstructured source (e.g., PDFs, text files and word documents) mining of symptoms becomes tiresome.

This paper shows how content can be extracted from unstructured data, e.g., a word document or portable document format. Using PubMed and Google Scholar as a data source and pre-processing (data cleaning) techniques gained from information retrieval discipline combined with ontologies to extract content automatically from published conference research projects and journals.

The results of this work can highly benefit the domain experts in biological fields in identification of symptoms of meningitis which in turn can help in combating and eradicating meningitis from Ethiopia as stated by the Millennium Goals.

Keywords: Meningitis; Text Mining; Ontologies

1. Introduction

The proliferation of large amounts of data available on the Web, on corporate Intranets, on news wires and on Life-science journals is overwhelming. Life-science journal publishing has undergone a digital revolution in the last decade. These life-science publications embody a store of knowledge and information of interactions and relations among biological entities, which is very important for the understanding of biological processes. With biomedical literature increasing at a rate of several thousand research projects per week, it is impossible to keep abreast of all developments, although the amount of data available is constantly increasing, ability to absorb and process this information remains constant. Search engines only exacerbate the problem by making more and more documents available in a matter of a few key strokes. Text mining (TM), Natural language processing (NLP) and Information Extraction (IE) have shown the most promising techniques in making biological literature more accessible and easy to retrieve information and associations from thousands of documents [5].

Text mining, NLP and Information Extraction as a new and exciting research area [1], try to solve the information overload problem by using techniques from data mining, machine learning, information retrieval (IR) and knowledge management. This shared basic techniques involve the pre-processing of document collections (text categorization, information extraction, term extraction), the storage of the intermediate representations, the techniques to analyse these intermediate representations (such as distribution analysis, clustering, trend analysis, and association rules), and visualization of the results.

On the other hand according to WHO [1], Ethiopia is at higher rate of outbreak of meningitis. Because of this there is an active research in Alert Hospital to eradicate meningitis from Ethiopia. Every year, bacterial meningitis epidemics affect more than 400 million people living in 21 countries of the "African meningitis belt" (from Senegal to Ethiopia). In this area over 800 000 cases were reported in the last years (1996–2010). Of these cases, 10% resulted in deaths, with another 10–20% developing neurological squeal. During the 2010 epidemic season (weeks 1–26) 22,831 cases were recorded in 14 countries under enhanced surveillance. Among these 22,831 cases there were 2,415 deaths [2, 3]. The most affected countries in the region are Burkina Faso, Chad, Ethiopia, and Niger. Burkina Faso, Ethiopia, and Niger were accountable for 65% of all cases in Africa. In major epidemics, the attack rate range is 100 to 800 people per 100,000. However, communities can have attack rates as high as 1,000 per 100,000. During these epidemics, young children have the highest attack rates. The meningitis in these regions has caused many deaths every year at an estimated economic cost of huge amounts of money [4].

Consequently, this paper is aimed to help biological researchers in Ethiopia by giving the tool, to quickly and efficiently to find and extract information (the symptoms of meningitis) from published conference research projects and journals. For this purpose the tools and techniques of text mining and information extractions have been used.

The problem with unstructured documents is even worse in biological literatures, if we look at Medline database [7], which maintains the abstracts of research projects in the field of biomedical research, had a growth of 500,000 new research projects in 2004 per year. In 2010, this has become two research projects per minute [7]. This availability of huge textual resources provides the scientist with the chance to search for correlations or associations such as protein–protein interactions, gene–disease associations, disease symptom association, disease-cause association and new findings about the research area [8]. Nevertheless these huge number of documents with more and more information in them are kept untouched because there is hardly a tool that can mine the knowledge that is hidden, many researchers currently are extracting information manually and their discovery is as good as their processing power.

Currently an active research is being held on meningitis symptoms at Alert Hospital in Addis Ababa, as WHO states that Ethiopia is currently at the pick of meningitis outbreak [1] and researchers at Alert Hospital are doing a research on the symptoms of meningitis.

The problem with unstructured documents is also faced by Alert Hospital researchers. Using techniques and tools of text mining, this paper will contribute its share by extracting symptoms of meningitis from published conference research projects and journals. Text mining is the process of discovering new, previously unknown information, by automatically extracting information from a usually large amount of different unstructured textual resources by computer

The general objective of this paper is to identify the symptoms of meningitis disease from meningitis literatures, using the basic techniques and algorithms of information retrieval combined with ontologies and to produce a prototype.

2. Related Work

The volume of published biomedical research, and therefore the underlying biomedical knowledge base, is expanding at an increasing rate. While scientific information in general has been growing exponentially for several centuries, the absolute numbers specific to modern medicine are very impressive. The MEDLINE 2004 database contains over 12.5 million records and the database is currently growing at the rate of 500,000 new citations each year. With such explosive growth, it is extremely challenging to keep up to date with all of the new discoveries and theories even within one's own field of biomedical research.

Starting with a collection of documents, a text mining tool would retrieve a particular document and preprocess it by checking format and character sets. Then it would go through a text analysis phase, sometimes repeating techniques until information is extracted. Three text analysis techniques are presented: collection, preprocessing and extraction or analyzing, but many other combinations of techniques could be used depending on the goals. The resulting information can be placed in a management information.

Many attempts have been made to get the most out of the increasing biological literatures and different scholars have been proposing and implementing their work on their respective field. One pioneer work is the work of Don Swanson. This paper uses the preprocessing of documents the same as Swanson since some books place the works of Swanson as of concept extraction or text mining and some put it as different technologies [9].

The work of Swanson is categorized in the Concept Linkage, because in his work he actually found a cure. Swanson pioneered the research of knowledge discovery from text by exploring the benefits of inferring associations in a series of experiments using simple semi-automated methods to aid human discovery. Titles from MEDLINE were used to make connections between seemingly dissociated arguments: the connection between migraine and magnesium deficiency, which has been subsequently validated experimentally; between indomethacin and Alzheimer's disease, and between Curcuma longa and retinal diseases, Crohn's disease and disorders related to the spinal cord.

Swanson's work is different from the work in this paper due to 1) event though both works use basic preprocessing techniques the category is different; the work in this paper is more of information extraction but Swanson's work is Concept Linkage. 2) There is help of visualization in the work in this paper which holds visualization of the final work. 3) The tools, procedures and algorithms are different. This work is inclined to information extraction, since only symptoms are extracted from the collected documents.

Another work in the area of text mining is the work of Mathiak and Eckstein [9]. They stated that their work of text mining has five parts: text gathering, text preprocessing, data analysis, visualization, and evaluation. The aim was to analyse the different methods applicable to the five steps and to add their own results if possible [9], and to present the most feasible way of text mining. It presents a framework from collecting the texts to visualizing the result. In the procedures used and the overall steps involved, their aim was to see different methods and check the applicability of the methods to the five steps.

3. The Proposed Solution

Data collection is the first step in information extraction, by choosing Google scholar and PubMed as source and tool. The reason behind is accessing a journal or conference proceeding from Google Scholar is quite different from PubMed.

First, a keyword was selected that can represent the theme of the documents that was needed for the research project and following were selected "meningitis", "meningitis symptoms", "new findings of meningitis" and "meningitis 2013". The keyword "meningitis" is believed to represent the documents that are related to meningitis and "New findings of meningitis" is selected because in the new findings of meningitis the symptoms are also included. Since we are looking for new findings regarding symptoms, the keyword "meningitis symptoms" is self-explanatory. As it can be seen from Table 1, it is the central theme of the documents that are used since 40% of the documents are gathered form this keyword search. With "meningitis 2013" some results are displayed on both search engines. This term is suitable for the desired search because symptoms up to 2012 are already known and what is needed is the 2013 findings on meningitis focusing on symptoms.

Table 1: Search results from Google scholar and PubMed

Keyword	*Website*	*Results*
Meningitis	PubMed	61,475
	Google Scholar	672,000
New findings of Meningitis	PubMed	2,687
	Google Scholar	252,000
Meningitis Symptoms	PubMed	32,770
	Google Scholar	283,000
Meningitis 2013	PubMed	1,710
	Google Scholar	131,000
Total		*1,436,642*

Using the above keywords in PubMed, PubMed central (PubMed for only full articles) and Google scholar the following result was obtained.

As can be seen from Table 1, Google Scholar presents a large amount of result when compared to PubMed with the same keyword but this doesn't mean that all the results are relevant since some junk results are also included. Therefore, the next step was to filter the above results to more relevant documents by limiting the number of search pages and restricting results from only certain trusted websites and also can provide freely for review the document. Some trusted websites for biological research were added after searching the web like CDC, WHO, and the Red Cross foundation. When using this websites for the given search dramatically, the Google search results were decreased as shown in Table 2.

Table 2: Google Scholar result using specific websites

Keyword	*Result*
Meningitis	15,000
New findings of Meningitis	8,000
Meningitis Symptoms	15,600
Meningitis 2013	10,000
Total	*48,600*

Finally results that are presented from the first to the third pages are considered since unrelated contents are usually displayed after the third page.

At last 6,553 papers that qualify the above filter methods were downloaded both from Google Scholar and PubMed. During the gathering process from Google Scholar and PubMed 2,210 articles have been removed due to redundancy from both sources. A total of 4,343 papers were collected and made available for the next step of preprocessing.

To successfully employ text mining on PDF encoded paper, it would be advantageous to start with customizing the conversion process in order to be able to optimize on all levels. A java API was implemented for the conversation process to meet the requirement. A free API provided by Apache called PDFBox is used to convert the PDFs. Using Apache PDFBox API, the collected 4,343 pdf files where converted to text file.

Text Tokenization is the processing of breaking or chopping a continuous character stream into meaningful constituents called tokens. In text mining specifically in term extraction, the presence of words/terms and their statistical distribution play a significant role rather than the sequence of the terms, this is called the bag-of-words approach. In bag-of-word approach to make a statistics of words, tokenization is applied.

After tokenization the tokens (words) were about 4,893,520. These tokenized words have too much redundancy either by having the same word in different form as in the past and the future or having words that are irrelevant called stop words. In order to decrease the dimensionality of the words, terms that are grammatically close to each other (like "cell" and "cells") are mapped to one term via word stemming. Some authors like [6] put stemming after tokenization in order to decrease the dimensionality of the tokenized words.

Stemming is used to map a word to its root word. Porter's algorithm is implemented for stemming since it is a well-developed and maintained stemming algorithm for English language. A free Java API by Porter is used to stem the tokenized words.

The tokenized words were successfully stemmed modifying Porter's algorithm. During stemming about 600,000 words have been stemmed and has decreased the dimensionality of the words from 4,893,520 to 4,293,520. But in a bag-of-words approach tokenization and stemming are not enough. Still there are some stop words so a stop word dictionary is used which is available in the national center for text mining (NCTM) and is believed to be a standard dictionary for stop words. However the dictionary has its downside; it doesn't include stop words for biological documents About 15 stop words have been added using word count frequency. The list and their frequency is shown in Table 3.

Table 3: Biological Added stop words

No.	*Word*	*Count/frequency*
1	Cell	3451
2	Tissue	3008
3	DNA	2972
4	RNA	2900
5	Portion	2900
6	mRNA	2680
7	rRNA	2500
8	Mitotic	2322
9	Cyclin	2010
10	Yeast	1800
11	Genome	1525
12	receptor	1010
13	Gene	896

After selecting the stop words list, a simple Java class was built that can match between words that are presented as stop words and words that are stemmed. The implemented program matches the words and removes similar words. In this process almost a third of the stemmed words have been removed due to stop word and redundancy.

In all from the initial of 4,893,520 word preprocessing (stemming and stop word removal) greatly reduced the dimensionality by 46% and the cleaned words are 2,512,300. Then comes term extraction. Term extraction first removes white spaces and commas and put the words in a collected form. When the documents are closely seen only symptoms and other relevant words remain in the list.

The ultimate goal of this paper is to extract the symptoms from literature. This is where ontologies came in play. Ontologies as a conceptual framework possess different concepts in a tree structure, and these concepts are expressed using a term. If ontology of symptoms is used all symptom terms are presented on a given clinical symptoms ontology. So that ontology of clinical symptoms from open biological ontology (obo) which is a free and trusted ontological provider is downloaded and then the file is exported to plain text to resolve the issue of filtering only symptoms from the pre-processed files by matching and cross referencing with the ontology just like stop word removal. The overall task at this point is the symptom terms that are filtered using ontologies are the symptoms of meningitis and can be presented as one, but further filtering is required since this work is intended to present the new symptoms rather than presenting all (old and new) symptoms together.

The filtering of terms using ontology to get a collection of words that are used to describe symptoms gives excellent result. But further analysis is required since the end users need the recent symptom not a collection of symptoms.

Phrase is used for an expression that consists of one or more words. Sometimes symptoms can be phrases that constitute more than one word, for example, "stiff neck" is a symptom that consists of two words. This work uses bag of words approach. In this approach the position of words is ignored and focuses on the word level techniques. The problem is if a symptom is a phrase and if tokenization is applied the meaning is lost. Therefore a mechanism should be implemented to incorporate phrases.

A couple of techniques have been implemented to deal with phrases.

To lower the ontology to the word level or actual clinical names, e.g., hair loss will be mapped to alopecia (correct clinical term for hair loss). There are many phrases that fall in to this category. Out of 767 types of clinical symptoms 278 phrases can be mapped to the original medical term.

After conversation and before tokenizing the collected documents other phrase symptoms were collected that much the ontology phrases. These are around 62 out of 489 symptoms. If tokenized the relation between the two words will be lost and never discovers the symptoms.

Using the above two methods the symptoms that are phrases and words of symptoms were extracted. Still there are some issues that are not covered by the above two techniques, e.g., if papers use nonscientific terms to describe the symptoms, the matching will be ineffective but this is a rare case since the publications are for the scientific community.

4. Discussion

The objective of Information Extraction (IE) is recognizing and extracting certain types of information from unstructured or semi structured documents. Ontology based information extraction is more precise in extracting because the machine is not extracting blindly rather by using definition of all the words provided by the ontology. Information extraction with the guide of ontology has three parts:

- Process natural language text documents.
- Present the output using ontologies.
- The information extraction process is guided by the ontology to extract things such as classes, properties, instances and terms.

In this paper, Java open source tools and preprocessing technique is used to gain information retrieval and also build a system more like a prototype that can take input of texts and convert, tokenize, stem, remove stop words and cross reference with the ontology provided and then display the result. The program has successfully extracted information from nearly 4,343 conference and journal research papers. However these results are very much dependent on the ontology provided. The ontology used is from open biological ontology (OBO). Therefore the result is expected to incorporate all clinical symptoms.

The evaluation for this work is done in two ways. First is checking the relevance of the tools and techniques. Then we have to check if the results are as expected. Choosing and using the right tools and techniques can lead to the right outcomes. The goal of this work is to identify the symptoms of meningitis from meningitis literature using content mining techniques and algorithms and to develop a prototype. Alert Hospital as a testing environment for this work conformed the results as valid. The results will soon be considered as known symptoms when another vaccine is issued for meningitis. The symptoms that were discovered manually were 13 in number and with this work they were 10. Two symptoms were displaced when dealing with phrases and it was a tolerable error as a starting work.

The prototype is developed using Java and coded using NetBeans IDE. The prototype can potentially convert the given pdf files, preprocess (tokenizing, stemming, stop word removal), load ontology in text file to match and display the result of the finding. The User Interface consists of file buttons to browse PDF files, to pre-process the documents, load ontology and load known and an extract button to load known symptoms and extract.

Symptoms of meningitis are essential in the domain of biology, medical research and patient diagnosis. Knowing the symptom of a disease helps researchers in understanding the underlining principles of specific diseases and this powerful knowledge can help in developing a cure or vaccine for the specific disease or even for further research on combating the existing one.

The prototype, the underneath principles and the results were explained to the domain experts in Alert Hospital, then they were free to test the prototype with their own data. More than 95% of the results matched with the experts test. The 5% error was caused by the data pre-processing step which skips some redundant words.

5. Conclusion and Future Work

This research attempted to show the possible application of term extraction using IR data pre-processing steps and ontology to increase the precision of the terms (symptoms) extraction. The data collection followed by file conversion from PDF to text file for more flexible preprocessing was then cleaned using the IR data preprocessing steps that include tokenization, stop word removal and then after term extraction using ontologies stemming.

The data collection and preparation were major tasks due to the uncleanness of the data collected from PubMed and Google Scholar. This is also due to higher volume or size of the data in the database. After keywords were selected by consulting domain experts the two websites were queried for any relevant documents. The search results were numerous and some unwanted junk were in it. Therefore some cross referencing and trusted website source filtering were applied to limit the search result. After successfully filtering and acquiring the desired documents, the documents were converted to text file to successfully preprocess using open Java API called PDFBox.

Here the data is ready for preprocessing so, tokenization, stop word removal and stemming were applied to clean the data and make available for the general objective which is finding the symptoms of meningitis from published research papers. Since the data is clean and only relevant terms are there in the documents, but clinical symptoms are hard to find in the mixed words. Therefore, using clinical symptoms ontology, all technical terms of symptoms were extracted from the cleaned documents.

The work of this research can benefit researchers in Alert Hospital on finding the current symptoms of meningitis. Also scientist who investigate the treatment of any disease will use it for reference to cross check effect of meningitis for their patients.

References

[1] Pubmed, NCBI, http://www.ncbi.nlm.nih.gov/pubmed, Last Accessed on 18 September 2013.

[2] "wikipedia.org," 12 September 2013, Available at http://en.wikipedia.org/wiki/Meningitis.

[3] "webmd," 12 September 2013, Available at http://children.webmd.com/vaccines/tc/meningitis-topic-overview.

[4] "Mafricar," menafricar.org, 15 September 2013, Available at http://www.menafricar.org/meningitis-and-africa, Last accessed on 17 September 2013.

[5] "WHO," WHO, 16 September 2013, Available at http://www.who.int/mediacentre/factsheets/fs141/en/, Last accessed on 18 September 2013.

[6] "chealth.canoe.ca," canoe, 12 September 2013, Available at tp://chealth.canoe.ca/, Last accessed on 15 September 2013.

[7] R. Feldman, "The Text Mining Handbook," Israel: Bar-Ilan University, Israel, 2007.

[8] WHO, Control of epidemic meningococcal disease, Vol. 3, No. 3, pp. 70-80, 1998.

[9] Wiki, "http://en.wikipedia.org/wiki/Google_Scholar", Available at http://en.wikipedia.org/wiki/Google_Scholar.