Template based Medical Reports Summarization

Ahmed Y. Abu El-Qumsan
Gaza, Palestine

Alaa M. Elhalees
The Islamic University of Gaza
Gaza, Palestine

ABSTRACT
The torrential information in the medical records is considered a great problem because it difficult to distinguish the needed and necessary information from the huge quantity of data. As a result, the importance of summarize medical reports is growing day after day. Medical information extraction is one of the important topics that aim to identify medical information and detect hidden relations. This topic is considered one of the most important topics in the field of text mining where is used to process unstructured texts and extract meaningful information which is hidden in the unstructured texts.

The information extracted from medical reports is very useful to medical staff to detect hidden relations between medical information, and making decisions that will improve the medical service for patients, in addition to saving time and effort.

In our paper, an approach that use template based medical reports summarization has been developed to transfer medical reports from semi structured and unstructured form to structured form. It classifies the identified entities then extracts important information such as diseases, medical procedures, and drugs. After that, it can discovery hidden relationship between medical information by using association rules. The dataset used in this paper was collected from the Palestinian Ministry of Health.

To evaluate the performance and effectiveness of our model, human expert has been used as a reference to measure the degree of acceptance of the extracted association rules which have been extracted from the dataset. So, Likert's scale has been used for evaluation. After the data analysis obtained from the questionnaire. It shows us that the proportion of accuracy association rules, which have been extracted is about 80%.

General Terms
Named Entity Recognition, Text Mining, Template Based Summarization, Text Summarization, Association Rules.

Keywords
Template Based Summarization, Text Mining, Information Extraction, Medical Reports, Named Entity Recognition, Association Rules.

1. INTRODUCTION
Due to the rapid growth of the information in the world, users have to face the information overload. The information overload either leads to wastage of significant time in browsing all the information or some useful information could be missed out [1]. To overcome this problem text summarization can be used.

Automatic text summarization is maturing and may provide a solution to the information overload problem and a very powerful tool to save time and resources, and optimize availability for an expert in any domain area [1].

Another reason to create automated summarization system is: when there are two different people may create different summaries of the same article based on what they think is most important and how their perceive the article [1].

Radev and Mckeown [2] defined a summary as "*a text that is* produced from one or more texts, that conveys important information in the original text(s), and that is no longer than half of the original text(s) and usually significantly less than that".

In term of tasks, there are two tasks in text summarization: content (sentence) selection and template based summarization. Content (sentence) selection "extraction summary" is a selection of sentences or phrases from the original text with the highest score and put it together to a new shorter text without changing the original document. While template based summarization "information extraction" is the task of automatically extracting structured information from unstructured and/or semi-structured documents [3].

The rapid growths of medical records motivates and lead hospitals, primary care centers, and health organizations to use technology to reduce the effort, extract important information, and to speed up the process of analyzing and linking information to discover and predict diseases, drugs, and medical procedure for patients.

The main purpose of this paper is to convert text medical records to structured form using template summarization. After that this paper creates a new model for extract important information and detects and discovers diseases, drugs, and medical procedures using unstructured text and data mining techniques. This model will be used in the last stage of an investigation to help medical staff to efficiently detect hidden relations between medical information, and making decisions that will improve the medical service for patients.

The dataset used in this paper were collected from the Palestinian Ministry of Health, especially from government hospitals, which is estimated at 2200 medical report. The report contains an overview of the status of the patient and symptoms, diagnosis and other information.

Understating relationships between medical information can help medical staff to detect hidden information in order to identify and predict diseases and medications and procedures. For example, Figure 1 has an archive of medical reports unstructured $R = (R_1, R_2, R_3, ..., R_N)$ where n is the number of medical reports. Where, the first report contains one diagnosis of acute ischemic CVA, DM, and HTN, a medical procedure is CT. The second report contains a number of diagnoses, such as Cervical disc bulge, and CVA, medical procedures is MRI. The third report contains the Fibrillation, hemiparesis, and CVA, medical procedures such as neurological examination, the drug is Crestor.

It is noted that for the same disease (CVA), there are different diagnoses and different procedures.

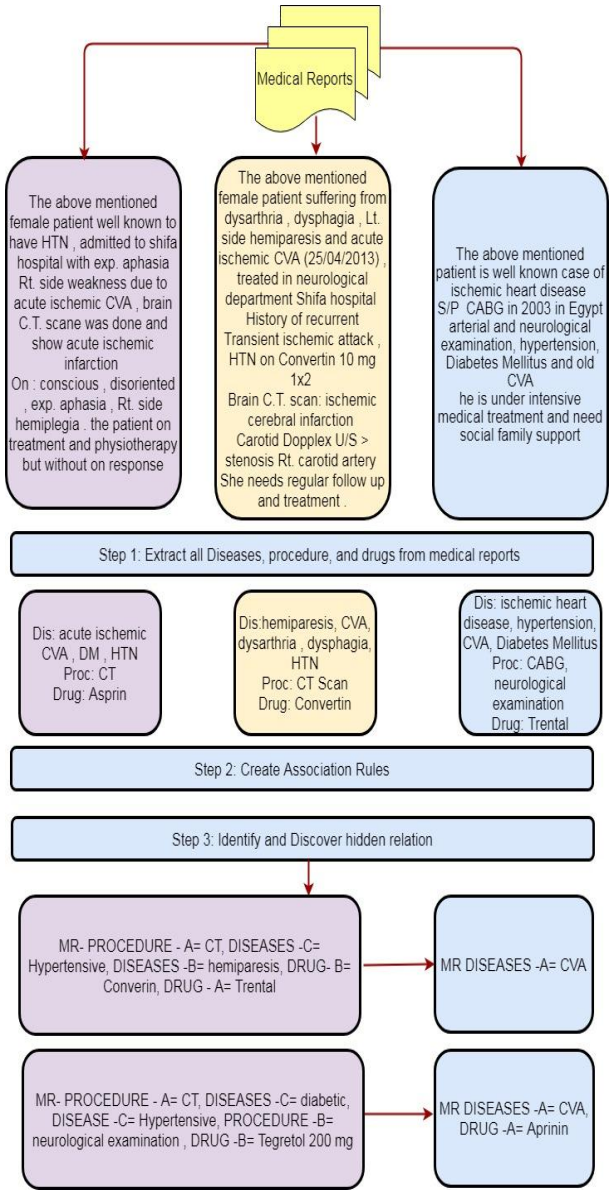

Figure 1: Example of relationship between medical reports

The extracted information will help doctors and staff medical predict diseases and medical procedures and drugs for patients.

In the paper model, a three steps approach to detect and predict disease and medical procedures and medications for patients is suggested. In the first step, extract medical information from medical reports include diseases and medical procedures and drugs. In the second step, create association rules from of the data that has been extracted from the medical reports. In the final step, Extract and discover hidden relationships between the medical reports and the analysis and prediction of diseases and medical procedures and medications for patients. The experimental results will demonstrate the effectiveness of the proposed approach.

2. RELATED WORK

Many researchers gave a great attention to text summarization, but a very few of them invented models for text summarization by template in medical domain. In this section, a number of research works that focused on text summarization (template based summarization, sentence selection summarization), and Association rules in medical domain. This literature review is divided work into three sections: Sentence selection summarization, template based summarization, and association rules.

2.1 Template Based Summarization

There are some researches in Template Based Summarization in medical domain such as:

Bunescu, et al. [4] proposed a technique for extracting information from biomedical text. The focus was on the initial stage of identifying information on interacting proteins, specifically the problem of recognizing protein and gene names with high precision. To determine the names of the protein was used protein tagger where the success of a protein tagger depends on how well it captures the regularities of protein naming and name variations. They used Medline abstracts in order to extract the names of protein of them. They used the standard measures of precision and recall, the results were promising, where the precision was 93%, and at recall was 82%.

Jung, et al. [5] developed that method automated medication extraction system, which can accurately extract medication names and signatures from medical records. The proposed system consists of three main steps for perform information extraction, namely: pre-processing is to determine the sentence boundaries in a medical record, a semantic tagger to break an input sentence into tokens and label proper words or phrases with a semantic category, and parsing component of system uses a context-free grammar to parse textual sentences into structured forms. They evaluated system performance using two types of datasets: discharge summaries and clinic visit notes. Results showed that system can extract drug names and signature information such as strength, route, and frequency from discharge summaries and clinic visit notes with over 90% F-measure.

Delégeret, et al. [6] proposed system aims to a development corpus and a priori knowledge for automatic extraction of medical information such as the drug names and associated information (mode, dosage, etc.) from narrative patient records, relies on a semantic lexicon and extraction rules. Also, they showed that controlled modifications (lexicon filtering and rule refinement) were the improvements that best raised the performance. They evaluated system performance and showed good results (global F-measure of 77%). Further testing of different configurations substantially improved the system (global F-measure of 81%), performing well for all types of information (e.g., 84% for drug names and 88% for modes).

Jonnagaddala, et al. [7] developed a system for determining and extract coronary artery disease (CAD) risk factors from unstructured electronic health records. Using clinical text mining and to calculate 10-year coronary artery disease risk scores in a cohort of diabetic patients. After that, they developed a rule-based system to extract risk factors: age, gender, total cholesterol, HDL-C, blood pressure, diabetes history and smoking history. Unstructured electronic health records (EHRs) were obtained from the i2b2 2014 shared task 2 which deals with identifying risk factors for heart disease over a period of time. The results showed that the output from the text mining system was reliable, but there was a significant amount of missing data to calculate the Framingham risk score. A systematic approach for understanding missing data

was followed by implementation of imputation strategies. An analysis of the 10-year Framingham risk scores for coronary artery disease in this cohort has shown that the majority of the diabetic patients are at moderate risk of CAD.

2.2 Sentence Selection Summarization

In addition, there are some researches in Sentence Selection Summarization in medical domain such as:

Chen & Verma [8] proposed a new user query based text summarization technique that makes use of unified medical language system, an ontology knowledge source from National Library of Medicine. They compared their proposed method with keyword-only approach, and this ontology-based method performs clearly better. The proposed method also showed potential to be used in other information retrieval areas. They used ontology to expand query words and assign scores to sentences based on number of original keywords (query words) and expanded keywords.

Sarkar [9] developed summarization method to find the relevant information on the Medical Literatures on the web. The main approach of their paper is basically based on combining several domain specific features with some other known features such as term frequency, title and position to improve the summarization performance in the medical domain. The author identified a list of cue terms and phrases specific to the medical domain. The idea is that the phrases like "We report", "We present", "World Health Organization", "This study is", "Prevention of" … etc., considered as cue terms of summarization. The system had three phases: First, document preprocessing component deals with formatting the input document, segmentation and stop removal. The second phase sentence ranking component assigns scores to the sentences based on the domain knowledge, word level and sentence level features. The summary generation component selects top n sentences based on scores. Finally, the sentences included in to the summary are reordered to increase the readability. Results showed that the incorporation of domain specific features improves the summarization performance.

Xu, et al. [10] developed a system called MedEx. The system extracts medication names and signatures (e.g., dose, route, and frequency) from clinical narratives. MedEx was initially developed using discharge summaries. The goal of their system to develop a medication parser that can accurately extract drug names, signatures, and contextual information. They built the medication parser using a semantic-based approach. The dataset that was used: discharge summaries and clinic visit notes. The results showed the system performed well on identifying not only drug names (F-measure 93.2%), but also signature information, such as strength, route, and frequency, with F-measures of 94.5%, 93.9%, and 96.0% respectively.

Gold, et al [11] proposed method that extracts medication information such as drug names and signature information such as dose, route, and frequency from discharge summaries. Their parser relies on a library of regular expressions and a lexicon of drug names to identify medication information. Both the lexicon and the parsing rules are flexible, and can be easily customized for other types of clinical notes, or other discharge summaries with different writing styles. Evaluation on a data set of 26 discharge summaries showed that drug names were identified with a precision of 94.1% and a recall of 82.5%, but other signature information such as dose and frequency had much lower precisions.

2.3 Association rules in medical domain

There are some researches in Association Rules in medical domain such as:

Doddi, et al. [12] used approach to analyze a large database containing medical record data. The main goal of their system was to discover relationships between medical procedures performed on a patient and the reported diagnoses and the purpose of their system is to demonstrate its applicability to medical data. The researchers used common method to discover and predicting such relationships is association rules between procedures and diagnoses. Where they considered association rules useful to measuring joint frequencies for common combinations of medical procedures and the corresponding diagnoses. Most of the discovered rules in their paper can be potentially very revealing and beneficial to medical professionals.

Rashid, et al. [13] built a system for to find out relations among the primary disease and other secondary diseases. Where was used association rule mining to extract knowledge from clinical data for predicting correlation of diseases carried by a patient the researchers developed a system for Clinical State Correlation Prediction (CSCP) which extracts data from patients' healthcare database, transforms the online transaction processing (OLTP) data into a Data Warehouse by generating association rules. Their system is more generic version of CSCP system that can work for all diseases in similar fashion and generate correlations depending on the input dataset. The drawback of their paper was the use of a small dataset plus it is not real data.

Ordonez, et al. [14] The contribution of their paper was to find and discover new association rules in medical data to predict heart disease and validating rules used by an expert system to aid in diagnosing coronary heart disease. The authors of this paper focused on two aspects in this work. First, mapping medical data to a transaction format suitable for mining association rules. Second; identifying useful constraints to aid in diagnosing coronary heart disease correctly. The researchers worked on improved algorithm to discover constrained association rules.

Nahar, et al. [15] Used three different rule mining algorithms - Apriori, Predictive Apriori and Tertius - to identify the sick and healthy factors which contribute to heart disease for males and females. The researchers focused on the identify of coronary heart disease based on gender and significant risk factors. The dataset used in their research is Cleveland dataset, a publicly available dataset and widely popular with data mining researchers. Two experiments have been performed. The first experiment sets out extracting rules to indicate healthy and sick conditions. The gender of a person has been found to be an important factor influencing heart disease. Second, experiment is so performed to discover rules based on gender.

3. PROBLEM DESCRIPTION

Hospitals, primary care centers, and health organizations need a system to handle a huge number of text medical reports which are produced per day to reduce effort and time in order to extract necessary information and find hidden relationships and discover diseases, medical procedures, and medications for patients. These relations are very important to detect links between diseases, procedures, and drugs and extract useful information from medical reports.

4. PROPOSED IDEA

In this section, proposed Template Based Medical Reports Summarization is presented as seen in Figure 2.

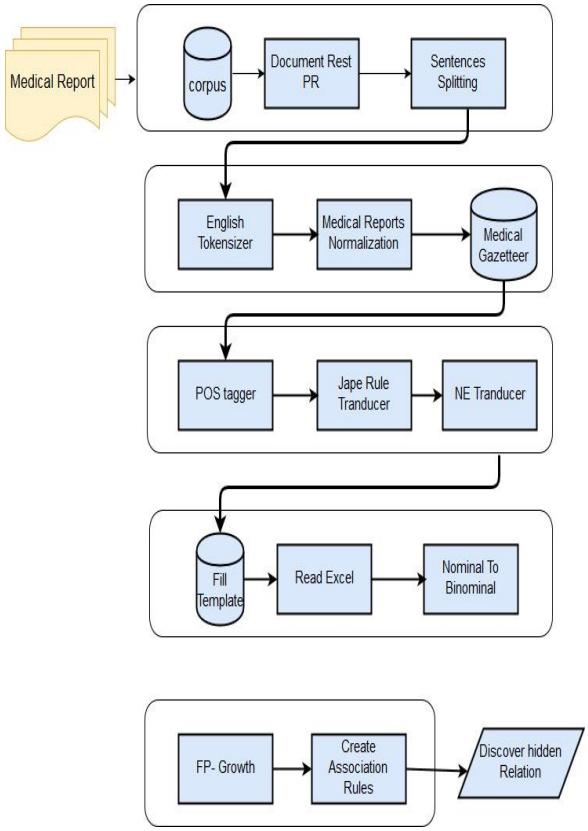

Figure 4: System Architecture

Initial Preparation Stage Architecture

This stage contains four main components: data gathering, data preprocessing, tokenization, normalization. Each component is described as follows:

4.1 Data Acquisition

The first step in our proposed approach is data gathering. The dataset is conducted in order to build a corpus. A corpus used to collect documents in one place and allow run analysis in all documents at the same time. The data set is collected from medical reports from the Palestinian Ministry of Health for the proposed system that contains approximately 2200 medical report of English text. These reports contain a comprehensive summary of the medical condition of the patient from diseases, procedures medical, and medicine, the creation of this report based on the patient's request is used for several purposes, including exterior referral application for treatment or to request for help from the authorities concerned to provide assistance to patients. The data collected from 2009 to 2015, a total of 2300 medical reports were collected.

4.2 Data Preprocessing

To use text mining, data preparation is needed to prepare our data to be ready for applying the mining methods. It aims to transform the medical reports to a form that is suitable for the text and data mining techniques.

4.2.1 Document Reset PR

Used to remove annotations from any previous processing. This is mostly needed for testing to ensure that documents are returned to their initial state before processing [16].

4.2.2 Sentences Splitter

The sentence splitter segments the input text into several sentences. Moreover, the boundaries of the sentence can be recognized by a full stop, punctuation, end of line, etc. As a result of this segmentation the output will be annotations for each sentence and annotations for each boundary.

4.2.3 Tokenization

Is the process of breaking up the text into units called tokens. The tokens may be words or number or punctuation mark. Tokenization does this task by locating word boundaries. The ending point of a word and beginning of the next word is called word boundaries [17]. Generally, tokenization occurs at the word level.

4.2.4 Normalization

Is very important and critical in our paper, it is frequently used when converting text to numbers, dates, acronyms, and abbreviations are non-standard words that need to be pronounced [18]. In medical reports, it is possible to write Diabetes Mellitus type 2 in several different ways:

- DM type II.
- Diabetes mellitus type 2.
- D.M. type II.
- DM type two.

Another example, it is possible to write hypertension disease in several different ways:

- Hypertension.
- HTN.

Therefore, to make the data more consistent, this process is applied. All these forms were adopted.

4.2.5 Part of Speech (POS)

Tagging is the procedure that assigns a category for each word in the text such as noun, pronoun, verb, preposition, adverb, etc. As there are many words that have different meanings based on contextual information. POS information is essential to disambiguate the meaning of words. For any task that involves semantic analysis, assigning POS information to the token words becomes the primary task. However, POS tagger is a basic tool for various applications in NLP field such as information retrieval (IR), information extraction (IE), etc. Moreover, POS tagger is necessary as a tool to build up any language corpus [19, 20].

4.3 The Proposed Template-Based

Before discuss the extract medical information from medical reports, template Based has been identified that will be used in the information extraction process which is consists of three medical attributes, namely: diseases, medical procedures and drugs as shown in Table 1. The reason these attributes are chosen is that these attributes are the most important thing in the report. In addition, most medical reports contain these attributes.

Table 1: Template Based Medical Reports Summarization example

Attributes	Example
Disease	Breast Cancer, Metastatic lymph;
Procedure	Radiation, ACT, Mammography; Axillary clearance, Quadrendectomy;
Drug	Zoladex, Valodex;

4.3 Extract Medical Information Stage

Information extraction stage is the basis of our paper, where our main goal is to extract medical information and then identify and discover hidden relationships between diseases, medical procedures, and drugs from unstructured medical reports. There are many tools and methods on the Internet to extract named entity recognition from text such as OpenNLP [21]. Up to the researcher's knowledge, A few researchers addressed this topic to extract medical information from real medical reports and then do some data mining processes to discover hidden relationships between diseases. However, this stage is implemented using GATE tools to identify extract diseases, medical procedures, and drugs names using two methods: firstly, a predefined list is used or Gazetteers and added new diseases to it. Secondly, many rule-based approaches to develop were adopted.

4.3.1 Gazetteers

The gazetteer consists of lists specific information such as names of cities, organizations, locations etc. These are usually used when the number of instances of a particular class of named entities is finite and could be stored in a database. For example, it is easy to identify the Months' names in the text by referring to an existing list rather than writing complex rules to identify these entities.

4.3.1.1 Gazetteer Normalization

As shown in Table 2 three Gazetteer lists that were used in our paper, two copies was made for each Gazetteer list, the first copy, uppercase letters, second copy, lowercase letters, because the GATE tool does not support case sensitive.

Table 2: Number of records of Gazetteer lists

#	List	# Of Records
1	disease.lst	52000
2	disease_lower.lst	52000
3	procedure.lst	2916
4	procedure_lower.lst	2916
5	drug.lst	11391
6	drug_lower.lst	11391

4.3.2 Rule-based approach

The rule-based approach applies a set of rules is either manually defined or automatically learned. The text is then compared against the rules and a rule is fired if a match is found. A pattern is usually represented as a regular expression to relies on linguistic knowledge in order to extract pattern base for a location name, person name, organization, etc.

When this pattern matches a sequence of tokens, the specified action is fired [22]. An action can be labeling a sequence of tokens as an entity. For example, to label any sequence of tokens of the form "Mr. X" where X is a capitalized word as a person entity, the following rule can be defined:

(token = "Mr." orthography type = *FirstCap*) →person name.
Manually creating the rules for named entity recognition requires human expertise and is labor intensive [22].

Many JAPE rule-based algorithms were implemented using GATE tool to improve nominating the correct medical information from unstructured text. So, the rule base objective is divided into three sections.

4.3.2.1 Rules for Diseases Names Extractor

The JAPE rule-based algorithm was used besides the gazetteers in order to improve information extraction process from medical reports. The main goal of these rules is to discover non-existent diseases in gazetteers used in our program. For example:

- The first rule: Any word comes after the "**chronic"** word shall be considered a disease.
- The second rule: Any word comes before the "**pain"** word shall be considered a disease.
- The third rule: Any word comes before the "**cancer"** word shall be considered a disease.
- The fourth rule: Many diseases share the same suffixes, like **arthritis**, **colitis**, and **bronchitis** all shares a common suffix "**-itis**".
- The fifth rule: Many diseases share the same suffixes, like **adenopathy**, **allopathy**, and **arthropathy** all shares a common suffix "**-pathy**".

4.3.2.2 Rules for Drugs Names Extractor

The JAPE rule-based algorithm was used besides the gazetteers in order to improve information extraction process from medical reports. The main goal of these rules is to discover non-existent drugs in gazetteers used in our program. For example:

- The first rule: Any word or number comes before the regular expression like this "**1x1",** shall be considered a drug.
- The second rule: Any word or number comes before the "**mg", "inh", and "tab"** word shall be considered a drug.

4.3.2.3 Rules for Medical Procedures Names Extractor

JAPE rule-based algorithms were used besides the gazetteers in order to improve information extraction process from medical reports. The main goal of these rules is to discover non-existent medical procedures in gazetteers used in our program. For example:

- Many medical procedures share same prefix or suffix, like Adrenalectomy, Sclerotomy, and Osteotomy all shares a common suffix "**-tomy**".

4.4 Filling Template

After extracting the medical information, the next step dumps all this information that has been extracted from the medical reports in the template is equipped with advance. The goal of this step is to perform some data mining techniques that could benefit the medical staff and data analysts in the Palestinian Ministry of Health. To do this step, JAPE rule has been processed to extract medical information and carried over into the Excel file pre-equipped.

4.5 Create Association Rule

Considered association rules are a common approach to discover information and identify relationships among different items. This approach was used to analyze a large

database containing medical reports data [23]. Our aim is to obtain association rules indicating relationships between diseases, medical procedures, and drugs.

4.7 Expert Evaluation

After generating the association rules, manual evaluation was performed to ensure the accuracy the rules that have been extracted from the data. Rules were classified into seven categories: Cardiothoracic, Thoracic, General Surgery, Neurology, Endocrinology, Urology, and Orthopedic. From three to four doctors have been chosen to carry out the evaluation of the rules in each category that have been extracted from the data in order to measure the rules accuracy. So, Likert's scale was used for evaluation.

4.7.1 Likert's scale

A psychometric response scale primarily used in questionnaires to obtain participant's preferences or degree of agreement with a statement or set of statements. Likert scales are a non-comparative scaling technique and are unidimensional (only measure a single trait) in nature. Respondents are asked to indicate their level of agreement with a given statement by way of an ordinal scale (Bertram).

Since Likert's scale of 5 point was used which would result in the interval from (1) to (5) was distributed into (5) interval, each interval had a length of ((5-1)/5) = 0.8. Therefore, for the average (mean) score the intervals were defined as:

Factors scoring in average of 3.40 or more shall be considered as high importance [24].

5. EXPERIMENTS AND RESULTS

In this section, the experimental results were presented and analyzed to provide evidence that our approach can identify medical information such as diseases, medical procedures, and drugs from medical reports. Also, it illustrates the association rules that have been extracted from medical information, then it discusses some of these the rules. Finally; the association rules that have been extracted to discover the hidden relation between medical were evaluated.

Table 3: Likert Scale

Degree of Agreement	From	To
Very low	1.00	1.79
Low	1.80	2.59
Medium	2.60	3.39
High	3.40	4.19
Very high	4.20	5.00

5.1 Medical Reports Corpus

Real medical reports were used as a source of the corpus, where we got medical reports from the Palestinian Ministry of Health, in particular from the Shifa Medical Complex. The data is about 2200 medical reports. Each report contains an overview of the status of the patient, diseases, symptoms, medical procedures, drugs, and other information.

Medical reports are comprehensive reports; include most medical departments such as Cardiothoracic, Thoracic, General Surgery, Neurology, Orthopedic, Urology and Endocrinology and so on. The data used targeting patients from the Gaza strip only, and it a new somewhat from the year 2009 – 2016. The average size of medical report per word is about 60 words/report. Table 4 is shown number of medical records.

Table 4: Number of records from each department

Medical departments	# Of Records
Cardiothoracic	550
Thoracic	370
General Surgery	280
Neurology	400
Orthopedic	220
Urology	235
Endocrinology	145

5.2 Data Preprocessing Stage

GATE Developer tools have a collection of operation that is suitable for text mining. In this phase, medical report corpus is prepared to make them standardized format for the text mining process. However, a number of preprocessing techniques were applied to deal with noisy, missing, and inconsistent data. There are many of preprocessing techniques such as: sentences splitter, document normalization, tokenization and part of speech tagger. For more details about in Figure 3 show preprocessing methods used in our system using GATE tools.

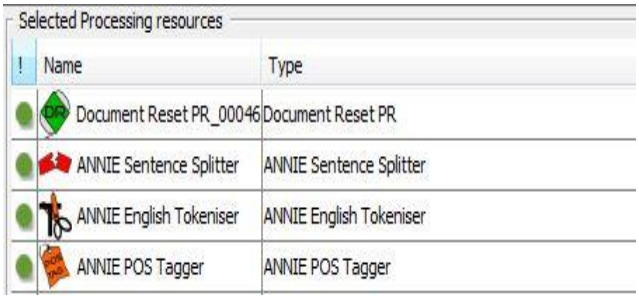

Figure 3: Preprocessing techniques

5.3 Name Entity Recognition

Most researchers in NLP use GATE to create their own programs and pipelines. GATE comes with pre-load plugins handle many fields and Multilanguage. In this phase, ANNIE application (A Nearly-New Information Extraction system) used to tag previous medical reports corpus with named entities to extract medical information from medical reports.

JAPE Rule (Extrac_to_template)

JAPE rule was created to convert JAPE annotation to excel file. This rule is used at the end of our work in the GATE program. It works to extract information that has been extracted from the medical reports and dump it into an external file often with .txt extension. After this process, txt file to file Excel was converted to data dump in a predefined template as example shown in Table 5. In order to use this data exists in the Excel file in the process of data mining and discover interesting and hidden patterns in the data. In this JAPE rule, some constraints were inserted in exporting medical information out of GATE tools to satisfy our goal in discovery hidden relationship and to get the best knowledge.

Table 5: Template Based Medical Reports Summarization example

Entity Class	Example
Disease	Breast Cancer, Metastatic lymph;
Procedure	Radiation, ACT, Mammography; Axillary clearance, Quadrendectomy;
Drag	Zoladex,Valodex;

5.4 Association Rules Results

Using statistical tools and modeling techniques, one can discover interesting and hidden patterns in the data. These patterns may not be easily detected using traditional methods. Therefore, next step in the experiment is to use the association rules to reveal relationships among different medical information and identify the indirect relationship between different medical information and discover the hidden relation between individual.

However, not all the association rules that have been extracted from the dataset due to paper limits were shown. In this part, diseases can be considered as an antecedent item set, and the other diseases can be considered as a consequent item set is as shown in Table 6. Medical procedures can be considered as an antecedent item set, and the other diseases can be considered as a consequent item set is shows the 12 association rules between medical procedures (antecedent) and their associated diseases (consequent), and drugs can be considered as an antecedent item set, and the diseases can be considered as a consequent item set is shows the 12 association rules between drugs (antecedent) and their associated diseases (consequent). The minimum support value is usually determined by the users.

5.4.1 Diseases → Diseases

Table 5 shows the 12 association rules between diseases (antecedent) and their associated other diseases (consequent), some of these rules are:

Rule No. 1 says that if the patient has a "Sclerosing Cholangitis", and "Esophageal Varices" then he has a high probability of having "Cirrhosis" disease. The number of patients for this rule is low, but when conditions hold the disease probability will be high.

Rule No. 2 if a person has "Recurrent Chest Infection" then it is almost sure that person has "Shortness of breath (SOB)", and "dyspnea".

Rule No. 3 relates "Lower Respiratory Tract Infection" disease with a chance of having a "Cystic Fibrosis" disease. Observing was concluded that according to medical knowledge the "Cystic Fibrosis" has a higher chance of being diseased than the other lung disease. As can be seen the rules that involve the lung disease confirm this fact since they have higher support almost 100% confidence.

Rule No. 4 shows that the disease "Cirrhosis" him a direct relationship with "Congestive Heart Failure (CHF)" disease and "hypertension (HTN)".

Rule No.5 also patients who are diagnosed "Parasthesia" disease usually suffer from "Lower Back Pain" disease.

Rule No. 6 show that people who are infected with "Hemiplegia", and "Hypertensive" are more people likely to suffer from "Cerebrovascular Accident (CVA)".

Table 6: Association Rules Obtained - Diseases to Diseases

NO.	Association Rule	Antecedent	Consequent
1	{Sclerosing Cholangitis, Esophageal Varices} → Cirrhosis	Sclerosing Cholangitis, Esophageal Varices	Cirrhosis
2	{Recurrent Chest Infection} → Shortness of Breath (SOB), Dyspnea	Cirrhosis	Shortness of Breath (SOB), Dyspnea
3	{Lower respiratory Tract Infection} → Cystic fibrosis	Lower respiratory Tract Infection	Cystic fibrosis
4	{Cirrhosis} → Congestive Heart Failure (CHF), Hypertension (HTN)	Cirrhosis	Congestive Heart Failure (CHF), Hypertension (HTN)
5	{Low back pain} → Parasthesia	Low back pain	Parasthesia
6	{Hemiparesis, Hypertensive} → Cerebrovascular Accident (CVA)	Hemiparesis, Hypertensive	Cerebrovascular Accident (CVA)

5.5 System Subjective Evaluation

System evaluation is a hard task especially in the field of text and data mining. To ensure that the system works well with association rules, human expert is used as a reference to measure the degree of acceptance of the association rules which have been extracted from the dataset. So, Likert's scale is used for evaluation.

Association rules are divided into three sections based on medical information extracted such as: diseases, procedures, and drugs. Where each of these sections has been classified into several departments medical, such as Cardiothoracic, Thoracic, General Surgery, Neurology, Orthopedic, Urology and Endocrinology. 25 doctors were selected from the ministry of health from Shifa Medical Hospital in particular - from different departments such as Cardiology Department, Neurosurgery, Orthopedics and other departments. Three doctors from each department were selected to fill out the questionnaire manually and determine the degree of acceptance the rules that have been extracted.

After the data analysis obtained from the questionnaire. It shows us that the proportion of accuracy association rules, which have been extracted it is about 80%, as shown in Table 7.

The best results have been noted in the Department of Neurology and followed by thoracic section. Also, the best

results were noted at the level of diseases and followed by medical procedures and finally drugs.

It is concluded that some names of drugs that have been extracted from the medical reports had been written by the brand name and not a medical name, this is the effect on the process of extracting information correctly from the medical reports.

Table 7: The results of expert evaluation for association rules

Department	Diseases	Procedures	Drugs	Average
Thoracic	84.7%	83%	74.5%	80.7%
Cardiothoracic	76.6%	82%	80%	79.5%
Neurology	95%	89%	80%	88%
Endocrinology	80%	74%	66%	73.3%
Surgery	90%	80%	70%	80%
Urology	96.6%	68%	64%	76.2%
Orthopedic	80%	83%	91%	84.6%
Average	86.1%	79.8%	75%	80.3%

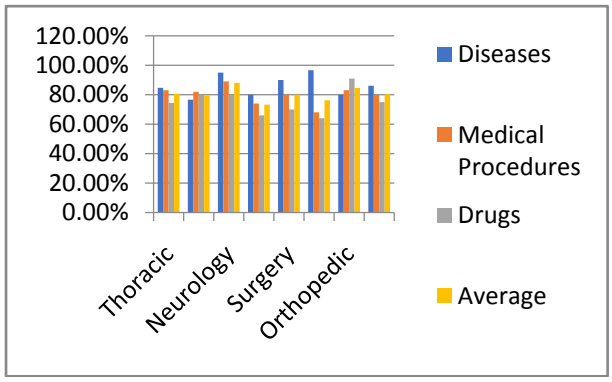

Figure 4: The results of expert evaluation for association rules

6. CONCLUSION AND FUTURE WORK

Summary

Text mining play a vital role in information extraction, where used to extract particular information form unstructured text. This information may discover a new knowledge and help in making decision. The important of this field has been growing because difficult mining a great data is stored as free text. The abundance of medical records has increased the amount of data available in hospitals, primary care centers and health organizations. There is an urgent need for intelligent tools to deal with such data.

A theoretical foundation for this research was presented. Then the Text summarization (TS) and Template Based Summarization (TBS) were discussed. Then, the Information Extraction (IE) was defined and explained the basic tasks that must follow to build information extraction system.

In this paper, a new approach to extract important information was created and detect diseases, drugs, and medical procedures using text and data mining techniques. Ministry of health dataset was used in this work. All of them came from the previous medical reports in the period from 2009 to 2016. The data set included 2200 records.

The approach consists of several stages: preparing the corpus, Extraction of Medical Information, Fill the Templates, and Create Association Rules.

The approach has the ability to achieve the following task:

- Extract diseases, procedures, and drugs from medical reports.
- The system able to discover the unlimited hidden relationship between information medical.

Explained the experimental setup was presented the corpus characteristics, and data preprocessing stage and implementation of the Name Entity Recognition (NER) using GATE tools. Also, it predicts the hidden relationships between medical information by association rules. Finally, we presented the results of association rules and degree of acceptance.

According to questionnaire results, we found that the proportion of accuracy association rules, which have been extracted it are the 80%.

Good number of rules was not known to doctors, so some doctors have evaluated them (normal, unacceptable) and after searching for the accuracy of these rules it found that some of these rules very accurate.

Recommendations

There are some of the recommendations can be formulated to adopt the goal of this paper, like the following:

- The hospitals, primary care centers and health organizations should computerize the all medical reports to extract more knowledge leads to help in improving the health service.
- The top management in health centers should support information technology field and developing systems used text and data mining process and artificial intelligent to help the medical staff to discover and predict diseases and also improve medical services.
- Circulate the association rules in their respective sections to help doctors link certain diseases to other related diseases, diseases related to certain medical procedures or linking certain diseases to certain drugs.
- A good number of diseases relationships have been discovered with other diseases that have not been known to the doctors who have evaluated.
- Do more research in this field to use machine learning to extract medical information in order to enhance the accuracy of the system. In addition to use medical records with Arabic language.

Future Work

According to the results of experiment and the limitations that we faced in our paper, this work can be improved in multiple directions:

- Use machine learning to extract medical information in order to enhance the accuracy of the system.
- Expand the circle of extracted information such as finding, substance, situation etc. from the existing data that can help us improve and enhance the

discovery and prediction of diseases, medicines and medical procedures for patients.

- Use medical records with Arabic language or mixed Arabic and English.
- Extending our approach to work on extract information from images such as X-ray.
- Use medical records with hand written after handwriting recognition.
- Contribution of building ontology to improve the process of extracting medical information from medical records.
- Use other techniques for data mining such as clustering and outlier analysis.

7. ACKNOWLEDGMENTS

This research was supported by Qatar Charity under Ibhath project for research grants, which is funded by the Cooperation Council for the Arab States of the Gulf throughout Islamic Development Bank.

8. REFERENCES

[1] lahari, E., Kumar, D., and Ubale, M. 2014. A Comprehensive Survey on Feature Extraction in Text Summarization. Int.j. computer Technonlogy & Application, Vol 5 (1), 248-256.

[2] Radev, D. and Mckeown, K. 2002. Introduction to the special issue on summarization. Computational Linguistics, 28 (4), 339 – 408.

[3] Mani, I., and Maybury, M. 2001. Advaces in Automatic Text Summarization. The MIT Cambridge, Massachusetts london, England.

[4] Bunescu, M., Ge R., Mooney, R., Marcotte, E., and Ramani, A. 2002. Extracting Gene and Protein Names from Biomedical Abstracts. Unpublished Technical Note.

[5] Jung, J., and Jo, G. 2003. Template-Based E-mail Summarization for Wireless Devices. computer and information sciences - ISCIS, LNCS 2869, 99–106.

[6] Deléger, L., Grouin, C., & Zweigenbaum, P. 2010. Extracting medical information from narrative patient records: the case of medication-related information. J Am Med Inform Assoc, 17(5): 555–558.

[7] Jonnagaddala, J., Liaw, S., Ray, P., Kumar, M., Chang, N., and Dai, H. 2015. Coronary artery disease risk assessment from unstructured electronic health records using text mining. Journal of Biomedical Informatics 58, S203–S210.

[8] Chen, A., and Verma, R. 2006. A query-based medical Information summarization system Using Ontology Knowledge. In the proceedings of the 19th IEEE Symposium on Computer based Medical Systems.

[9] Sarkar, K. 2009. Using Domain Knowledge for Text Summarization in Medical Domain. International Journal of Recent Trends in Engineering, 1 (1).

[10] Xu, H. Stenner, S., Doan, S., Johnson, K., Waitman, L., & Denny, J. 2010. MedEx: a medication information extraction system for clinical narratives. Journal of the American Medical of Informatics Associations, 17(1):19-24.

[11] Gold, S., Elhadad, N., Zhu, X., Cimino, J., & Hripcsak, G. 2008. Extracting Structured Medication Event Information from Discharge Summaries. AMIA Annual Sumposium Proceeding Archive., 237–241.

[12] Doddi, S., Marathe, A., Ravi, S., & Torney, D. 2001. Discovery of Association Rules in Medical Data. Med. Inform. Internet. Med., 26, 25–33.

[13] Rashid, M., Hoque, M., & Sattar, S. 2014. Association Rules Mining Based Clinical Observations. Bioinformation, 9(11): 555–559.

[14] Ordonez, C., Omiecinski, E., Braal, L., Santana, C., Ezquerra, N., Taboada, J., rawczynska, E. 2001. Mining Constrained Association Rules to Predict Heart Disease. Proceeding ICDM '01 Proceedings of the 2001 IEEE International Conference on Data Mining, 433-440.

[15] Nahar, J., Imam, T., Tickle, K., & Chen, Y. 2013. Association rule mining to detect factors which contribute to heart disease in males and females. Expert Systems with Applications 40, 1086–1093.

[16] Greenwood, M., Roberts, A., Aswani, N., & Gooch, P. 2012. Initial prototype for semantic annotation of the Khresmoi literature. project deliverable Khresmoi.

[17] Spasić, I., Zhao, B., Jones, C., & Button, K. 2015. KneeTex: an ontology–driven system for information extraction from MRI reports. Journal of Biomedical Semantics.

[18] Sproat, R., Black, A., Chen, S., Kumar, S., Ostendorfk, M., & Richards, C. 2001. Normalization of non-standard words. Computer Speech and Language 15, 287–333.

[19] Elsebai, A. 2009. A rules based system for named entity recognition in modern standard. University of Salford.

[20] Mukund, S., Srihari, R., & Peterson, E. 2010. An Information-Extraction System for Urdu—A Resource-Poor Language. ACM Transactions on Asian Language Information Processing, 9 (4).

[21] Al-Zaidy, R., Fung, B., Youssef, A., & Fortin, F. 2012. Mining criminal networks from unstructured text documents. Digital Investigation, 8 (3-4), 147-160.

[22] Jiang, J. 2012. Information extraction from text. C.C. Aggarwal, C. Zhai (Eds.), Mining text data, Springer, 11–41.

[23] Doddi, S., Marathe, A., Ravi, S., & Torney, D. 2001. Discovery of Association Rules in Medical Data. Med. Inform. Internet. Med., 26, 25–33.

[24] Ozen, G., Yaman, M., & Acar, G. 2012. Determination of the employment status of graduates of recreation department. The Online Journal of Recreation and Sport, 1 (2).