The Islamic University – Gaza
Research and Postgraduate Affairs
Faculty of Information Technology
Master of information technology

الجامعة الإسلامية – غزة
شئون البحث العلمي والدراسات العليا
كلية تكنولوجيا المعلومات
ماجستير تكنولوجيا المعلومات

Template Based Medical Reports Summarization

(Ministry of Health Case Study)

تلخيص التقارير الطبية معتمدة على قوالب وزارة الصحة دراسة حالة

Ahmed Y. Abu El-Qumsan

Supervised by

Dr. Alaa EL-Halees

Prof. of Computer Science

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of information technology

June / 2017

Abstract

The torrential information in the medical records is considered a great problem because it

إقــــــرار

أنا الموقع أدناه مقدم الرسالة التي تحمل العنوان:

Template Based Medical Reports Summarization

(Ministry of Health Case Study)

تلخيص التقارير الطبية معتمدة على قوالب وزارة الصحة دراسة حالة

أقر بأن ما اشتملت عليه هذه الرسالة إنما هو نتاج جهدي الخاص، باستثناء ما تمت الإشارة إليه حيثما ورد، وأن هذه الرسالة ككل أو أي جزء منها لم يقدم من قبل الآخرين لنيل درجة أو لقب علمي أو بحثي لدى أي مؤسسة تعليمية أو بحثية أخرى. وأن حقوق النشر محفوظة للجامعة الإسلامية غزة –فلسطين

Declaration

I hereby certify that this submission is the result of my own work, except where otherwise acknowledged, and that this thesis (or any part of it) has not been submitted for a higher degree or quantification to any other university or institution. All copyrights are reserves toIslamic University – Gaza strip paiestine

Student's name:	أحمد يونس أبو القمصان	اسم الطالب:
Signature:	أحمد أبو القمصان	التوقيع:
Date:	2017/7/22	التاريخ:

بسم الله الرحمن الرحيم

الجامعة الإسلامية-غزة

The Islamic University of Gaza

مكتب نائب الرئيس للبحث العلمي والدراسات العليا **هاتف داخلي:** 1150

الرقم: ج س غ/35/ Ref:

التاريخ: 2017/06/14 Date:

نتيجة الحكم على أطروحة ماجستير

بناءً على موافقة شئون البحث العلمي والدراسات العليا بالجامعة الإسلامية بغزة على تشكيل لجنة الحكم على أطروحة الباحث/ **احمد يونس محمد ابوالقمصان** لنيل درجة الماجستير في **كلية *تكنولوجيا المعلومات*** برنامج تكنولوجيا المعلومات وموضوعها:

تلخيص التقارير الطبية معتمدة على قوالب وزارة الصحة دراسة حالة

Template Based Medical Reports Summarization Ministry Of Health Case Study

وبعد المناقشة التي تمت اليوم الأربعاء 19 رمضان 1438هـ، الموافق 2017/06/14م الساعة الحادية عشر صباحاً ، اجتمعت لجنة الحكم على الأطروحة والمكونة من:

أ.د. علاء مصطفى الهليس	مشرفاً و رئيساً	....
د. أشرف يونس مغاري	مناقشاً داخلياً	....
د. محمد أحمد غزال	مناقشاً خارجياً	....

وبعد المداولة أوصت اللجنة بمنح الباحث درجة الماجستير في **كلية *تكنولوجيا المعلومات*/** برنامج تكنولوجيا المعلومات.

واللجنة إذ تمنحه هذه الدرجة فإنها توصيه بتقوى الله ولزوم طاعته وأن يسخر علمه في خدمة دينه ووطنه.

والله ولي التوفيق ،،،

نائب الرئيس لشئون البحث العلمي والدراسات العليا

أ.د. عبدالرؤوف علي المناعمة

شئون البحث العلمي والدراسات العليا
الجامعة الإسلامية غزة
The Islamic University
Research & Graduate Affairs

Abstract

The torrential information in the medical records is considered a great problem because it is 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 topic that aims 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 research, we propose to develop an approach that use template based medical reports summarization 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, we can discovery hidden relationship between medical information by using association rules. The dataset we used in this research was collected from the Palestinian Ministry of Health.

To evaluate the performance and effectiveness of our extracted association rules, we used human experts as a reference to measure the degree of acceptance of the extracted association rules which have been extracted from the dataset to assess the accuracy of information extracted from medical reports. So, we used Likert's scale 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%.

Keywords: *Template based summarization, Medical Report, Information Extraction, Named Entity Recognition, Text Mining, Association Rules.*

الملخص

تعتبر المعلومات الغزيرة في السجلات الطبية مشكلة كبيرة بسبب صعوبة التمييز بين المعلومات اللازمة والضرورية من الكم الهائل من البيانات.

ونتيجة لذلك، تتزايد أهمية تلخيص التقارير الطبية يوما بعد يوم. استخراج المعلومات الطبية هي واحدة من المواضيع الساخنة في الابحاث العلمية التي تهدف إلى التعرف على المعلومات الطبية والكشف عن العلاقات الخفية بين هذه المعلومات. ويعتبر هذا الموضوع واحد من أهم المواضيع في مجال استخراج النصوص حيث يستخدم لمعالجة النصوص غير منتظمة واستخلاص المعلومات المفيدة التي كانت مخبأة في النصوص غير منتظمة. المعلومات المستخرجة من التقارير الطبية مفيدة جدا للعاملين في المجال الطبي وذلك للكشف عن العلاقات الخفية بين المعلومات الطبية، واتخاذ القرارات التي من شأنها تحسين الخدمات الطبية للمرضى، بالإضافة إلى توفير الوقت والجهد.

في بحثنا، نقترح تطوير نهج لتلخيص التقارير الطبية المعتمد على قالب لتحويل التقارير الطبية من شكل شبه منظم وغير منظم إلى شكل منظم، وتصنيف الكيانات المحددة واستخراج المعلومات الهامة مثل الأمراض والإجراءات الطبية، والأدوية. بعد ذلك، اكتشاف علاقة خفية جديدة بين المعلومات الطبية من خلال قواعد الترابط.
تم جمع مجموعة البيانات المستخدمة في هذا البحث من وزارة الصحة الفلسطينية.

لتقييم أداء وفعالية النموذج المقترح، استخدمنا الخبير البشري كمرجع لقياس درجة قبول قواعد الترابط التي تم استخراجها من مجموعة البيانات. لذلك، استخدمنا مقياس ليكرت للتقييم.
بعد تحليل البيانات التي تم الحصول عليها من الاستبيان. فإنه يدل لنا أن نسبة دقة قواعد الترابط، والتي تم استخراجها هو حوالي 80٪.

الكلمات المفتاحية: التلخيص المعتمد على القالب، السجلات الطبية، استخراج المعلومات، التنقيب عن النص، قواعد الترابط، التعرف على الكيانات المسماة.

بِسْمِ اللَّهِ الرَّحْمَٰنِ الرَّحِيمِ

﴿٣١﴾ قَالُوا۟

سُبْحَٰنَكَ لَا عِلْمَ لَنَآ إِلَّا مَا عَلَّمْتَنَآ ۖ إِنَّكَ أَنتَ ٱلْعَلِيمُ ٱلْحَكِيمُ

﴿٣٢﴾

صَدَقَ اللَّهُ الْعَظِيمُ

Dedication

To my beloved parents

To my brothers and sisters

To those who gave me support

To all of them I dedicate this work

Acknowledgements

First of all, I thank Allah for all knowledge and education I gain which leads to achievement of this thesis.

Second, I would like to thank my advisor **Dr. Alaa EL-Halees** for his continued encouragement, unlimited efforts, persistent motivation, support, and great knowledge throughout my thesis, without his help, guidance, and follow-up, this thesis would never have been.

I would also like to thank my parents, and my brothers and sisters for their constant support and encouragement.

Thank for all the colleagues in the Ministry of Health who helped me to accomplish this research, especially the engineers of the IT unit.

Thanks to anyone who participated in all the achievement of this thesis either directly or indirectly.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

TM	Text Mining
IE	Information Extraction
MR	Medical Report
TS	Text Summarization
TBS	Template Based Summarization
NLP	Natural Language Processing
NER	Named Entity Recognition
NE	Named Entity
CO	CO-reference resolution
GATE	General Architecture for Text Engineering
ANNIE	A Nearly-New Information Extraction system
JAPE	Java Annotation Patterns Engine
PR	Processing Resources
LHS	Left hand side
RHS	Right hand side
POS	Part-of-Speech
CSCP	Clinical State Correlation Prediction
OLTP	Online Transaction Processing
SL	Supervised Learning
SSL	Semi-supervised learning
UL	Unsupervised learning

Chapter 1
Introduction

Chapter 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 (lahari et al., 2014). To overcome this problem text summarization can be used.

Automatic text summarization is maturing and may provide a solution to the information overload problem. In addition, a very powerful tool to save time and resources, and optimize availability for an expert in any domain area (lahari et al., 2014).

Radev and Mckeown (2002) defined 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*".

The rapid growth 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.

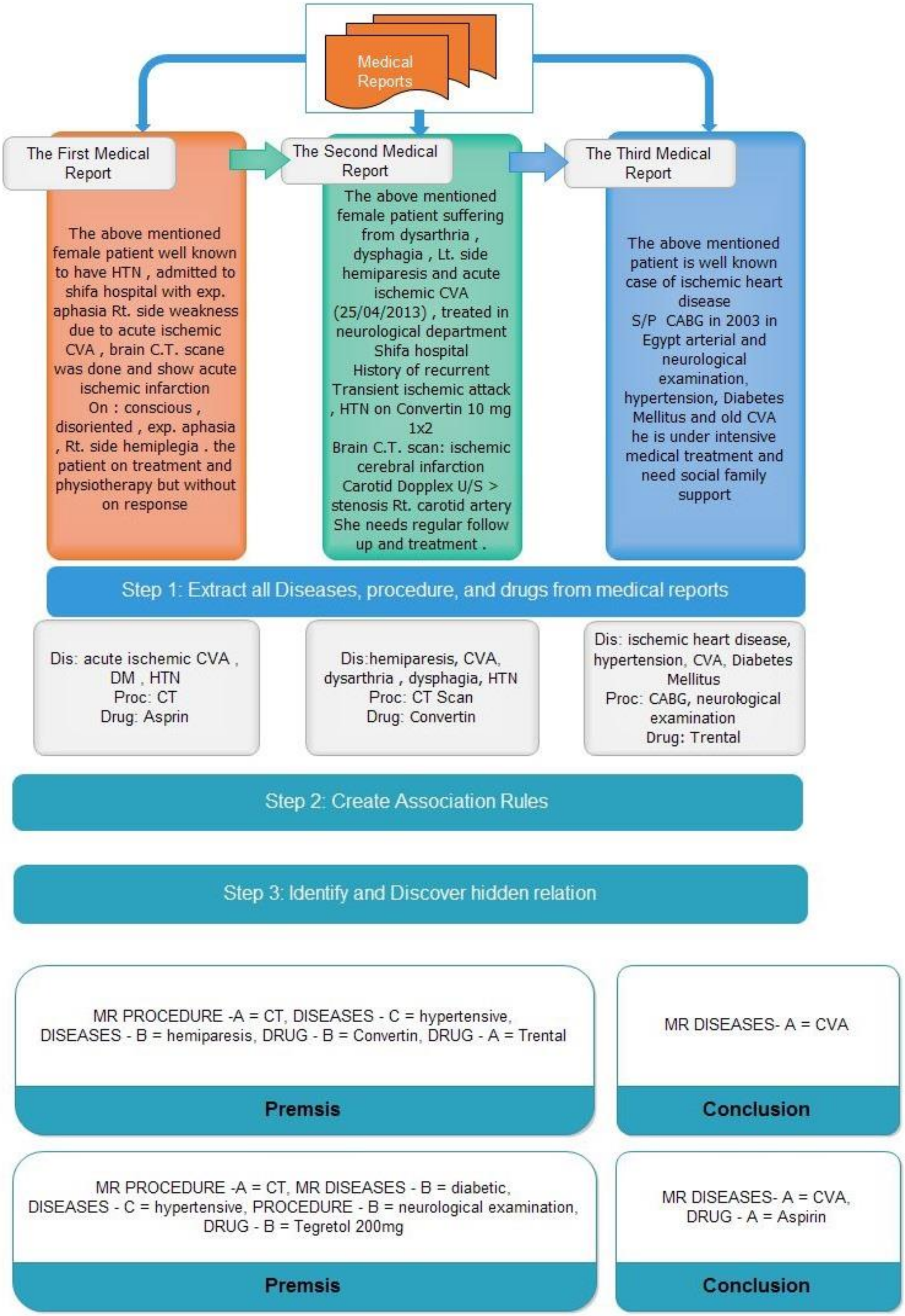
The main purpose of this thesis is to convert text medical records to structured form using template based summarization. After that this thesis create a new approach for extract important information and detect and discover diseases, drugs, and medical procedures using unstructured text and data mining techniques. The focus on solving the problem of how to discover relations between medical information from unstructured reports and extract useful information using text mining methods. This model will be used 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 research 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 drugs and procedures. For example, suppose in Figure 1.1 we have 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 a number of diagnosis of acute ischemic CVA, DM, and HTN, a medical procedure is CT, the drug Aspirin. The second report contains a number of diagnoses, such as hemiparesis, CVA, dysarthria, dysphagia, and HTN, medical procedures is CT Scan, the drug is Convertin. The third report contains the ischemic heart disease, hypertension, CVA, and Diabetes Mellitus, the medical procedures such as CABG, neurological examination, the drug is Trental.

We note that for the same disease (CVA), there are different diagnoses and different procedures and different drugs.

In our research, we bind these reports and information extracted that will help doctors and medical staff identify and discover diseases and medical procedures and drugs for patients.



Medical Reports

The First Medical Report

The above mentioned female patient well known to have HTN , admitted to shifa hospital with exp. aphasia Rt. side weakness due to acute ischemic CVA , brain C.T. scane was done and show acute ischemic infarction On : conscious , disoriented , exp. aphasia , Rt. side hemiplegia . the patient on treatment and physiotherapy but without on response

The Second Medical Report

The above mentioned female patient suffering from dysarthria , dysphagia , Lt. side hemiparesis and acute ischemic CVA (25/04/2013) , treated in neurological department Shifa hospital History of recurrent Transient ischemic attack , HTN on Convertin 10 mg 1x2 Brain C.T. scan: ischemic cerebral infarction Carotid Dopplex U/S > stenosis Rt. carotid artery She needs regular follow up and treatment .

The Third Medical Report

The above mentioned patient is well known case of ischemic heart disease S/P CABG in 2003 in Egypt arterial and neurological examination, hypertension, Diabetes Mellitus and old CVA he is under intensive medical treatment and need social family support

Step 1: Extract all Diseases, procedure, and drugs from medical reports

Dis: acute ischemic CVA , DM , HTN
Proc: CT
Drug: Asprin

Dis:hemiparesis, CVA, dysarthria , dysphagia, HTN
Proc: CT Scan
Drug: Convertin

Dis: ischemic heart disease, hypertension, CVA, Diabetes Mellitus
Proc: CABG, neurological examination
Drug: Trental

Step 2: Create Association Rules

Step 3: Identify and Discover hidden relation

Premsis	Conclusion
MR PROCEDURE -A = CT, DISEASES - C = hypertensive, DISEASES - B = hemiparesis, DRUG - B = Convertin, DRUG - A = Trental	MR DISEASES- A = CVA
MR PROCEDURE -A = CT, MR DISEASES - B = diabetic, DISEASES - C = hypertensive, PROCEDURE - B = neurological examination, DRUG - B = Tegretol 200mg	MR DISEASES- A = CVA, DRUG - A = Aspirin

Figure (1.1): Example of relationship between medical reports

In the thesis approach, three steps approach have been proposed to detect and predict disease and medical procedures and drugs for patients is suggested. In the first step, we extracted medical information from medical reports include diseases and medical procedures and drugs. In the second step, we created the association rules of the data extracted from the medical reports. In the final step, we discovered hidden relationships between the medical reports and the analysis and prediction of diseases and medical procedures and drugs for patients. The experimental results will demonstrate the effectiveness of the proposed approach.

1.1 Problem Statement

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. In addition, find hidden relationships and discover diseases, medical procedures, and drugs for patients. These relations are very important to detect links between the medical information extracted.

1.2 Objectives

1.2.1 Main Objective

The main objective of this thesis is to develop an efficient approach template based medical reports summarization to classify the entities and extract important information from unstructured medical reports. This helps to identify direct and indirect relationships between medical reports items in order to assist medical staff to detect and discover diseases, medical procedures, and drugs necessary for patients.

1.2.2 Specific Objectives

The specific objectives of the thesis are to:

- Search for real data medical reports from government hospitals.
- Perform preprocessing phases and classification efficiently.
- Find a suitable algorithm to extract medical named entities from

unstructured text.

- Analyze medical reports and trace patterns of behaviors which help uncover relationships between seemingly unrelated data in a relational database.
- Evaluate the association rules extracted using human experts.

1.3 Significance of the Thesis

1. It focuses on the medical domain in the technology area because it contains challenging tasks such as medical entity recognition, abbreviation disambiguation, inter-conceptual relationship detection, anaphora resolution, and text summarization.
2. Saving efforts and time by helping the users to find which useful information fits with his/her interest based on the produced summary by the template.
3. Increase friendly use of a computer so make the computer more intelligent.
4. Help data analysts, statisticians, and staff medical to generate a template for a medical report.
5. Detect and predict diseases, medical procedures, and drugs, extract relations from medical reports.

The topic of information extraction in medical domain little work have been investigated, on the other side, other domains such as newswire text, which has been researched heavily by researchers.

Challenges of this Thesis

There are several challenges in this area:

- Data set is written in without grammar rules and unstructured English language.
- Most of the medical reports ignore capitalization and tokenization.
- Collecting sensitive data about medical reports from government hospitals in Gaza strip.
- Due to the sensitive nature of real medical datasets, they are not easily available

for academic research because they involve problems and difficulties.

1.4 Scope and Limitations

- The work focuses on extracts automatic text summarization by template.
- We deal with single medical report summarization.
- We deal with English medical report summarization.
- We deal with the medical reports that have been produced from the government hospitals.
- Medical reports are comprehensive reports, include all the medical departments such as reports reception and emergency, reports surgical ward, reports pediatrics, reports obstetrics, and gynecology and so on.
- The data used in this research are from year 2009 - 2016.
- The data used targeting patients from the Gaza strip only.
- We will not consider post-processing summarization.

1.5 Methodology

The research methodology, as seen in Figure 1.2, employed in this thesis is described and summarized in the following points:

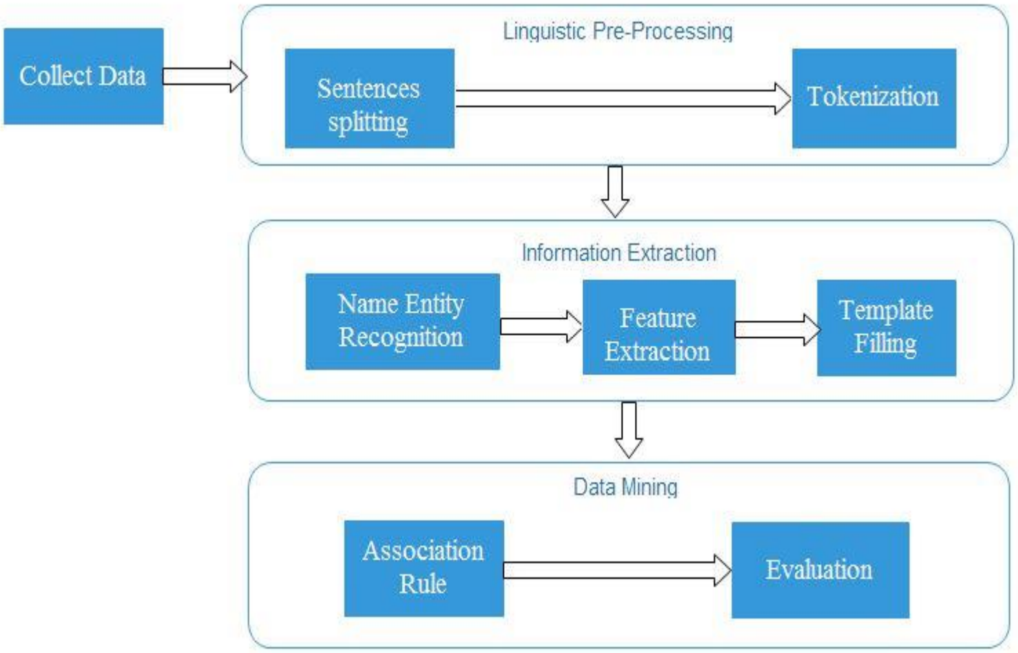


Figure (1.2): The Research Methodology

1. Collect Data: The dataset used in this research from the Palestinian Ministry of Health in Gaza strip, which is estimated at 2200 medical report. The report contains an overview of the status of the patient and symptoms, diagnosis, drugs, and other information.

2. **Linguistic Per-processing:** For each sentence in the medical report, a **pre-processing** operation is applied as:
   a. **Sentences splitting**: is the core of many NLP activities.
   b. **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 (Spasić et al., 2015).

3. **Name Entity Recognition:** It is considered the main step to the process of extracting information that aims to locate and classify elements in medical reports

into pre-defined categories such as the names of diseases, procedures, finding, etc., consists of several operations (Goel and Yadav, 2016):

3.1 Feature Extraction: A group of features was extracted from each word in the medical report and assign a score for each feature, these features as (Wang and Patrick, 2009):

- ➢ Orthographic Features: A set of conventions for writing a language. It includes rules of spelling, hyphenation, capitalization, word breaks, emphasis, and punctuation.
- ➢ Affixes: is part of word formation process in the English language, consists of prefixes and suffixes (Wangand Patrick, 2009).
- ➢ Abbreviations and Acronyms: The abbreviation lists were constructed from two resources: abbreviations from SNOMED CT terminology, abbreviations & acronyms from the hospital (Wangand Patrick, 2009).
- ➢ POS Features: The use of POS helps to determine the boundaries of named entities (Wangand Patrick, 2009).

4. **Template Filling:** is a final step in the extraction of information from the medical records and fill out the information in the template.
5. **Create Association Rules:** It will help to detect, analyze, and predict for diseases and medical procedures and drugs for patients.
6. **Evaluation:** evaluate the effectiveness of the proposed approach, we measure the precision of both direct and indirect relation between diseases, medical procedures, and drugs. By manually expecting the result where obtained from experiment data set; where expert human can verify the results and evaluate the accuracy of the association rules.

1.6 Thesis Format

The thesis consists of six chapters organized as follows:

Chapter 2: Discusses the state of the art and literature survey.

Chapter 3: Theoretical foundation of the thesis.

Chapter 4: Presents the proposed Template Based Medical Reports Summarization approach.

Chapter 5: presents the experimental results and evaluation.

Chapter 6: presents the conclusions, recommendation and future work.

Chapter 2

Related Works

Chapter 2
Related Works

Many researchers gave a great attention to text summarization, but a very few of them proposed 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 chapter 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. (2002) 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 at precision was 93%, and at recall was 82%.

Jung, et al. (2010) developed method automated medication extraction system, which can accurately **extract medication names** and signatures from medical records. The proposed system consists of three main steps to 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.

Deléger, et al. (2010) 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).

Chun, et al. (2006) built system that is used to extract relation between the disease and gene names from MedLine abstracts. They constructed a dictionary for disease and gene names from six public databases and extracted relation candidates by dictionary matching. But dictionary matching produces a large number of results not accurate. To improve the precision of recognizing disease and gene names, they used maximum entropy model to filter out results incorrect. Maximum entropy models exhibited the best performance in the CoNLL-2003 Shared Task of NER, and are widely used in classification problems in natural language processing. Therefore, the researchers have building an annotated corpus is to construct the training data for machine learning that filter out results incorrect from the dictionary-based results. The researchers identified three types of relations between the disease and gene. First, Pathophysiology, or the mechanisms of diseases, containing etiology, or the causes of diseases. Second, Therapeutic significance of the genes or the gene products, more specifically classified to their therapeutic use and their potential as therapeutic targets. Final, the used the genes and the gene products as markers for the disease risk, diagnosis, and prognosis. The results showed the performance of relation extraction is heavily dependent upon the performance of NER filtering and that the filtering improves the precision of relation extraction by 26.7% at the cost of a small reduction in recall. But the limitation of his study is not address the problem of polysemous terms, which would cause difficulty in linking such terms with database entries.

Hristovski, et al. (2003) developed an interactive biomedical discovery support system (BITOLA). The system extract relations between diseases and genes using background knowledge about the chromosomal location of the starting disease as well as the chromosomal location of the candidate genes from resources such as LocusLink, HUGO and OMIM. They used Medline as the source of the known relations between biomedical concepts. The goal of the system was to discover new, potentially meaningful relations between diseases and genes which do not occur together in the same published article. If concept X and concept Y are related to each other, the systems assume that concepts Z and X have some relationship if Z is relevant to Y. Finally, the systems check whether X and Z appear together in the medical literature. If they do not appear together, this pair (X and Z) is considered as a potentially new relation.

Parth, et al. (2014) proposed hybrid approach based on **CRF and SVM** to find out disorder mentions from clinical documents and a dictionary look-up approach on a customized UMLS meta-thesaurus to find corresponding CUI. Whereas **CRF** based sequencing algorithm was used to find different medical entities and binary **SVM** classifier was used to find relationship between entities. In the experiment were used MIMIC database, the system did produce competitive results and achieved F-score of 0.714 and accuracy of 0.599.

Jonnagaddala, et al. (2015) 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.

Meystre, et al. (2006) developed system to automate the problem list using NLP to extract potential medical problems from free-text medical documents. The goal of their system was to improve the problem lists quality by increasing its completeness, accuracy, and timeliness. Their system is made of two main components: background application and the problem list management application. The background application does all the text processing and analysis, and stores extracted medical problems in the central clinical database. These problems can then be accessed by the problem list management application integrated in Electronic Health Record. They describe an evaluation of the background application responsible for processing the medical documents and detecting problems. The NLP part of their system uses the UMLS MetaMap Transfer (MMTx) application and a negation detection algorithm called NegEx to extract 80 different medical problems selected for their frequency. When using MMTx with its default data set, we measured a recall of 0.74 and a precision of 0.756. A custom data subset for MMTx was created, making it faster and significantly improving the recall to 0.896 with a non-significant reduction in precision.

Adupa, et al. (2009) proposed **an information extraction**-based approach that converts unstructured text into a structured form. The structured data are then compared against a list of eligibility criteria using a **rule-based** system to determine which patients qualify for enrollment in a heart failure clinical trial. The researchers used dataset collected from the EPIC EHR used by Northwestern Memorial Group. As a result, the proposed approach recall and precision values of 0.95 and 0.86, respectively

2.2 Sentence Selection Summarization

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

Chen & Verma (2006) 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 assigns scores to sentences based on number of original keywords (query words) and expanded keywords.

Sarkar (2009) 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. (2010) developed a system called MedEx. The system **extract 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. (2008) 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.

Aramaki, **et al. (2009)** tried to solve the problem of how to **extract medical information** are written in natural language. The authors developed system (TEXT2TABLE) that converts medical text into a table structure. The system's core technologies are: First, medical event recognition modules, second a negative event identification module and used SVM-based classifier that identified whether an event has actually occurred or not. The proposed SVM-based classifier uses both BOW information and dependency parsing results.

The experimental results demonstrated that of the system has 85.8% F-measure and revealed that syntactic information can contribute to the method's accuracy.

Result, the proposed approach recall and precision values of 0.95 and 0.86, respectively.

2.3 Association Rules in Medical Domain

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

Doddi, et al. (2001) 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. (2014) built a system 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. (2001) 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. (2013) 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.

To conclude, none of the above-mentioned papers handled extraction of medical information from the medical reports and then apply association rules to the data that has been extracted. In addition, researches that aimed to extract medical information did not depend on real dataset. Also, the current thesis focuses on unstructured English slang language texts and obtained real data sets from ministry of Health. Therefore, the new approach extracts medical information and applies association rules using medical reports. In addition, the new approach provides more features about medical information that can help medical staff to (1) extraction of medical information include: diseases, procedures, and drugs, (2) determine direct and indirect relationships between disease and other diseases, as well as relationships between procedures and diseases as well as relationships between drugs and diseases.

2.4 Summary

In this chapter, we presented a review of some related works in the field of template based summarization, and identified the advantage and drawbacks of existing approaches. We classified the previous works into three categories: The first category includes approaches used sentence selection summarization. The second category includes approaches used template based summarization. The final category includes approaches used association rules. We explained that the data used in the thesis are real data. Additionally, how the medical staff will benefit from the information extracted.

In the next chapter, we provide a detailed description of the proposed template based medical reports summarization approach.

Chapter 3

Theoretical Foundation

Chapter 3

Theoretical Foundation

This chapter gives an overview of the background theory of our thesis. First we discussed types of text summarization (sentence selection, template based), followed by providing an overview to Information extraction tasks include Lexical Analysis, Named Entity recognition, Syntactic Structure, Coreference Analysis Pattern Matching, and Events Merging. After that, we talked about development environment used in this thesis. Finally, we present an overview to association rules.

3.1 Text Summarization

Text summarization (TS) is the process of identifying the most important meaningful information in a document or set of related documents and compressing them into a shorter version preserving its overall meanings (Gonnade, 2015). The goal of automatic text summarization is to present the source text into a shorter version with semantics. Another goal is when there are two different users may have created different summaries of the same article based on what they think is most important, preferences, backgrounds, and how they perceive the article (Fajer & Omar, 2014). Also, the important parts of the document depend on the structure of the source documents. Since information that the user already knows should not be included in a summary and at the same time information that is important for one user may not be for another. It is very difficult to achieve consistent judgments about summary quality from human judges. For this reason, it is difficult to evaluate automatic summarization (Erkan & Radev, 2004).

There are several types of text summarization. According to the number of documents, single document summarization takes a single document as an input to perform summarization and produce a single output document. Whereas multi-document summarization it takes numerous documents as an input to perform summarization and deliver a single output document (Bharti et al., 2017).

On the Internet, one can find a lot of examples for single/multi automatic text summarization system such as: Google News, Blog summarization tool, Sakher Arabic summarization, MS Word summarizer and Personal Digital Assistant (PDA) summarizer (Das & Martins, 2007).

Depending on the nature of text representation in the summary can be classified into extractive and abstractive summarization. An extractive is a summary consisting of a number of salient sentences selected from the original document. An abstractive summarization is an understanding of the main concepts in a document and then express those concepts in clear natural language (Gupta & Lehal, 2010). An abstraction is summary mainly human activity. Therefore, in this thesis we concentrate on extraction summary.

There are several methods for extractive text summarization including supervised and unsupervised. In the first method, unsupervised algorithms is used statistical and linguistic information obtained from the text (Asgari et al., 2014). Statistics-based approaches derive weights of key terms and determine the sentence importance by the total weight the sentence contains, whereas linguistics-based approaches identify term relationship in the document through part-of-speech tagging, grammar analysis, thesaurus usage, and extract meaningful sentences (Chang &Hsiao, 2008). The advantage of this method is its simplicity. As the disadvantage of the method, it may be frequented some of the words are not so important which may cause a deviation in text summarization (Asgari et al., 2014).

In the second method, supervised algorithms used the data set which are labeled by human. In other words, there is a set of input text and its summary. The sentences is initially broken into segments. Each segment is represented by a set of predefined features (e.g. the number of term-frequency, location of the segment, number of title words in the segment). Then supervised learning algorithms are used to train the summarizer to extract important sentence segments, based on the feature vector (Yang & Chuang, 2000). These algorithms include decision trees, Bayes rule, neural networks and fuzzy logic. The disadvantage of these methods is the accuracy and speed reduction when working on large documents, due to the large number of comparisons (Asgari et al., 2014). Also, we need labeled data for training phase. Since we do not have labeled data in our dataset, we plan to use unsupervised learning.

There are two tasks in text summarization: content (sentence) selection and template based summarization. Content (sentence) selection "extraction summary" is 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 machine-readable documents (Mani & Maybury, 2001). Since template based can be used in application such as template based summarization.

3.2 Template Based Summarization

Template based summarization is the process of putting useful text present in a document in a condensed format. Here user has the freedom of choosing what should be present in the summary. In other words, user prepares template based on which summary is generated (Desai et al., 2015).

Template based summarization is identified specific pieces of information in unstructured or semi-structured text documents and transforms this unstructured information from a corpus of documents or web pages into a structured database. For example, in a financial transaction, an information extraction system could extract the transaction type, date, customer, principal, currency, and interest rate, which would usually be formatted as a database record suitable for subsequent processing such as data trend analyses, summaries, and report generations (Jung et al., 2010). Therefore, information extraction is the main part of template based summarization.

3.3 Information Extraction

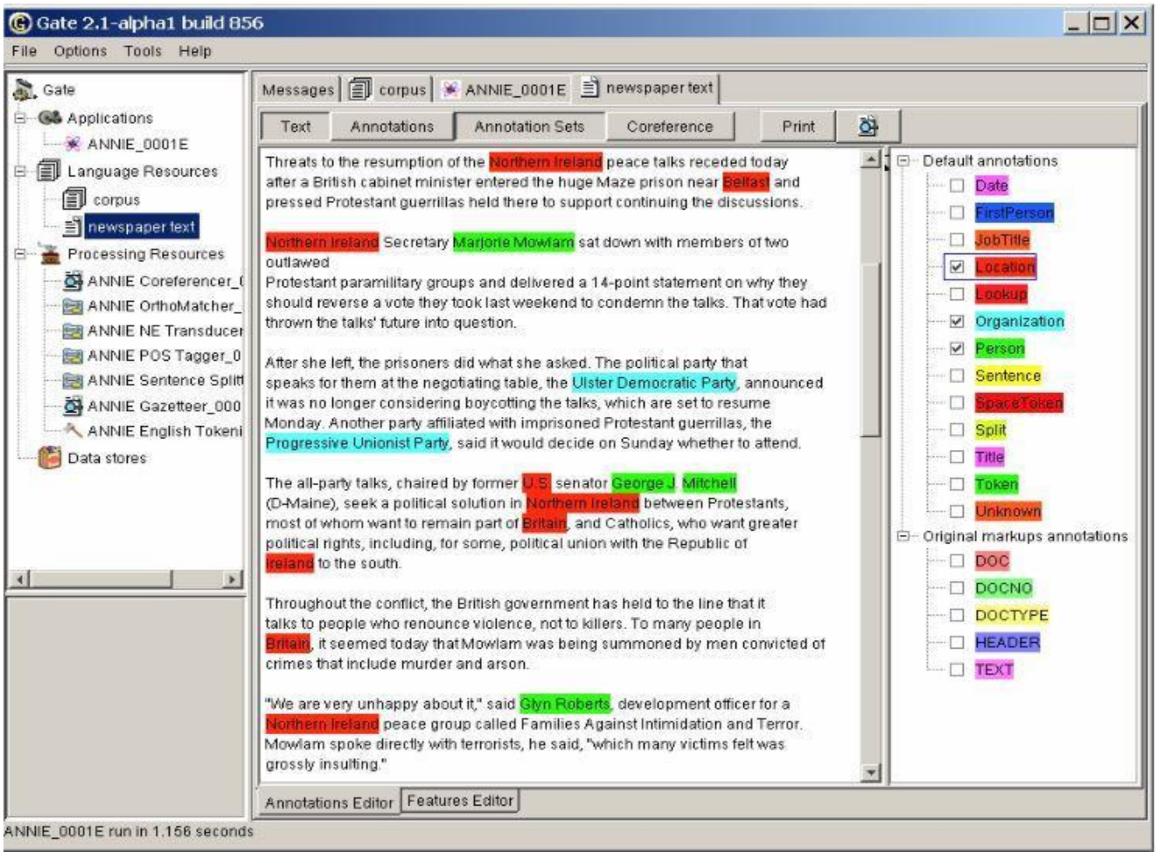
With an increase in data growth available electronically, it caused us a difficult problem to extract and find facts. Usually this data happens the unstructured texts and available in the human language form. Furthermore, we spend a long time in reading and analyze texts to extract useful information. The main objective of NLP is to understand the implicit meaning of texts and transform them into intelligible form. Thus, it is possible to represent the knowledge in a simple format like tables or templates. Information Extraction (IE) is considered an effective way to achieve the goals of NLP, where locates the required information. IE reduces facts in text into a structured form such as templates or tables (Elsebai, 2009). The following is the main information extraction tasks:

3.3.1 Lexical Analysis

It is the first phase of extracting information: text is divided into sentences and then sentences are divided into tokens. Each token is looked up in the dictionary to determine it is possible parts of speech and features.

3.3.2 Named Entity Recognition

The concept of NER was born in MUC (Message Understanding Conference) in the 1990s. It is a very important task of information extraction that aims to find and classify names in unstructured forms such as persons, organizations, locations, dates and time expressions and monetary amounts (percent, money, weight, etc.). An example of a named entity annotation is shown in Figure 3.1 Named Entity (NE) extraction is an essential tool for term extraction that is important in various NLP applications. For instance, text summarization systems, IE, automatic speech recognition, Machine translation, and question answering (Elsebai, 2009 & Elyazjy, 2015).



Threats to the resumption of the Northern Ireland peace talks receded today after a British cabinet minister entered the huge Maze prison near Belfast and pressed Protestant guerrillas held there to support continuing the discussions.

Northern Ireland Secretary Marjorie Mowlam sat down with members of two outlawed
Protestant paramilitary groups and delivered a 14-point statement on why they should reverse a vote they took last weekend to condemn the talks. That vote had thrown the talks' future into question.

After she left, the prisoners did what she asked. The political party that speaks for them at the negotiating table, the Ulster Democratic Party, announced it was no longer considering boycotting the talks, which are set to resume Monday. Another party affiliated with imprisoned Protestant guerrillas, the Progressive Unionist Party, said it would decide on Sunday whether to attend.

The all-party talks, chaired by former U.S. senator George J. Mitchell (D-Maine), seek a political solution in Northern Ireland between Protestants, most of whom want to remain part of Britain, and Catholics, who want greater political rights, including, for some, political union with the Republic of Ireland to the south.

Throughout the conflict, the British government has held to the line that it talks to people who renounce violence, not to killers. To many people in Britain, it seemed today that Mowlam was being summoned by men convicted of crimes that include murder and arson.

"We are very unhappy about it," said Glyn Roberts, development officer for a Northern Ireland peace group called Families Against Intimidation and Terror. Mowlam spoke directly with terrorists, he said, "which many victims felt was grossly insulting."

ANNIE_0001E run in 1.156 seconds

Figure (3.1): Named Entity Recognition in GATE

There is three methods of learning Named-Entity recognition: Supervised Learning (SL), Semi-supervised Learning (SSL), and Unsupervised Learning (UL). The main imperfection of SL is the requirement of a large annotated corpus. The lack of such resources and the high cost of creating them lead to two other alternative learning methods (Nadeau, 2007) as cited in (Elyazji, 2015). In our work, we will concentrate on (UL) because not have labeled data in our dataset.

- **Supervised Learning**

The idea of supervised learning is to learn automatically from a large amount of training data and then supervised by human (Graliński et al., 2009). It is more generalisable and adaptable to new classes of words (Takeuchi and Collier, 2005). The main problem with (SL) is that a large amount of tagged data is needed to implement an effective system, and the accuracy of the models in a given domain is dependent on the training corpus. For this reason, some more recent references have experimented with the inclusion of knowledge-based techniques (Azpeitia et al., 2014). Examples of models that are based on this approach of supervised learning techniques include Conditional Random Fields (CRF), Decision Trees, Hidden Markov Models (HMM), Support Vector Machines (SVM), and Maximum Entropy Models (ME) (Sekine and Ranchhod, 2009).

- **Semi-supervised learning**

Semi-supervised learning exploits both labeled and unlabeled data. It proves useful when labeled data is scarce and hard to construct while unlabeled data is abundant and easy to access (Liu et al., 2011). Moreover, it needs some small data which are used as a seed for the training (Konkol and Konop, 2011). The most effective systems are based on linguistic features. The famous techniques used for this approach is called bootstrapping, that only requires minimal supervision, namely, a set of seeds in order to initiate the learning process (Althobaiti et al., 2013).

- **Unsupervised Learning**

The unsupervised learning do not need any training data. The techniques based on existing semantic lexical resources such as WordNet, on lexical patterns and statistics computed on a large unannotated corpus (Nadeau and Sekine, 2007).

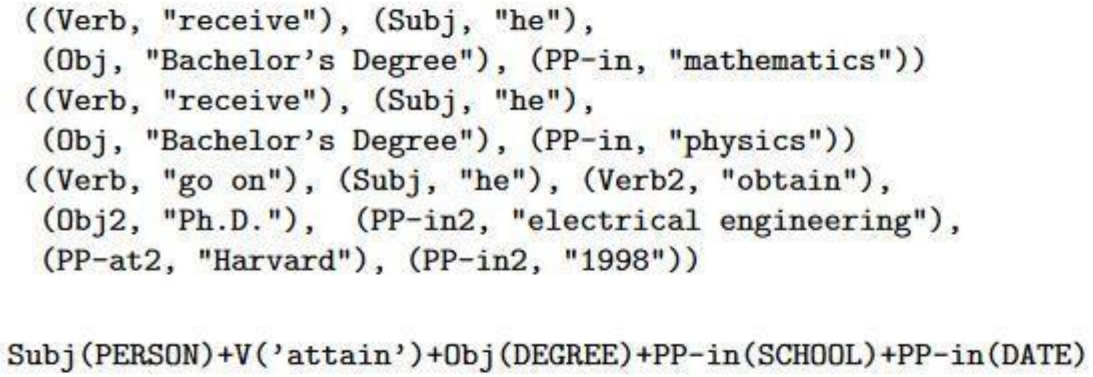
3.3.3 Syntactic Structure

Syntactic Structure aims it helps to understand the roles played by different words in a body of text. Processing a sentence syntactically involves determining the subject, predicate, nouns, verbs, pronouns, etc. The words themselves are not enough, focus on Syntactic analysis be on sentences and not words, because a word may function as different parts of speech in different contexts sometimes (noun, verb). (Redd & Hanumanthappa, 2014).

2.3.4 Pattern Matching

All previous phases have been in a sense preparatory for the scenario pattern matching. A key phase of any IE system is its set of Patterns Matching (or extraction rules) that is used to extract from each document the information relevant to a particular extraction task. As writing useful extraction patterns is a difficult, time-consuming task. It is used only to process documents that contain grammatical, plain text. Such extraction rules are based on syntactic and semantic constraints that help identify the relevant information within a document (Muslea, 1999).

Each extraction rule is made up of three patterns: 1) a pre-filler pattern that must match the text immediately preceding the slot-filler, 2) a slot-filler pattern that must match the actual slot-filler, and 3) a post-filler pattern that must match the text immediately following the filler. The extraction rules also contain information about what template and slot they apply to (Mialtz, 2010).



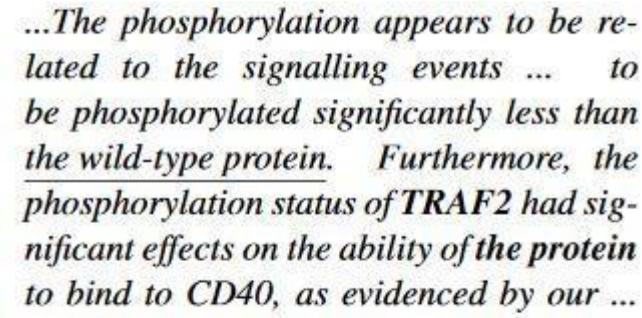
```
((Verb, "receive"), (Subj, "he"),
 (Obj, "Bachelor’s Degree"), (PP-in, "mathematics"))
((Verb, "receive"), (Subj, "he"),
 (Obj, "Bachelor’s Degree"), (PP-in, "physics"))
((Verb, "go on"), (Subj, "he"), (Verb2, "obtain"),
 (Obj2, "Ph.D."),  (PP-in2, "electrical engineering"),
 (PP-at2, "Harvard"), (PP-in2, "1998"))

Subj(PERSON)+V(’attain’)+Obj(DEGREE)+PP-in(SCHOOL)+PP-in(DATE)
```

Figure (3.2): Pattern Match Example

The figure means that the main verb must be (a form of) “attain", the subject NPmust be of type PERSON, the object NP must be of type DEGREE, and prepositional phrases headed by \in" either designate the SCHOOL or the DATE slots of the relation (Mialtz, 2010).

3.3.5 Coreference Analysis

Coreference Analysis is an important stage in the process of extracting information because it can help resolve ambiguous cases of semantic types and it can automatically provide links between entities and as well can facilitate better indexing for information search with rich semantic information. This phase occurs when two or more expressions in a text refer to the same object, activity as shown in Figure 3.3. The following passage includes an interacting relation; the binding event between the anaphoric mention the protein and a cell entity CD40 is implied in the text. The mention, the protein, refers to the specific protein name, TRAF2, previously introduced in the same text (Choi et al., 2014).



...The phosphorylation appears to be related to the signalling events ... to be phosphorylated significantly less than the wild-type protein. Furthermore, the phosphorylation status of ***TRAF2*** *had significant effects on the ability of* ***the protein*** *to bind to CD40, as evidenced by our ...*

Figure (3.3): Conferences analysis

3.3.6 Inferencing and Events Merging

In many situations, partial information about an event may be spread over several sentences, this information needs to be combined before a template can be generated. In other cases, some of the information is only implicit and needs to be made implicit through an inference process (Choi et al., 2014).

3.4 Integrated Development Environments

To generate NE from the text, a tool can be used this tool called Integrated Development Environments, common Environments are:

3.4.1 Stanford CoreNLP

An extensible pipeline that provides core natural language analysis. This toolkit is quite widely used, both in the research NLP community and also among commercial and government users of open source NLP technology. Stanford CoreNLP is based on linear-chain conditional random fields. Stanford CoreNLP can work with any character encoding, it supports most languages, including English, Arabic, Chinese and European languages. It provides a lightweight framework, using plain Java objects (rather than something of heavier weight, such as XML or UIMA's Common Analysis System (CAS) objects). StanfordCoreNLP provides many essential annotators such as tokenize: Sentence split, cleanXML, Truecase, POS tag, Lemma, Gender, NER, RegexNER, Parse, and Sentiment. The output formats include XML, inlineXML, slashTags and character of sets, the latter being of special interest for this implementation. Therefore, many researchers used Stanford CoreNLP tools such as (Kumar et al., 2005) and (Finkel et al., 2005) and other.

3.4.2 GATE the General Architecture for Text Engineering

This is one of the most popular tools used to dealing with NLP. GATE is an infrastructure for developing and deploying software components that process human language (Cunningham, 2000). It is free and open source tools developed at the University of Sheffield in 1996.GATE is built based on JAVA used by the researcher as infrastructure for developing and deploying software components that process human languages such as NER projects, coreference resolution, and others (Shaalan, 2014). GATE handle with Multilanguage such as Arabic, English, Chinese, Hindi, etc. GATE support many text file formats such as XML, HTML, PDF, RTF, email, and plan text. Moreover, GATE provides a set of NLP tools including tokeniser, gazetteer, POS tagger named entity recognition, chunker and parsers which are important for any development of natural language systems (Cunningham, 2000). Also, GATE has features to build rule-based NER system which helps the researcher and development to build their grammatical rules as a finite state transducer using JAPE (a Java Annotation Patterns

Engine). Therefore, many researchers used GATE tools such as (Elsebai, 2009), (Shoaib, 2011), and other.

We used **GATE** in our work to perform tokeniser, gazetteer, POS tagger, named entity recognition, and migrate all data extracted from medical records to an external file.

3.4.3 OpenNLP

OpenNLPis a machine learning based toolkit for natural language processing. It supports the most common NLP tasks, such as sentence segmentation, tokenization, part-of-speech tagging, named entity extraction, chunking, coreference resolution, and parsing. OpenNLP also includes maximum entropy and perception based machine learning. OpenNLP is a Java based library for various natural language processing task (Azpeitia et al., 2014). Therefore, many researchers used OpenNLP tools such as (Sharma and Goyal, 2011) and (N.V et al., 2010) and other.

3.4.4 LingPipe

LingPipeis a tool kit for processing text using computational linguistics, it is a set of Java libraries developed by Alias-I for natural language processing. It supports different natural language processing such as POS tagging, NE recognition, spelling correction, it is also offering additional functionalities such as orthographic correction and text classification in English. It offers a user interface and various demos through which it is possible to test texts (Marrero, 2009). NER in LingPipe components based on hidden Markov model interface and the learned model can be evaluated using k-fold cross validation over annotated dataset. LingPipe is multi-lingual such as Arabic, English, Chinese (Carpenter, 2006). It is open-source and free of charge for research causes, but it is possible to purchase it for commercial use. Therefore, many researchers used LingPipe tool such as (Atdag and Labatut, 2013) and others.

3.5 Association Rule

In the recent era, medical science has revealed that the occurrence of one disease can lead to several associated diseases. For example, Heart-Block can lead to the occurrences of other diseases like Hypertension, Cardiac - Arrest and so on. It is, however, still an interesting problem, to see how far the medical philosophy holds from statistical point of view (Rashid et al., 2014).

Formerly, statistical tools and modeling techniques were used to discover interesting and hidden patterns in the data. These patterns may not be easily detected using traditional methods. So, it was the most convenient way to discover the relationship between the elements are **association rules** (Doddi et al., 2001).

Launch the task of association rule mining was in 1993. It has received a great deal of attention. Today the mining of such rules is still one of the most popular pattern discovery methods in Knowledge Discovery in Databases (KDD) (Hipp et al., 2000). It finds interesting association or correlation relationships among a large set of data items. The discovery of interesting association relationships among huge amounts of transaction records can support many decision-making processes. Since then, association rule mining has been studied and applied in many domains (e.g. economic and financial time series, medical, etc.). Association rules have been widely used by the retail industry under the name “market-basket analysis”.

In this thesis, we use this approach to analyze a large database containing medical-record data and extracting useful information. Our aim is to obtain association rules indicating relationships between diseases and other diseases, procedures performed on a patient and the diagnoses, and drugs and their relationship with diseases (Doddi et al., 2001).

Measures of Association rules

To select interesting rules from the set of all possible rules, constraints on various measures of significance and interest can be used. The best known constraints are minimum thresholds on support and confidence.

1. Support

Support means the frequency that the data fields “A” and “B” involved in association rules occur together in the data set. Only the association rules appear frequently in the itemsets, when it gets high accuracy. When the frequency of “A” and “B” occurring at the same time is equal to or greater than the designated minimum support threshold, “A” and “B” meet frequent itemsets (Bhargava & Shukla, 2016). Support can be expressed as shown in Equation 3.1

$$s\,(A \longrightarrow B) = P\,(AB) = \frac{N\,(AB)}{|D|}, \qquad (1)$$

where N(AB) is the record number of “A” and “B” that appeared together, and |D| is the total record number of transactions in data sets.

2. Confidence

Confidence is the statistics of probability P (B | A) that subsequent events occur under the condition of occurrence of the precursor events in trading data sets. It is used to measure the reliability of the rules. Formula is (Bhargava & Shukla, 2016). Confidence can be expressed as shown in Equation 3.2

$$c\,(A \longrightarrow B) = P\,(B \mid A) = \frac{P\,(AB)}{P\,(A)}. \qquad (2)$$

3.6 Summary

In this chapter, we have presented a theoretical foundation for this research. We discussed the Text summarization (TS), and Template Based Summarization (TBS). Then, we defined the Information Extraction (IE) and explained the basic tasks that must follow to build information extraction system which includes: Lexical Analysis, Named Entity recognition, Syntactic Structure, Scenario Pattern Matching, Coreference Analysis and Inferencing and Events Merging. After that, we describe integrated development environment. Finally, we described Association Rules.

In the next chapter, we discuss research methodology.

Chapter 4

Research Methodology

Chapter 4

Research Methodology

In this chapter, we proposed Template Based Medical Reports Summarization as seen in figure 4.1. Section 4.1 describes data gathering and some details about the medical reports collected. Section 4.2 describes the corpus collection and preprocessing stage, which is comprised of three components: data gathering, tokenization and normalization. Section 4.3 presents proposed template based. Section 4.4 presents the steps of extracting medical information from medical reports. Section 4.5 describes information discharge that has been extracted from the medical reports in the predefined template. Section 4.6 describes some data mining processes, which were used in our research. Finally, Section 4.7 presents expert evaluation.

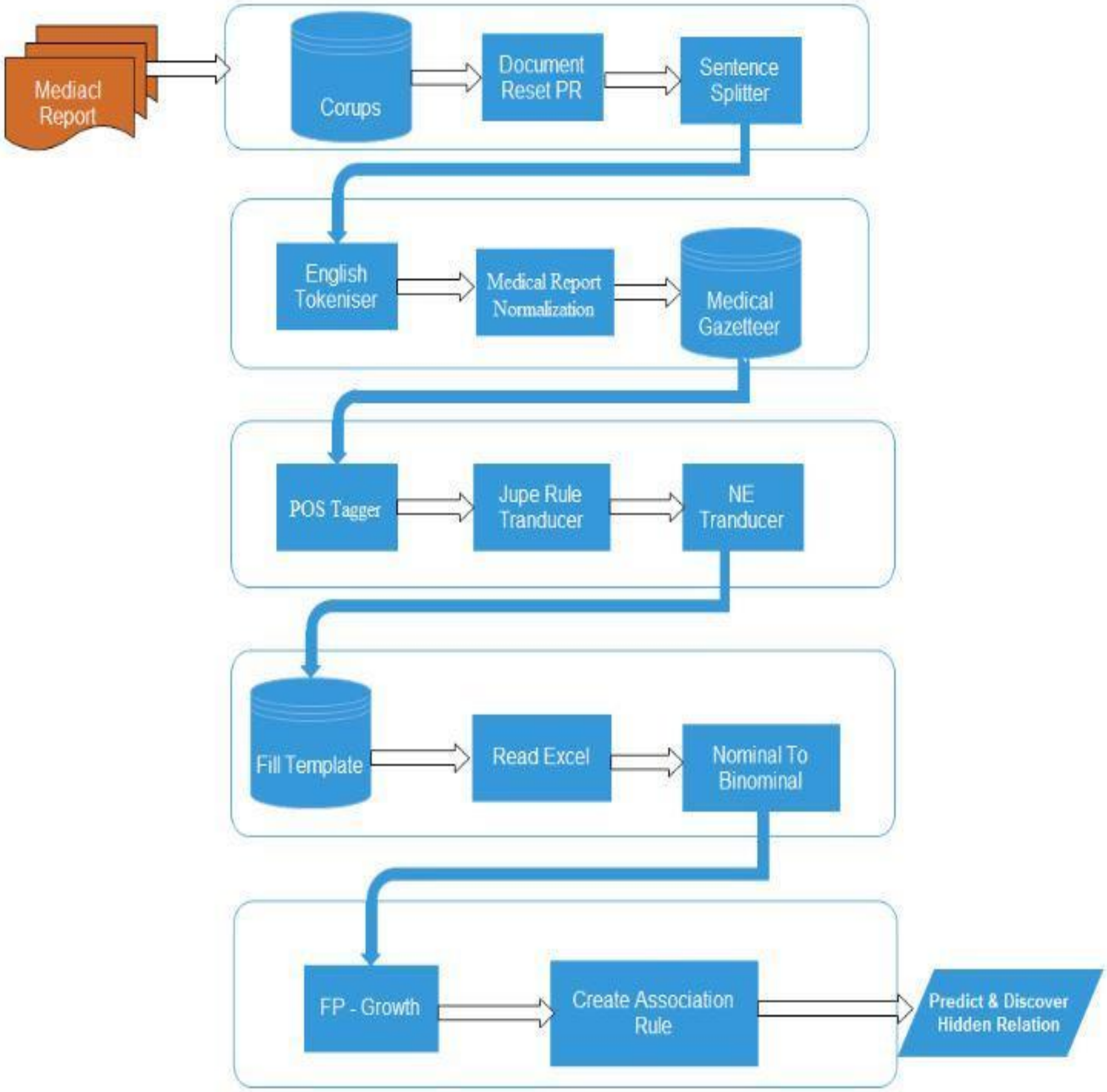


Figure (4.1): Apporach Architecture

4.1 Data Acquisition

The first step in our proposed system 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. We got the data set 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 2016, we collected a total 2300 medical reports.

4.2 Data Preprocessing

To use text mining we need to prepare our data to be ready for applying the mining methods. We aim to transform the medical reports to a form that is suitable for the text and data mining techniques. As shown in Figure 4.2 preprocessing includes the following steps:

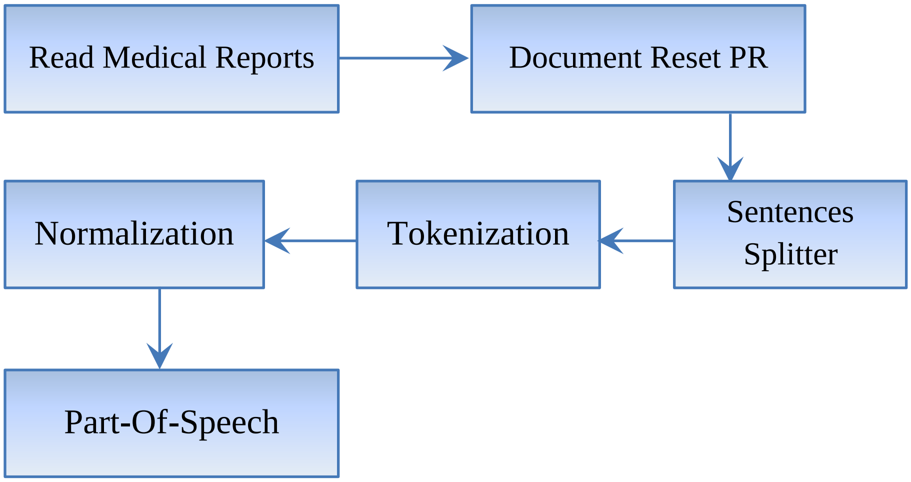


Figure (4.2): Preprocessing Process

4.2.1 Document Reset PR

Document reset PR is 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 (Greenwood et al., 2012).

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

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 (Spasić et al., 2015). Generally, tokenization occurs at the word level.

4.2.4 Normalization

Normalization is very important and critical in our research, it is frequently used when converting text to numbers, dates, acronyms, and abbreviations are non-standard words that need to be pronounced (Sproatet al., 2001). 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. We adopt one form of all these forms.

4.2.5 Part of Speech (POS)

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. (Elsebai, 2009). 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 (Mukund et al., 2010). 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 (Elsebai, 2009).

4.3 The Proposed Template-Based

Before discuss the extract medical information from medical reports, we identified the template Based we will use in the information extraction process where consists three medical attributes, namely: diseases, medical procedures and drugs as shown in Table 4.1. The reason we choose these attributes is that these attributes are the most important thing in the report. In addition, most medical reports contain these attributes.

Table (4.1): Template Based Medical Reports Summarization example

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

4.4 Extract Medical Information Stage

Information extraction stages is the basis of our research, where our main goal is to extract medical information and then predict 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 (Al-Zaidy et al., 2012). 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, we have implemented this stage using GATE tool to identify extract diseases, medical procedures, and drugs names we used two methods: firstly, used a predefined list or Gazetteers and added new diseases to it. Secondly, we adopt many rule-based approaches to develop our system.

4.4.1 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. This type of gazetteer is built manually. Therefore, for the extracting proper name we implement the following preprocessing steps to increase the quality of the result:

- **Create Gazetteer list**

We have created most of the lists of the gazetteers from SNOMED CT - which is the most comprehensive and precise clinical health terminology product in the world, owned and distributed around the world by SNOMED International (SNOMED International, 2016). SNOMED CT has been developed collaboratively to ensure it meets the diverse needs and expectations of the worldwide medical profession and is now accepted as a common global language for health terms. Concepts are organized into 19 hierarchies such as body structures, clinical findings, events and procedures- (Lee et al., 2010).

Where we have created three categories:

  - Diseases names list
  - Medical procedures names list
  - Drugs names list

- **Gazetteer Normalization**

As shown in Table 4.2 three Gazetteer lists that were used in our research, we made two copies for each Gazetteer list, the first copy, uppercase letters, second copy, lowercase letters, because the GATE tool does not support case sensitive.

Table (4.2): 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

4.4.2 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 (Jiang, 2012). 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 (Jiang, 2012).

4.4.2.1 JAPE: Regular Expressions over Annotations

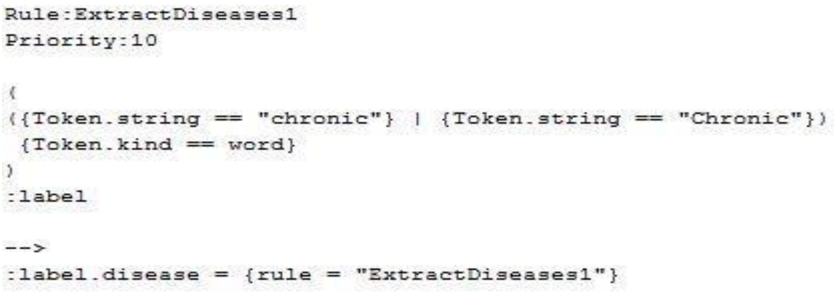
A JAPE grammar consists of a set of phases, each of which consists of a set of pattern/action rules. The phases run sequentially and constitute a cascade of finite state transducers over annotations. The left-hand-side (LHS) of the rules consist of an annotation pattern that may contain regular expression operators (e.g. *, ?, +). The right-hand-side (RHS) consists of annotation manipulation statements (Bordea et al., 2015).

4.4.2.2 Rules for Diseases Names Extractor

We used the JAPE rule-based algorithm 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**".

Figure 4.3 gives an example of disease rule.



```
Rule:ExtractDiseases1
Priority:10

(
({Token.string == "chronic"} | {Token.string == "Chronic"})
 {Token.kind == word}
)
:label

-->
:label.disease = {rule = "ExtractDiseases1"}
```

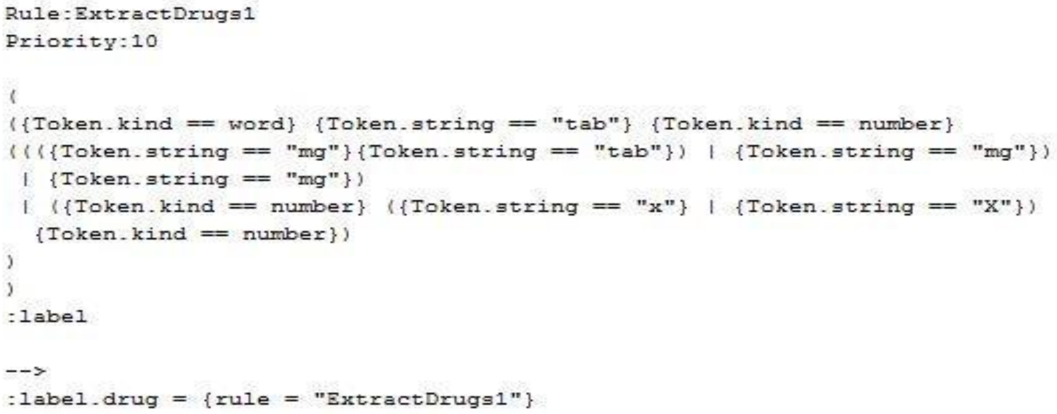
Figure (4.3): Diseases Extraction

4.4.2.3 Rules for Drugs Names Extractor

We used the JAPE rule-based algorithm 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.

Figure 4.4 gives an example of drug rule.



```
Rule:ExtractDrugs1
Priority:10

(
({Token.kind == word} {Token.string == "tab"} {Token.kind == number}
((({Token.string == "mg"}{Token.string == "tab"}) | {Token.string == "mg"})
 | {Token.string == "mg"})
 | ({Token.kind == number} ({Token.string == "x"} | {Token.string == "X"})
  {Token.kind == number})
)
)
:label

-->
:label.drug = {rule = "ExtractDrugs1"}
```

Figure (4.4): Drugs Extraction

The complete implementation of medical information extractor is listed in Appendix A.

4.4.2.4 Rules for Medical Procedures Names Extractor

We used the JAPE rule-based algorithm 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**".

4.5 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 and some statistical process that could benefit the medical staff and data analysts in the Palestinian Ministry of Health. To do this step, we have processed JAPE rule so as to extract medical information and carried over into the Excel file pre-equipped.

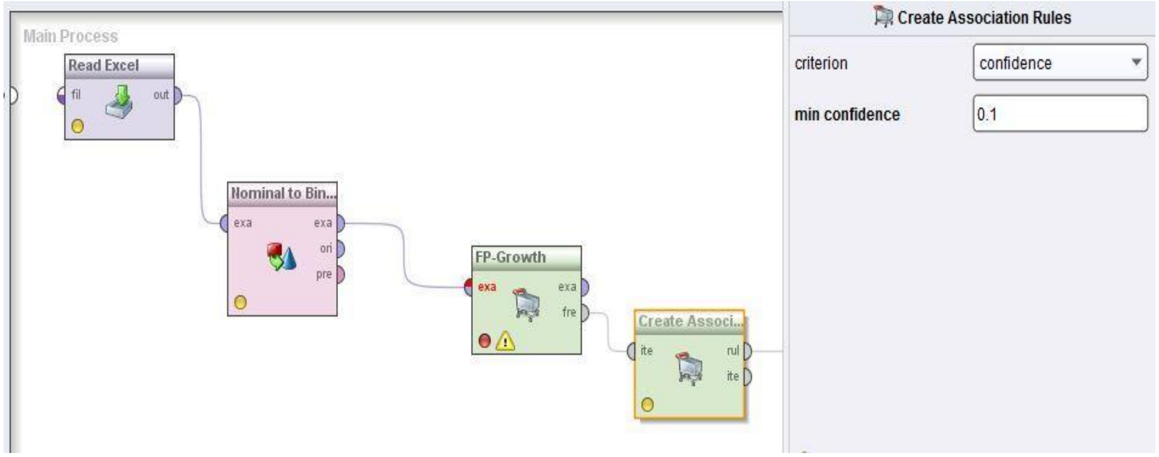
4.6 Association Rules Generation

To evaluate our approach, we constructed association rule. Stages of generate of the association rules contain four components: Read Excel, Convert Nominal to Binominal, FP-Growth, and Create Association Rules.

4.6.1 Read Excel

This is the first stage to create the association rules: we perform load dataset collected in the previous stage in Excel file pre-equipped. Usually the file is Excel sheet. The table must have a format such that each row is an example and each column represents an attribute. We note that the first row of the Excel sheet might be used for attribute names which can be indicated by a parameter.

Missing data values in Excel should be indicated by empty cells or by cells containing only \?\ (Akthar & Hahne, 2012).



Create Association Rules

criterion: confidence

min confidence: 0.1

Figure (4.5): Create Association Rules

4.6.2 Convert Nominal to Binomial

Since association rules algorithm only accept binomial values, we convert nominal values in our data to binomial. At this stages all nominal values converted to binomial. For example, if a nominal attribute with name "costs" and possible nominal values "low", "moderate", and "high" is transformed, the result is a set of three binominal attributes "costs = low", "costs = moderate", and "costs = high". Only one of the values of each attribute is true for a specific example: the other values are false (Akthar & Hahne, 2012).

4.6.3 FP-Growth

In the previous stage, we convert the nominal values to binomial values to be accepted at this stage. Where allows frequent itemset discovery without candidate itemset generation. It contains a two-step: The first step, build a compact data structure called the FP-tree. The second step, Extracts frequent itemsets directly from the FP-tree. A major advantage of FPGrowth compared to Apriori is that it uses only two data scans and is therefore often applicable even on large data sets. Disadvantage of FPGrowth is FP-Tree may not fit in memory and FP-Tree is expensive to build (Verhein, 2008).

4.6.4 Create Association Rule

Considered association rules are a common approach to discover information and identify relationships among different items. We use this approach to analyze a large database containing medical reports data (Doddi et al., 2001). Our aim is to obtain association rules indicating relationships between diseases, medical procedures, and drugs.

Association rules are if/then statements that help uncover relationships between seemingly unrelated data in a relational database or other information repository. An example of an association rule would be "If a customer buys a dozen eggs, he is 80% likely to also purchase milk." (Saxena et al., 2012). We can simulate this example on medical reports.

An association rule has two parts, an antecedent (if) and a consequent (then). An antecedent is an item found in the data. A consequent is an item that is found in combination with the antecedent. Association rules are created by analyzing data for frequent if/then patterns and using the criteria support and confidence to identify the most important relationships. Support is an indication of how frequently the items appear in the database. The confidence indicates the number of times the if/then statements have been found to be true (Saxena et al., 2012).

4.7 Expert Evaluation

After generating the association rules, we performed manual evaluation to ensure the accuracy of the rules that have been extracted from the data. Rules were classified into seven categories: Cardiothoracic, Thoracic, General Surgery, Neurology, Endocrinology, Urology, and Orthopedic. We chose from three to four doctors 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, we used Likert’s scale for evaluation.

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, 2015).

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:

Table (4.2): 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

Factors scoring in average of 3.40 or more shall be considered as high importance (Ozen et al., 2012).

4.8 Summary

In this chapter, we presented the proposed template based medical reports summarization approach. We presented the stages of approach beginning from data acquisition, then we presented the extract medical information from unstructured text using GATE tools, after that, we explained the process of dump the extracted information into an external file. Finally, we presented association rules to extract hidden relationships between information extracted from medical reports.

In the next chapter, we present and discuss the experiments carried out to realize and evaluate the proposed approach.

Chapter 5

Experiments and Results

Chapter 5

Experiments and Results

In this chapter, we present and analyze the experimental results to provide evidence that our approach can identify medical information such as diseases, medical procedures, and drugs from medical reports (as describe in Section 4.3). Also, it illustrates the association rules that have been extracted from medical information, then it discusses some rules that have been extracted (as describe in Section 4.5). Finally; we evaluate the association rules that have been extracted to discover the hidden relation between medical information (as describe in Section 4.7).

5.1 *Experiments* Setup

In this section, a description of the experimental environment and tools used in the experiments.

5.1.1 Experimental Environment

We applied experiments on a machine with properties that shown in Table 5.1

Table (5.1): Machine environment properties

System Model	HP ProBook
Processor	2.60 GHz Intel Core i5-3230M
Memory Modules	8 GBDDR2 RAM
Hard disk	750 GB
Operating System	Windows 7

5.1.2 Experiments Tools

To implement our work, we need different components at different stages for extract medical information. So, various kinds of software tools have been used, which are Microsoft Excel 2010, GATE and Java, Eclipse for code editing, and RapidMiner. These tools are described below:

- **Microsoft Excel 2010:** We get medical reports as Excel file, where each row has one report.
- **GATE:** Used to manipulate natural language processing techniques in our approach, and conduct experiments practical and extracting the medical information from medical reports.
- **Java:** Used to write code for the process of extracting medical information from GATE into an external file.
- **RapidMiner:** Used to apply indirect relationship discovery through association rules.

5.2 Medical Reports Corpus

We used real medical reports 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 5.2 is shown number of medical records and Figure 5.1 gives sample of medical report we used in our experiments.

Table (5.2): 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

The above mentioned patient 51 years old female, Known case of Rt. Breast cancer with Quadrendectomy and axillary clearance. Patient given radiation because metastatic lymph nodes 26/8/2009 then given ACT protocol in Tel Hashomer Hospital under Zoladex + Valodex. Patient under follow up U/S abdomen + pelvis done 22/5/2012 show normal study both ovaries small size although Zoladex treatment. Patient invited for follow up date: 27/3/2013 in oncology department Tel Hashomer Hospital

Patient needs social and medical supports and follow up.

Figure (5.1): Sample of medical report are unstructured form

5.3 Importing Data

To import data, we create one corpus in GATE as shown in Figure 5.2, the medical reports were compiled in Excel file, then we divided the file into five sections, each section containing 440 reports, due to the size of the reports is somewhat significant. After that, the 440 records were converted to 440 worksheets, then it the 440 worksheets was converted to 440 Excel files, in other words, each medical report was separated into its own Excel file. Finally; these Excel files have been uploaded to the GATE program to extract medical information.

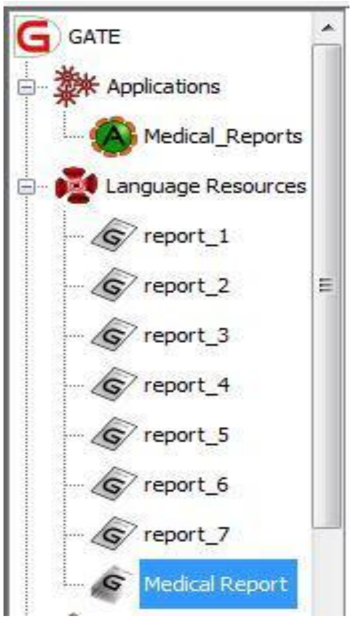
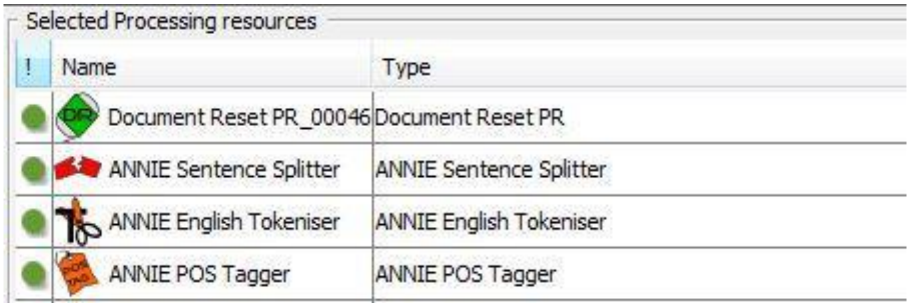


Figure (5.2): Medical Reports corpus

5.4 Data Preprocessing Stage

GATE Developer tools have a collection of operation that is suitable for text mining. In this phase, medical report corpus that identifies in previous section 5.3 are prepared to make them standardized format for the text mining process. There are many of preprocessing techniques such as: sentences splitter, document normalization, tokenization and part of speech tagger. For more details about in Figure 5.3 show preprocessing methods used in our system using GATE tools in section 4.2.



Selected Processing resources

!	Name	Type
●	Document Reset PR_00046	Document Reset PR
●	ANNIE Sentence Splitter	ANNIE Sentence Splitter
●	ANNIE English Tokeniser	ANNIE English Tokeniser
●	ANNIE POS Tagger	ANNIE POS Tagger

Figure (5.3): Preprocessing techniques

- **Document Reset**

Enables the document to be reset to its original state, by removing all the annotation sets and their contents, apart from the one containing the document format analysis (Elyazji, 2015).

- **Sentence Splitter**

Assigns annotations of type Split to sentence boundaries in text based on punctuation and “\t” and“\r” keys.

- **Normalization**

The normalization was done manually. For example, some medical abbreviations, acronyms, and numbers have been converted into text. As shown in the following example:

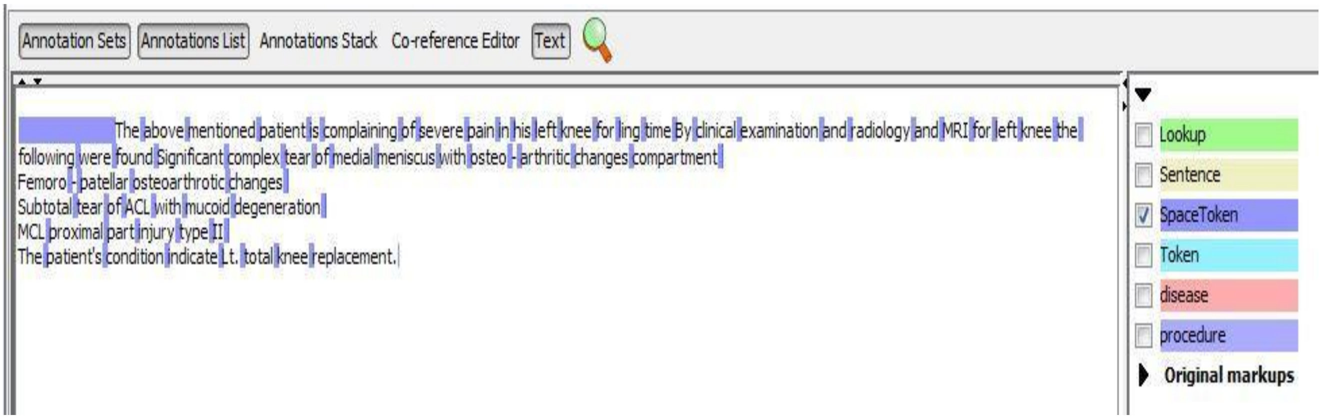
Convert (**IHD**) to ischemic heart disease.

Convert (**LBP**) to low back pain.

And so on.

- **Tokeniser**

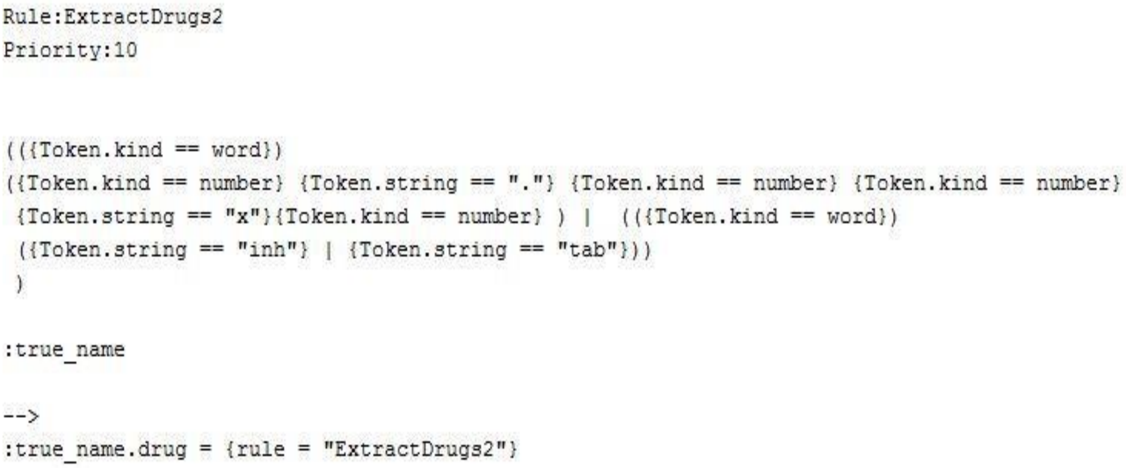
Tokenization is the process of breaking up the text into units called tokens. The tokens may be words or number or punctuation mark, a word is considered as a sequence of connected letters either upper or lower case, a number of the sequence of digits, a symbol represented as @, #, etc. The rest of the tokens are considered as a gap between words and are represented as white space as shown in Figure 5.4, describe in Section 4.2.3.



The above mentioned patient is complaining of severe pain in his left knee for ling time By clinical examination and radiology and MRI for left knee the following were found Significant complex tear of medial meniscus with osteo - arthritic changes compartment
Femoro - patellar osteoarthrotic changes
Subtotal tear of ACL with mucoid degeneration
MCL proximal part injury type II
The patient's condition indicate Lt. total knee replacement.

Figure (5.4): An example of a text after tokenization

Here, the tokenization was used in the JAPE rules to extract some medical information from the medical reports in a specific formula as shown an example in Figure 5.5.



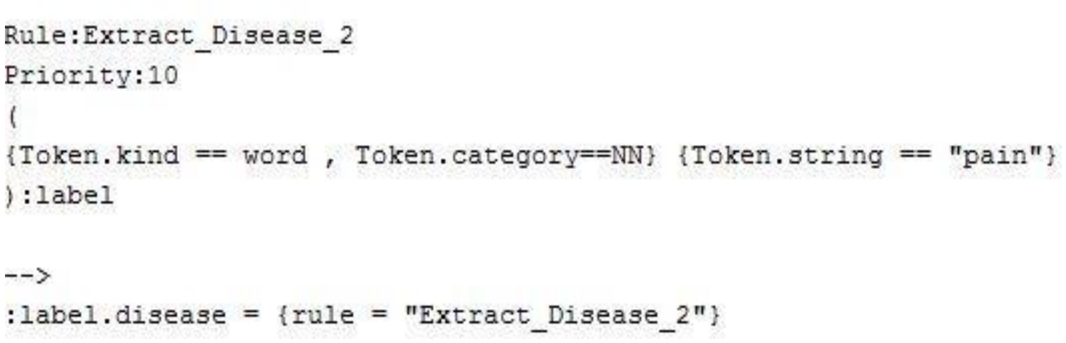
```
Rule:ExtractDrugs2
Priority:10


(({Token.kind == word})
({Token.kind == number} {Token.string == "."} {Token.kind == number} {Token.kind == number}
 {Token.string == "x"}{Token.kind == number} ) |  (({Token.kind == word})
 ({Token.string == "inh"} | {Token.string == "tab"}))
 )

:true_name

-->
:true_name.drug = {rule = "ExtractDrugs2"}
```

Figure (5.5): JAPE Rule to extract drugs.

- **Part of Speech Tagger**

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. Moreover, POS tagger is necessary as a tool to build up any language corpus.

It was used in the JAPE rule to identify certain diseases and medical procedures as shown in Figure 5.6 as an example.



```
Rule:Extract_Disease_2
Priority:10
(
{Token.kind == word , Token.category==NN} {Token.string == "pain"}
):label

-->
:label.disease = {rule = "Extract_Disease_2"}
```

Figure (5.6): JAPE Rule to extract diseases

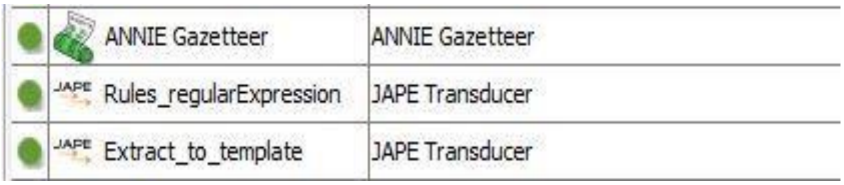
5.5 Name Entity Recognition

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, we used ANNIE application (A Nearly-New Information Extraction system) to tag previous medical reports corpus that identified in section 5.2 with named entities to extract medical information from medical reports.

5.5.1 A Nearly-New Information Extraction system

A family of Processing Resources for language analysis is included in the shape of A Nearly-New Information Extraction system (ANNIE). These components use finite state techniques to implement various tasks from tokenization to semantic tagging or verb phrase chunking (Bordea et al., 2015).

In this research, we created a new ANNIE to handle our medical reports corpus describe in section 5.2. ANNIE components from a pipeline as shown in Figure 5.7 as part of ANNIE used to extract named entity recognition.



● ANNIE Gazetteer	ANNIE Gazetteer
● Rules_regularExpression	JAPE Transducer
● Extract_to_template	JAPE Transducer

Figure (5.7): Name Entity Extraction

Gazetteer

All Gazetteers used in our thesis were compiled from SNOMED CT (Lee et al., 2010). It is a database that contains a significant number of medical terminology including The International Classification of Diseases (ICD10), medical procedures, finding, and so on.

The gazetteer lists used are plain text files, with one entry per line. We add as ANNIE Gazetteer. Each list represents a set of names, such as names of diseases, medical procedures, drugs, more details describe in Section 4.4.1. Table 5.3 gives a sample of the list of diseases, procedures, and drugs used in the project:

Table (5.3): Sample of the list of diseases, procedures, and drugs

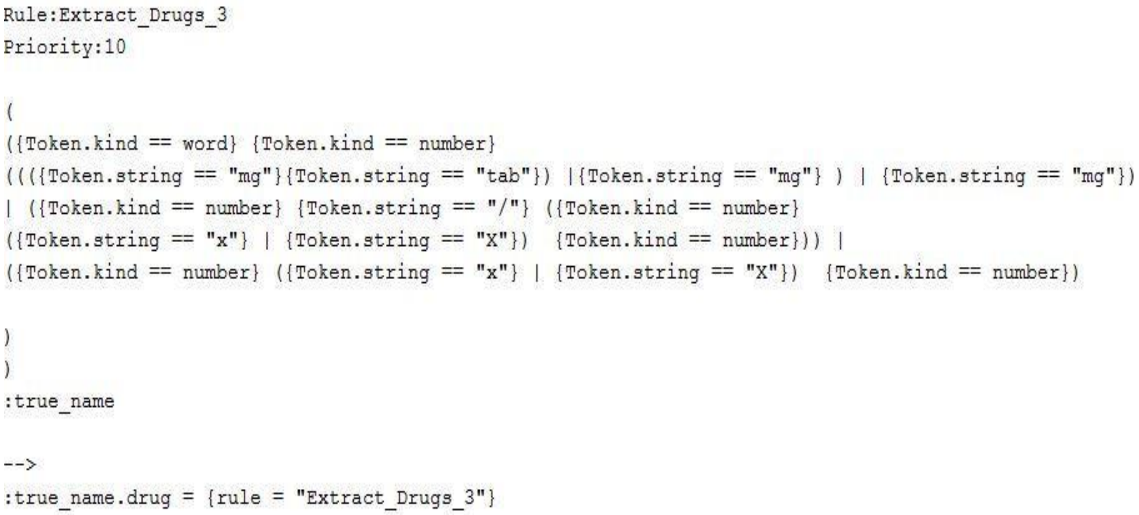
Diseases List	Procedures List	Drugs List
Amyloidosis	Phototerapy	Infanrix
Arthritis	Biopsy	Imuran
Alopecia aceata	Radiation	Inapsine
Amyloidosis	Repetitive strain injury	Lac-hydrin
Botulism	Cutdown placement	Lacosamide
Breast cancer	Peritoneal lavage	Enskycetablets
Hepatitis	Bone marrow collecting	Lipidex
Rheumatoid arthritis	Cardiac ablation	Cardiloc
Peripheral neuropathy	Computerized tomography	Ciprofloxacin
Scleritis	Diaphragm	Murelax

The number of records in the list of diseases is 52000 records, while in the number of records in the list medical procedures is 2916 records, either the number of records in the list of medicine is 11391 records.

JAPE Rule (Rules_RegularExpression)

Java Annotation Patterns Engine (JAPE) is a pattern-matching language. We used JAPE to implement regular expression base on rules. The Left Hand Side (LHS) of each rule contains patterns to be matched, and the Right Hand Side (RHS) contains details of annotations to be created (Elsebai, 2009). For instance, this rule could be written in JAPE as shown in Figure 5.8. Once these JAPE rules are called, the Extract_Drugs_3, will be extracted all sentences from medical reports that have the following patterns:

Baby Aspirin 100mg 1x1
lasix 500mg tab 1/4x1
Aldactone 50mg 1x1
Amicor 5mg 1x1



```
Rule:Extract_Drugs_3
Priority:10

(
({Token.kind == word} {Token.kind == number}
((({Token.string == "mg"}{Token.string == "tab"}) |{Token.string == "mg"} ) | {Token.string == "mg"})
| ({Token.kind == number} {Token.string == "/"} ({Token.kind == number}
({Token.string == "x"} | {Token.string == "X"})  {Token.kind == number})) |
({Token.kind == number} ({Token.string == "x"} | {Token.string == "X"})  {Token.kind == number})

)
)
:true_name

-->
:true_name.drug = {rule = "Extract_Drugs_3"}
```

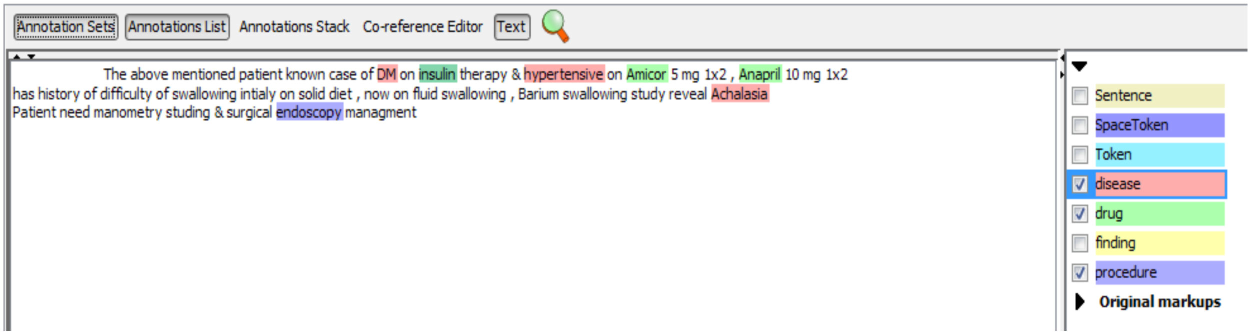
Figure (5.8): JAPE Rule to extract drugs

We identify many JAPE rules to satisfy high accuracy in medical information extract for more details about research JAPE rule list describe in Section 4.4.2.

After selecting the Processing Resources (PR), for the pipeline, the application running and the result display as annotations.

Annotations

The considered annotations one of the key features in GATE is that it gives the opportunity to represent information about the text. However, different processing module such as tokenizer and POS tagger running over text, represent as shown in Figure 5.9 using annotations features.



The above mentioned patient known case of DM on insulin therapy & hypertensive on Amicor 5 mg 1x2 , Anapril 10 mg 1x2
has history of difficulty of swallowing intialy on solid diet , now on fluid swallowing , Barium swallowing study reveal Achalasia
Patient need manometry studing & surgical endoscopy managment

Figure (5.9): Various annotations of medical reports

JAPE Rule (Extrac_to_template)

We create a JAPE rule as give in Figure 5.10 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. An example is shown in Appendix A List 05 – often with .txt extension. After this process, we convert txt file to file Excel to data dump in a predefined template as example shown in Table 5.4. 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, we put some constraints in exporting medical information out of GATE tools to satisfy our goal in discovery hidden relationship and to get the best knowledge.

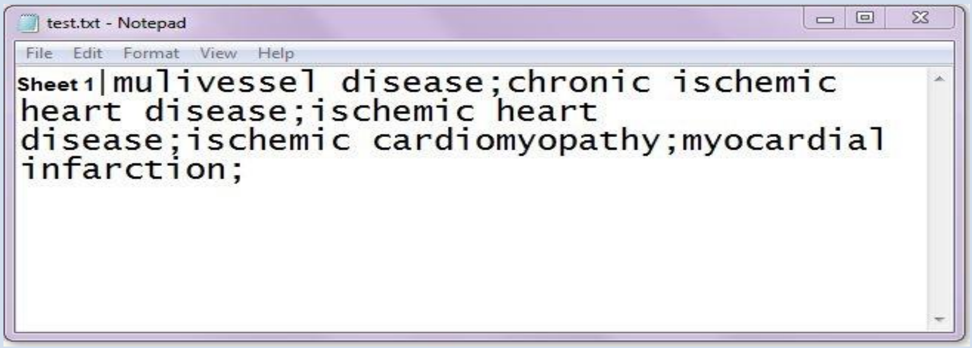
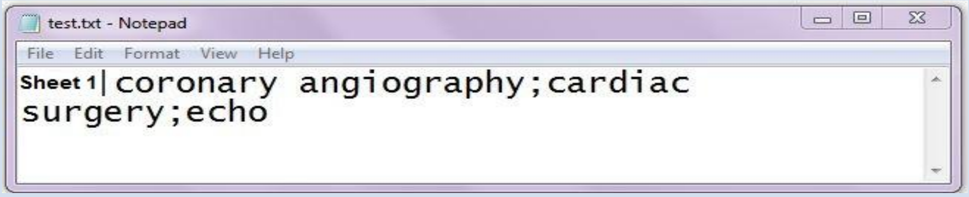
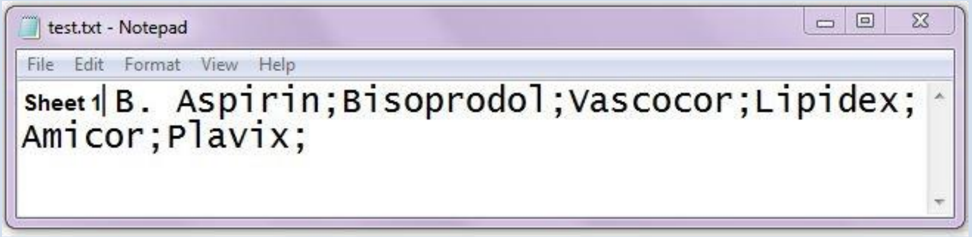
Medical Report	The above mentioned male patient is suffering from chronic ischemic heart disease, old myocardial infarction and ischemic cardiomyopathy He underwent coronary angiography, which revealed mulivessel disease, and he needs for cardiac surgery, by echo his last EF: 36 % He is maintaining on: B. Aspirin 100 mg 1x1 Bisoprodol 2.5 mg 1x1 Vascocor 20 mg 1x1 Lipidex 40 mg 1x1 Amicor 5 mg 1x1 Plavix 75 mg 1x1 The patient under follow up in out clinic of coronary.
Diseases	 Sheet 1|mulivessel disease;chronic ischemic heart disease;ischemic heart disease;ischemic cardiomyopathy;myocardial infarction;
Procedures	 Sheet 1|coronary angiography;cardiac surgery;echo
Drugs	 Sheet 1|B. Aspirin;Bisoprodol;Vascocor;Lipidex;Amicor;Plavix;

Figure (5.10): The resulting file of the process execution JAPE Rule (Extrac_to_template)

Table (5.4): Template Based Medical Reports Summarization example

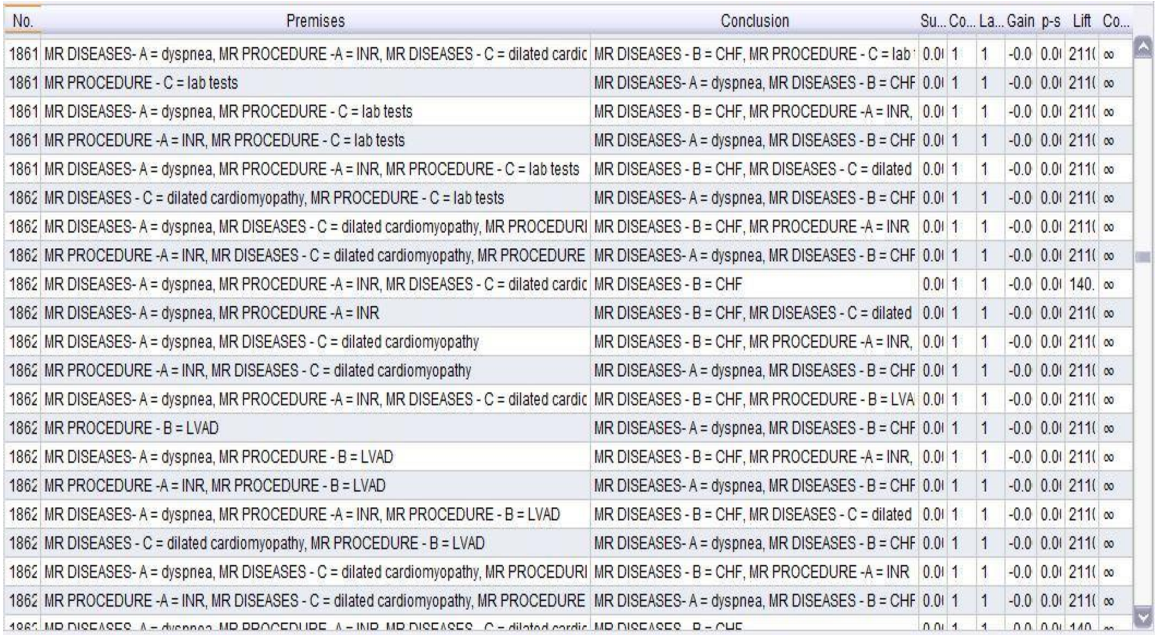
Entity Class	Example
Disease	Breast Cancer, Metastatic lymph;
Procedure	Radiation, ACT, Mamography; Axillary clearance,Quadrendectomy;
Drug	Zoladex,Valodex;

5.6 Association Rules Results

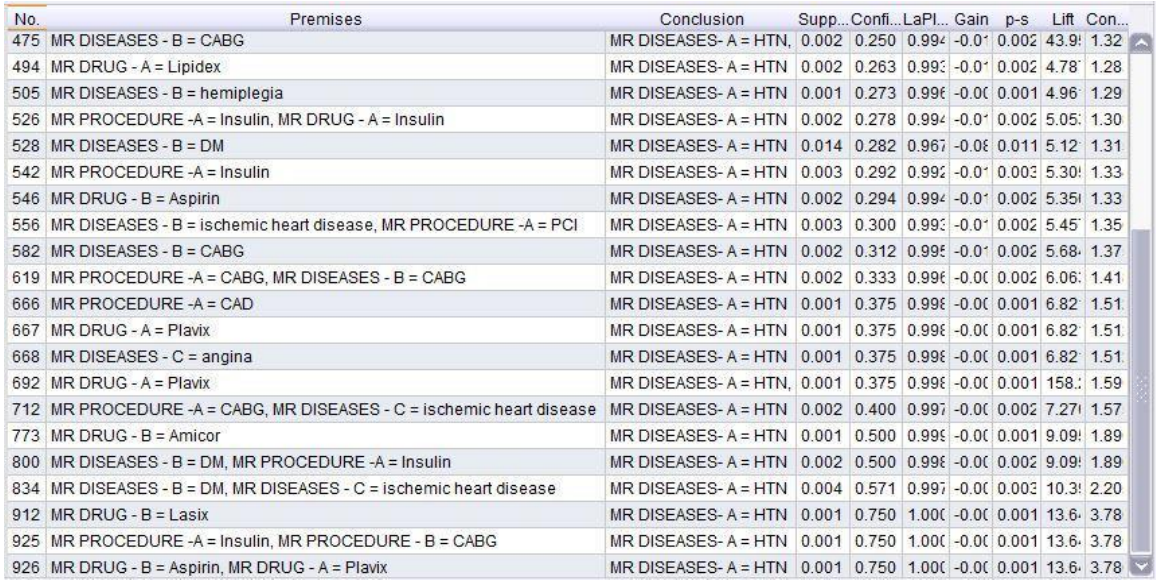
Using statistical tools and modeling techniques, one can discover interesting and hidden patterns in the data. These patterns may not be easily detected using traditional methods. Therefore, next step in the experiment is to use the association rules to reveal relationships among different medical information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

However, we did not show all the association rules that have been extracted from the dataset due to thesis limits Figure 5.9 shows some of the rules of association that have been created. 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 5.5. Medical procedures can be considered as an antecedent item set, and the other diseases can be considered as a consequent item set is as show in Table 5.6 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 as show in Table 5.7 shows the 12 association rules between drugs (antecedent) and their associated diseases (consequent). The minimum support value is usually determined by the users.

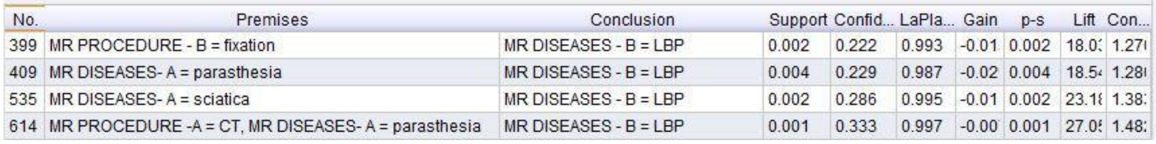
Using Association rules we extracted the following rules:



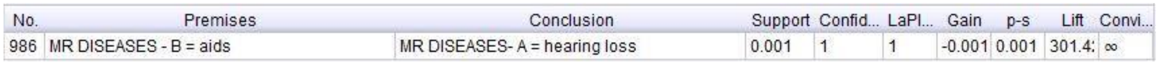
No.	Premises	Conclusion	Su...	Co...	La...	Gain	p-s	Lift	Co...
1861	MR DISEASES- A = dyspnea, MR PROCEDURE -A = INR, MR DISEASES - C = dilated cardic	MR DISEASES - B = CHF, MR PROCEDURE - C = lab	0.0	1	1	-0.0	0.0	211	∞
1861	MR PROCEDURE - C = lab tests	MR DISEASES- A = dyspnea, MR DISEASES - B = CHF	0.0	1	1	-0.0	0.0	211	∞
1861	MR DISEASES- A = dyspnea, MR PROCEDURE - C = lab tests	MR DISEASES - B = CHF, MR PROCEDURE -A = INR,	0.0	1	1	-0.0	0.0	211	∞
1861	MR PROCEDURE -A = INR, MR PROCEDURE - C = lab tests	MR DISEASES- A = dyspnea, MR DISEASES - B = CHF	0.0	1	1	-0.0	0.0	211	∞
1861	MR DISEASES- A = dyspnea, MR PROCEDURE -A = INR, MR PROCEDURE - C = lab tests	MR DISEASES - B = CHF, MR DISEASES - C = dilated	0.0	1	1	-0.0	0.0	211	∞
1862	MR DISEASES - C = dilated cardiomyopathy, MR PROCEDURE - C = lab tests	MR DISEASES- A = dyspnea, MR DISEASES - B = CHF	0.0	1	1	-0.0	0.0	211	∞
1862	MR DISEASES- A = dyspnea, MR DISEASES - C = dilated cardiomyopathy, MR PROCEDURI	MR DISEASES - B = CHF, MR PROCEDURE -A = INR	0.0	1	1	-0.0	0.0	211	∞
1862	MR PROCEDURE -A = INR, MR DISEASES - C = dilated cardiomyopathy, MR PROCEDURE	MR DISEASES- A = dyspnea, MR DISEASES - B = CHF	0.0	1	1	-0.0	0.0	211	∞
1862	MR DISEASES- A = dyspnea, MR PROCEDURE -A = INR, MR DISEASES - C = dilated cardic	MR DISEASES - B = CHF	0.0	1	1	-0.0	0.0	140.	∞
1862	MR DISEASES- A = dyspnea, MR PROCEDURE -A = INR	MR DISEASES - B = CHF, MR DISEASES - C = dilated	0.0	1	1	-0.0	0.0	211	∞
1862	MR DISEASES- A = dyspnea, MR DISEASES - C = dilated cardiomyopathy	MR DISEASES - B = CHF, MR PROCEDURE -A = INR,	0.0	1	1	-0.0	0.0	211	∞
1862	MR PROCEDURE -A = INR, MR DISEASES - C = dilated cardiomyopathy	MR DISEASES- A = dyspnea, MR DISEASES - B = CHF	0.0	1	1	-0.0	0.0	211	∞
1862	MR DISEASES- A = dyspnea, MR PROCEDURE -A = INR, MR DISEASES - C = dilated cardic	MR DISEASES - B = CHF, MR PROCEDURE - B = LVA	0.0	1	1	-0.0	0.0	211	∞
1862	MR PROCEDURE - B = LVAD	MR DISEASES- A = dyspnea, MR DISEASES - B = CHF	0.0	1	1	-0.0	0.0	211	∞
1862	MR DISEASES- A = dyspnea, MR PROCEDURE - B = LVAD	MR DISEASES - B = CHF, MR PROCEDURE -A = INR,	0.0	1	1	-0.0	0.0	211	∞
1862	MR PROCEDURE -A = INR, MR PROCEDURE - B = LVAD	MR DISEASES- A = dyspnea, MR DISEASES - B = CHF	0.0	1	1	-0.0	0.0	211	∞
1862	MR DISEASES- A = dyspnea, MR PROCEDURE -A = INR, MR PROCEDURE - B = LVAD	MR DISEASES - B = CHF, MR DISEASES - C = dilated	0.0	1	1	-0.0	0.0	211	∞
1862	MR DISEASES - C = dilated cardiomyopathy, MR PROCEDURE - B = LVAD	MR DISEASES- A = dyspnea, MR DISEASES - B = CHF	0.0	1	1	-0.0	0.0	211	∞
1862	MR DISEASES- A = dyspnea, MR DISEASES - C = dilated cardiomyopathy, MR PROCEDURI	MR DISEASES - B = CHF, MR PROCEDURE -A = INR	0.0	1	1	-0.0	0.0	211	∞
1862	MR PROCEDURE -A = INR, MR DISEASES - C = dilated cardiomyopathy, MR PROCEDURE	MR DISEASES- A = dyspnea, MR DISEASES - B = CHF	0.0	1	1	-0.0	0.0	211	∞
186	MR DISEASES- A = dyspnea, MR PROCEDURE -A = INR, MR DISEASES - C = dilated cardic	MR DISEASES - B = CHF	0.0	1	1	-0.0	0.0	140	∞



No.	Premises	Conclusion	Supp...	Confi...	LaPl...	Gain	p-s	Lift	Con...
475	MR DISEASES - B = CABG	MR DISEASES- A = HTN,	0.002	0.250	0.994	-0.01	0.002	43.9	1.32
494	MR DRUG - A = Lipidex	MR DISEASES- A = HTN	0.002	0.263	0.993	-0.01	0.002	4.78	1.28
505	MR DISEASES - B = hemiplegia	MR DISEASES- A = HTN	0.001	0.273	0.996	-0.00	0.001	4.96	1.29
526	MR PROCEDURE -A = Insulin, MR DRUG - A = Insulin	MR DISEASES- A = HTN	0.002	0.278	0.994	-0.01	0.002	5.05	1.30
528	MR DISEASES - B = DM	MR DISEASES- A = HTN	0.014	0.282	0.967	-0.08	0.011	5.12	1.31
542	MR PROCEDURE -A = Insulin	MR DISEASES- A = HTN	0.003	0.292	0.992	-0.01	0.003	5.30	1.33
546	MR DRUG - B = Aspirin	MR DISEASES- A = HTN	0.002	0.294	0.994	-0.01	0.002	5.35	1.33
556	MR DISEASES - B = ischemic heart disease, MR PROCEDURE -A = PCI	MR DISEASES- A = HTN	0.003	0.300	0.993	-0.01	0.002	5.45	1.35
582	MR DISEASES - B = CABG	MR DISEASES- A = HTN	0.002	0.312	0.995	-0.01	0.002	5.68	1.37
619	MR PROCEDURE -A = CABG, MR DISEASES - B = CABG	MR DISEASES- A = HTN	0.002	0.333	0.996	-0.00	0.002	6.06	1.41
666	MR PROCEDURE -A = CAD	MR DISEASES- A = HTN	0.001	0.375	0.998	-0.00	0.001	6.82	1.51
667	MR DRUG - A = Plavix	MR DISEASES- A = HTN	0.001	0.375	0.998	-0.00	0.001	6.82	1.51
668	MR DISEASES - C = angina	MR DISEASES- A = HTN	0.001	0.375	0.998	-0.00	0.001	6.82	1.51
692	MR DRUG - A = Plavix	MR DISEASES- A = HTN,	0.001	0.375	0.998	-0.00	0.001	158.	1.59
712	MR PROCEDURE -A = CABG, MR DISEASES - C = ischemic heart disease	MR DISEASES- A = HTN	0.002	0.400	0.997	-0.00	0.002	7.27	1.57
773	MR DRUG - B = Amicor	MR DISEASES- A = HTN	0.001	0.500	0.999	-0.00	0.001	9.09	1.89
800	MR DISEASES - B = DM, MR PROCEDURE -A = Insulin	MR DISEASES- A = HTN	0.002	0.500	0.998	-0.00	0.002	9.09	1.89
834	MR DISEASES - B = DM, MR DISEASES - C = ischemic heart disease	MR DISEASES- A = HTN	0.004	0.571	0.997	-0.00	0.003	10.3	2.20
912	MR DRUG - B = Lasix	MR DISEASES- A = HTN	0.001	0.750	1.000	-0.00	0.001	13.6	3.78
925	MR PROCEDURE -A = Insulin, MR PROCEDURE - B = CABG	MR DISEASES- A = HTN	0.001	0.750	1.000	-0.00	0.001	13.6	3.78
926	MR DRUG - B = Aspirin, MR DRUG - A = Plavix	MR DISEASES- A = HTN	0.001	0.750	1.000	-0.00	0.001	13.6	3.78



No.	Premises	Conclusion	Support	Confid...	LaPla...	Gain	p-s	Lift	Con...
399	MR PROCEDURE - B = fixation	MR DISEASES - B = LBP	0.002	0.222	0.993	-0.01	0.002	18.0	1.27
409	MR DISEASES- A = parasthesia	MR DISEASES - B = LBP	0.004	0.229	0.987	-0.02	0.004	18.5	1.28
535	MR DISEASES- A = sciatica	MR DISEASES - B = LBP	0.002	0.286	0.995	-0.01	0.002	23.1	1.38
614	MR PROCEDURE -A = CT, MR DISEASES- A = parasthesia	MR DISEASES - B = LBP	0.001	0.333	0.997	-0.00	0.001	27.0	1.48



No.	Premises	Conclusion	Support	Confid...	LaPl...	Gain	p-s	Lift	Convi...
986	MR DISEASES - B = aids	MR DISEASES- A = hearing loss	0.001	1	1	-0.001	0.001	301.4	∞

Figure (5.9): Some of the generated rules

5.6.1 Diseases → Diseases

Table 5.5 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. 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. We conclude observing 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)”. **Rule No. 7** also illustrates a logical relationship between the “Ischemic Heart Disease (IHD)”, and “Myopathy” disease and the diagnosis of “Cardiomegaly.”

Rule No. 8 shows Number of patients had who have been diagnosed “Microcephaly” disease, and “Ischemic Encephalopathy” disease, they are susceptible “Epilepsy” disease.

The confidence level of 100% suggests that for virtually all the patients who were suffering from “Microcephaly” disease and “Ischemic Encephalopathy” disease, they are more likely to have “Epilepsy” disease.

Rule No. 9 when a patient infected both of “Paresthesia” disease, and “Spinal Canal Stenosis” disease, there would be a high probability that he infected with “Lower Back Pain (LBP)” disease.

Rule No. 10 is somewhat perplexing. Indicates an unclear correspondence between two diseases “Low Limb Ischemia” disease, and “Gangrene”, with disease “Urinary Tract Infection (UTI)”. But there is no relationship between these diseases and the “disease Urinary Tract Infection (UTI)”. One plausible explanation for such a rule is that perhaps among older patients these problems occur concurrently.

Rule No. 11 also illustrates a predictable set of diseases such as “Diarrhea, Crohn’s” disease related to “Anal Fissure”.

Rule No. 12 shows that patients who have “Hydronephrosis” disease usually suffering from a “Urinary Tract Infection (UTI)” disease.

We notice that most of the former rules good for discovering of diseases from another disease and could be useful to identify the class, even if have a weak support and rarely happen.

Also note that a good number of rules were not known to doctors, so some doctors have evaluated them (normal, unacceptable) and after searching for the accuracy of these rules we found some of these rules very accurate. An example of these rules:

One of the rules, the relationship of AIDS to the disease of hearing loss, since the three doctors who evaluated the rules unanimously found the rule inaccurate. But after searching for this rule, I found that in a 2013 American study, it was confirmed that most AIDS sufferers suffer from hearing loss (Assuiti et al., 2013).

Table (5.5): 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)
7	{Ischemic Heart Disease (IHD), Myopathy}→ Cardiomegaly	Ischemic Heart Disease (IHD), Myopathy	Cardiomegaly
8	{Microcephaly, Ischemic Encephalopathy} → Epilepsy	Microcephaly, Ischemic Encephalopathy	Epilepsy

9	{Paresthesia, spinal canal stenosis} → Low back pain (LBP)	Paresthesia, spinal canal stenosis	Low back pain (LBP)
10	{Low Limb Ischemia, Gangrene} → Urinary Tract Infection (UTI)	Low Limb Ischemia, Gangrene	Urinary Tract Infection (UTI)
11	{Diarrhea, Crohn’sDisease} → Anal Fissure	Diarrhea, Crohn’sDisease	Anal Fissure
12	{Urinary Tract Infection (UTI)} → Hydronephrosis	Urinary Tract Infection (UTI)	Hydronephrosis

5.6.2 Procedures → Diseases

Most of the rules in Table 5.6 indicate a rationalizable correspondence between a set of medical procedures and diseases. Below we discuss the rules in the order in which they are presented in the table.

Rule No. 1 shows that the number of patients had undergone the two procedures, namely Echo, and angiography and were diagnosed with “Shortness of breath (SOB).” The confidence level of 1 suggests that for virtually all the patients who had undergone the two procedures, the diagnosis was “Shortness of breath (SOB).”

Rule No. 2 illustrates a logical relationship between the procedures pertaining to Chemotherapy, Hormonal therapy, and Radiotherapy and the diagnosis of “Breast Cancer”.

Rule No. 3 is somewhat perplexing. One can see a reasonable correspondence between two procedures electrocardiogram (ECG), and X-ray and disease Cardiomegaly but Coronary artery bypass graft (CABG) seem to have no direct relationship with Cardiomegaly. One plausible explanation for such a rule is that perhaps among older patients these problems occur concurrently. In fact, this points out a weakness of the association rule approach. As the approach does not use any knowledge of the underlying domain, not all the rules generated are meaningful. This reminds us that data mining is just a tool that provides businesses with a method of generating hypotheses. It does not verify the hypothesis; nor does it provide any information regarding the value of that hypotheses to the business. These hypotheses must be analyzed and verified by people with domain knowledge and expertise.

Rule No. 4 provides a meaningful correspondence between disease “Low Back Pain (LBP)” and procedures such as “Magnetic Resinance Imaging (MRI)”, “Fixation”, and “Discectomy”.

Rule No. 5 contains a number of procedures are all reasonable for diagnosing patients with “Ischemic Heart Disease”.

However, **Rule No. 6** indicates an unclear correspondence between the diagnosis of “Congestive Heart Failure (CHF)” and “Continuous Positive Airway Pressure Therapy (CPAP)” procedure. We believe that this rule was generated because a significant fraction of the patients who were diagnosed with “Congestive Heart Failure (CHF)” disorder shared “Continuous Positive Airway Pressure Therapy (CPAP)” but there is no relationship between this procedure and the disease “Congestive Heart Failure (CHF)”.

Rule No. 7 and No. 8 are very similar. The procedures listed in these rules are all reasonable for diagnosing patients with “Nasal Congestion”, and “Hydronephrosis”.

Rule No. 9 also illustrates a predictable set of procedures such as “Neurological Examination”, “(CT)”, “Percutaneous Coronary Intervention (PCI)” related to “Cerebrovascular Accident (CVA)”.

Rule No. 10 shows that patients who have “Gun Shot” usually undergo to three procedures namely, “Diaphragm”, “Fixation”, “X-ray”.

Rule No. 11 We observe in this rule that procedure “In Vitro Fertilization (IVF)” that has a direct relationship with "Infertility" disease.

Rule No. 12 also illustrates a logical relationship between the procedures "Electrocardiogram (ECG)" and the diagnosis of "Asthma."

Table (5.6): Association Rules Obtained – Procedure to Diseases

NO.	Association Rule	Antecedent	Consequent
1	{Echo, Angiography} → Shortness of breath (SOB)	Echo, Angiography	Shortness of breath (SOB)
2	{Chemotherapy, Hormonal Therapy, Radiotherapy} → Breast Cancer	Chemotherapy, Hormonal Therapy, Radiotherapy	Breast Cancer
3	{Electrocardiogram (ECG), X-ray, Coronary Artery Bypass Graft (CABG)} → Cardiomegaly	Electrocardiogram (ECG), X-ray, Coronary Artery Bypass Graft (CABG)}	Cardiomegaly
4	{Magnetic Resinance Imaging (MRI), Fixation, Discectomy}→ Low Back Pain (LBP)	Magnetic Resinance Imaging (MRI), Fixation, Discectomy	Low Back Pain (LBP)
5	{Electrocardiogram (ECG), Coronary Artery Bypass Grafting (CABG)} → Ischemic Heart Disease	Electrocardiogram (ECG), Coronary Artery Bypass Grafting (CABG)	Continuous Positive Airway Pressure Therapy (CPAP), Pacemaker
6	{Continuous Positive Airway Pressure Therapy (CPAP), Pacemaker} → Congestive Heart Failure (CHF)	Continuous Positive Airway Pressure Therapy (CPAP), Pacemaker	Congestive Heart Failure (CHF)
7	{Septoplasty} → Nasal Congestion	Septoplasty	Nasal Congestion

8	{Kidney Function, CT} → Hydronephrosis	Kidney Function, CT	Hydronephrosis
9	Neurological Examination, (CT), Percutaneous Coronary Intervention (PCI)} → Cerebrovascular Accident (CVA)	Neurological Examination, (CT), Percutaneous Coronary Intervention (PCI)}	Cerebrovascular Accident (CVA)
10	{Diaphragm, Fixation, X-ray} → Gun Shot	Diaphragm, Fixation, X-ray	Diaphragm, Fixation, X-ray
11	{In Vitro fertilization (IVF)} → Infertility	In Vitro fertilization (IVF)	Infertility
12	{ECG} → Asthma	ECG	Asthma

Overall, all the rules except No. 3 and No. 6 indicate a clear correspondence between procedures and diagnoses. The quantitative information included in the rules can be potentially very revealing and beneficial to medical professionals.

We noted that some rules that may seem relevant and useful, like rules1,2 below, because it gives an indication of a serious situation, but it has a very weak support and confidence, after all, this means that most of the association rules created from this dataset are useful for predicting low.

5.6.3 Drugs → Diseases

Most of the rules in Table 5.7 indicate a rationalizable correspondence between a set of drugs and diseases. Below we discuss the rules in the order in which they are presented in the table.

Rule No. 1 shows that the number of patients taking the three drugs, namely “Lipidex”, “Ciprofloxacin”, and “Cardiloc” and were diagnosed with “Ischemic Heart Disease

(IHD)”. The confidence level of 1 suggests that for virtually all the patients who are taking the three drugs, the diagnosis was “Ischemic Heart Disease (IHD)”.

Rule No. 2 illustrates a logical relationship between the drug pertaining to “Methotrexate” and the diagnosis of “Breast Cancer”.

Rule No. 3 There is no logical or direct relationship between “Aspirin”, and “Pednisone” drugs and “Dyslipiemia” disease.

Rule No. 4 provides a meaningful correspondence between drug “Enalapril” and disease “Congestive Heart Failure (CHF)”.

Rule No. 5 also illustrates a predictable set of drugs such as “Tental”, and “Convertin” related to “Cerebrovascular accident (CVA)”.

Rule No. 6 patients who take drugs such as “Trental”, “Crestor”, and “Convertin”, are usually suffer from “Hemiparesis” disease.

However, **Rule No. 7** indicates an unclear correspondence between the diagnosis of “Hydronephrosis” and “Lasix” drug. We believe that this rule was generated because a significant fraction of the patients who were diagnosed with “Hydronephrosis” disorder shared “Lasix” but there is no relationship between this drug and the “disease Hydronephrosis”.

Rule No. 8 shows that patients who suffer from “Anal Fissure” disease usually give them “Infliximab” drug.

Rule No. 9 also illustrates a logical relationship between the drug “Thalidomide” and the diagnosis of “Amyloidosis”.

Table (5.7): Association Rules Obtained – Drugs to Diseases

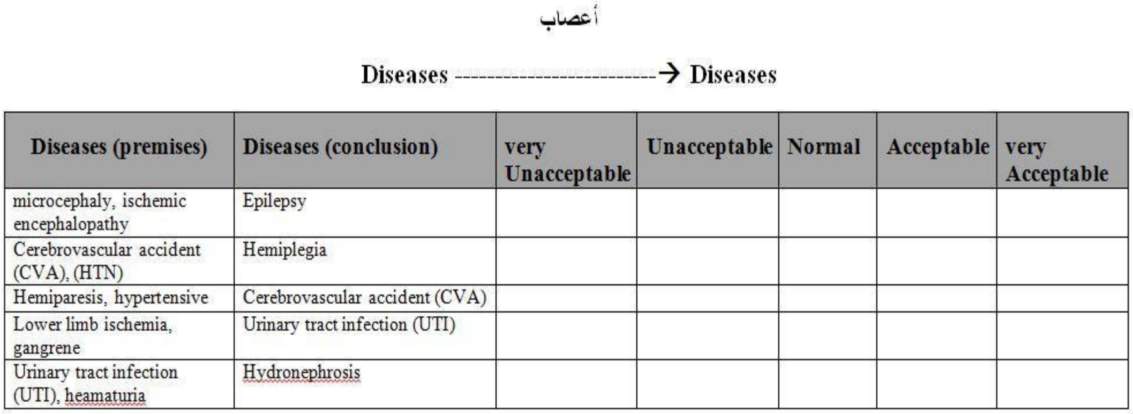
NO.	Association Rule	Antecedent	Consequent
1	{Lipidex, Ciprofloxacin, Cardiloc} → Ischemic Heart Disease	Lipidex, Ciprofloxacin, Cardiloc	Ischemic Heart Disease
2	{Methotrexate} → Breast Cancer	Methotrexate	Breast Cancer
3	{Aspirin, Pednisone} → Dyslipiemia	Aspirin, Pednisone	Dyslipiemia
4	{Enalapril} → Congestive Heart Failure (CHF)	Enalapril	Congestive Heart Failure (CHF)
5	{Tental, Convertin} → Cerebrovascular Accident (CVA)	Tental, Convertin	Cerebrovascular Accident (CVA)
6	{Trental, Crestor, Convertin} → Hemiparesis	Trental, Crestor, Convertin	Hemiparesis
7	{Lasix} → Hydronephrosis	Lasix	Hydronephrosis
8	{Infliximab} → Anal Fissure	Infliximab	Anal Fissure
9	{Thalidomide}→ Amyloidosis	Thalidomide	Amyloidosis

The above discussion shows how the application of association rules to medical data may be of interest to physicians. In this thesis, we focused our attention on finding associations between procedures, diseases, and drugs. A physician who is new to the field may benefit substantially by knowing the set of commonly performed procedures for a particular diagnosis. Association rules can also provide an indication of collections of diagnoses that are correlated and are likely to occur together.

5.7 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, we used human expert as a reference to measure the degree of acceptance of the association rules which have been extracted from the dataset. So, we used Likert's scale for evaluation as describe in Section 4.7.

We have divided association rules 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. We have selected twenty-one doctors. Doctors names shown in Appendix A List 06 - 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. As shown in Figure 5.12, 5.13, 5.14 for example.

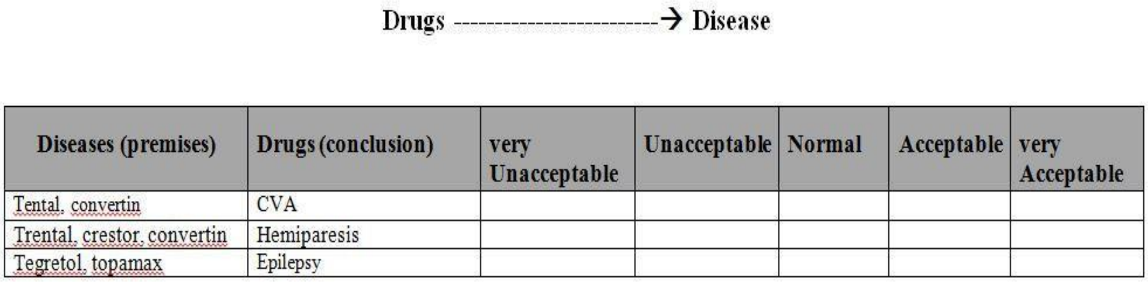


أعصاب

Diseases -------------------------→ Diseases

Diseases (premises)	Diseases (conclusion)	very Unacceptable	Unacceptable	Normal	Acceptable	very Acceptable
microcephaly, ischemic encephalopathy	Epilepsy					
Cerebrovascular accident (CVA), (HTN)	Hemiplegia					
Hemiparesis, hypertensive	Cerebrovascular accident (CVA)					
Lower limb ischemia, gangrene	Urinary tract infection (UTI)					
Urinary tract infection (UTI), heamaturia	Hydronephrosis					

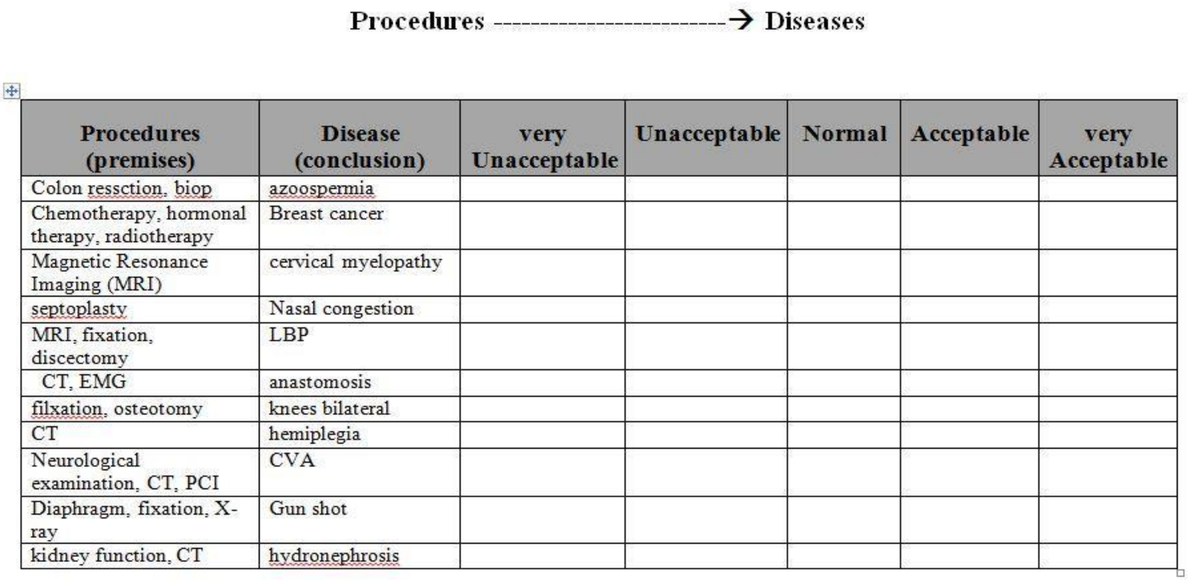
Figure (5.10): A questionnaire to measure the accuracy acceptance association rules



Drugs -------------------------→ Disease

Diseases (premises)	Drugs (conclusion)	very Unacceptable	Unacceptable	Normal	Acceptable	very Acceptable
Tental, convertin	CVA					
Trental, crestor, convertin	Hemiparesis					
Tegretol, topamax	Epilepsy					

Figure (5.11): A questionnaire to measure the accuracy acceptance association rules



Procedures -------------------------→ Diseases

Procedures (premises)	Disease (conclusion)	very Unacceptable	Unacceptable	Normal	Acceptable	very Acceptable
Colon ressction, biop	azoospermia					
Chemotherapy, hormonal therapy, radiotherapy	Breast cancer					
Magnetic Resonance Imaging (MRI)	cervical myelopathy					
septoplasty	Nasal congestion					
MRI, fixation, discectomy	LBP					
CT, EMG	anastomosis					
filxation, osteotomy	knees bilateral					
CT	hemiplegia					
Neurological examination, CT, PCI	CVA					
Diaphragm, fixation, X-ray	Gun shot					
kidney function, CT	hydronephrosis					

Figure (5.12): A questionnaire to measure the accuracy acceptance association rules

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 5.8.

We note that the best results have been in the Department of Neurology and followed by thoracic section. Also, we note that the best results were at the level of diseases and followed by medical procedures and finally drugs.

Our conclusion 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 (5.8): The results of expert evaluation for association rules

Department	Diseases	Medical 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%

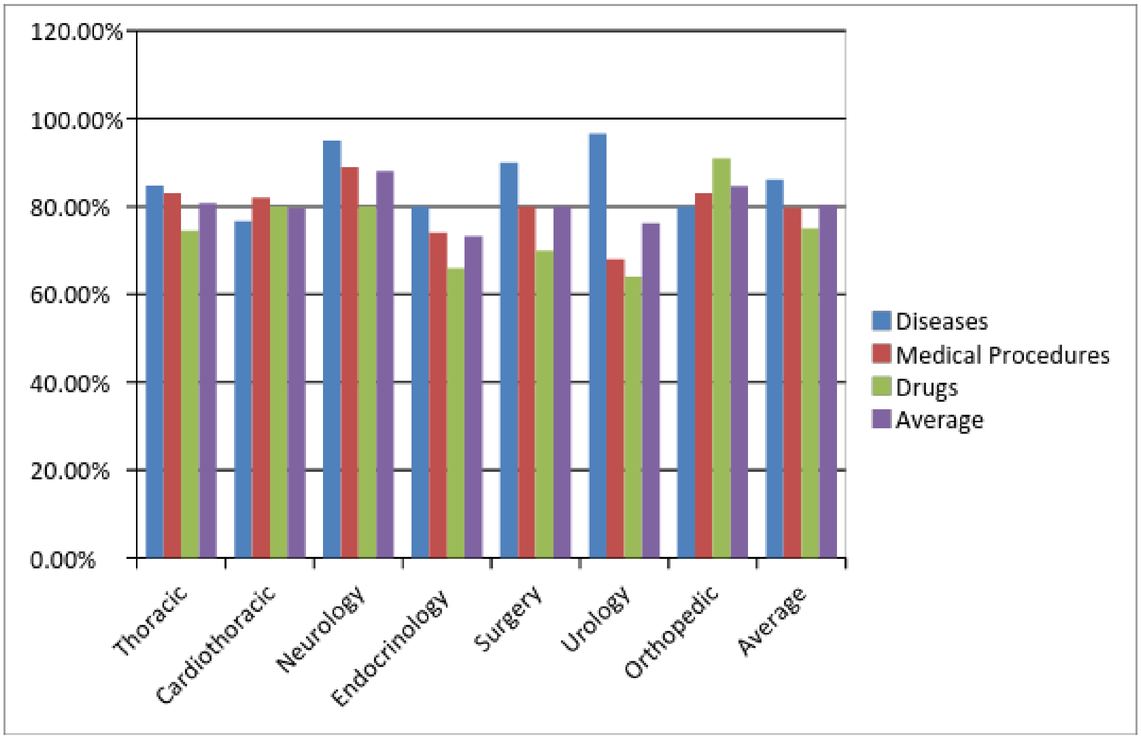


Figure (5.15): The results of expert evaluation for association rules

5.8 *Summary*

This chapter presented and analyzed the experimental results. It explained the experimental setup were 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, filled by doctors in each field, we found that the proportion of accuracy association rules, which have been extracted it is about 80%.

Chapter 6

Conclusion and Future Work

Chapter 6
Conclusion and Future Work

This chapter concludes the thesis, then presents contribution of the thesis, after that, describes some recommendations, finally; gives some suggestions for future work.

6.1 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 grow 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.

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

In this research, we created a new approach to extract important information and detect and predicted diseases, drugs, and medical procedures using text and data mining techniques. Ministry of health dataset were 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.

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

We have shown that our 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 were 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%.

We note that a good number of rules were not known to doctors, so some doctors have evaluated them (normal, unacceptable) and after searching for the accuracy of these rules we found some of these rules very accurate.

6.2 Recommendations

There are some of the recommendations can be formulated to adopt the goal of this thesis, 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 relationship 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.

6.3 Future Work

According to the results of experiment and the limitations that we faced in our thesis, 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.

Bibliography

Bibliography

(2017, January 23). Retrieved from SNOMED International: http://www.snomed.org/snomed-ct

Adupa, A., Garg, R., Corona-Cox, J., Shah, S., & Jonnalagadda, S. (2016). An Information Extraction Approach to Prescreen Heart Failure Patients for Clinical Trials. *CoRR abs/1609.01594*.

Akthar, F., & Hahne, C. (2012). *RapidMiner 5*. Dortmund: Rapid-I GmbH, Stockumer Str. 475. www.rapid-i.com.

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

Aramaki, E., Miura, Y., Tonoike, M., Ohkuma, T., Mashuichi, H., & Ohe, K. (2009). Medical Text Summarization System based on Named Entity Recognition and Modality Identification. *Proceedings of the Workshop on BioNLP*.

Asgari, H., Masoumi, B., & sheijani, O. (2014). Automatic Text Summarization Based on Multi-Agent Particle Swarm Optimization. *Iranian Conference on Intelligent Systems (ICIS)*.

Assuiti, L., Lanzoni, G., Santos, F., Erdmann, A., & Meirelles, B. (2013). Hearing loss in people with HIV/AIDS and associated factors: an integrative review. *Brazilian Journal of Otorhinolaryngology, 79*(2).

Atdag, S., & Labatut, V. (2013). A Comparison of Named Entity Recognition Tools Applied to Biographical Texts. *2nd International Conference on Systems and Computer Science, Villeneuve d'Ascq (FR)*, 228-233.

Azpeitia, A., Cuadros, M., Gaines, S., & Rigau, G. (2014). Nerc-fr: Supervised named entity recognition for French. *17th International Conference on Text, Speech and Dialogue (TSD).*

Bertram, D. (2017, January 25). *Likert Scales are the meaning of life*. Retrieved from poincare: http://poincare.matf.bg.ac.rs/~kristina/topic-dane-likert.pdf

Bhargava, N., & Shukla, M. (2016). Survey of Interestingness Measures for Association Rules Mining: Data Mining, Data Science for Business Perspective. *IRACST - International Journal of Computer Science and Information Technology & Security (IJCSITS),* ISSN: 2249-9555 6 (2).

Bharti, S., Babu, K., & Jena, S. (2017). Automatic Keyword Extraction for Text Summarization: A Survey. *National Institute of Technology*.

Bordea, G., Stefan, T., & Handschuh, S. (2015). First steps toward semi-automatic extraction of claims from scientific publications.

Bunescu, R., Ge, R., Mooney, R., Marcotte, E., & Ramani, A. (2002). Extracting Gene and Protein Names from Biomedical Abstracts. *Unpublished Technical Note*.

Chang, T., & Hsiao, W. (2008). A hybrid approach to automatic text summarization. *Computer and Information Technology. CIT 2008. 8th IEEE International Conference on*.

Chen, P., & 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.*

Choi, M., Verspoor, K., & Zobel, J. (2014). Analysis of Coreference Relations in the Biomedical Literature. *Analysis of Coreference Relations in the Biomedical Literature. In Proceedings of Australasian Language Technology Association Workshop*, 134.

Chun, J., Tsuruoka, Y., Kim, J., Shiba, R., Nagata, N., Hishiki, T., & Tsujii, J. (2006). Extraction of Gene-Disease Relations from Medline Using Domain Dictionaries and Machine Learning. *Pacific Symposium on Biocomputing*, 11:4-15.

Cunningham, H., Maynard, D., & Tablan, V. (2000). *JAPE: a Java Annotation Patterns Engine (Second Edition).* Sheffield: Department of Computer Science, University of Sheffield.

Das, D., & Martins, A. (2007). A Survey on Automatic Text Summarization. *Language Technologies Institute. Carnegie Mellon University.*

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.

Desai, P., H, S., & Chiplunkar, N. (2015). Template Based Algorithm for Automatic Summarization and Dialogue Management for Text Documents. *IJRET: International Journal of Research in Engineering and Technology*, 04 (11).

Desai, P., H, S., & Chiplunkar, N. (2015). Template Based Algorithm for Automatic Summarization and Dialogue Management for Text Documents. *IJRET: International Journal of Research in Engineering and Technology,* 04 (11).

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

Elsebai, A. (2009). A rules based system for named entity recognition in modern standard. *University of Salford.*

Elyezjy, N., & El-Halees, A. (2015). Investigating crimes using text mining and network analysis. *International Journal of Computer Applications*, 126 (8), 19-25.

Fajer, H., & Omar, N. (2014). Automatic Arabic Text Summarization Using Clustering and Keyphrase Extraction. *International Conference on Information Technology and Multimedia (ICIMU)*.

Finkel, J., Grenager, T., & Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*.

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.

Gonnade, P. (2015). Automatic Text Summarization System Using Extraction Based Technique. *International Journal of Emerging Technologies and Innovative Research (www.jetir.org)*, 2 (7).

Greenwood, M., Roberts, A., Aswani, N., & Gooch, P. (2012). Initial prototype for semantic annotation of the Khresmoi literature. *project deliverable Khresmoi*.

Gupta, V., & Lehal, G. (2010). A Survey of Text Summarization Extractive Techniques. *Journal of Emerging Technologies in Web Intelligence*, 2 (3), 258-268.

Hipp, J., Guntzer, U., & Nakhaeizadeh, G. (2000). Algorithms for Association Rule Mining a General Survey and Comparison. *ACM SIGKDD* , 2 (1), 58-64.

Hristovski, D., Peterlin, B., Mitchell, J., & Humphrey, S. (2003). Improving Literature Based Discovery Support by Genetic Knowledge Integration. *Institute of Biomedical Informatics*.

Ježek, K., & Steinberge, J. (2007). Automatic Text Summarization (The state of the art 2007 and new challenges). *in the proceeding of Document Understanding Confernce (DUC), Rochester, New York USA*.

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

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

Jung, J., & Jo, G. (2003). Template-Based E-mail Summarization for Wireless Devices. *computer and information sciences - ISCIS, LNCS 2869*, 99–106.

Konkol, M., & Konop, M. (2011). Maximum Entropy Named Entity Recognition for Czech language. *14th International Conference*.

lahari, E. K. (2014). A Comprehensive Survey on Feature Extraction in Text Summarization. *Int.j. computer Technonlogy & Application, Vol 5 (1)*, 248-256.

Lee, H., Lau, F., & Quan, H. (2010). A method for encoding clinical datasets with SNOMED CT. *BMC Medical Informatics and Decision Making*.

Liu, X., Zhang, S., Wei, F., & Zhou, M. (2011). Recognizing Named Entities in Tweets. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 359–367.

Mani, I., & Maybury, M. (2001). Advaces in Automatic Text Summarization. *The MIT Cambridge, Massachusetts london, England*.

Marrero, M., Sánchez-Cuadrado, S., Morato, J., & Andreadakis, G. (2009). Evaluation of Named Entity Extraction Systems. *Advances in Computational Linguistics*, 47-58.

Meystre, S., & Haug, P. (2006). Natural language processing to extract medical problems from electronic clinical documents. *Performance evaluation. Journal of Biomedical Informatics 39*, 589–599.

Mialtz, M. (2010). Information Extraction from Wikipedia Using Pattern Learning. *Acta Cybernetica 19*, 677-694.

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).

Muslea, I. (1999). Extraction Patterns for Information Extraction Tasks: A Survey. *In AAAI-99 Workshop on Machine Learning for Information Extraction*.

N.V, S., Mitra, P., & Ghosh, S. (2010). Conditional Random Field Based Named Entity Recognition in Geological Text. *International Journal of Computer Applications (0975 – 8887),* 1 (3).

Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigations*, 30, 3-26.

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.

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.

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).

Radev, D., & Mckeown, K. (2002). Introduction to the special issue on summarization. *Computational Linguistics*, 28 (4), 339 – 408.

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

Redd, M., & Hanumanthappa, M. (2014). Semantical and Syntactical Analysis of NLP. *(IJCSIT) International Journal of Computer Science and Information Technologies,*, 5 (3).

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

Saxena, R., Shrivastava, S., & Mathur, A. (2012). Association Rules Mining using Modified Genetic Algorithm. *International Journal of Scientific Engineering and Technology*, 1 (4), 35-38.

Sharma, R., & Goyal, V. (2011). Name Entity Recognition Systems for Hindi Using CRF Approach. *Volume 139 of the series Communications in Computer and Information Science*, 31-35.

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

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.

Sun, B. (2011, November 2). *Named entity recognition Evaluation of Existing Systems*. Retrieved from http://daim.idi.ntnu.no/masteroppgave.pdf

Takeuchi, K., & Collier, N. (2005). Bio-Medical Entity Extraction using Support Vector Machines. *Artificial Intelligence in Medicine*.

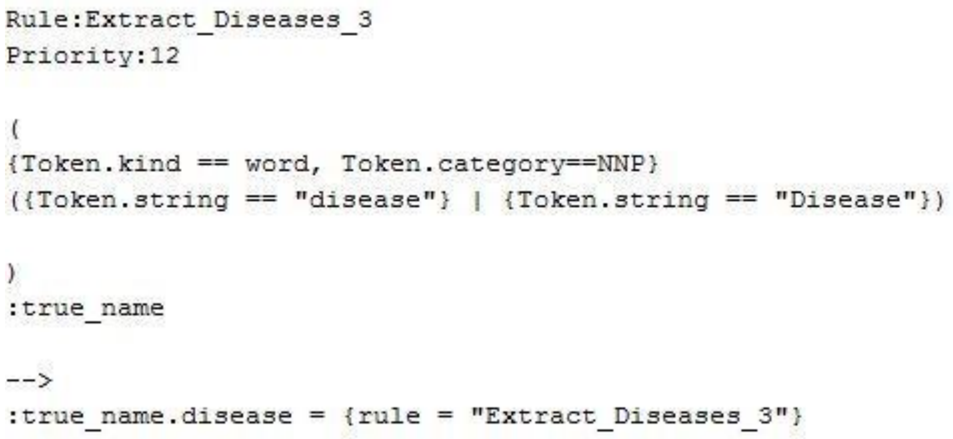
Wang, Y., & Patrick, J. (2009). Cascading Classifiers for Named Entity Recognition in Clinical Notes. *Workshop Biomedical Information Extraction - Borovets, Bulgaria*, 42–49.

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.

Yang, J., & Chuang, W. (2000). Text Summarization by Sentence Segment Extraction Using Machine Learning Algorithms. *PADKK '00 Proceedings of the 4th Pacific-Asia conference on Knowledge Discovery and Data Mining, Current Issues and New Applications*, 454-457.

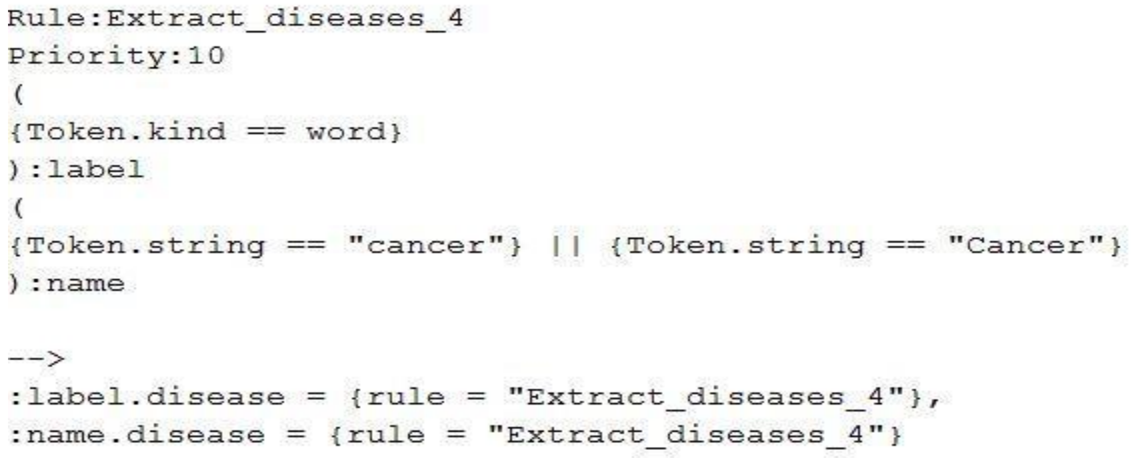
Appendix

Appendix A



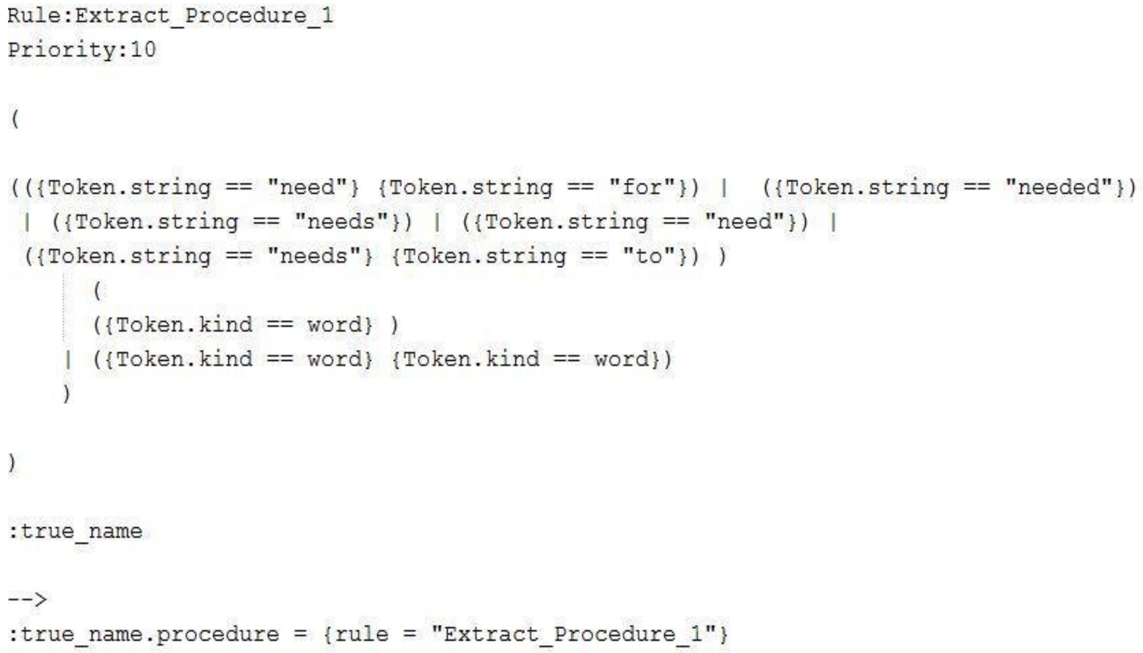
```
Rule:Extract_Diseases_3
Priority:12

(
{Token.kind == word, Token.category==NNP}
({Token.string == "disease"} | {Token.string == "Disease"})

)
:true_name

-->
:true_name.disease = {rule = "Extract_Diseases_3"}
```

List 0.1: JAPE Rule to extract diseases



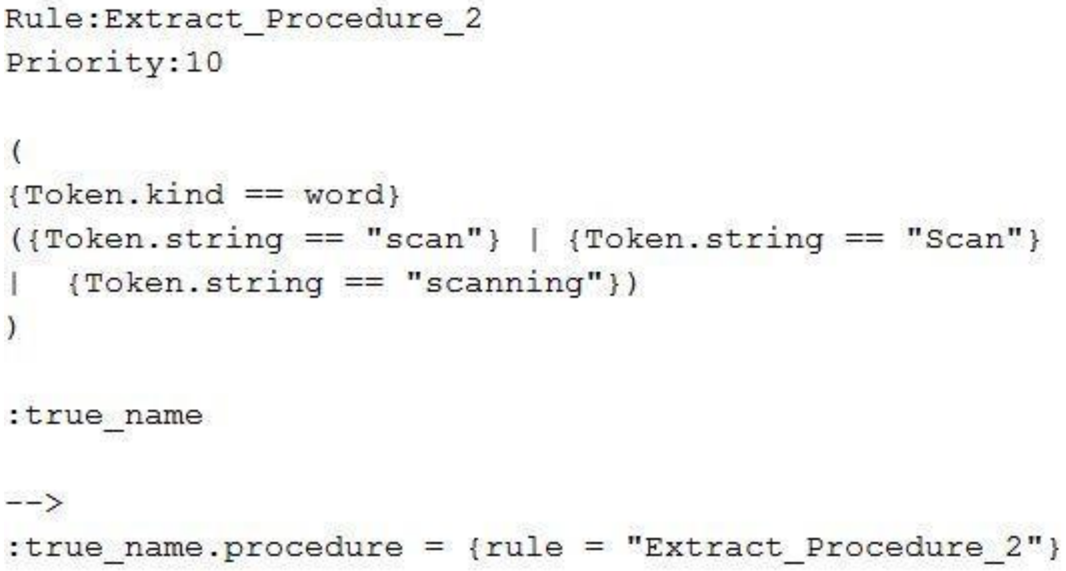
```
Rule:Extract_diseases_4
Priority:10
(
{Token.kind == word}
):label
(
{Token.string == "cancer"} || {Token.string == "Cancer"}
):name

-->
:label.disease = {rule = "Extract_diseases_4"},
:name.disease = {rule = "Extract_diseases_4"}
```

List 0.2: JAPE Rule to extract cancers disease



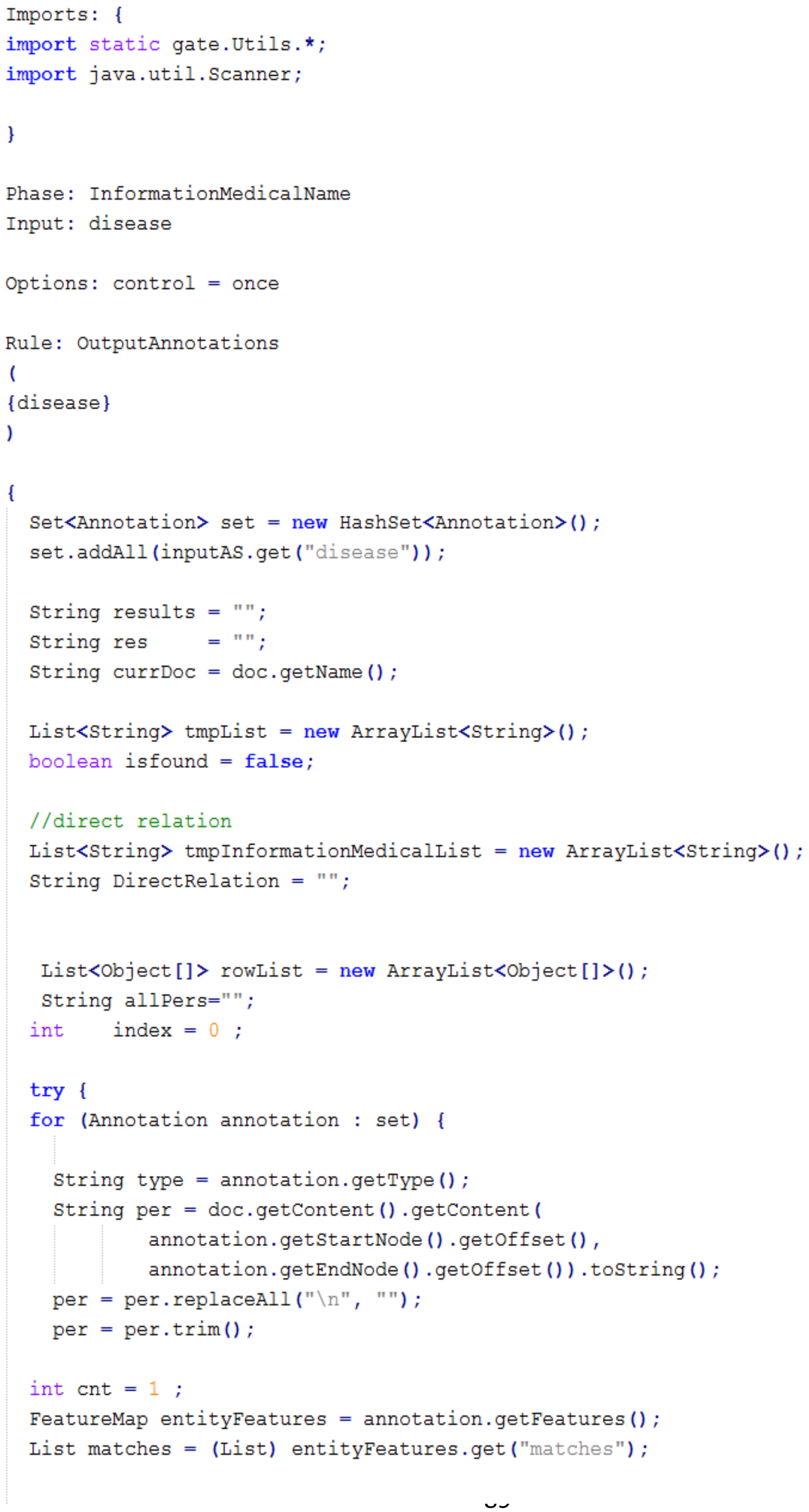
```
Rule:Extract_Procedure_1
Priority:10

(

(({Token.string == "need"} {Token.string == "for"}) |  ({Token.string == "needed"})
 | ({Token.string == "needs"}) | ({Token.string == "need"}) |
 ({Token.string == "needs"} {Token.string == "to"}) )
        (
        ({Token.kind == word} )
      | ({Token.kind == word} {Token.kind == word})
      )

)

:true_name

-->
:true_name.procedure = {rule = "Extract_Procedure_1"}
```

List 0.3: JAPE Rule to extract Procedures

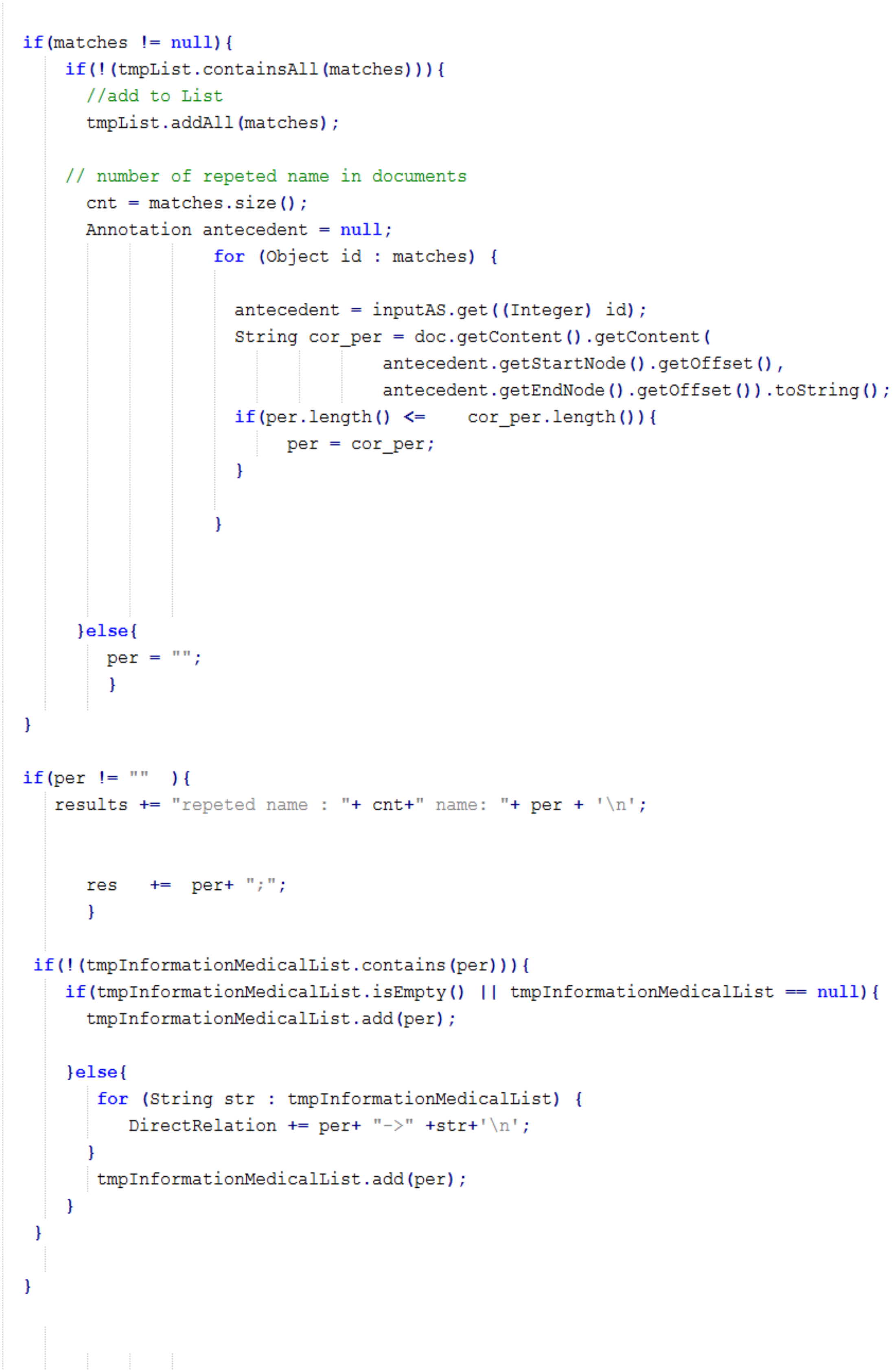


```
Rule:Extract_Procedure_2
Priority:10

(
{Token.kind == word}
({Token.string == "scan"} | {Token.string == "Scan"}
|  {Token.string == "scanning"})
)

:true_name

-->
:true_name.procedure = {rule = "Extract_Procedure_2"}
```

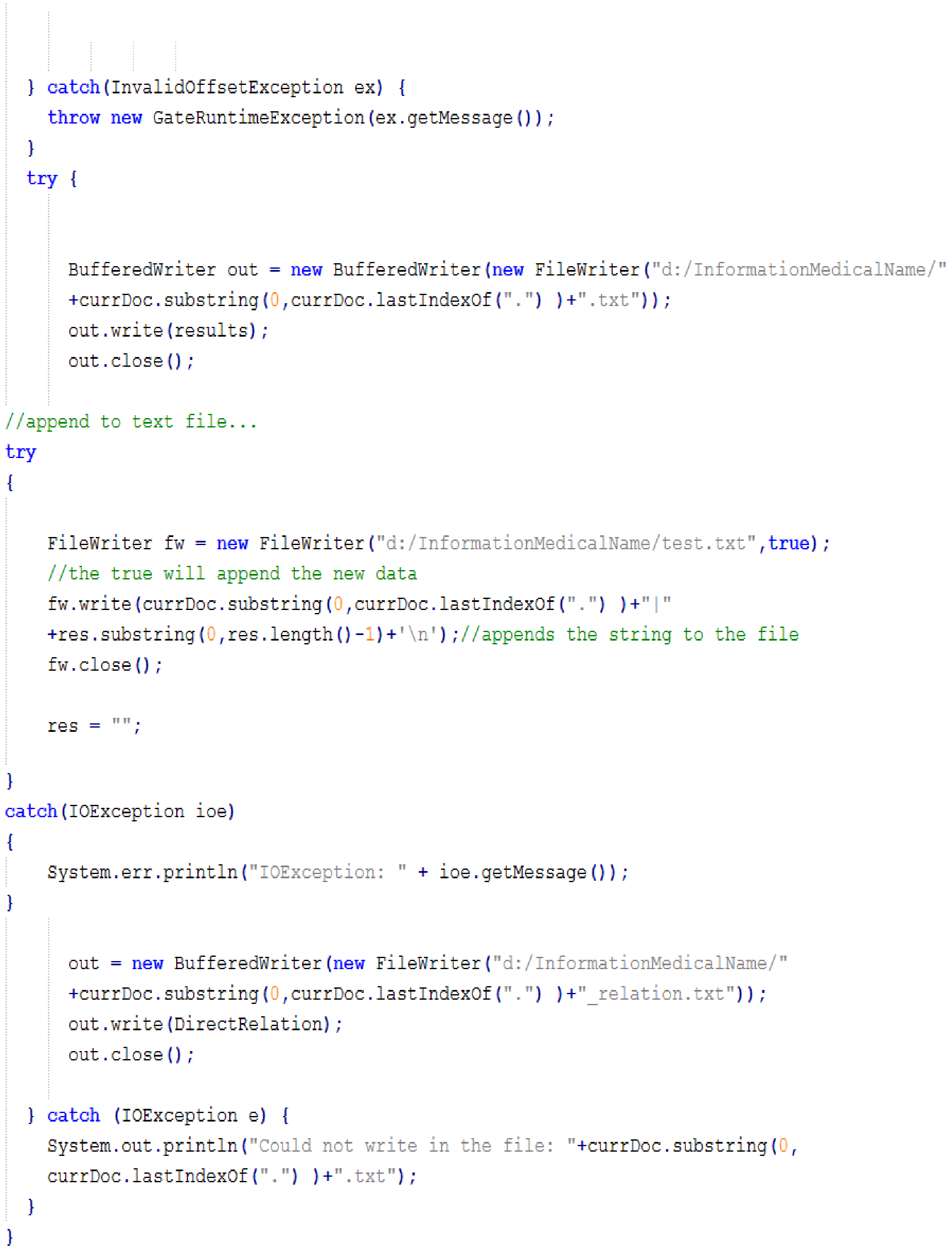
List 0.4: JAPE Rule to extract Procedures



```
Imports: {
import static gate.Utils.*;
import java.util.Scanner;

}

Phase: InformationMedicalName
Input: disease

Options: control = once

Rule: OutputAnnotations
(
{disease}
)

{
  Set<Annotation> set = new HashSet<Annotation>();
  set.addAll(inputAS.get("disease"));

  String results = "";
  String res     = "";
  String currDoc = doc.getName();

  List<String> tmpList = new ArrayList<String>();
  boolean isfound = false;

  //direct relation
  List<String> tmpInformationMedicalList = new ArrayList<String>();
  String DirectRelation = "";


   List<Object[]> rowList = new ArrayList<Object[]>();
   String allPers="";
  int     index = 0 ;

  try {
  for (Annotation annotation : set) {

    String type = annotation.getType();
    String per = doc.getContent().getContent(
            annotation.getStartNode().getOffset(),
            annotation.getEndNode().getOffset()).toString();
    per = per.replaceAll("\n", "");
    per = per.trim();

  int cnt = 1 ;
  FeatureMap entityFeatures = annotation.getFeatures();
  List matches = (List) entityFeatures.get("matches");
```



```
if(matches != null){
    if(!(tmpList.containsAll(matches))){
      //add to List
      tmpList.addAll(matches);

    // number of repeted name in documents
      cnt = matches.size();
      Annotation antecedent = null;
                  for (Object id : matches) {

                    antecedent = inputAS.get((Integer) id);
                    String cor_per = doc.getContent().getContent(
                                  antecedent.getStartNode().getOffset(),
                                  antecedent.getEndNode().getOffset()).toString();
                    if(per.length() <=    cor_per.length()){
                         per = cor_per;
                    }

                  }



     }else{
        per = "";
        }
}

if(per != ""  ){
   results += "repeted name : "+ cnt+" name: "+ per + '\n';


      res   +=  per+ ";";
      }

 if(!(tmpInformationMedicalList.contains(per))){
    if(tmpInformationMedicalList.isEmpty() || tmpInformationMedicalList == null){
      tmpInformationMedicalList.add(per);

    }else{
       for (String str : tmpInformationMedicalList) {
          DirectRelation += per+ "->" +str+'\n';
      }
       tmpInformationMedicalList.add(per);
    }
 }

}
```



```
  } catch(InvalidOffsetException ex) {
    throw new GateRuntimeException(ex.getMessage());
  }
  try {


      BufferedWriter out = new BufferedWriter(new FileWriter("d:/InformationMedicalName/"
      +currDoc.substring(0,currDoc.lastIndexOf(".") )+".txt"));
      out.write(results);
      out.close();

//append to text file...
try
{

    FileWriter fw = new FileWriter("d:/InformationMedicalName/test.txt",true);
    //the true will append the new data
    fw.write(currDoc.substring(0,currDoc.lastIndexOf(".") )+"|"
    +res.substring(0,res.length()-1)+'\n');//appends the string to the file
    fw.close();

    res = "";

}
catch(IOException ioe)
{
    System.err.println("IOException: " + ioe.getMessage());
}

      out = new BufferedWriter(new FileWriter("d:/InformationMedicalName/"
      +currDoc.substring(0,currDoc.lastIndexOf(".") )+"_relation.txt"));
      out.write(DirectRelation);
      out.close();

  } catch (IOException e) {
    System.out.println("Could not write in the file: "+currDoc.substring(0,
    currDoc.lastIndexOf(".") )+".txt");
  }
}
```

List 0.5: JAPE Code to Extract Medical Information (Diseases - Procedures, Drugs)

#	Doctors Names	Department
1	Baha Alden jabber al-ataawna	Cardiothoracic
2	Mohammed hessian Ali Habib	Cardiothoracic
5	Mohammed Abd-Alhamed Mohammed abu hasseira	Cardiothoracic
12	Mohammed Mohammed Mohammed shahwan	Cardiothoracic
19	Mohammed Ibrahim Nasar	Cardiothoracic
2	Waleed Abdulsalam Daoud	Thoracic
3	Mahmoud Hashem nomaan alkozendar	Thoracic
4	Hessian Ismail Ibrahim alattar	Thoracic
14	Ayman Tewfik Abu el-awf	General Internal
6	Mohammed Abdel Raheem Ahmed zaqout	General Internal
7	Sami Abdulsalam Tewfik alissawi	Endocrinology
8	Mohammed Ali marouf	Endocrinology
9	Marwan Hasan Mahmoud Abu saada	General Surgery
10	Mohamoud qassem Mohammed abu khater	Neurology
11	Mona Yousef Mohamoud kaskeen	Neurology
15	Read Saleh fares seyam	Orthopedic
16	Shams Alden Samir el-dejany	Orthopedic
17	Mohamoud Adnan matter	Orthopedic
18	Amal Saleh Abu daya	Urology
13	Osama Mohammed Salman Abu jabal	Urology
20	Fayez Fawaz Abdelrahman zedan	Urology

List 0.6: Names of Doctors Who Evaluated the Rules