國立臺灣大學電機資訊學院資訊網路與多媒體研究所

碩士論文

Graduate Institute of Networking and Multimedia

College of Electrical Engineering and Computer Science

National Taiwan University

Master Thesis

肝癌病患文字病歷報告之錯別字辨識與更正方法

A method for detecting misspelled words in medical narrative reports: A case study on the patients with liver cancer

劉子華

Tzu-Hua Liu

指導教授：賴飛羆 博士

Advisor: Feipei Lai, Ph.D.

中華民國 101 年 6 月

June, 2012

誌謝

轉眼間，即將從臺大畢業了，從一開始踏入研究所的懵懵懂懂，到現在可以大聲的宣告我完成了人生一個階段性的目標，在這不止學習到學業方面的知識，更結交到不少志同道合的朋友，更真正發掘了自我真正的興趣與目標。

在這時刻，我將藉由這份論文來說明碩士生涯中的研究成果，這份論文能夠順利完成，首要感謝的便是我的指導教授 賴飛羆老師，在老師的實驗室中，老師不僅指導我們研究的方向，更看重我們對於研究與學習方式與態度，老師鼓勵我們不斷大膽的嘗試各種可能，也感謝老師的支持，讓我能在除了研究上精進，更有機會到公司實習與至美國進行交換學生計畫。

平曉鷗 學姐更是這份論文功不可沒的幕後推手，學姐不僅擁有豐富的學識與經驗，更擁有驚人的耐心，常常不厭其煩的叮嚀我常遺漏的小細節，更細心且詳盡的回答我任何的問題，讓我在研究的路上有著一盞指引方向的明燈。也感謝小組成員意儒學姐、子軍學長、景崴學姐、亞霖學長及相茹學妹，很懷念那為了找尋問題的答案而不眠不休討論的日子。

而有了實驗室成員的陪伴，讓我在這兩年的研究生活一點也不感到孤單，謝謝偉昕學長、鎮宇學長、家平學長、俐瑾學姐、逸帆學長、峰生學長、煌仁、冠仲、緯志、時廷，及實驗室的學長姐與學弟妹們。

感謝 莊立民教授、沈榮麟教授、陳澤雄教授、鐘玉芳教授願意於百忙之中抽空前來擔任我的口試委員，並細心審閱指導我的論文，使得論文更加完成。

更感謝在美國進行交換學生計畫時，所認識的女友 怡廷，除了幫我校正論文裡英文的錯誤，也在我交換期間非常照顧我，申請學校時也不停的為我加油打氣，

接下來即將到美國攻讀博士班，希望之後的日子都有妳的陪伴。

我想把最後的感謝獻給我的父母與家人，沒有他們的支持，我無法全力於學業上衝刺，更不可能進入臺灣最高之學術殿堂，很感恩我能出生在如此和樂溫馨的家庭，父母即使肩上擔著不小的財務壓力，但心中只想到要給孩子最好的生活。無論自己作了什麼決定，雖然父母總是因為擔心孩子而嘮叨幾句，但仍是在背後全力支持著我的選擇。父親也是本校的校友，於 2000 年進修森林系博士班，經過十年的努力，終於在 2010 年取得了森林系博士學位，從中我看到了父親承擔了課業與工作雙方面的壓力，研究的路途中雖有沮喪，但卻不斷努力堅持地走下去，最終達成目標。同年之際，我接續的父親的任務，繼續在臺大攻讀碩士學位，一直以來我都以父親為榮，如今我希望能成為父母的驕傲。也期待我未來的兒女能接下此任務延續下去。

中文摘要

病歷資料擁有豐富的疾病、醫療程序和治療結果等資訊。在之前的研究裡，我們實做一資訊擷取系統提取肝癌病人文字報告裡肝癌相關資訊。資訊擷取系統提取的結果將用於建立預測肝癌復發的模型。然而，由於這些文字的醫療報告為醫護人員手動輸入，其中難免會有錯別字，這些因素造成醫療擷取系統抽取資訊上的困難，而因此遺漏掉無法被抽取出的珍貴醫療資訊，所以如何減少報告中錯別字對於未來的研究是非常重要的。

本研究的目的在於提供一個有效率的方式去辨識醫療報告中之錯別字並加以更正，以幫助醫療擷取系統能抽取出更多醫療資訊。我們設計一套錯別字辨正與標準化系統，用於校正報告中之錯別字。透過這套方法，能在系統抽取資訊前，將文章中內之錯別字更正，以幫助醫療擷取系統能從中擷取到先前因錯別字與不同表示法的原因而無法被擷取出來之資訊，以提高系統之資訊抽取率。由於醫療報告過於龐大，使用人為方式尋找錯誤是非常耗費人力與時間的，透過這個方法，可有效減少醫療報告之錯別字，並改善病歷之品質。

關鍵字：病歷資料、資訊擷取系統、錯別字、標準化、醫療資訊。

ABSTRACT

Textual medical records contain valuable information about diseases, medical procedures and treatment results. In our previous work, we implemented the information extraction system for extracting the desired information from liver cancer patients' textual reports. These extracted results produced by information extraction system are used for supporting the development of recurrence predictive model. However, these narrative reports are made by human manually. Therefore, improving the correctness of medical reports is very important for further research.

In the study, we already implemented the information extraction system which can extract medical information for liver cancer recurrence predicting model. But, detecting and correcting the misspelling words of all medical reports manually would be a time-consuming and labor-intensive task. Therefore, the aim of this study is to provide an efficient way for facilitating the process of checking and correcting the misspelling words. We implemented the error handling system for correcting the misspelled words of each medical report. After the preprocessing procedure executed by the error handling system, the information extraction system can extract out those information which cannot be found due to the misspelling words. In this way, it can highly promote the accuracy of the medical extracted results.

Keywords: textual medical records, information extraction system, misspelling, normalization, medical informatics.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1 Introduction

1.1 Motivation

Electronic Medical Record (EMR) about liver cancer contains great quantity of precious medical information about Admission, Discharge, Summary and Surgery. In our previous work we developed the Medical Information Extraction System (MIES), which helps us to extract out the medical information from those EHRs. The medical information is used for our disease recurrent predict model. However, there are some misspelling words in the reports which lead to some information cannot be extracted out. Therefore, to reduce misspelling words in the medical report is very important to further application.

In this case, we focus on the textual report about the liver cancer. However, the most common method to correct the misspelling words is checking and correcting the medical reports manually. But, the process of manually reviewing these medical reports is a stubborn task that is usually long and laborious. In the medical reports from the Electronic Medical Record (EMR), the correct words occupy a larger proportion compared to the misspelling words. Therefore, reviewers waste most of their time on those correct words. According to this problem, we aim to design an error handling system to find those misspellings. With the help of the system, the IE system is able to extract out the valuable information which cannot be extracted out in previous work.

1.2 Purpose

Misspellings are common in medical documents and can be an obstacle to information retrieval which has been around for many years [2]. Although significant progress has been made, there is still room for improvement. A common information retrieval (IR) task includes identification of documents that contain some or all members of a set of semantic concepts provided by the user. Many narrative medical documents are created under significant time constraints and are not proofread afterwards, resulting in frequent misspellings [21].

In this study, we developed a Misspellings Handling System (MHS) for correcting the misspellings in the Electronic Medical Record (EMR), in order to increase the information can be extracted out by the Medical Information Extraction System we developed before.

1.3 General Procedure

In this study we processed the textual medical records about patients with liver cancer in National Taiwan University Hospital (NTUH). Before the textual medical records are sent to the IE system. The records will be passed to Misspellings Handling System (MHS). The MHS will use different algorithms to calculate the similarity between every word in the report and the word in the corpus we predefined before. We defined a certain range of distance. When the calculated distance is in the range, we considered these words may be misspelled. The MHS will replace the misspelled word with correct word in corpus.

After the preprocessing of misspellings correcting, the corrected report will be sent to the IE system to generate a lot of extracted results according to the concepts we defined previously. The extracted results are grouped into different kinds of information, such as tumor, tumor size and tumor location by the IE system. The extracted information will be represented in structured format and used for supporting the development of recurrence predictive model. In order to evaluate the performance of the MHS, we divided our data into training set and testing set.

.

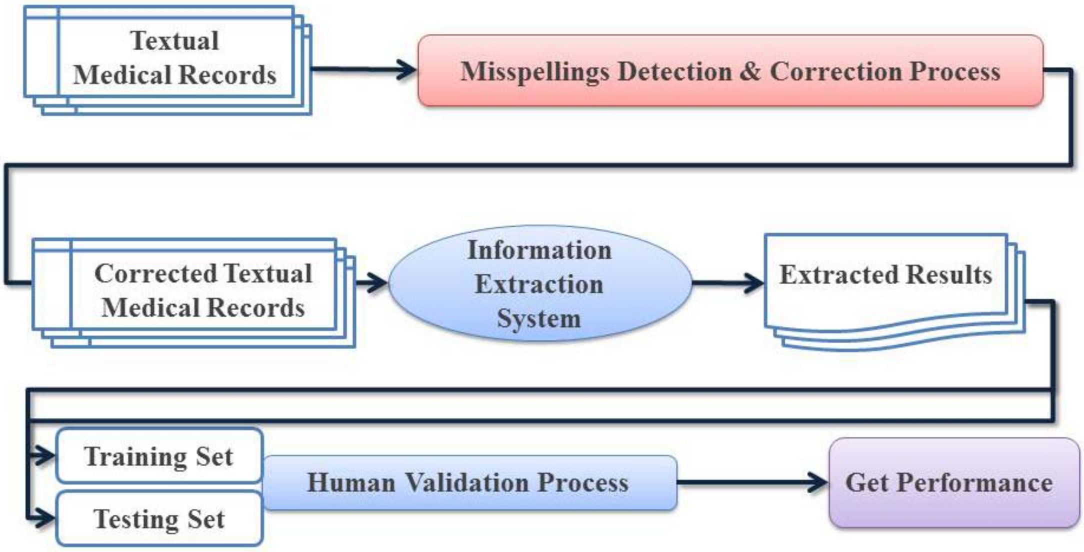

Textual Medical Records
Misspellings Detection & Correction Process
Corrected Textual Medical Records
Information Extraction System
Extracted Results
Training Set
Testing Set
Human Validation Process
Get Performance

Figure 1 Overview of the general steps of this research.

1.4 Thesis Organization

The rest of this thesis is divided into four chapters as follows. Chapter 2 provides introduction to the use of information extraction system on medical domain, the related domain knowledge of liver cancer, the review of related works and the introduction to the datasets used in this thesis. Chapter 3 covers the methods we employed to develop this application such as the information extraction process, the manually reviewing of extracted information and the misspellings handling system to correct the misspellings in the report. Chapter 4 presents the results and discussions of this application. Finally, the conclusions and future works are given in the last chapter.

Chapter 2 Background

2.1 Electronic Clinical Report System

In the past, most of the clinical reports were stored in paper-based records. However, it is not convenient for further data analysis and transfer between medical institutions. Those problems can be solved by using electronic clinical report system. The number of hospitals adopting electronic clinical report system is increasing progressively. In most of systems, the clinical data are stored in database in free text format. Even though it solved the transfer problem, analyzing those free text format data still an issue need to be resolved. Because the free text report contains unstructured context which can not be understood by a computer without human intervention. Another solution to this problem is to record clinical information by formatted structure. Compared to the free text method, formatted structure can strongly improve the data structuralizing. Those well-formatted data can be analyzed, queried and handled by a computer directly. The electronic clinical report system recording clinical information in formatted structure can help medical experts analyze those medical data more efficiently.

2.2 Information Extraction

In order to transform a free text report into a structured report, information extraction is an adequate way to accomplish this goal. Information Extraction applied several technologies such as natural language processing (NLP) for handling non-structuralized documents automatically. The automatic processing of medical data has become an attractive issue for the past decades. In the medical domain, information extraction helps medical experts to extract out useful information and worthy medical knowledge.

2.2.1 Rule-Based Information Extraction System

Compared with statistical processing (machine learning methods), Natural Language Processing techniques are kind of rule-based methods. Due to the machine learning methods are not applicable to be adopted directly for analyzing complicated structures of medical reports, which are often used to process the cases for simple outcome such as certain diseases whether occur or not. In contrast, the rule-based methods can deal with this kind of issue more rapidly, and creating the rule-based system requires a lot of domain knowledge and time.

A rule-based information extraction system was developed for extracting medical information from Polish medical documents about mammography reports and Diabetic patient records[3].

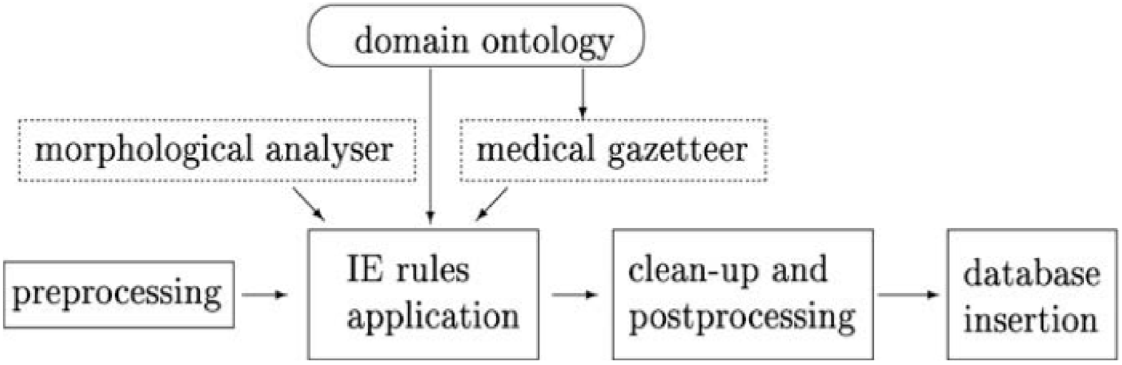


Figure 2 Overall process of rule-based information extraction system

2.2.2 Ontology

Ontologies represent knowledge about an interesting domain [4]. The ontology defines the concepts in the domain and also the relationships that hold between those concepts. The most recent development in standard ontology languages is ontology web language (OWL) from the World Wide Web Consortium (W3C) [5].

In the IE system, we use regular expressions to recognize the information in the textual medical records. It is inconvenient to update the current version of regular expression in the IE system. The open source Protégé ontology editor was adopted to construct the ontology model which provides a friendly user interface [6]. After reviewers had added the additional regular expressions to the corresponding ontology classes in the human-labeling process, the protégé exported the ontology to an OWL/XML file. The IE system would read this OWL/XML file and add the regular expression to the corresponding concept[7].

The advantage of using the ontology is that it provides a lot of elasticity to our IE system, if we want to add additional regular expressions in the extraction process, we do not have to modify the regular expression in our IE system but to add additional regular expression in the ontology model[7].

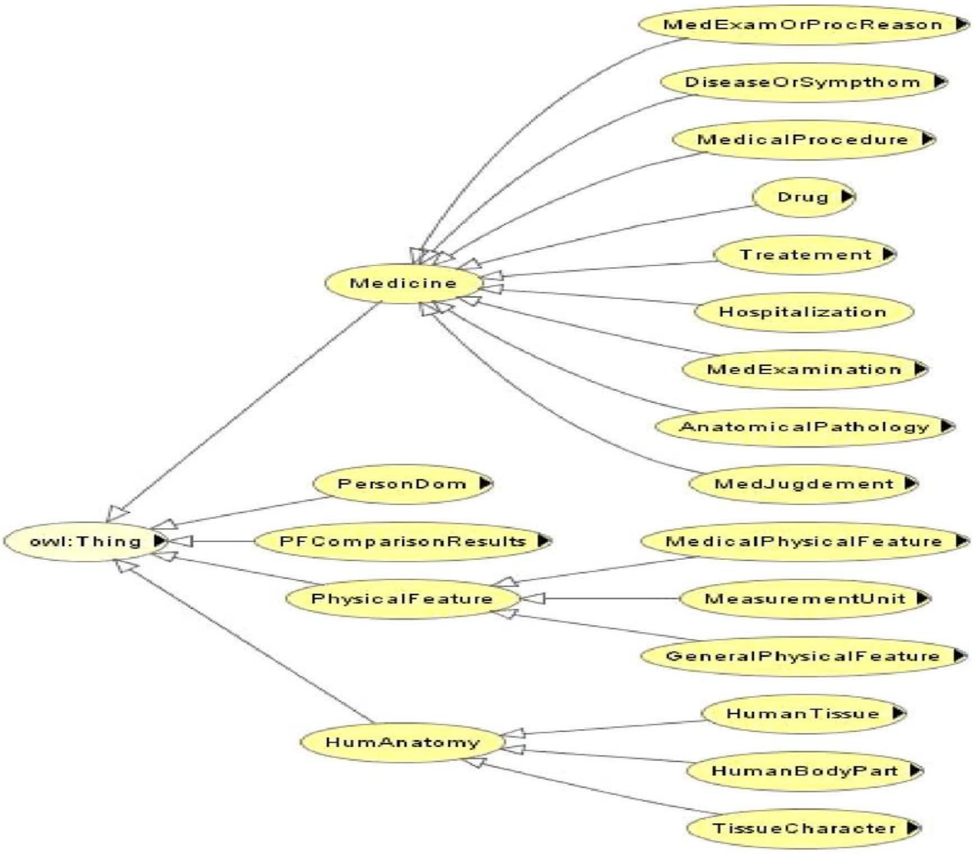
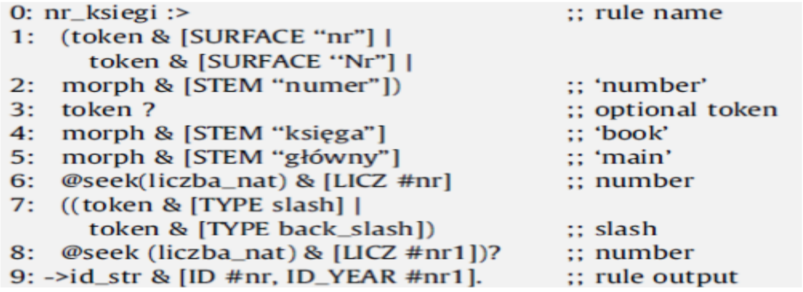


Figure 3 Overview of top level of medical ontology

The authors in the study [8] mentioned above defined various information extraction rules for processing clinical documents. These rules are used to extract specific information and the extracted data are filled into the predefined templates according to the ontology[3]. These extraction rules are defined by the Regular Expression based on the ontology and medical lexicon in Polish. The following figure is an extraction rule defined in the rule-based information extraction system, which is used to distinguish patient’s identification number. The Regular Expression body, placed after the “:>” character which describes input sequences, for example, elements that must be identified in the text. While the “->” character points out the structure of the resulting output.

There were experts to evaluate the accuracy of extracted result manually. For most of the extracted templates, precision and recall are above 80%. This result proves the rule-based information extraction system can be reliable and useful for automatic clinical data extraction.



```
0: nr_ksiegi :>                               ;; rule name
1:   (token & [SURFACE "nr"] |
       token & [SURFACE "Nr"] |
2:   morph & [STEM "numer"])                  ;; 'number'
3:   token ?                                  ;; optional token
4:   morph & [STEM "księga"]                  ;; 'book'
5:   morph & [STEM "główny"]                  ;; 'main'
6:   @seek(liczba_nat) & [LICZ #nr]           ;; number
7:   ((token & [TYPE slash] |
       token & [TYPE back_slash])             ;; slash
8:   @seek (liczba_nat) & [LICZ #nr1])?       ;; number
9: ->id_str & [ID #nr, ID_YEAR #nr1].         ;; rule output
```

Figure 4 Extraction rule for patient's identification number

2.3 Information Extraction System on Medical Records

In the medical domain, the Electronic Medical Record (EMR) is becoming a daily experience in most of the practices and hospitals worldwide. However, much of the available data is in free text form, a convenient way of expressing concepts and events but especially challenging if one wants to perform automatic searches, summarization or statistical analyses. In order to make full use of the information contained in EMRs for improving the quality of medical care through decision support, evidence-based medicine, disease surveillance, the Information Extraction (IE) System is designed to process textual clinical records and to perform automatic and accurate mapping of free text reports onto a structured representation. Thus, there seems to be an increased demand for information extraction tools applied to those medical narratives records. The use of the IE system can help us efficiently to acquire the information hidden in the large amount of medical records [9].

We developed an IE system for extracting the valuable information from liver cancer patients’ textual reports. There are three issues should be concerned when we try to extract medical information by the IE system.

1. The reusability and expansion of the IE system.
2. The data quality of extracted results from the IE system
3. The coverage of extracted results from the IE system

First of all, the IE system uses Regular expression to match the target information in the EHR. If the IE system is designed for a specific cancer domain, it will be a big disaster when we are required to rewrite all regular expressions defined previously in the IE system so that can be fit on other cancer domain. Besides, it’s an annoying task to modify the codes that define the regular expression in the IE system when there is a requirement to modify the current version of regular expression. In our previous work, we addressed two approaches to cope with this issue. First, we introduced an ontology model as external regular expression to our IE system. The ontology model defines the regular expressions that have to be imported into the IE system, which separates the domain knowledge from the IE system so as to make our application more flexible. The use of ontology provides convenience to our IE system when we want to modify the current regular expressions or create a new ontology that helps the IE system to fit other specific cancer domain [7].

Secondly, the quality of extracted results of medical information is crucial to the IE system. Although the IE system extracted a large number of medical information, the wrong information cannot be used. Narrative medical records store patient related information and extracting this information correctly is an important mission for further practical application [7]. Data quality issues that related to the secondary use of EHR are discussed in [10]. The authors address three categories of data quality issues:

Incompleteness (missing information), Inconsistency (information mismatches the data source), Inaccuracy (non-specific, non-standards-based, inexact, incorrect, or imprecise information). Mostly, the traditional ways of examining the quality of extracted information is reviewing manually. However, the reviewing process is time consuming and prone to error. Previously, we proposed a validation system to predict the correctness of extracted results after the IE process. A confidence score label is given by our validation system. Confidence score label stands for the correctness of the extracted result. The reviewer can concentrate on the information with low confidence score [7].

Thirdly, as we mentioned before, the IE system extracts out medical information from the report by using regular expression. Only the information completely matched with regular expression defined in the IE system can be extracted out. However, these narrative reports are made by human manually. Therefore, improving the correctness of original medical reports is very important for further research. There are two main factors which will affect the extract result:

1. Misspelling words in the reports
2. Different representations to the same meaning.

In this study, we developed a Misspellings Handling System (MHS) for correcting the misspellings in the Electronic Medical Record (EMR), in order to increase the information extracted out by the IE system. We maintain our own corpus which contains all the medical terms we used in the IE system. The MHS has several algorithms for calculating the word distance of words in corpus and target word (word in report). Word distance represents similarity between two words. According to the word distance, the MHS can replace the misspelled word in the report with a corrected word in corpus. So that the IE system can extract those misspelled information which cannot be extracted out previously.

2.4 Liver Cancer

The liver is made up of different cell types, such as bile ducts, blood vessels, fat-storing cells and hepatocytes which make up 70-80% of the liver tissue [11]. The type of liver cancer arising from the liver is known as primary liver cancer. The proportion of primary liver cancers arising from hepatocytes is about 90% to 95% and this kind of primary liver cancer is also called hepatocellular carcinoma [11].

The other type of liver cancer is called secondary liver cancer or metastatic liver cancer [11]. Because blood from all parts of the body must pass through the liver for filtration, cancer cells from other organs and tissues easily reach the liver. The liver cancer may originate from other organs such as the colon, stomach, rectum, esophagus, pancreas, breast, lung or skin. The term liver cancer actually can refer to either primary liver cancer or secondary liver cancer [12].

2.5 Methods of Diagnosis for Liver Cancer

2.5.1 Blood Test

The most widely used biochemical blood test is alpha-fetoprotein (AFP), which is a protein normally made by the immature liver cells in the fetus [12]. At birth, infants have relatively high blood serum levels of AFP, which fall to normal adult levels by the first year of life. But a high level of AFP cannot be used to confirm a diagnosis of liver cancer, because cirrhosis or chronic hepatitis can also produce high AFP levels. In the AFP tests, about 50-75% of primary liver cancer patients have abnormally high levels of AFP [12]. Therefore, a normal AFP does not exclude liver cancer. Also, an abnormal AFP does not imply that a patient has liver cancer. It is important to note that patients with cirrhosis and an abnormal AFP, despite having no documentable liver cancer, still are at very high risk of developing liver cancer or actually already have an undiscovered liver cancer.

2.5.2 Imaging Studies

Imaging studies is a very important part in the diagnosis of liver cancer. Even though imaging studies cannot tell the difference between a hepatoma and other abnormal masses, lesion or nodules in the liver, a good imaging study can provide information as to the size of the tumor, the number of tumors, and whether the tumor has involved major blood vessels locally or spread outside of the liver [11]. There are several types of studies, each having its advantages and disadvantages. Practically, several studies combined often complement each other.

Angiography: The angiography is helpful in deciding whether the tumor can be removed by surgery.

Ultrasound (Echo): The ultrasound is used to show any abnormal growths in your liver. Ultrasound is the most sensitive imaging study for diagnosing and characterizing liver cancer, according to studies from Japan and Taiwan reports [11]. Ultrasound examination is usually the first study ordered if liver cancer is suspected in a patient. Morever, the price of an ultrasound is quite low as compared to the other types of scans [11].

Computed tomography scan (CT): CT scans is a very common imaging study used in the Unite State for workup of tumors in the liver. The CT scans show how large the tumor is and whether it has spread to other organs or to the lymph nodes. And CT is much less operator-dependent than is ultrasound. However the CT is more expensive, and requires the use of contrast material, which has the potential risks of an allergic reaction and other effects on kidney function [11].

Magnetic resonance imaging (MRI): The MRI providing very clear images of the body which can help clinicians to determine whether the liver cancer is primary type or secondary type. Its advantage over CT is that MRI can provide partial views of the body in different planes [11].

2.5.3 Liver Biopsy

Liver biopsy is considered to provide the definite diagnosis of liver cancer [13]. The procedure of liver biopsy involves using an ultrasound or CT scan. This allows the doctor to guide the needle into the right place of abnormal tissue, then cutting out a small piece of liver tissue and sending it to a laboratory. A pathologist in the laboratory will look at the tissue under a microscope to see whether it has cancerous cells [13].

2.6 Treatment of Liver Cancer

2.6.1 Radiofrequency Ablation

Radiofrequency ablation (RFA) is a medical procedure to ablate tumor or other abnormal tissue. The treatment procedure is done without opening the abdomen by just using ultrasound or CT scan for visual guidance. It is a good option for patients with small liver tumors who cannot have their tumors removed by surgery. The ideal size of a liver cancer tumor for RFA is less than 5 cm.

2.6.2 Liver Resection

The goal of liver resection is to completely remove the tumor without leaving any tumor behind [14]. This option is limited to patients with one or two small (3 cm or less) tumors and no major blood vessels involved, ideally without cirrhosis. There are two biggest concerns after liver resection [14]. First, the patient can develop liver failure due to the remaining portion of the liver is unable to provide the necessary support for life. Second, the cancer could come back in the future after a liver resection. No test can guarantee that cancer cells had not spread before it was removed. That's why this type of treatment is only used for small liver tumors where there is less chance of spread.

2.7 Datasets

The data for training and testing our misspellings handling system came from the National Taiwan University Hospital liver cancer patients' medical records. There are six types of medical records in the data set, admission report, discharge summary, echo report, operation report, radiology report, pathology report. Different medical reports contain different information about the patients.

The training set contains 74 liver cancer patients who took RFA to treat their cancer. These 74 liver cancer patients have totally 3195 narrative medical records.

The testing set contains 78 liver cancer patients who took RFA to treat their cancer. These 78 liver cancer patients have totally 2539 narrative medical records.

After the misspellings handling process, the corrected reports in the training set and testing set will transfer into the IE system for information extraction process. Then, the medical results will be extracted out by the IE system.

2.7.1 Target Information

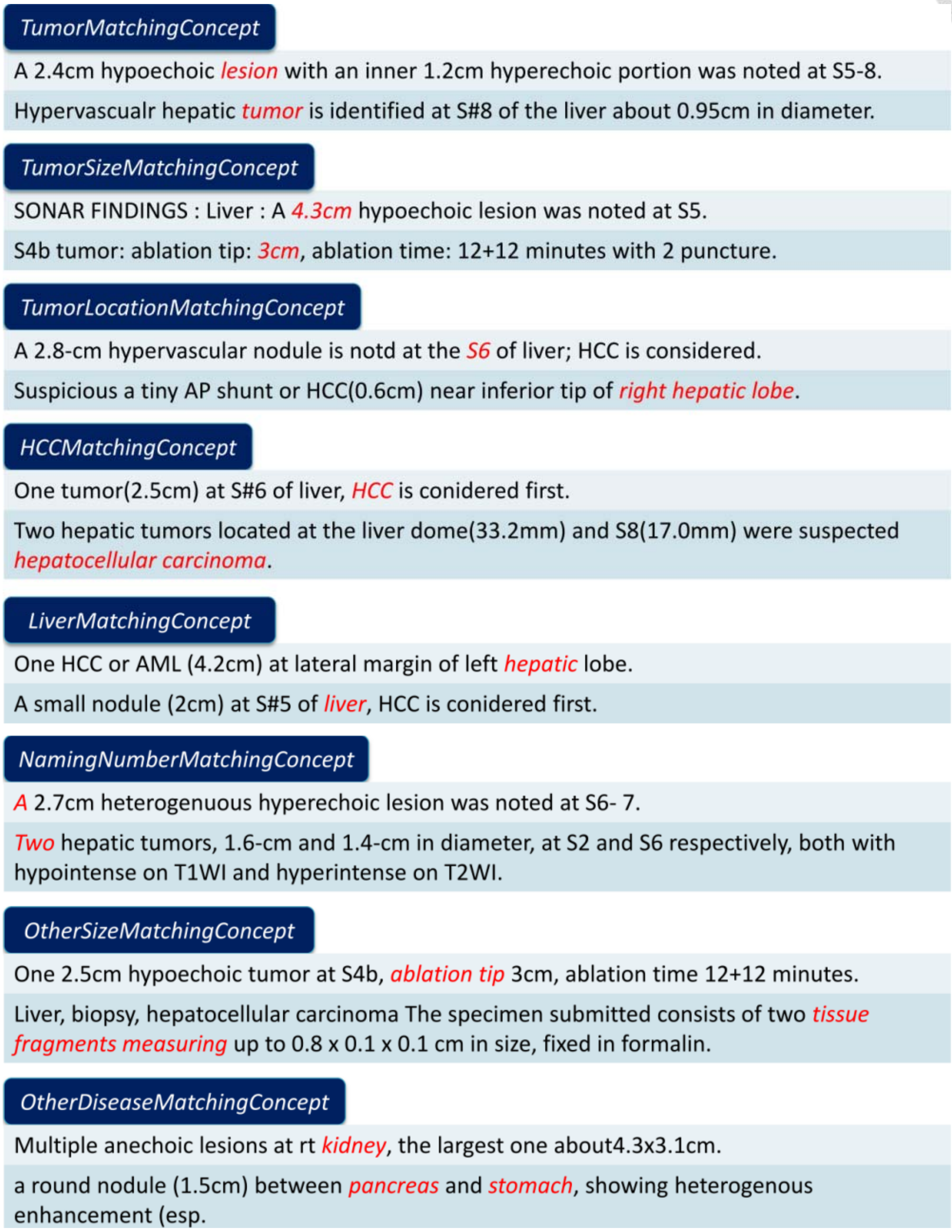
Textual medical records are used for describing the medical related information, including diagnosis, tumor information, cancer staging, treatment and recurrent status. In this study, we focus on the tumor information in the liver cancer patients' textual medical reports. We implemented an information extraction system to extract and gather desired information in structured format. The tumor information about tumor, tumor size, tumor location was extracted and grouped. The other tumor related information was also collected in the IE process. In the following, we list 9 concepts that should be extracted from the narrative medical reports in the IE process.

Table 1 9 concepts defined in our Information Extraction System

Name of Concept	***Description***
TumorMatchingConcept	This concept indicates the finding of abnormal tissues described in the textual medical records.
TumorSizeMatchingConcept	This concept indicates the figure information about size of any object in the textual medical records.
TumorLocationMatchingConcept	This concept indicates liver anatomy information that used to describe the location of abnormal tissues in the textual medical records.
HCCMatchingConcept	This concept indicates the terms hepatocellular carcinoma or HCC in the textual medical records.
LiverMatchingConcept	This concept indicates the terms liver and hepatic in the textual medical records.
NamingNumberMatchingConcept	This concept indicates the quantity information of abnormal tissues in the textual medical records.
OtherSizeMatchingConcept	This concept indicates the other object (exclude abnormal tissues) which may also have figure

	information in the textual medical records.
OtherDiseaseMatchingConcept	This concept indicates the other organ (exclude liver) or other disease information in the textual medical records
SpecialMatchingConcept	This concept was added in the human-labeling process which usually appears in the original sentence of uncertain-extracted information or wrong-extracted information.

Figure 2 shows the example of the 9 concepts appearing in the textual medical records that should be extracted by the IE system. Table 1 shows the detailed information about the occurrence of each concept in the training set and testing set, the amount of concept per patient and per report.



TumorMatchingConcept

A 2.4cm hypoechoic *lesion* with an inner 1.2cm hyperechoic portion was noted at S5-8.

Hypervascualr hepatic *tumor* is identified at S#8 of the liver about 0.95cm in diameter.

TumorSizeMatchingConcept

SONAR FINDINGS : Liver : A *4.3cm* hypoechoic lesion was noted at S5.

S4b tumor: ablation tip: *3cm*, ablation time: 12+12 minutes with 2 puncture.

TumorLocationMatchingConcept

A 2.8-cm hypervascular nodule is notd at the *S6* of liver; HCC is considered.

Suspicious a tiny AP shunt or HCC(0.6cm) near inferior tip of *right hepatic lobe*.

HCCMatchingConcept

One tumor(2.5cm) at S#6 of liver, *HCC* is conidered first.

Two hepatic tumors located at the liver dome(33.2mm) and S8(17.0mm) were suspected *hepatocellular carcinoma*.

LiverMatchingConcept

One HCC or AML (4.2cm) at lateral margin of left *hepatic* lobe.

A small nodule (2cm) at S#5 of *liver*, HCC is conidered first.

NamingNumberMatchingConcept

A 2.7cm heterogenuous hyperechoic lesion was noted at S6- 7.

Two hepatic tumors, 1.6-cm and 1.4-cm in diameter, at S2 and S6 respectively, both with hypointense on T1WI and hyperintense on T2WI.

OtherSizeMatchingConcept

One 2.5cm hypoechoic tumor at S4b, *ablation tip* 3cm, ablation time 12+12 minutes.

Liver, biopsy, hepatocellular carcinoma The specimen submitted consists of two *tissue fragments measuring* up to 0.8 x 0.1 x 0.1 cm in size, fixed in formalin.

OtherDiseaseMatchingConcept

Multiple anechoic lesions at rt *kidney*, the largest one about4.3x3.1cm.

a round nodule (1.5cm) between *pancreas* and *stomach*, showing heterogenous enhancement (esp.

Figure 5 Example of concept occurs in the sentence of textual medical records.

Chapter 3 Method

3.1 Information Extraction System

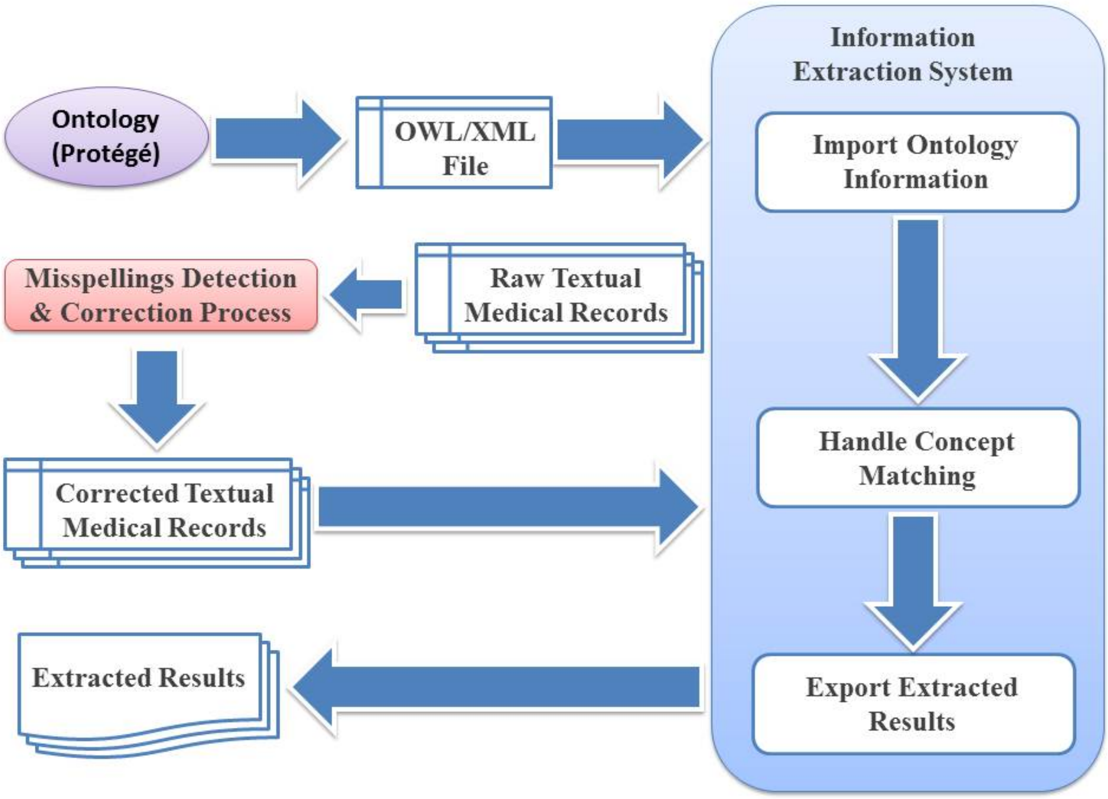

Ontology
(Protégé)
OWL/XML File
Information Extraction System
Import Ontology Information
Misspellings Detection & Correction Process
Raw Textual Medical Records
Handle Concept Matching
Corrected Textual Medical Records
Extracted Results
Export Extracted Results

Figure 6 An overview of general procedure of the Information Extraction System.

The Information Extraction System was helping us to extract the target information in the free text Medical records and export the extracted results in a structured form for medical experts do the further analysis. Figure 3 shows the general steps of the Information Extraction System to process the misspellings correction and extraction of the textual medical records and obtain external knowledge source from the ontology.

First, The Information Extraction System will import external knowledge resource from the ontology file OWL/XML. The ontology was edited by Protégé software which is an open source ontology editor and knowledge-based framework developed by Stanford University. The additional regular expression information defined in the ontology is imported into the IE system.

Second, the raw textual medical reports from National Taiwan University Hospital will be transferred into The Misspellings Handling System (MHS) for misspellings detection and correction process. The system has several corpus for different concepts for generating suggestion list for misspellings by the phonetic algorithms (such as Metaphone algorithm, Double Metaphone algorithm and Soundex algorithm). After the suggestion list for the misspellings was created, the MHS will select a better selection through ranking algorithm, than replace the misspellings with the correction word.

Third, after processing the raw textual medical records, the IE system starts to handle concepts matching and extract information from corrected textual medical records. Finally, the IE system outputs an Excel file which contains information about all extracted results, source sentences and features that corresponding to each extracted result.

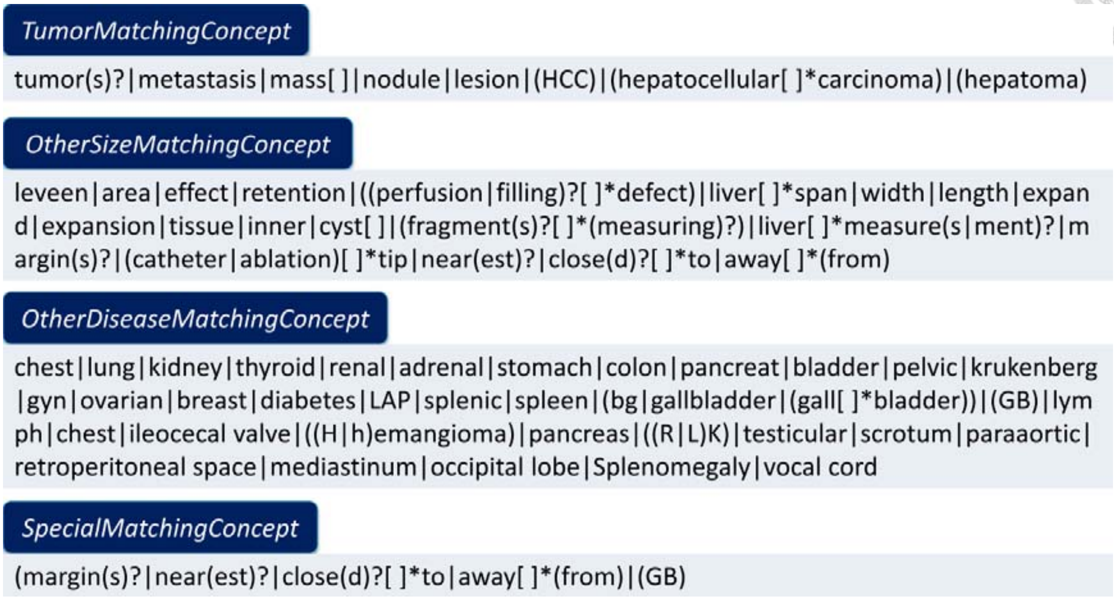
3.1.1 Regular Expression

A regular expression is a set of pattern matching rules encoded in a string according to certain syntax rules. Regular Expressions are used when users want to search for specific lines of text which containing a particular pattern. In the IE application, regular expression provides a concise and flexible means for matching, specifying and recognizing strings of text, such as particular characters, words, or patterns of characters [15]. With the help of regular expression, it is simple to search for a specific word or string of characters.

In our research, the regular expression was defined in two places, the original version of regular expression in the IE system and the additional regular expressions in the ontology model that added by reviewers in the human-labeling process. The regular expression defined in the ontology model was used as external knowledge source to expand the regular expression in the IE system.

In the IE system, regular expression was used to match the target information in the liver cancer patients' textual reports. We define the regular expression according to the desired information that should be extracted from medical records. Figure 4 lists the example of 4 concepts with their regular expression. Here is their brief introduction:

- TumorMatchingConcept: This concept was to filter the sentences that contain the information about tumor.
- OtherSizeMatchingConcept: This concept was used to estimate the figure information about size belonging to the tumor or other objects.
- OtherDiseaseMatchingConcept: This concept was used to estimate the tumor information belonging to the liver, other organs or other diseases.
- SpecialMatchingConcept: The string or pattern defined in this concept was due to their frequent occurrence in the wrongly extracted results. We aim to use this concept to help the validation system to distinguish these wrongly extracted results.



TumorMatchingConcept

tumor(s)?|metastasis|mass[ ]|nodule|lesion|(HCC)|(hepatocellular[ ]*carcinoma)|(hepatoma)

OtherSizeMatchingConcept

leveen|area|effect|retention|((perfusion|filling)?[ ]*defect)|liver[ ]*span|width|length|expand|expansion|tissue|inner|cyst[ ]|(fragment(s)?[ ]*(measuring)?)|liver[ ]*measure(s|ment)?|margin(s)?|(catheter|ablation)[ ]*tip|near(est)?|close(d)?[ ]*to|away[ ]*(from)

OtherDiseaseMatchingConcept

chest|lung|kidney|thyroid|renal|adrenal|stomach|colon|pancreat|bladder|pelvic|krukenberg|gyn|ovarian|breast|diabetes|LAP|splenic|spleen|(bg|gallbladder|(gall[ ]*bladder))|(GB)|lymph|chest|ileocecal valve|((H|h)emangioma)|pancreas|((R|L)K)|testicular|scrotum|paraaortic|retroperitoneal space|mediastinum|occipital lobe|Splenomegaly|vocal cord

SpecialMatchingConcept

(margin(s)?|near(est)?|close(d)?[ ]*to|away[ ]*(from)|(GB)

Figure 7 Example of regular expression defined in the IE system.

3.1.2 Ontology

Ontologies are used to describe knowledge about the interesting domain [4]. The ontology defines the concepts in the domain and also the relationships that hold between those concepts. The most recent development in standard ontology languages is ontology web language (OWL) from the World Wide Web Consortium (W3C) [5].

In the IE system, we use regular expressions to recognize the information in the textual medical records. It is inconvenient to update the current version of regular expression in the IE system. Therefore, we define the additional regular expression in the ontology based on OWL. In the ontology model, we had created 9 classes that corresponding to the target information and 4 classes that are subclass of the TumorLocationMatchingConcept. Each class has a property called "hasRegex" which represents the attribute of regular expression. Figure 5 shows the hierarchy structure of the ontology model. The open source Protégé ontology editor was adopted to construct the ontology model which provides a friendly user interface [6]. Figure 6 shows the user interface of the protégé editor and the property "hasRegex" with the regular expression attribute of the class "OtherSizeMatchingConcept". After reviewers had added the additional regular expressions to the corresponding ontology classes in the human-labeling process, the protégé exported the ontology to an OWL/XML file. The IE system would read this OWL/XML file and add the regular expression to the corresponding concept. Figure 7 shows the example of OWL/XML file exported by the protégé editor.

The advantage of using the ontology is that it provides a lot of elasticity to our IE system, if we want to add additional regular expressions in the extraction process, we do not have to modify the regular expression in our IE system but to add additional regular expressions in the ontology model.

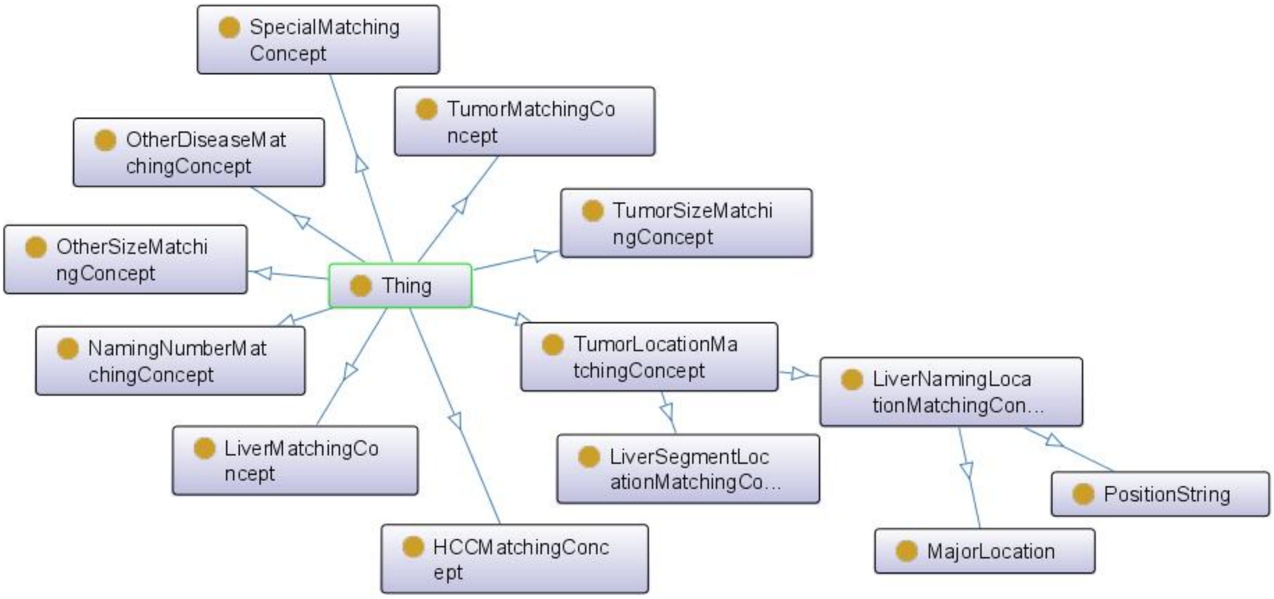
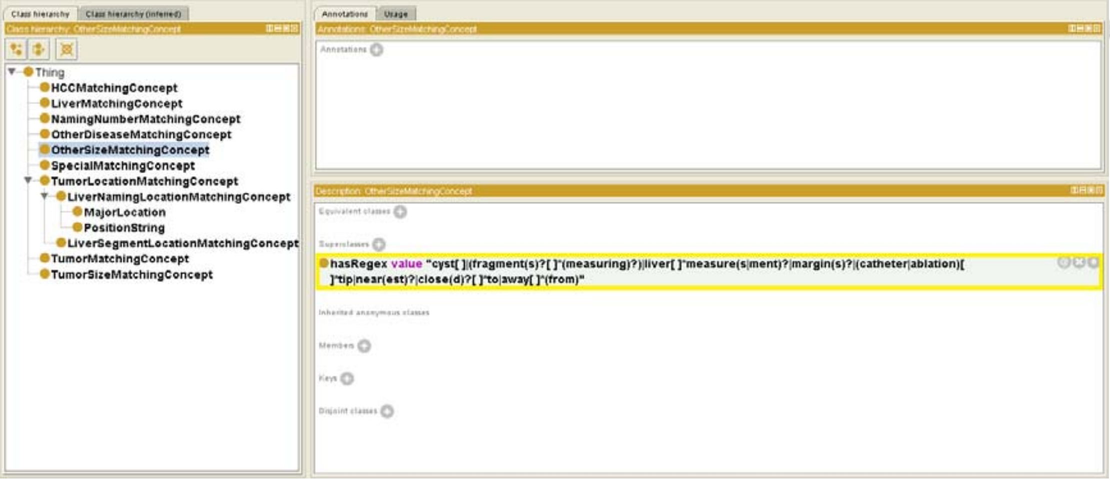


Figure 8 Hierarchy structure of the ontology model.

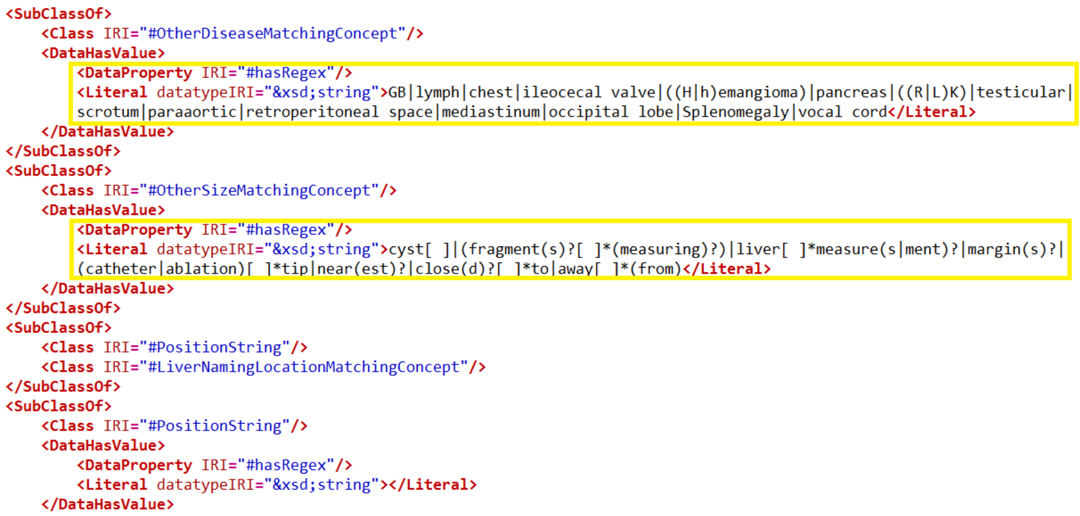


Class hierarchy
Class hierarchy (inferred)
Class hierarchy: OtherSizeMatchingConcept
Annotations
Usage
Annotations: OtherSizeMatchingConcept
Annotations
Description: OtherSizeMatchingConcept
Equivalent classes
Superclasses
Inherited anonymous classes
Members
Keys
Disjoint classes

```
Thing
    HCCMatchingConcept
    LiverMatchingConcept
    NamingNumberMatchingConcept
    OtherDiseaseMatchingConcept
    OtherSizeMatchingConcept
    SpecialMatchingConcept
    TumorLocationMatchingConcept
        LiverNamingLocationMatchingConcept
            MajorLocation
            PositionString
        LiverSegmentLocationMatchingConcept
    TumorMatchingConcept
    TumorSizeMatchingConcept
```

```
hasRegex value "cyst[ ]|(fragment(s)?[ ]*(measuring)?)|liver[ ]*measure(s|ment)?|margin(s)?|(catheter|ablation)[
 ]*tip|near(est)?|close(d)?[ ]*to|away[ ]*(from)"
```

Figure 9 User interface of protégé editor.



```
<SubClassOf>
    <Class IRI="#OtherDiseaseMatchingConcept"/>
    <DataHasValue>
        <DataProperty IRI="#hasRegex"/>
        <Literal datatypeIRI="&xsd;string">GB|lymph|chest|ileocecal valve|((H|h)emangioma)|pancreas|((R|L)K)|testicular|
        scrotum|paraaortic|retroperitoneal space|mediastinum|occipital lobe|Splenomegaly|vocal cord</Literal>
    </DataHasValue>
</SubClassOf>
<SubClassOf>
    <Class IRI="#OtherSizeMatchingConcept"/>
    <DataHasValue>
        <DataProperty IRI="#hasRegex"/>
        <Literal datatypeIRI="&xsd;string">cyst[ ]|(fragment(s)?[ ]*(measuring)?)|liver[ ]*measure(s|ment)?|margin(s)?|
        (catheter|ablation)[ ]*tip|near(est)?|close(d)?[ ]*to|away[ ]*(from)</Literal>
    </DataHasValue>
</SubClassOf>
<SubClassOf>
    <Class IRI="#PositionString"/>
    <Class IRI="#LiverNamingLocationMatchingConcept"/>
</SubClassOf>
<SubClassOf>
    <Class IRI="#PositionString"/>
    <DataHasValue>
        <DataProperty IRI="#hasRegex"/>
        <Literal datatypeIRI="&xsd;string"></Literal>
    </DataHasValue>
```

Figure 10   The OWL/XML file produced by protégé.

3.1.3 Concept Matching

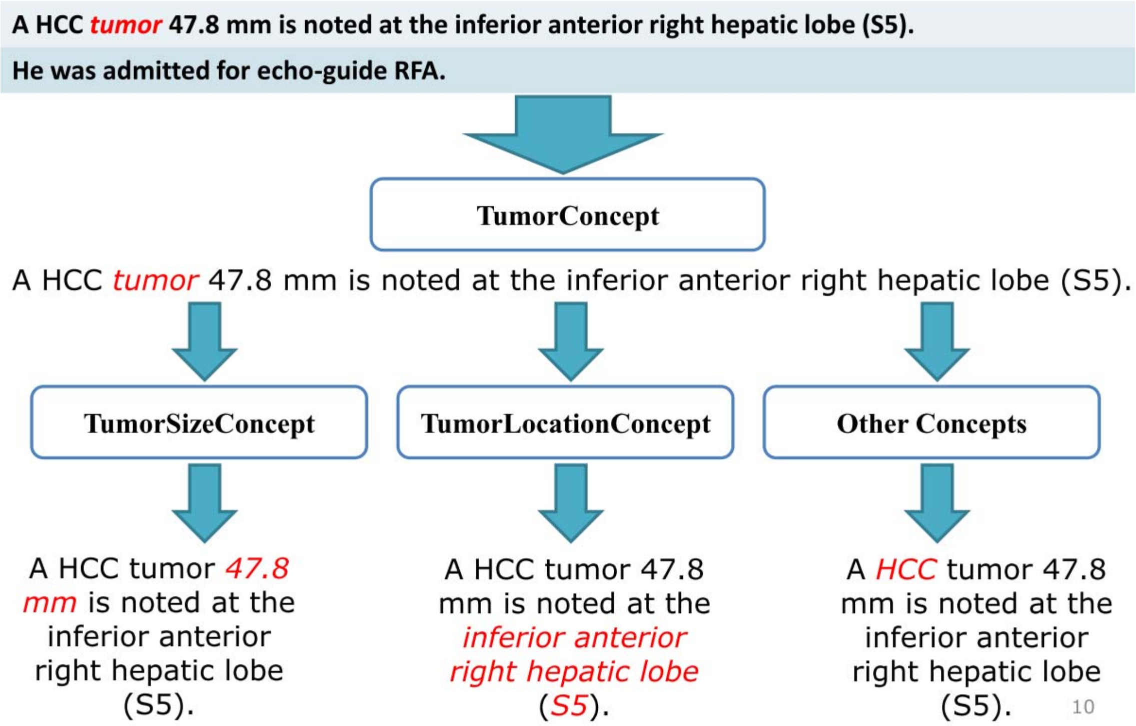

A HCC tumor 47.8 mm is noted at the inferior anterior right hepatic lobe (S5).
He was admitted for echo-guide RFA.
TumorConcept
A HCC tumor 47.8 mm is noted at the inferior anterior right hepatic lobe (S5).
TumorSizeConcept
TumorLocationConcept
Other Concepts
A HCC tumor 47.8 mm is noted at the inferior anterior right hepatic lobe (S5).
A HCC tumor 47.8 mm is noted at the inferior anterior right hepatic lobe (S5).
A HCC tumor 47.8 mm is noted at the inferior anterior right hepatic lobe (S5).
10

Figure 11 Example of two layer filters to extract the target information.

When textual medical reports enter the IE system, they are split into small piece of sentences and the IE system checks all sentences to find out whether the sentences contain target information.

The TumorMatchingConcept and other concepts cooperate like a two-layer filter to extract the sentences that contain the desired information. In the first layer, the sentence contains the information about the TumorMatchingConcept would pass the first layer. In the second layer, the sentence would be checked for searching other concepts. Figure 8 shows the procedure of two sentences to pass the two-layer filter and get the target information. The first sentence contains the TumorMatchingConcept, the second sentence does not contain this concept. Therefore, only the first sentence contains the TumorMatchingConcept can pass the first layer. In the second layer, the sentence would be checked several times to find other concepts. All extracted information would be collected and stored. Here, the first sentence contains the information about:

- TumorSizeMatchingConcept: 47.8 mm
- TumorLocationMatchingConcept: inferior anterior right hepatic and S5
- HCCMatchingConcept: HCC

3.2 Corpus

A corpus is a database contains a large number of texts. It is wildly used in Natural Language Processing. In this study, the corpus is a resource for misspellings detection and correction. We have to calculate similarity between the word in the report and the word in corpus to determine whether the word in the report is correct or misspelled, and select the correct word in the corpus to replace the misspelled word in the report. We had built our own corpus for our Misspellings Handling System. In addition to make corpus useful in our Misspellings Handling System, we have done extra part-of-speech tagging to every text in the corpus.

3.3 Approximate string matching algorithm

The regular expression only can find those information containing specific string which matched regular expression completely. However, misspellings result in a large majority of information not been extracted out by the regular expression defined in the Information Extraction System. Approximate string matching algorithm provides a solution to this issue by offering a fuzzy way for string matching which allows error [16]. The goal of this fuzzy algorithm is to perform a string matching of a pattern in a text which has suffered some kind of misspellings. There are several types of approximate string matching algorithm. We listed 2 types of algorithms to be used in this study: 1. Phonetic matching algorithms 2. Edit distance matching algorithms. The Phonetic Matching Algorithm is an algorithm which matches two different words with similar pronunciation to the same code, which allows phonetic similarity based word set comparison and indexing. We will introduce several approximate strings matching algorithms we tried to use in Misspellings Handling System as below:

3.3.1 Soundex algorithm

Soundex algorithm is one of the phonetic matching techniques which transform the given vocabulary into the encoding key according to the pronunciation of the word[17]. Soundex was announced in 1981 by Odell and Russel which is oldest one among phonetic matching algorithms. The principle of Soundex algorithm is to divide a set of letters into 7 disjoint sets, the letters in the same set which has similar pronunciation and each set has its own unique key. The Soundex algorithm will ignore the vowels and the letter considered to be silence (such w, h and y) during encoding[17]. The algorithm transforms all letters into the four letter code (except first letter of vocabulary), for instance the code of “international” is I-536 (I for first letter I, number 5 for letter N, number 3 for letter T, number 6 for letter R), and the code of “hypercellular” is H-162 (H for first letter H, number 1 for letter P, number 6 for letter R, number 2 for letter C). The advantages of Soundex algorithms are simple and small rule table size, which can perform extremely fast among other phonetic algorithms. However, the Limitation of Soundex algorithm is that the algorithm uses the rule-based way to encode the pronunciation code, so that Soundex algorithm does not allow the exception among the vocabulary during encoding process and it only can be used in English. Another limitation is that Soundex algorithm encoding vocabulary into four characters code which is difficult to tell the differences between long strings. We listed general steps of Soundex algorithm as below [17] :

1. Replace all letter (except first character) of string with its phonetic code.
2. Eliminate any consecutive repetition of codes.
3. Eliminate all occurrences of code 0 (eliminate vowels, and the letters H, W and Y).
4. Return the first four characters of the resulting string.

Table 2 Soundex coding rules

CODE	LETTERS
0	A, E, I, O, U, H, W, Y
1	B, F, P, V
2	C, G, J, K, Q, S, X, Z
3	D, T
4	L
5	M,N
6	R

3.3.2 PHONIX algorithm

PHONIX which is a phonetic retrieval technique developed for use with the URICA library system is similar to Soundex algorithm by mapping letters into set of code. [13] listed the general steps of Soundex algorithm as below:

1. Perform phonetic substitutions
2. Retain the first character for the retrieval code
3. Replace by ‘V’ if A, E, I, O, U or Y.
4. Where names end in ES, delete the E.

5. Append an E where names end in A, I, O, U and Y.
6. Drop the last character regardless.
7. Drop the new last character if not A, E, I, O, U or Y.
8. Repeat step 7 until a vowel (including Y) is found. This results in a word or name without its ending-sound.
9. Strip all occurrences of A, E, I, O, U, Y, H and W.
10. Remove one of all duplicate successive consonants.
11. Replace ALL consonants by their numeric values.
12. Prefix the retrieval code with the retained first character (may be a ‘V’)
13. Repeat steps 9, 10, 11 on the characters removed as stripped ending-sounds.

Table 3 PHONIX coding rules

CODE	LETTERS
0	A, E, H, I, O, U, W, Y
1	B, P
2	C, G, J, K, Q
3	D, T
4	L
5	M, N
6	R
7	F, V
8	S, X

3.3.3 Metaphone algorithm

As a result of Soundex algorithm encoding vocabulary into four characters code which is difficult to tell the differences between long strings, Metaphone algorithm is a variant of Soundex algorithm which improves on the Soundex algorithm by using information about variations and inconsistencies in English spelling and pronunciation to produce a more accurate encoding. Both of Metaphone and Soundex algorithms are work by encoding an input string to a Metaphone code which represents English pronunciation. Due to similar strings should share the same Metaphone code, different strings mapping into same Metaphone code are considered to have same pronunciation. The Metaphone algorithm processes string according to following principles:

1. Delete duplicate adjacent letters, except letter C.
2. Drop the first letter, if the string begins with ‘KN’, ‘GN’, ‘PN’, ‘AE’, ‘WR’.
3. Ignore ‘B’ if after ‘M’ at the end of the word.
4. ‘C’ encodes as ‘X’ if followed by ‘IA’ or ‘H’. ‘C’ encoded as ‘S’ if followed by ‘I’, ‘E’, or ‘Y’. Otherwise, ‘D’ encodes as ‘T’.
5. ‘D’ encodes as ‘J’ if followed by ‘GE’, ‘GY’, or ‘GI’. Otherwise, ‘D’ encodes as ‘T’.
6. Delete ‘G’ if followed by ‘H’ and ‘H’ is not at the end or before a vowel. Delete ‘G’ if followed by ‘N’ or ‘NED’ and is at the end.
7. ‘G’ encodes as ‘J’ if before ‘I’, ‘E’ or ‘Y’, and it is not in ‘GG’. Otherwise, ‘G’ encodes to ‘K’.
8. Delete ‘H’ if after vowel and not before a vowel.
9. ‘CK’ encodes as ‘K’.
10. ‘PH’ encodes as ‘F’.

11. ‘Q’ encodes as ‘K’.
12. ‘S’ encodes as ‘X’ if followed by ‘H’, ‘IO’ or ‘IA’.
13. ‘T’encodes as ‘X’ if followed by ‘IA’ or ‘IO’. ‘TH’ encodes to ‘0’. Delete ‘T’ if followed by ‘CH’.
14. ‘V’ encodes as ‘F’.
15. ‘WH’ encodes as ‘W’ if at beginning of the string. Delete ‘W’ if not followed by a vowel.
16. ‘X’ encodes as ‘S’ if at beginning of the string. Otherwise, ‘X’ encodes as ‘KS’.
17. Deleted ‘Y’ if not followed by a vowel.
18. ‘Z’ encodes as ‘S’.
19. Delete all vowels unless it is the beginning.

According to the principles listed above, Metaphone algorithm can encode string into Metaphone code. For instance, string ‘International’ encodes as ‘INTRNXNL’ (I for letter ‘I’, N for letter ‘N’, T for letter T, letter E will be deleted because E is a vowel but not in the beginning of the string, N for letter N, substring “TIO” transform to X because letter ‘IO’ followed by ‘T’, N for letter ‘N’, A will be deleted because E is a vowel, L for Letter ‘L’). Even though Metaphone provides higher accuracy by eliminating the limitation of the fixed 4 character code of Soundex, it still has some limitations.

3.3.4 Levenshtein Distance algorithm

Different from Soundex and Metaphone algorithms transforming the given vocabulary into the encoding key according to the pronunciation of the word, Levenshtein Distance proposed another way to identify the similarity of two words.

Levenshtein Distance is one of the dynamic programming algorithms which adapt the divide and conquer strategy to solve the complex problem by breaking them into smaller sub-problems. The principle of dynamic programming is that a given problem can be divided into different sub-problems, those sub-problems can be solved individually. Most of sub-problems are almost the same, after solving those sub-problems we can combine each solution to every sub-problems to get the final answer to the original problem. Two main factors of solving problem by dynamic programming is that the original problem has overlapping sub-problems and optimal substructure. A problem is said to have overlapping sub-problems if it can be broken down into sub-problems recursively which are reused multiple times or a recursive algorithm solves the same sub-problems rather than generating new sub-problem. For instance, Fibonacci sequence is one of the overlapping sub-problems. The *n*th Fibonacci number *F(n)* can be solved by calculating *F(n-1)* and *F(n-2)*, then sum up those two Fibonacci numbers. And the result of *F(n-1)* is from *F(n-2)* and *F(n-3)*. Another factor is that the problem needs to have optimal substructure. According to the “Principle of Optimality” purposed by Richard Bellman[18], a problem has optimal structure means if an optimal solution to a problem can be constructed efficiently from optimal solutions to its sub-problems. In order to reduce the number of computations, the dynamic programming algorithms store each result of every sub-problems in a table. If the same solution to the problem is needed in the future, it just has to look up the table instead of re-compute the solution again.

The Levenshtein distance algorithm uses a string metric to measure the similarity or dissimilarity of two strings. The Levenshtein uses numbers of edit for describe the similarity of two strings, there are 3 edit operations in Levenshtein distance algorithm.

1. Insertion: add one character in the string X in order to match the string Y.

   Ex. String X is “hypercllular” and string Y is “hypercellular”, string X needs to add a character “e” between “c” and “l” in order to match String Y.

2. Deletion: drop a character from the string X in order to match the string Y.

   Ex. String X is “hyperceallular” and string Y is “hypercellular”, string X needs to delete a character “a” between “e” and “l” in order to match String Y.

3. Substitution: exchange a character of the string X with a specific character in order to match the string Y

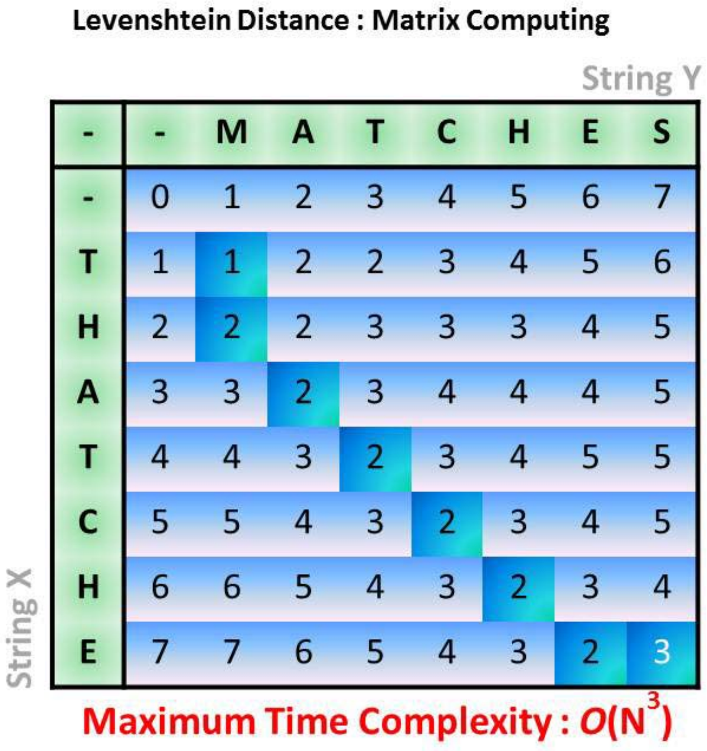
   Ex. String X is “hypercallular” and string Y is “hypercellular”, string X needs to exchange a character “a” between “c” and “l” with a character “e” in order to match String Y.

Assume given 2 strings X and Y, the Levenshtein distance *c[i, j]* defined as the similarity between the first *i* characters of string X and first *j* characters of string Y.

Equation 1 Recursive formula of Levenshtein distance algorithm

$$c[i,j] = \left\{ \begin{array}{ll} \quad 0 & if\ i = 0\ and\ j = 0 \\ c[i-1, j-1] & if\ i,j > 0\ and\ x_i\ =\ y_j \\ \max(c[i, j-1]+1, c[i-1, j]+1) & if\ i,j > 0\ and\ x_i\ \neq\ y_j \end{array} \right\}$$

According to the equation we list above, we can derive the following metric. The Levenshtein edit distance which is *c[m-1, n-1]* (where *m* is length of string X and *n* is length of string Y) represents the result of similarity between string X and string Y.



Levenshtein Distance : Matrix Computing

String Y

-	-	M	A	T	C	H	E	S
-	0	1	2	3	4	5	6	7
T	1	1	2	2	3	4	5	6
H	2	2	2	3	3	3	4	5
A	3	3	2	3	4	4	4	5
T	4	4	3	2	3	4	5	5
C	5	5	4	3	2	3	4	5
H	6	6	5	4	3	2	3	4
E	7	7	6	5	4	3	2	3

String X

Maximum Time Complexity : $O(N^3)$

Figure 12 String matrix of Levenshtein Distance Algorithm

In this example we calculate the Levenshtein distance between two strings “MATCHES” and “THATCHE”. As we can see the similarity between two strings is 3. By tracing from the most button right cell to the original we can obtain the minimum action list for modifying the string X to string Y. According to the tracing route, String X “THATCHE” wants to transfer to String Y needs to do following 3 edit actions. 1. Delete T from String X “THATCHE” will get “HATCHE” 2. Exchange first character “H” with character “M” will get “MATCHE”. 3. Add S in the end of the string will get “MATCHES”.

3.4 Misspellings Handling System

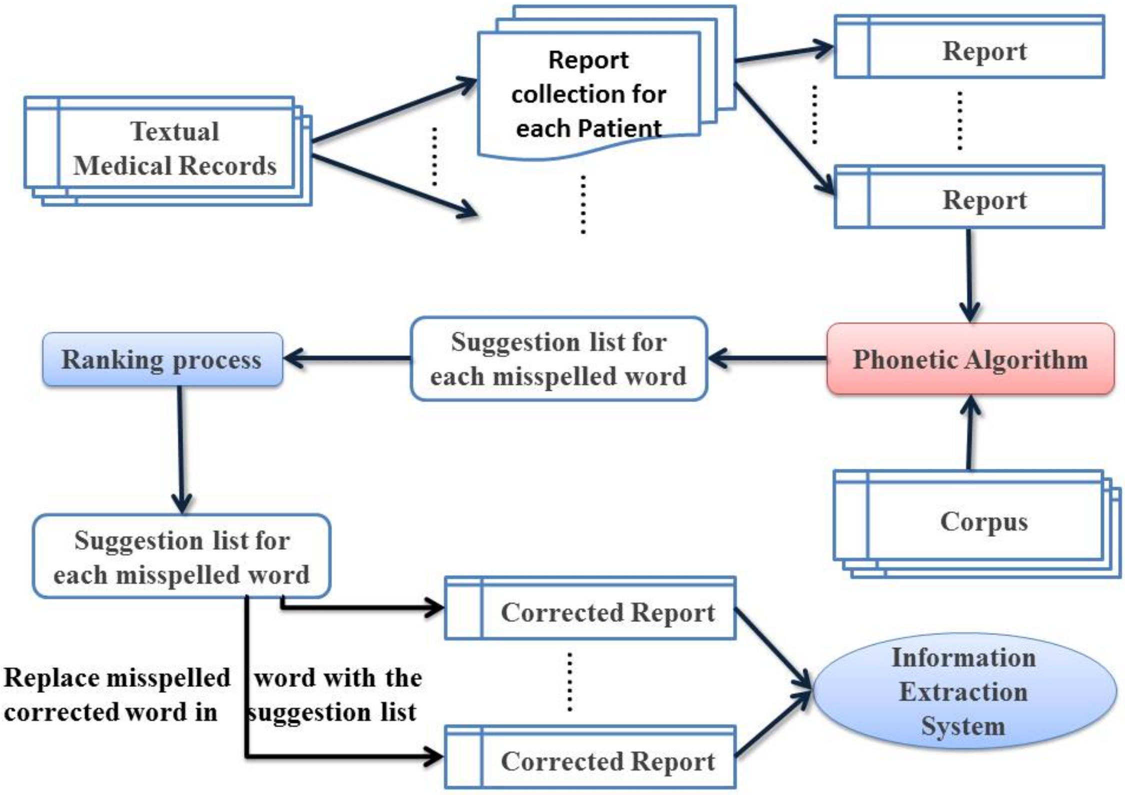

Textual Medical Records
Report collection for each Patient
Report
Report
Phonetic Algorithm
Corpus
Suggestion list for each misspelled word
Ranking process
Suggestion list for each misspelled word
Replace misspelled word with the corrected word in suggestion list
Corrected Report
Corrected Report
Information Extraction System

Figure 13 Overview of Misspellings Handling System

In order to correct the misspelling words in the report for the Information Extraction System to extract out the information which cannot be extracted out previously, we implemented a Misspellings Handling System to do this issue. After the processing of correcting misspellings, those missing information will be extracted out by our Information Extraction System.

The electronic medical reports from National Taiwan University Hospital have hierarchy, the data composed by many patients, each patient has his own report collection, and each report collection has numbers of report which has different types of report (admission report, discharge summary, echo report, operation report, radiology report, pathology report). After exposing text from the medical report, each text will be encoded by the phonetic algorithm. We built our own corpus by using the keyword in our regular expression. The words in corpus are also encoded by phonetic algorithm in order to compare words in the report. By comparing different between word in the report and word in the corpus in phonetic algorithm, system will generate a suggestion list for each misspelled word. However, there are many suggestions to a single misspelled word in the suggestion list. We have to select the best solution to the misspelled word. Thus, we developed a ranking process by using edit distance algorithm to calculate the distance between misspelled word and words in suggestion list. After the ranking process, we can get the most similar word from the suggestion list. We just need to replace the misspelled word with word in suggestion list with the highest ranking. And then we transfer the corrected report to the Information Extraction System. Those missing information which has misspelled word can be extracted out by the IE system after misspellings correction. In Figure 11, we illustrate the whole process of the Misspellings Handling System.

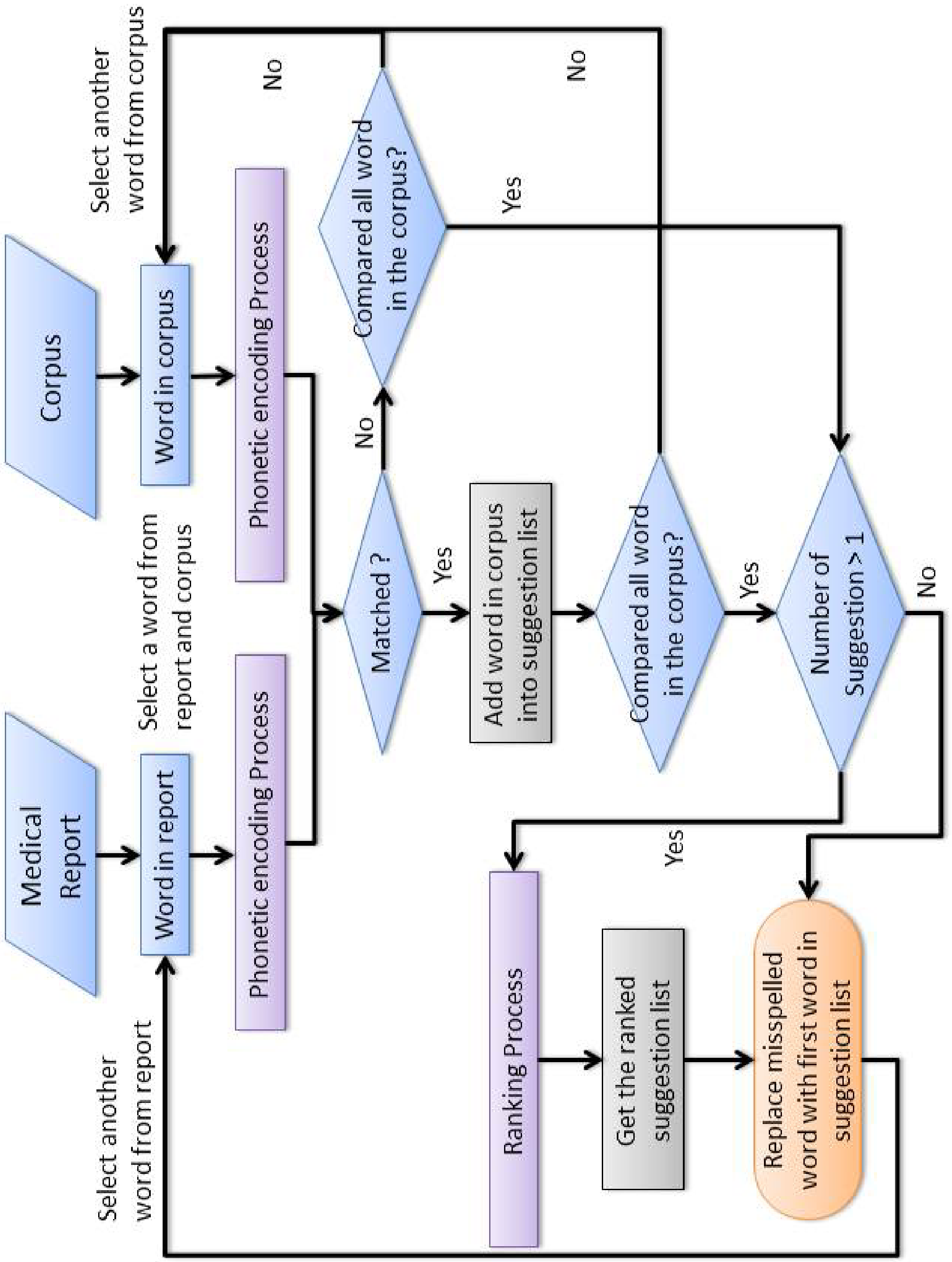

Medical Report
Word in report
Phonetic encoding Process
Select another word from report
Select a word from report and corpus
Corpus
Word in corpus
Phonetic encoding Process
Select another word from corpus
Matched ?
No
Yes
Compared all word in the corpus?
No
Yes
Add word in corpus into suggestion list
Compared all word in the corpus?
No
Yes
Number of Suggestion > 1
Yes
No
Ranking Process
Get the ranked suggestion list
Replace misspelled word with first word in suggestion list

Figure 14 Overview of Misspellings Handling Process

Chapter 4 Results and Discussions

4.1 Evaluation Methods

In order to evaluate the performance of Misspelling Handling System we used the following methods for further evaluation. We give the definitions to the following terms

Table 4 Definition of TP, TF, FP and FN

Name	Definition
True Positive (TP):	The misspelled words in the reports those were detected and corrected correctly.
True Negative (TN):	The correct words in the original report those were determined as correct words by Misspelling Handling System.
False Positive (FP):	The correct words in the original report those were determined as misspelled words or wrongly corrected by Misspelling Handling System.
False Negative (FN):	The misspelled words in the original report those were "not" determined as misspelled words by Misspelling Handling System.

After we got numbers of TP, TF, FP and FN, we used the following equations to evaluate our performance.

Precision: The fraction of detecting and correcting the misspelled words in the report correctly in our study. In other study, it measures how many of the items that the system identified were actually correct. The higher precision is, the better the system is at ensuring what has been identified is correct.[14]

Equation 2 Precision

$$Precision = \frac{tp}{tp + fp}$$

Recall: The fraction of correctly misspelling correction in all misspelling correction. In other study, it measures how many of the items that should have been identified actually were identified [14].

Equation 3 Recall

$$Recall = \frac{tp}{tp + fn}$$

F-measure: An evaluate score for the system which combines precision and recall is the harmonic mean of precision and recall. Evaluating a system performance only by its precision or recall result is not so fair. (For example, a system has high precision but has very low recall). In contrary, F-measure provides a better way for evaluating a system performance by combining both of precision and recall [14]. There are many types of F-measure (F_1, F_2 and $F_{0.5}$). We use F_1 in this study.

Equation 4 F_1-measure

$$F = 2 \cdot \frac{precision \ \cdot recall}{precision + recall}$$

4.2 Results

In this study we total processed 152 patients who have 5734 reports. We divide the data set in to 2 sets: Training set: 74 patients who have 3195 reports. Testing set: 78 patients who have 2539 reports.

Table 5 Performance of Misspellings Handling System

	Training Set		Testing Set	
Number of patients	*74*		*78*	
Number of reports	*3195*		*2539*	
Total # of error in the report	*203*		*357*	
Total # of error detected and corrected by MHS	*187*		*319*	
Total # of error correctly detected and corrected by MHS	*162*		*277*	
Total # of errors wrongly detected and corrected by MHS	*25*		*42*	
Score	*Precision*	*Recall*	*Precision*	*Recall*
	87%	*91%*	*86%*	*88%*
	F-measure		*F-measure*	
	89%		*88%*	

Table 5 shows the evaluation result of the performance of our Misspelling Handling System. First, we review the medical report manually to determine the total number of misspelled words in the original report. Second, we ran the Misspelling Handling System for processing those reports. Third, we reviewed the correcting result manually to check weather wrongly corrected or not. In the training set we found there are 203 errors in the original report, but only 187 misspelling candidates are found by our Misspelling Handling System in which 162 misspellings are detected and corrected correctly. Thus we got 87% in precision, 91% in recall and 89% in f-measure. In our testing set we found there are 357 misspellings in the original report, and there are 319 misspellings detected by our system in which 277 were detected and corrected correctly. Thus we got 86% in precision, 88% in recall and 88% in F-measure. In the following table we show some cases about wrong correction and correct correction.

Table 6 Examples of correctly correction

Target Word	Word in Corpus
hccso	hcc
stduy	study
brace	barce
clinico, clincal	clinic, clinical
heptocellular, hepatocellualr, hepatocelluler	hepatocellular
barcelonoa, barcilona, barcellona	barcelona
hepatooma	hepatoma
carcinomaa, carcionma	carcinoma
stauts, statis	status
lever, llliver, livero	liver
stagea	Stage
cancero	cancer

Table 7 Examples of wrongly correction

Target Word	Word in Corpus
stood, sided	state
duke	take
steady	study
folate, fluid, fleet	fold
thright	thyroid
cytoogy	stage
ketone	kidney
cotton	codeine
cheek	check
cleary	color
keelung	clinic

Chapter 5 Conclusions and Future Works

5.1 Conclusions

The raw medical report about patients with liver cancer contains lots of valuable medical related information. However, those information are stored in unstructured format which is hard for a computer system do the further analysis. We designed a Medical Information Extraction System for extracting the related information and store them in structured form. Because the information was input by human, there has some misspellings result in the difficulty for extracting information by the Medical Information Extraction System. Thus we implemented the Misspelling Handling System for detecting and correcting the misspellings in the medical report. After the preprocessing by theMisspelling Handling System, those missing information can be extracted out by the Medical Information Extraction System.

5.2 Future Works

Although we have implemented a Misspelling Handling System, the performance of the system still needs to be improved. Two main factors will affect the performance of the system. 1. The algorithm used in the System. Because different types of words (e.g. Medical Terminology and General English), the Medical Terms are suitable for using edit distance algorithm and General English are suitable for using phonetic algorithm. If we can apply different types of algorithm according to the word’s property which can help increase the performance of the Misspelling Handling System. 2. The words in the corpus. The more words in the corpus will help the Misspelling Handling System process more types of report. However, when we included too many words in the corpus which may result in the Misspelling Handling decreasing the recall and precision. We also need to consider the part of speech (such as noun, verb, adverb, adjective, etc.) and timing form (such as past, present, future) of the word. The words in the corpus are mostly noun and verb. When the words in the report are verb with different timing form (such as -ed, -ing, etc.), it will lead to Misspelling Handling System wrongly correcting and detecting the word. If we can solve those 2 main factors, the performance of the Misspelling Handling System will be better.

Another part need to be improved is the flexibility of the Misspelling Handling System. Our system includes a single corpus. If we can applied the ontology concept into the Misspelling Handling System, which can make our system process more diseases of report (breast cancer, kidney cancer, etc.) and more flexibile.

References

1. Mamlin, B.W., D.T. Heinze, and C.J. McDonald, *Automated extraction and normalization of findings from cancer-related free-text radiology reports.* AMIA Annual Symposium proceedings / AMIA Symposium. AMIA Symposium, 2003: p. 420-4.
2. Hripcsak, G., et al., *Unlocking Clinical Data from Narrative Reports: A Study of Natural Language Processing.* Annals of Internal Medicine, 1995. **122**(9): p. 681-688.
3. Wang, Z.-J., *The methodology of facilitating data analysis in medical informatics -information extraction from free-text data and structural data collection through the structure report interface*, in *Graduate Institute of Biomedical Electronics and Bioinformatics* 2010, National Taiwan University.
4. Spasic, I., et al., *Text mining and ontologies in biomedicine: Making sense of raw text.* Briefings in Bioinformatics, 2005. **6**(3): p. 239-251.
5. McGuinness, D.L. and F.v. Harmelen. Available from: http://www.w3.org/TR/owl-features/.
6. Protégé was developed by Stanford Center for Biomedical Informatics Research at the Stanford University School of Medicine. Available from: http://protege.stanford.edu/.
7. Wu, Y.-L., *A method for identifying confidence level of the extracted results from medical narrative reports: A case study focus on the patients with liver cancer*, master thesis, *Graduate Institute of Biomedical Electronics and Bioinformatics* 2012, National Taiwan University.
8. Mykowiecka, A., et al., *Rule-based information extraction from patients' clinical data.* J. of Biomedical Informatics, 2009. **42**(5): p. 923-936.
9. Cohen, A.M. and W.R. Hersh, *A survey of current work in biomedical text mining.* Briefings in Bioinformatics, 2005. **6**(1): p. 57-71.
10. Botsis, T., et al., *Secondary Use of EHR: Data Quality Issues and Informatics Opportunities.* AMIA Summits on Translational Science proceedings AMIA Summit on Translational Science, 2010. **2010**: p. 1-5.
11. Keith E. Stuart, M. and M. Melissa Conrad Stöppler. Available from: http://www.medicinenet.com/liver_cancer/article.htm.
12. Myo Thant, M.; This content was last reviewed August 15, 2010 by Dr. Reshma L. Mahtani.]. Available from: http://www.caring4cancer.com/go/liver/basics.
13. Gadd, T.N., *PHOENIX: the algorithm.* Program: Autom. Libr. Inf. Syst., 1990. **24**(4): p. 363-369.
14. Maynard, D.M.a.D., *Metrics for Evaluation of Ontology-based Information*, in *In WWW 2006 Workshop on Evaluation of Ontologies for the Web*2006.
15. Goyvaerts, J. 23 October 2011; Available from: http://www.regular-expressions.info/.
16. Martín, F.C., *Approximate string matching algorithms in art media archives* 2009, AGH University of Science and Technology.
17. Uzzaman, N., *A Bangla Phonetic Encoding for Better Spelling Suggestion*, in *Proc. 7th International Conference on Computer and Information Technology*2004.

18. Garaev, K.G., *A Remark on the Bellman Principle of Optimality.* Journal of The Franklin Institute, 1998. **335**(2): p. 395-400.