國立臺灣大學電機資訊學院生醫電子與資訊學研究所

碩士論文

Graduate Institute of Biomedical Electronics and Bioinformatics

College of Electrical Engineering and Computer Science

National Taiwan University

Master Thesis

肝癌病患病歷報告之辨識資訊萃取結果可信賴程度

A method for identifying confidence level of the extracted results from medical narrative reports: A case study focus on the patients with liver cancer

吳亞霖

Ya-Lin Wu

指導教授：賴飛羆 博士

Advisor: Feipei Lai, Ph.D.

中華民國 101 年 1 月

January, 2012

誌謝

能完成這篇論文，我要特別感謝我的指導教授賴飛羆老師在這兩年半的諄諄教誨，除了在課業上的指導外，老師平常的生活態度與言行舉止也非常值得學習。感謝實驗室的每一位成員對我的支持與鼓勵，但因要感謝的人實在太多了，因此直接感謝老天爺讓我有緣和大家相遇。

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。資訊擷取後，重要的是證明這些提取結果是可靠的。但沒有經由人為檢查的方式，

在這項研究中，我們兩個團隊成員已檢查所有提取信息。根據檢查結果可得到資訊擷取系統的準確度。人為檢查所有提取結果的方式，是一個耗時耗力的工作。因此，本研究的目的在於提供一個有效率的方式去檢查提取結果。我們設計一驗證系統，用於預測每個提取結果的正確性。據驗證系統預測的結果，檢查人員可以有效地檢查那些被驗證系統預測為錯誤資訊的提取結果並且進行校正，而不需檢查所有的提取結果。透過驗證系統可以提高檢查提取結果的效率。

關鍵字：病歷資料、資訊擷取系統、驗證系統。

ABSTRACT

Textual medical records constitute a rich source of information about diseases, medical procedures and treatment results. In our previous work, we implemented the information extraction (IE) system for extracting the desired information from liver cancer patients' textual reports. These extracted results produced by IE system are used for supporting the development of recurrence predictive model. After information was extracted by the IE system, it is important to prove these extracted results are reliable. However, we are not sure about the correctness of these extracted results without checking manually by the domain experts.

In the study, two of our team members had reviewed all extracted information. According to their reviews, the precision of the IE system can be analyzed. But, checking the correctness of all extracted results 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 all extracted results. We designed the validation system for predicting the correctness of each extracted result [1]. According to the prediction of the validation system, the reviewers can efficiently check the smaller part of extracted results predicted as low confidence extracted information by the validation system and correct them; instead of checking all extracted information. In this way, it can highly promote the efficiency of the future reviewing process.

Keywords: textual medical records, information extraction system, validation system.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1 Introduction

1.1 Motivation

The textual medical records about patients with liver cancer contain the medical related information about diagnosis, tumor information, cancer staging, treatment and recurrent status. In our previous work, we had developed the IE system for extracting the liver cancer related information. The extracted results produced by the IE system are used for developing the liver cancer recurrence predictive model. Therefore, the proving the reliability of extracted results is very important to further application.

In this case, we focus on the extracted results about the tumor information. However, the most common method to prove their reliability is checking the extracted information manually. But, the process of manually reviewing these extracted results is a stubborn task that is usually long and laborious. In the extracted information from the IE system, the correct extracted information occupies a larger proportion compared to the wrong or uncertain extracted information. Therefore, reviewers waste most of their time on those correct extracted information but actually they just have to correct and check the wrong or uncertain extracted information. According to this problem, we aim to design a system to find those wrong or uncertain extracted information. With the help of the system, reviewers can efficiently correct and check the wrong or uncertain extracted information that recognized by the validation system.

1.2 Purpose

Natural language processing technology has been around for many years [2]. Although significant progress has been made, there is still room for improvement.

For example, most implemented medical IE systems quote recall in the 80–85% range and precision of 95–99% [3]. This level of performance is not perfect. In order to provide a reliable data for further applications, the correcting of wrong extracted results is very important.

In this research, we develop a validation system and want to test the possibility of the validation system to predict the correctness of each extracted result after the IE process. According to the prediction of the validation system, the reviewers can efficiently check the uncertain-extracted information or wrong-extracted information labeled with a low confidence scores by the validation system and correct them; instead of checking all extracted results. With the help of the validation system, we hope to promote the efficiency of the manual reviewing process.

1.3 General Procedure

In this case we deal with the textual medical records of patients with liver cancer. The textual medical records pass the IE system to produce a lot of extracted results. The extracted results are the grouped information about tumor, tumor size, tumor location extracted by the IE system. There is other related information also extracted in the IE process. The information collected by the IE system was used to help the validation system to predict the correctness of each extracted result. Figure 1 shows the overview of the general procedure to process textual medical records, then check all extracted results manually and finally predict the correctness of each extracted result by the validation system.

In order to prepare a predictive training model and calculate the performance of the validation system, the extracted result was checked for their correctness manually in the human-labeling process. One of our team members gives a confidence score label to each extracted result. Then, the labeled training set and testing set are input into the validation system.

The support vector machine (SVM) classifier is employed for developing the validation system which is based on the information extracted by the IE system to give a confidence score label to each extracted result. After the step of validation system predicting the correctness of extracted results, each extracted results have two confidence scores, the human-labeling confidence score and the validation system predicted confidence score. Finally, the predicted confidence score of the validation system is compared with the human-labeling confidence score to calculate the performance of validation system. If the validation system has a good performance to predict the confidence score on each extracted result, reviewers can efficiently check the extracted results with low confidence label, instead of checking all extracted results.

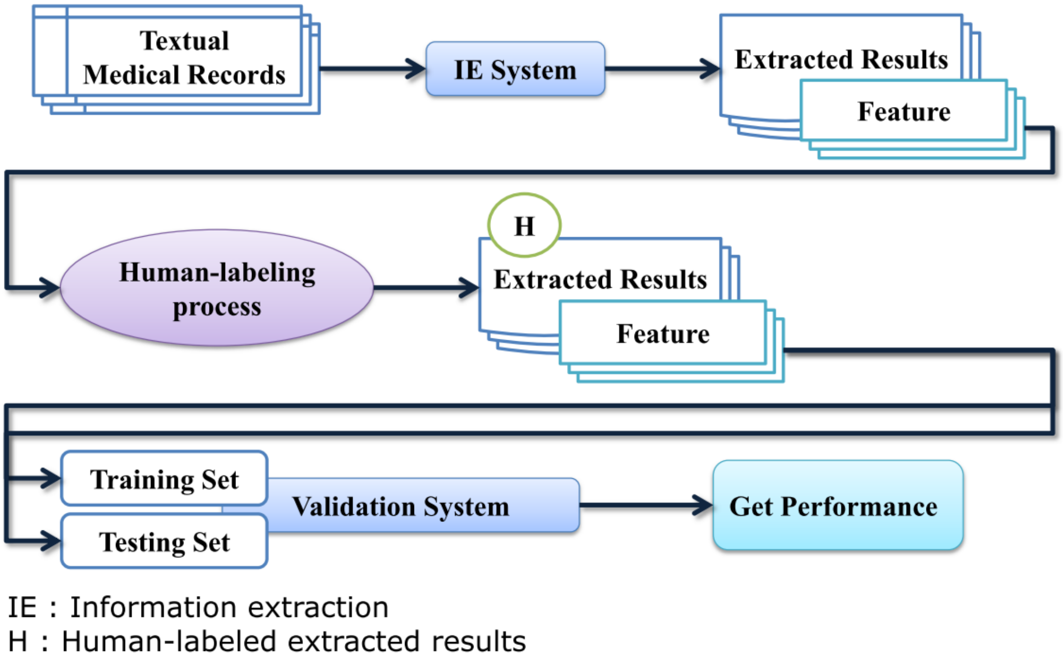


Figure 1 Overview of the general steps of this research.

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 validation system to predict the correctness of extracted information. 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 Information Extraction System on Medical Records

In the medical domain, there has been a surge of interest in using electronic medical record (EMR) systems to improve the quality of medical care through decision support, evidence-based medicine, disease surveillance, etc. In order to make full use of the information contained in EMRs, we have to deal with those unstructured textual data that make up a large portion of the medical records. Thus, there seems to be an increased demand for information extraction tools applied to those medical narratives records [4]. The use of the IE system can help us efficiently to acquire the information hidden in the large amount of medical records [5].

When we applied the IE system to those medical records, there are two big problems should be considered and we propose a proper solution. First, we have to concern the expansion and reusability of the IE system. In our case, we develop an IE system for extracting the desired information from liver cancer patients' textual reports. Regular expression was used by the IE system to match the target information in the textual medical records. If there is a requirement to add or remove the current version of regular expression, it is a trouble to modify the codes that define the regular expression in the IE system. On the other hand, if we want to apply this IE system to other specific cancer domain, such as breast cancer. It would be an annoying task to rewrite all regular expression defined in the IE system to fit the domain of breast cancer. Second, we have to concern the data quality of extracted results from the IE system. Data quality issues that related to the secondary use of EHR are discussed in [6]. The authors consider three categories of data quality issues:

1. Incompleteness: missing information.
2. Inconsistency: information mismatches the data source.
3. Inaccuracy: non-specific, non-standards-based, inexact, incorrect, or imprecise information.

Narrative medical records store patient related information and extracting this information correctly is an important mission for further practical application. The inconsistency is common to data collections. Currently, the main way to check the extracted information is thru manually reviewing. But the manual process is usually long, laborious and prone to errors. In [2], their suggestion is to develop software tools for automatic data validation and flexible data presentation in order to support information integrity.

In this research, we present two approaches to solve the two problems. First, we use the ontology model as an external regular expression source. The ontology model defines the regular expression that have to import into the IE system. These regular expressions are crucial to improve the performance of the IE system. The separation of algorithm in the IE system and domain knowledge in the ontology model makes our application have more elasticity. 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. Second, we propose a validation system to predict the correctness of extracted results after the IE process. The validation system will give a confidence score label to each extracted result. An extracted result would get a high confidence score label or a low confidence label. The confidence score label represents the correctness of each extracted result predicted by the validation system. According to our validation system, the reviewers can efficiently check the extracted results with a low confidence score label, instead of checking all extracted results produced by the IE system.

2.2 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 [7]. The type of liver cancer arising from the liver is known as primary liver cancer. The proportion of primary liver cancers arise from hepatocytes is about 90% to 95% and this kind of primary liver cancer is also called hepatocellular carcinoma [7].

The other type of liver cancer is called secondary liver cancer or metastatic liver cancer [7]. 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.

2.3 Diagnosis of Liver Cancer

2.3.1 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 [8]. 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 [8]. 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.

2.3.2 Imaging Studies

An imaging study can provide useful information as to the size, the location, the number of abnormal liver tissues. But, imaging studies cannot tell the difference between a hepatoma and other abnormal masses, lesion or nodules in the liver. There are several types of imaging studies:

Ultrasound (Echo): The ultrasound is used to show any abnormal growths in your liver.

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

Computed tomography scan (CT): CT scans can show how large the tumor is and whether it has spread to other organs or to the lymph nodes.

Magnetic resonance imaging (MRI): MRI can help clinicians to determine whether the liver cancer is primary type or secondary type.

2.3.3 Liver Biopsy

Liver biopsy is considered to provide the definite diagnosis of liver cancer [9]. In 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 [9].

2.4 Treatment of Liver Cancer

2.4.1 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.

2.4.2 Liver Resection

The goal of liver resection is to completely remove the tumor without leaving any tumor behind [10]. 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 [10]. 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.

2.5 Related Works

Information extraction aims to retrieve certain types of information from natural language text by processing them automatically. Ontology-based information extraction (OBIE) has recently emerged as a subfield of information extraction. In the OBIE systems, the IE process is guided by ontology to extract information related to the concepts defined in ontology [11]. The OBIE system defines a medical ontology to specify the information to be extracted (e.g. [12, 13]). In our approach, we did not use the ontology to structure textual information into template but to define the regular expression in the ontology which was used by the IE system to recognize target information in the textual medical records.

The data quality is very important for providing reliable applications. A research had stated their application [14], the validity of the breast cancer classifier that can help in early detection of malignancy depends on the quality of the underlying data. Unfortunately, most data suffer from inconsistencies, missing data, inter-observer variability and inappropriate term usage. Therefore, they aim to propose a more powerful IE system to get the reliable data. Other research [15] states that the negative results are an inevitable aspect of the NLP performance but they are partly due to the inconsistency, incompleteness and fragmentariness of the textual medical records. These negative results might be dangerous for further use but in their case the small percentage of negative results is statistically insignificant and practically negligible. In our research, we aim to provide a validation system to find these negative results and let the reviewers check and correct them. The combination of the IE system and the validation system can provide a more reliable data for further use of medical applications.

2.6 Datasets

The data for training and testing our validation 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 82 liver cancer patients who took RFA to treat their cancer and 120 liver cancer patients who took liver resection to treat their cancer. These 202 liver cancer patients have totally 5673 narrative medical records processed by the IE system.

The testing set contains 93 liver cancer patients who took liver resection to treat their cancer. These 93 liver cancer patients have totally 2067 narrative medical records processed by the IE system.

After the IE process, the extracted results in the training set and testing set obtain a confidence score label in the human-labeling process. Then, these manually labeled extracted results enter the validation system.

2.6.1 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 case, we focus on the tumor information in the liver cancer patients' textual medical reports. We implemented an information extraction system to extract and group desired information. 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.

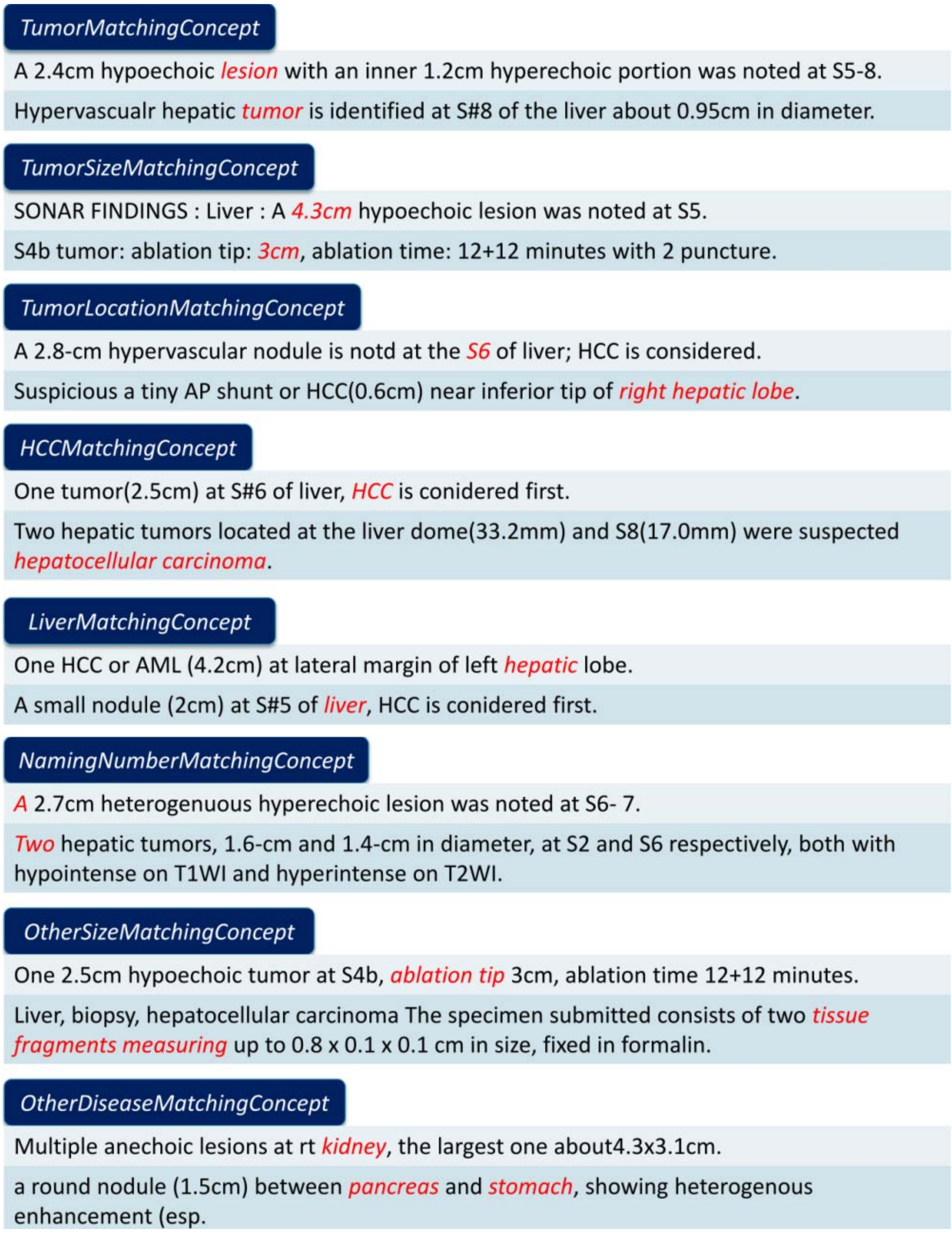
1. TumorMatchingConcept: This concept indicates the finding of abnormal tissues described in the textual medical records.
2. TumorSizeMatchingConcept: This concept indicates the figure information about size of any object in the textual medical records.

3. TumorLocationMatchingConcept: This concept indicates liver anatomy information that used to describe the location of abnormal tissues in the textual medical records.
4. HCCMatchingConcept: This concept indicates the terms hepatocellular carcinoma or HCC in the textual medical records.
5. LiverMatchingConcept: This concept indicates the terms liver and hepatic in the textual medical records.
6. NamingNumberMatchingConcept: This concept indicates the quantity information of abnormal tissues in the textual medical records.
7. OtherSizeMatchingConcept: This concept indicates the other object (exclude abnormal tissues) which may also have figure information in the textual medical records.
8. OtherDiseaseMatchingConcept: This concept indicates the other organ (exclude liver) or other disease information in the textual medical records.
9. 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.

Table 1 Occurrence information of the dataset

Concept Name	**Training Concept Number (RFA/OP)**			**Testing Concept Number (OP)**		
	Total	**Total 82/120 Patients**	**Total 2510/3163 Reports**	**Total**	**Total 93 Patients**	**Total 2067 Reports**
		Per Patient	**Per Report**		**Per Patient**	**Per Report**
Tumor	*1216/2091*	*14.8/17.4*	*0.5/0.7*	*1600*	*17.2*	*0.8*
TumorSize	*1100/1770*	*13.4/14.8*	*0.4/0.6*	*1426*	*15.3*	*0.7*
TumorLocation	*853/1381*	*10.4/11.5*	*0.3/0.4*	*1102*	*11.8*	*0.5*
HCC	*225/323*	*2.7/2.7*	*0.1/0.1*	*331*	*3.6*	*0.2*
Liver	*717/1141*	*8.7/9.5*	*0.3/0.4*	*927*	*10*	*0.4*
NamingNumber	*896/1330*	*10.9/11.1*	*0.4/0.4*	*956*	*10.3*	*0.5*
OtherSize	*384/494*	*4.7/4.1*	*0.2/0.2*	*452*	*4.9*	*0.2*
OtherDisease	*740/895*	*9/7.5*	*0.3/0.3*	*719*	*7.7*	*0.3*
Special	*206/331*	*2.5/2.8*	*0.1/0.1*	*269*	*2.9*	*0.1*



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

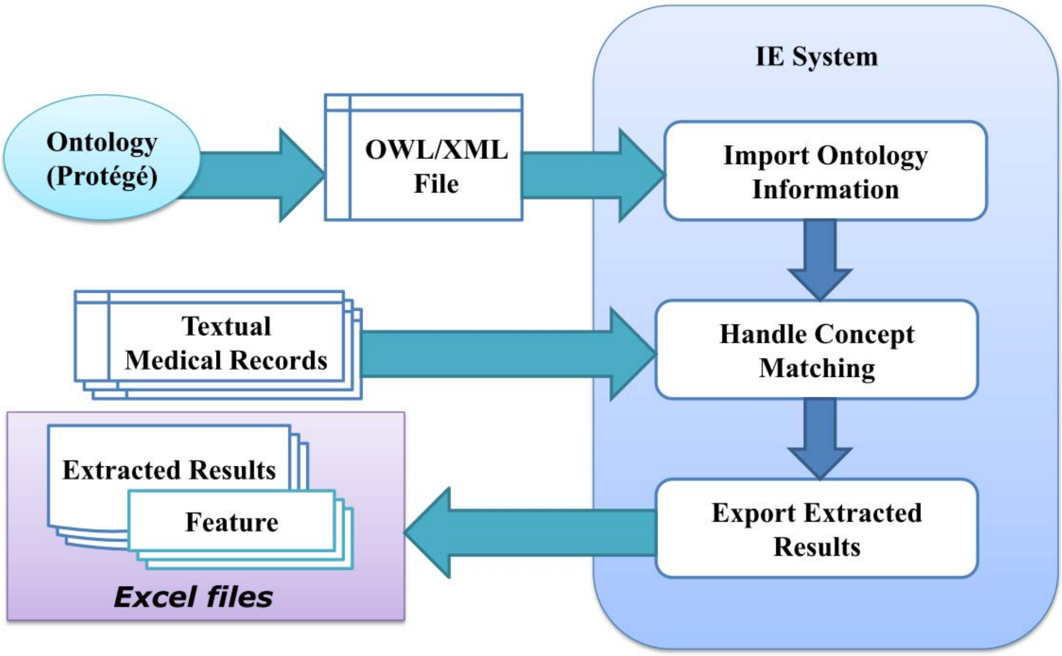
Multiple anechoic lesions at rt *kidney*, the largest one about4.3x3.1cm.

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

Figure 2 Example of concept occurs in the sentence of textual medical records.

Chapter 3 Method

3.1 Information Extraction System



IE：Information extraction

Figure 3 Overview of general steps of the IE system.

The information extraction system was designed to extract the target information in the textual medical records and export an Excel file that let the reviewers check the extracted results. Figure 3 shows the general procedure of the information extraction system to process the textual medical records and obtain external knowledge source from the ontology. When the IE system operates, it retrieves the supplement of regular expression from the OWL/XML file exported by the ontology editor, protégé. The additional regular expression information defined in the ontology is imported into the IE system. Then it retrieves admission reports, discharge summaries, echo reports, operation reports, radiology reports and pathology reports from the database. After retrieving textual medical records, the IE system starts to handle concepts matching and extract information from these 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

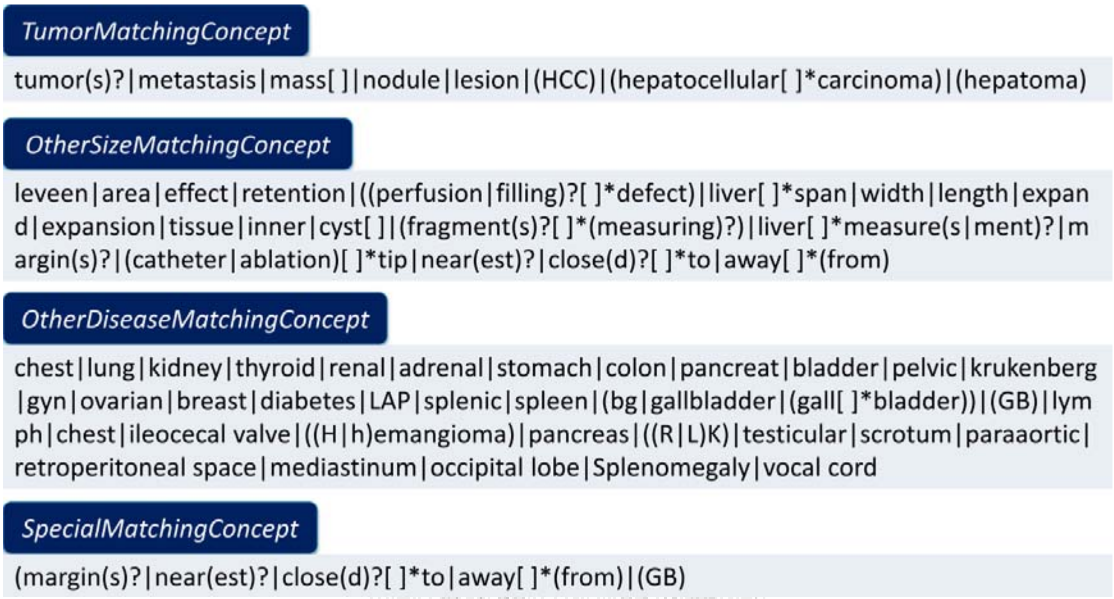
Regular expression is a pattern that specifies a set of strings which matched against a subject string from left to right. They are used when you want to search for specific lines of text which containing a particular pattern. In IE application, regular expression provides a concise and flexible means for matching, specifying and recognizing strings of text, such as particular characters, words, or patterns of characters [16]. 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 frequently occurrence in the wrong extracted results. We aim to use this concept to help the validation system to distinguish these wrong 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 4 Example of regular expression defined in the IE system.

3.1.2 Ontology

Ontologies are used to describe knowledge about the interesting domain [17]. 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) [18].

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 [19]. 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 expression in the ontology model.

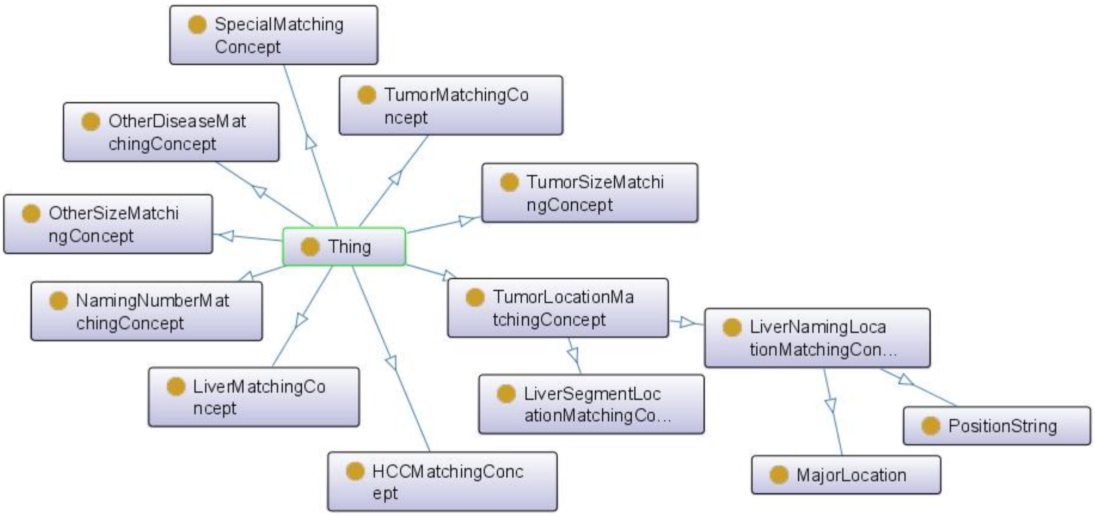
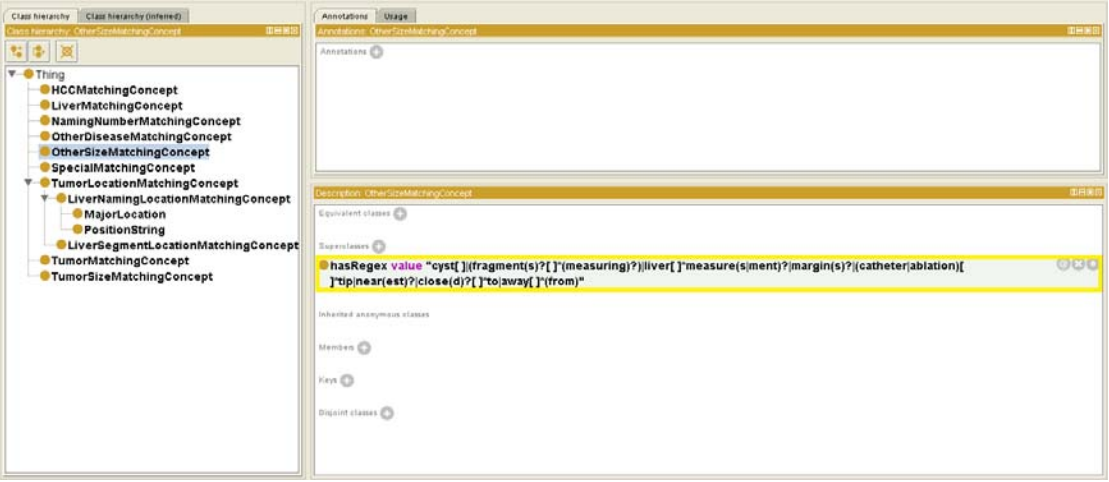


Figure 5 Hierarchy structure of the ontology model.



Class hierarchy | Class hierarchy (inferred)

Class hierarchy: OtherSizeMatchingConcept

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

Annotations | Usage

Annotations: OtherSizeMatchingConcept

Annotations

Description: OtherSizeMatchingConcept

Equivalent classes

Superclasses

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

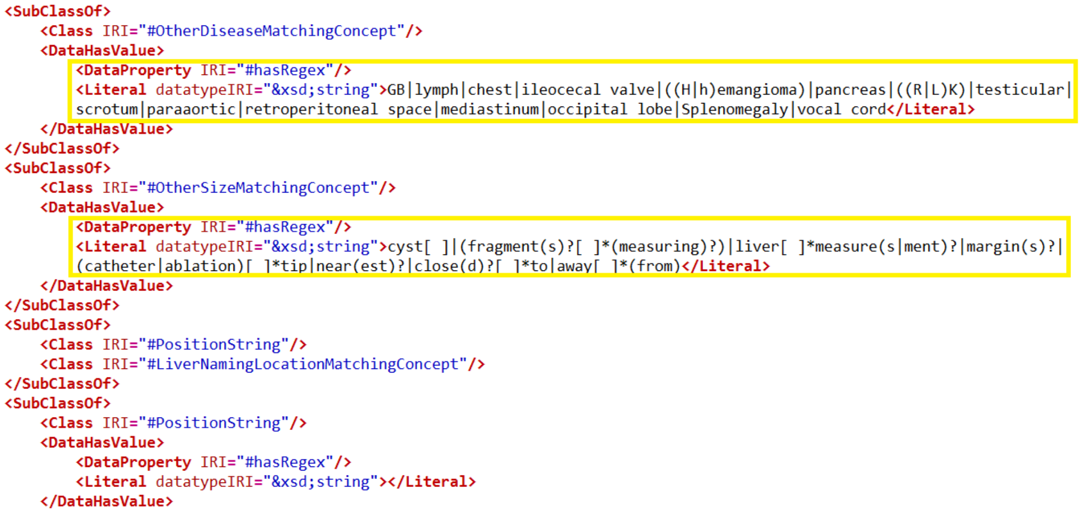
Inherited anonymous classes

Members

Keys

Disjoint classes

Figure 6 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 7 The OWL/XML file produced by protégé.

3.1.3 Concept Matching

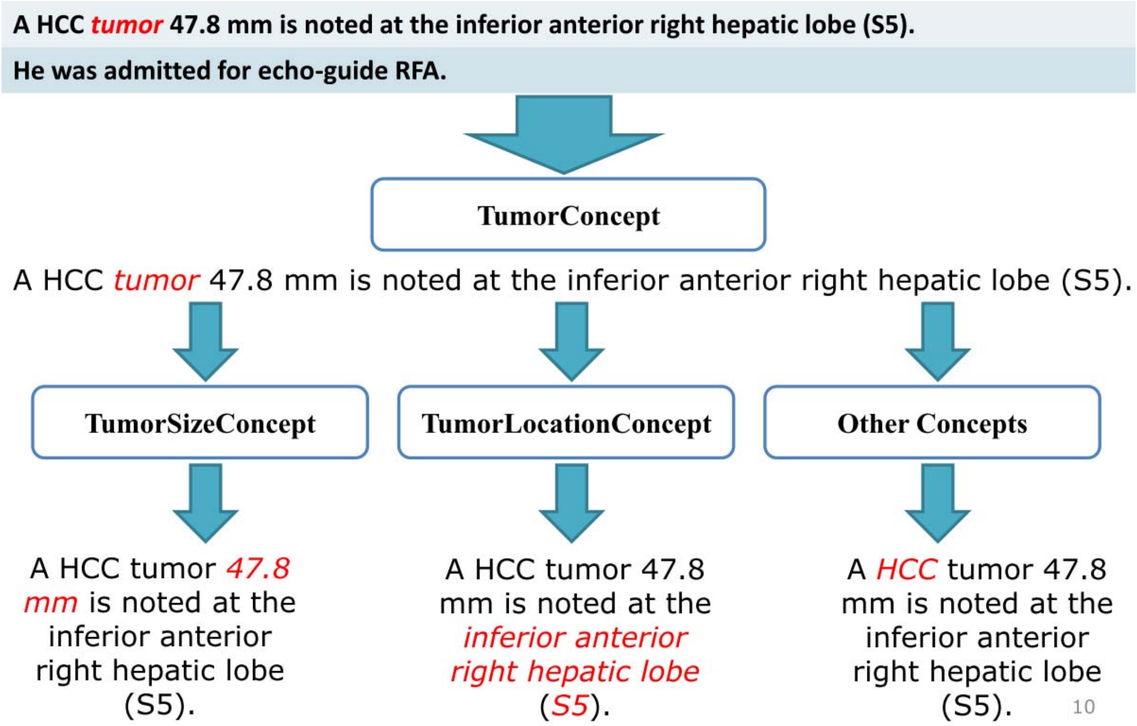

A HCC tumor 47.8 mm is noted at the inferior anterior right hepatic lobe (S5).
He was admitted for echo-guide RFA.
TumorConcept
A HCC tumor 47.8 mm is noted at the inferior anterior right hepatic lobe (S5).
TumorSizeConcept
TumorLocationConcept
Other Concepts
A HCC tumor 47.8 mm is noted at the inferior 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 8 Example of two layer filters to extract the target information.

When textual medical reports enter the IE system, it was split into small piece of sentences and the IE system check 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 Human-Labeling Process

The human reviewing step has two main tasks. In the first task, reviewers would check the correctness of extracted results and give a confidence score label to each extracted result. In the second task, reviewers may find some useful information in the reviewing process which can be used to improve the performance of the IE system. Reviewers add these key concepts into the ontology model. The regular expression defined in the ontology was used as external knowledge source to the IE system.

After the IE process, the IE system outputs an Excel file which contains all extracted information such as extracted results, the original sentence of each extracted result and other related features collected by the IE system. In the human-labeling process, reviewers check the correctness of each extracted result by comparing with their source sentences on the outputted Excel file from the IE system. The extracted result is the grouped information about 3 concepts.

- TumorMatchingConcept
- TumorSizeMatchingConcept
- TumorLocationMatchingConcept

If reviewers consider the extracted result is correct, reviewers would give a high confidence score label to the extracted result. On the contrary, if reviewers consider the extracted result is uncertain or wrong, reviewer would give a low confidence score label to the extracted result.

There are two functions of the human-labeling confidence score. First, it makes the validation system realize which type of extracted result is of high confidence or low confidence and let the validation system have the ability to predict the confidence label of each extracted result. Second, the human-labeling confidence score was used to compare with the validation system predicted confidence score to calculate the performance of the validation system. Figure 10 shows the example of correct extracted results labeled with high confidence score. Figure 11 shows the example of wrong extracted results labeled with low confidence score. Figure 12 shows the example of uncertain extracted results labeled with low confidence score.

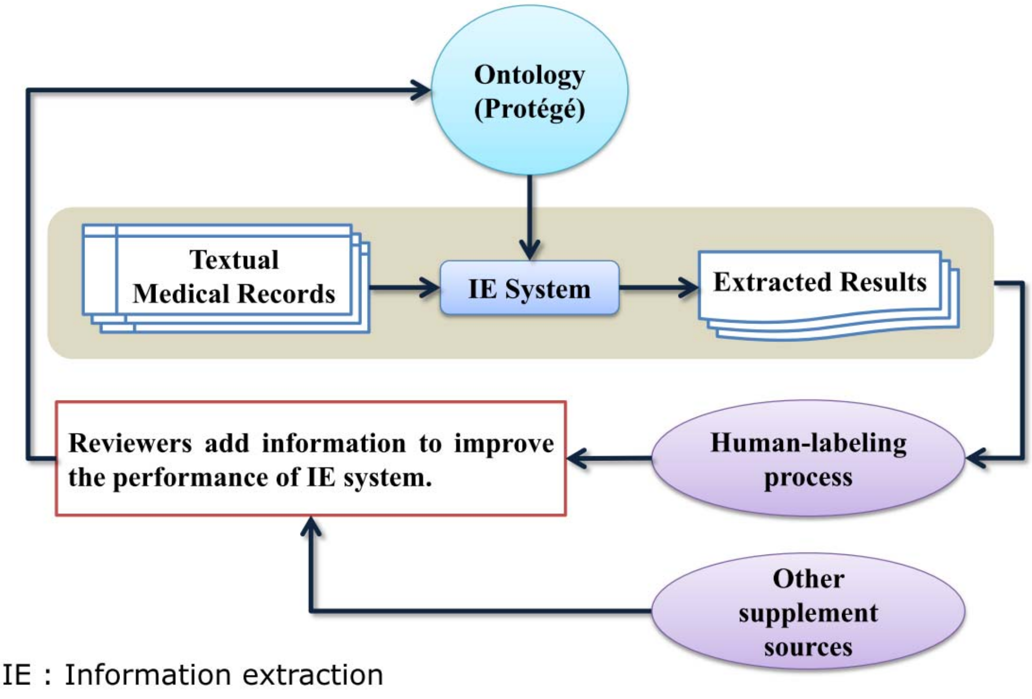
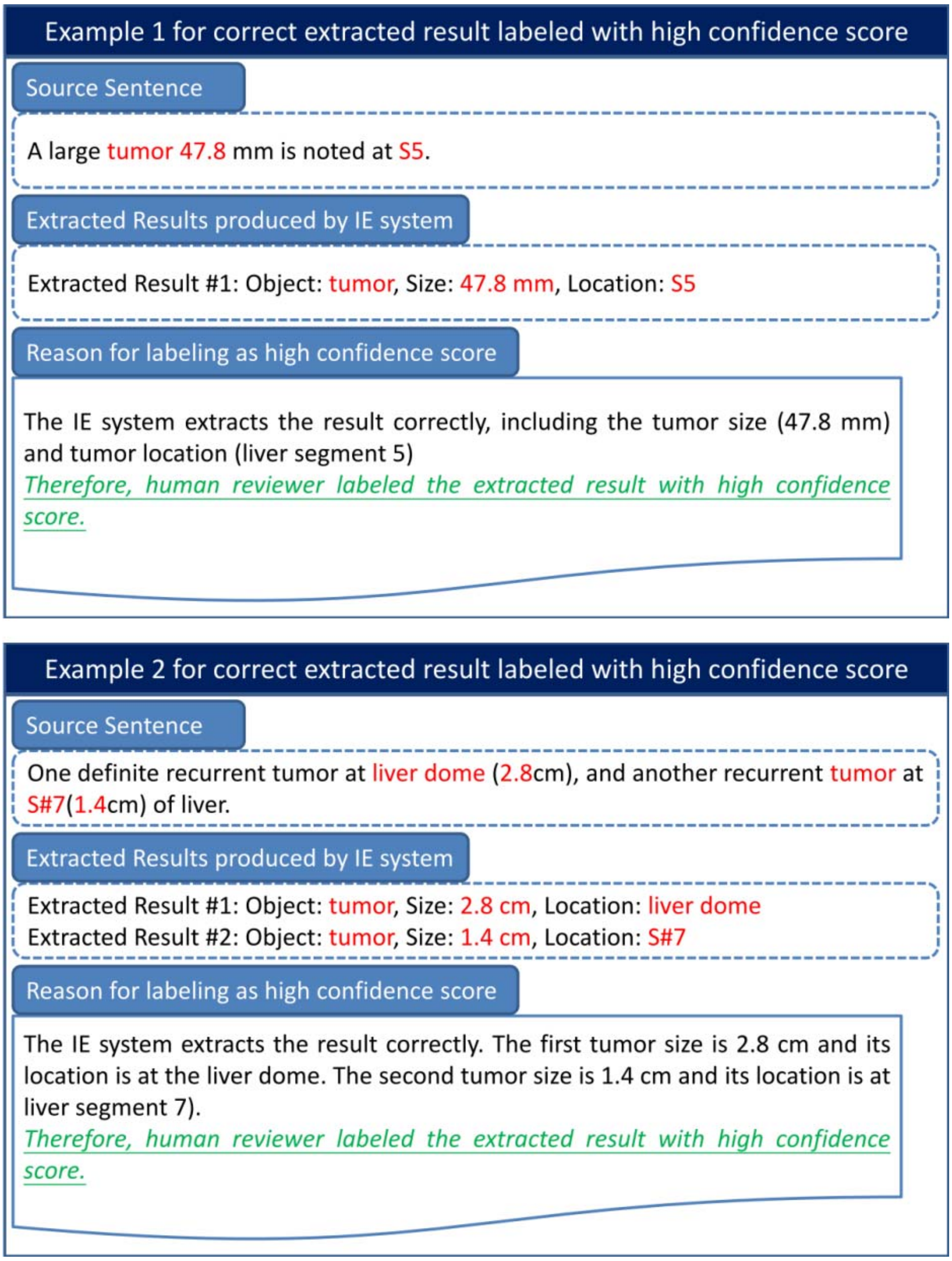


Figure 9 Overview of the supplement of knowledge source.

In the uncertain extracted results or wrong extracted results, reviewers may find some particular characters, words, or patterns of characters which are not specified in the regular expression and they may cause the IE system to extract the uncertain extracted results or wrong extracted results. Therefore, reviewers would add these particular characters, words or patterns of characters to the ontology model. The information defined in the ontology which updates the regular expression in the IE system. The change of regular expression would affect the IE system to extract information. Therefore, the textual medical records are processed by the IE system again as a result of the modification of regular expression in the IE system. The extracted results produced by the updated IE system are different from the previous one. Reviewers check and give a confidence score label to each extracted results again. If no more information can be added to the ontology model, the human-labeling extracted results enter the validation system for predicting the correctness of each extracted results. Figure 9 shows the general procedure of the reviewer to check the extracted results and add information to the ontology model in the human-labeling process. Beside the useful information found in the reviewing process was added to the ontology, the information such as disease terminology, human anatomy terminology etc. were also used to improve the performance of the IE system and added to the ontology.



Example 1 for correct extracted result labeled with high confidence score

Source Sentence

A large tumor 47.8 mm is noted at S5.

Extracted Results produced by IE system

Extracted Result #1: Object: tumor, Size: 47.8 mm, Location: S5

Reason for labeling as high confidence score

The IE system extracts the result correctly, including the tumor size (47.8 mm) and tumor location (liver segment 5)

Therefore, human reviewer labeled the extracted result with high confidence score.

Example 2 for correct extracted result labeled with high confidence score

Source Sentence

One definite recurrent tumor at liver dome (2.8cm), and another recurrent tumor at S#7(1.4cm) of liver.

Extracted Results produced by IE system

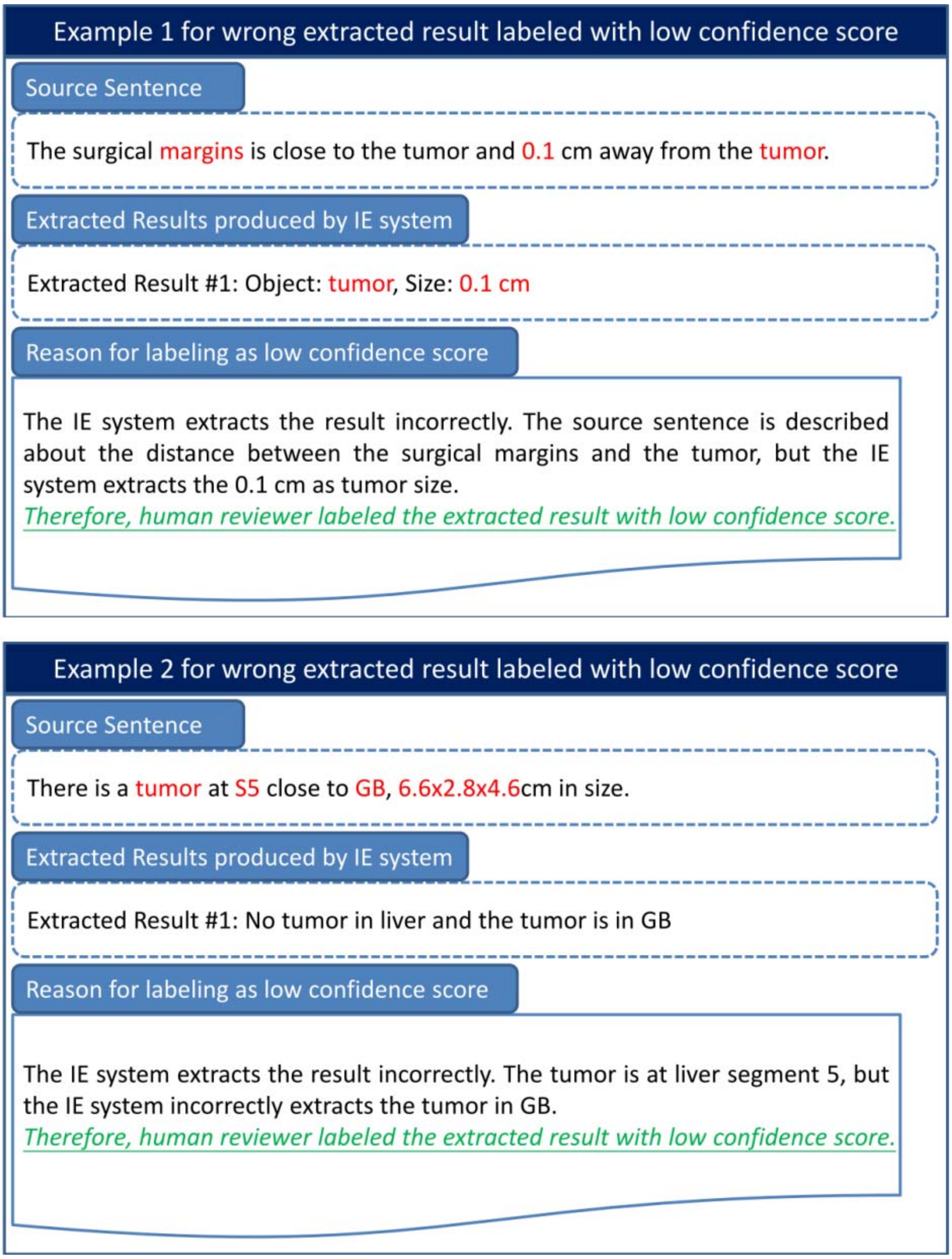
Extracted Result #1: Object: tumor, Size: 2.8 cm, Location: liver dome
Extracted Result #2: Object: tumor, Size: 1.4 cm, Location: S#7

Reason for labeling as high confidence score

The IE system extracts the result correctly. The first tumor size is 2.8 cm and its location is at the liver dome. The second tumor size is 1.4 cm and its location is at liver segment 7).

Therefore, human reviewer labeled the extracted result with high confidence score.

Figure 10 Examples of correct extracted results labeled with a high confidence score.



Example 1 for wrong extracted result labeled with low confidence score

Source Sentence

The surgical margins is close to the tumor and 0.1 cm away from the tumor.

Extracted Results produced by IE system

Extracted Result #1: Object: tumor, Size: 0.1 cm

Reason for labeling as low confidence score

The IE system extracts the result incorrectly. The source sentence is described about the distance between the surgical margins and the tumor, but the IE system extracts the 0.1 cm as tumor size.

Therefore, human reviewer labeled the extracted result with low confidence score.

Example 2 for wrong extracted result labeled with low confidence score

Source Sentence

There is a tumor at S5 close to GB, 6.6x2.8x4.6cm in size.

Extracted Results produced by IE system

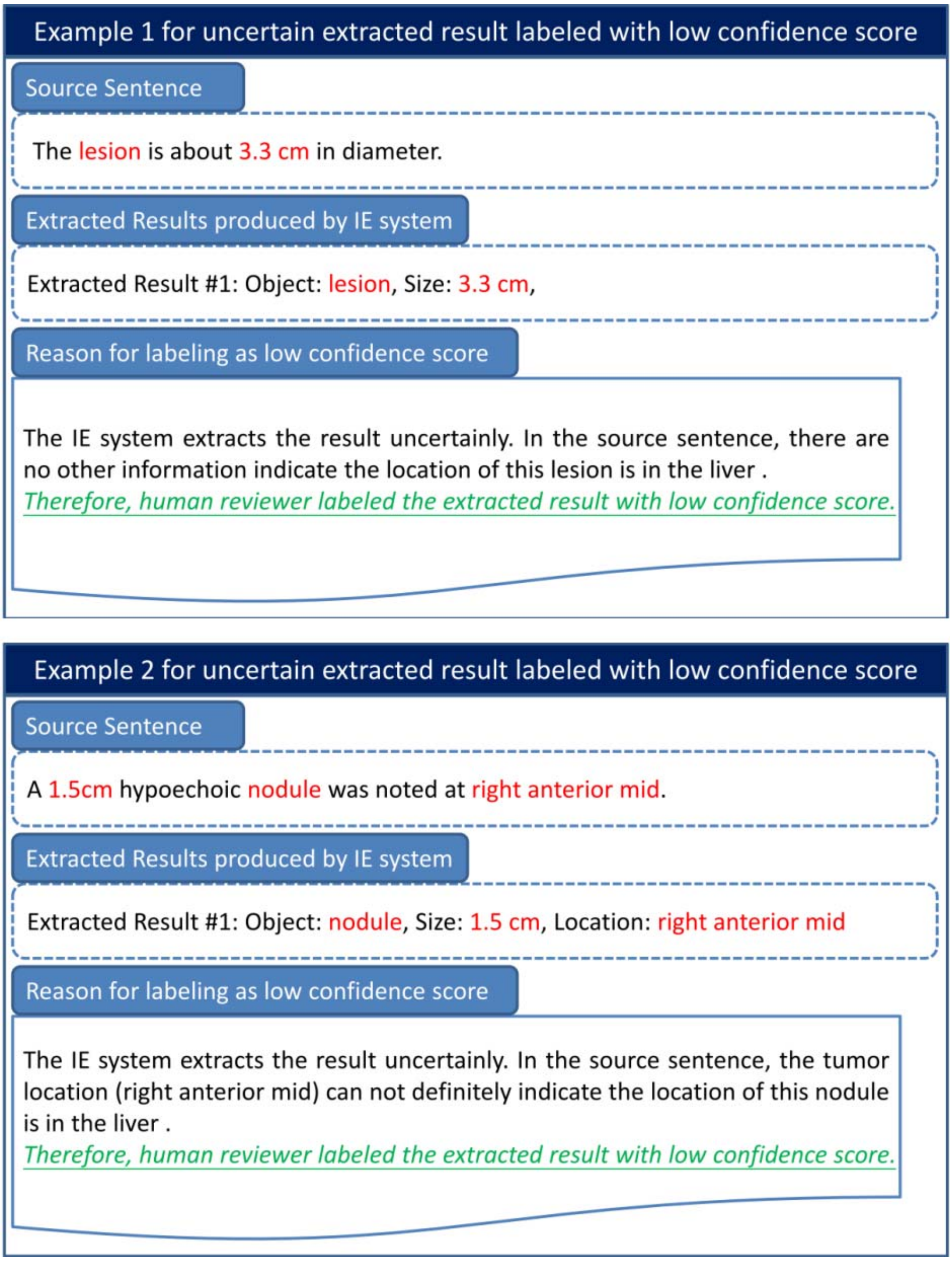
Extracted Result #1: No tumor in liver and the tumor is in GB

Reason for labeling as low confidence score

The IE system extracts the result incorrectly. The tumor is at liver segment 5, but the IE system incorrectly extracts the tumor in GB.

Therefore, human reviewer labeled the extracted result with low confidence score.

Figure 11 Examples of wrong extracted results labeled with a low confidence score.



Example 1 for uncertain extracted result labeled with low confidence score

Source Sentence

The lesion is about 3.3 cm in diameter.

Extracted Results produced by IE system

Extracted Result #1: Object: lesion, Size: 3.3 cm,

Reason for labeling as low confidence score

The IE system extracts the result uncertainly. In the source sentence, there are no other information indicate the location of this lesion is in the liver .
Therefore, human reviewer labeled the extracted result with low confidence score.

Example 2 for uncertain extracted result labeled with low confidence score

Source Sentence

A 1.5cm hypoechoic nodule was noted at right anterior mid.

Extracted Results produced by IE system

Extracted Result #1: Object: nodule, Size: 1.5 cm, Location: right anterior mid

Reason for labeling as low confidence score

The IE system extracts the result uncertainly. In the source sentence, the tumor location (right anterior mid) can not definitely indicate the location of this nodule is in the liver .
Therefore, human reviewer labeled the extracted result with low confidence score.

Figure 12 Examples of uncertain extracted results labeled with a low confidence score.

3.3 Validation System

The support vector machine (SVM) classifier is employed for developing the validation system [20]. A classification task usually involves with training and testing data which consist of some data instances. Each instance in the training set contains one class label and several attributes. The goal of SVM is to produce a model which predicts the class label of data instances in the testing set which are given with the attributes.

In the training set and testing set, each extracted result has its corresponding features collected by the IE system during the extracting process. The validation system bases on these collected features to predict the correctness of extracted results. Table 2 shows the total features used by the validation system to predict the correctness of extracted results. The features can be categorized into 3 groups:

1. The occurrences of concept in the source sentence. Figure 14 shows the example of this type of feature.
2. The order between two concepts in the source sentence. Figure 15 shows the example of this type of feature.
3. The distance between two concepts in the source sentence. Figure 16 shows the example of this type of feature.

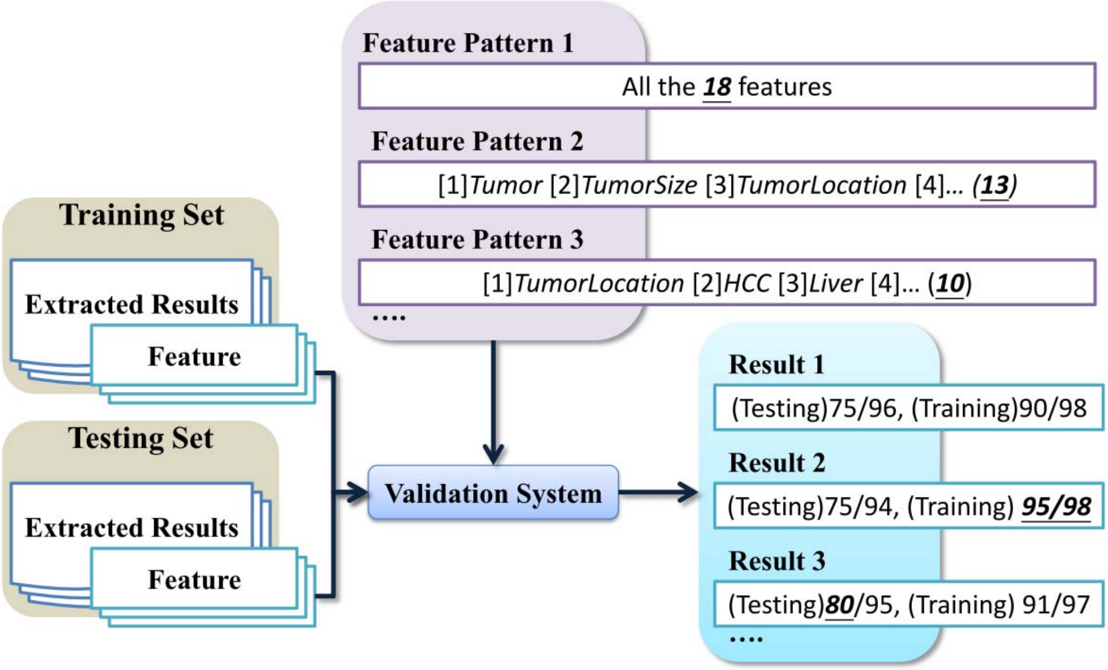


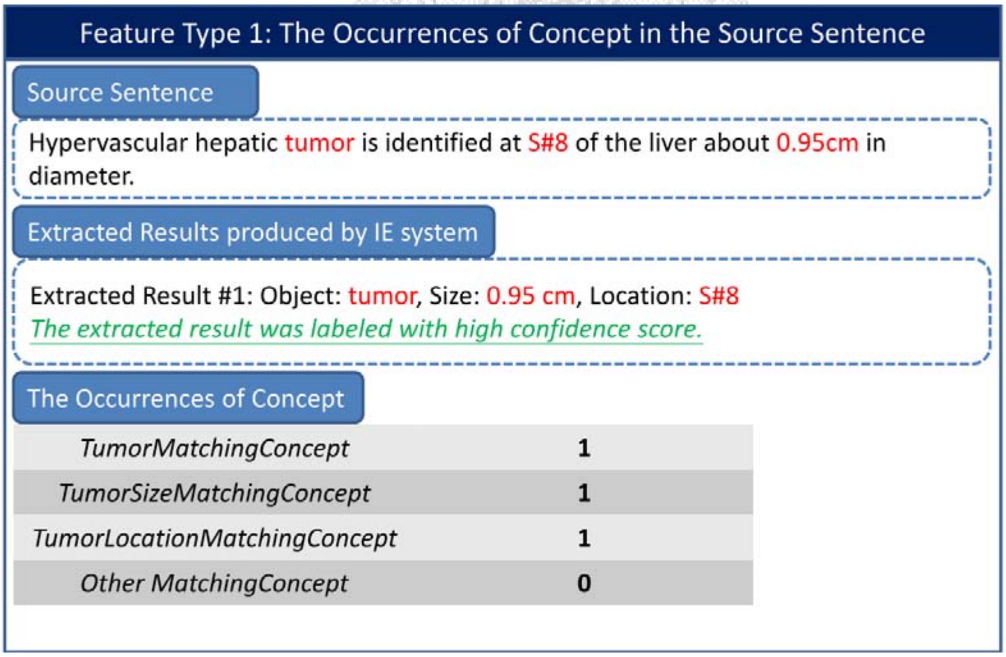
Figure 13 Overview of the validation system tries different feature patterns.

The IE system has collected 18 features that are used as input information to the validation system. According to the features of extracted results in training set, testing set and the selected feature patterns, the validation system can predict the confidence score of extracted results. Finally, the human-labeling confidence score would be compared with the validation system predicted confidence score to calculate the sensitivity and specificity. According to the prediction of the validation system, these extracted results predicted with low confidence are taken into the reviewing process for checking their correctness.

In the beginning, we use all the 18 features to the validation system and get 90% sensitivity, 98% specificity on predicting the training set; 75% sensitivity, 96% specificity on predicting the testing set. We make the validation system try other feature patterns which may have better performance on predicting the correctness of extracted results. Figure 13 shows the overview of the validation system which tries different feature patterns to get better performance on predicting the correctness of extracted results. The best performance on predicting the training set, 95% sensitivity, 98% specificity and use 13 features, is shown in Figure18. The best performance on predicting the testing set, 80% sensitivity, 95% specificity and use 10 features, is shown in Figure 19.

Table 2 The 18 features used by the validation system.

Feature type 1: The occurrences of concept in the source sentence (8)			
Tumor	*TumorSize*	*TumorLocation*	*HCC*
Liver	*NamingNumber*	*OtherSize*	*OtherDisease*
Feature type 2: The order between two concepts in the source sentence (5)			
Tumor & TumorSize	*TumorLocation&TumorSize*	*Special & TumorSize*	
TumorLocation & Tumor	*Special & Tumor*		
Feature type 3: The distance between two concepts in the source sentence (5)			
Tumor & TumorSize	*TumorLocation&TumorSize*	*Special & TumorSize*	
TumorLocation & Tumor	*Special & Tumor*		



Feature Type 1: The Occurrences of Concept in the Source Sentence

Source Sentence

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

Extracted Results produced by IE system

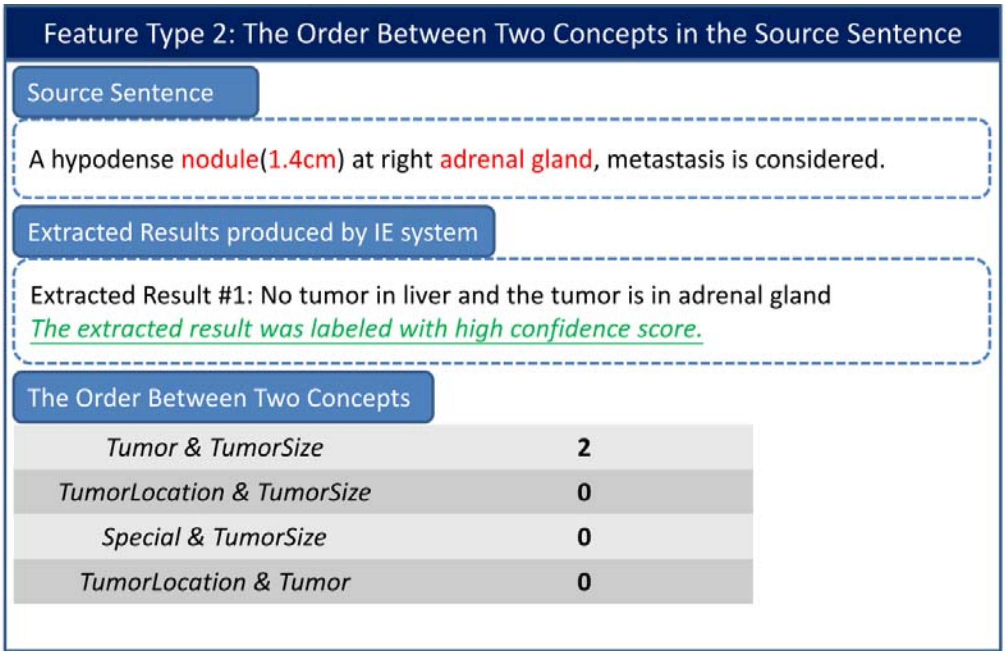
Extracted Result #1: Object: tumor, Size: 0.95 cm, Location: S#8

The extracted result was labeled with high confidence score.

The Occurrences of Concept

TumorMatchingConcept	1
TumorSizeMatchingConcept	1
TumorLocationMatchingConcept	1
Other MatchingConcept	0

Figure 14 Example the occurrences of concept in the source sentence.



Feature Type 2: The Order Between Two Concepts in the Source Sentence

Source Sentence

A hypodense nodule(1.4cm) at right adrenal gland, metastasis is considered.

Extracted Results produced by IE system

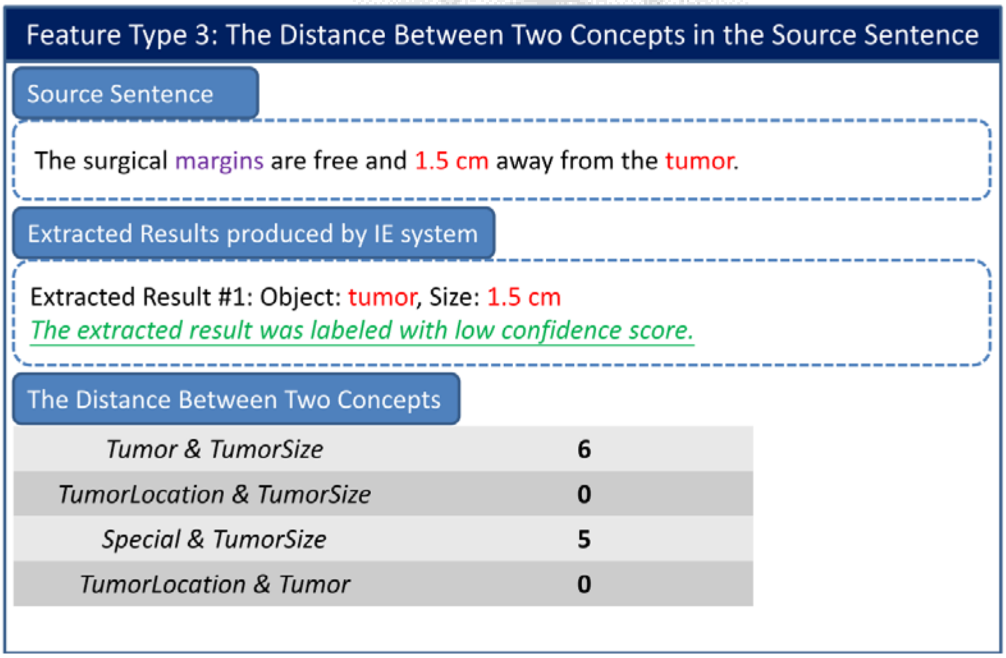
Extracted Result #1: No tumor in liver and the tumor is in adrenal gland

The extracted result was labeled with high confidence score.

The Order Between Two Concepts

Tumor & TumorSize	**2**
TumorLocation & TumorSize	**0**
Special & TumorSize	**0**
TumorLocation & Tumor	**0**

Figure 15 Example the order between two concepts in the source sentence.



Feature Type 3: The Distance Between Two Concepts in the Source Sentence

Source Sentence

The surgical margins are free and 1.5 cm away from the tumor.

Extracted Results produced by IE system

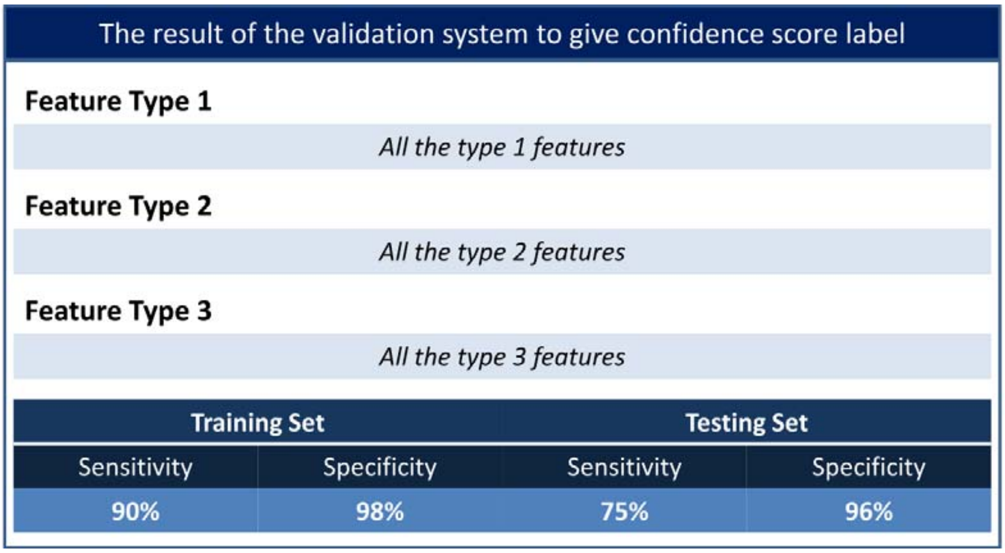
Extracted Result #1: Object: tumor, Size: 1.5 cm

The extracted result was labeled with low confidence score.

The Distance Between Two Concepts

Tumor & TumorSize	**6**
TumorLocation & TumorSize	**0**
Special & TumorSize	**5**
TumorLocation & Tumor	**0**

Figure 16 Example the distance between two concepts in the source sentence.



The result of the validation system to give confidence score label

Feature Type 1

All the type 1 features

Feature Type 2

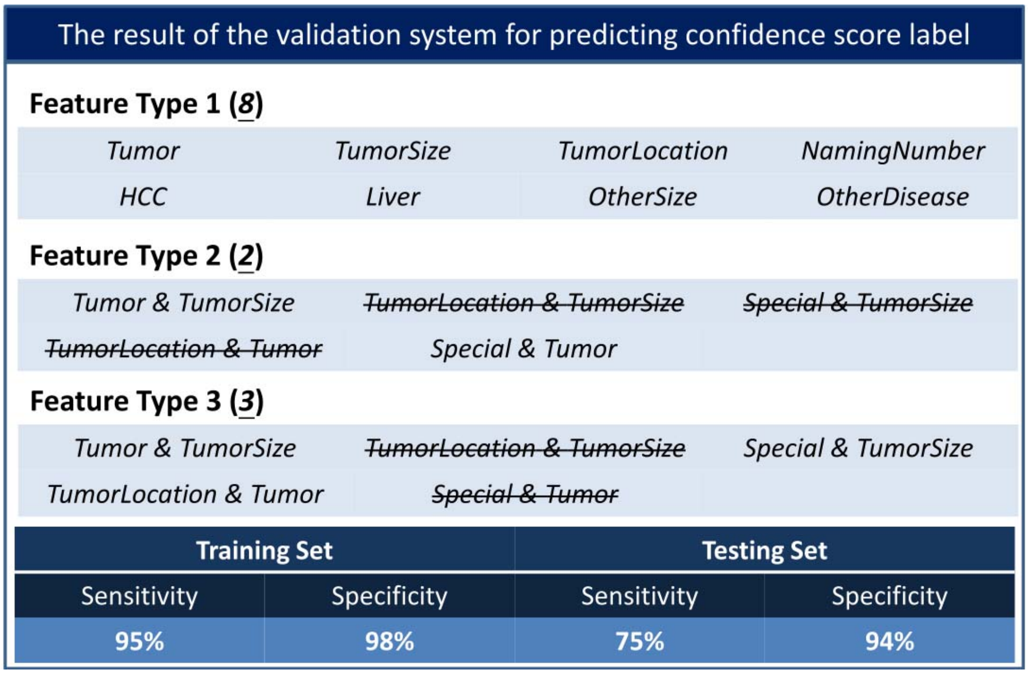
All the type 2 features

Feature Type 3

All the type 3 features

Training Set		Testing Set	
Sensitivity	Specificity	Sensitivity	Specificity
90%	98%	75%	96%

Figure 17 Sensitivity and specificity of using all 18 features.



The result of the validation system for predicting confidence score label

Feature Type 1 (8)

Tumor	*TumorSize*	*TumorLocation*	*NamingNumber*
HCC	*Liver*	*OtherSize*	*OtherDisease*

Feature Type 2 (2)

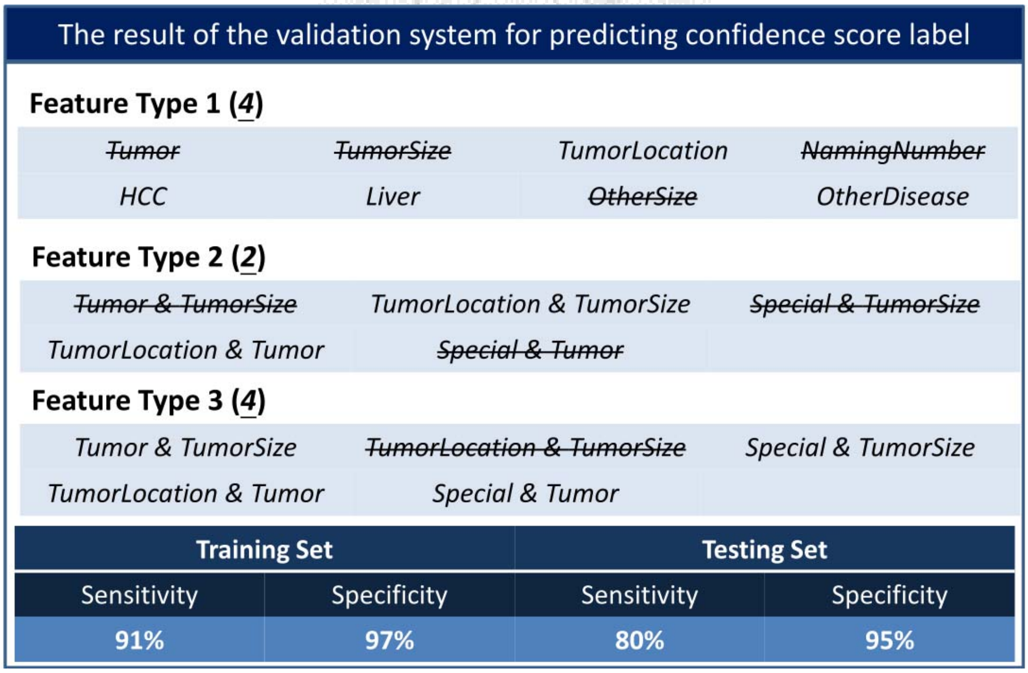
Tumor & TumorSize	~~*TumorLocation & TumorSize*~~	~~*Special & TumorSize*~~
~~*TumorLocation & Tumor*~~	*Special & Tumor*	

Feature Type 3 (3)

Tumor & TumorSize	~~*TumorLocation & TumorSize*~~	*Special & TumorSize*
TumorLocation & Tumor	~~*Special & Tumor*~~	

Training Set		Testing Set	
Sensitivity	Specificity	Sensitivity	Specificity
95%	**98%**	**75%**	**94%**

Figure 18 Sensitivity and specificity of using all 13 features.



The result of the validation system for predicting confidence score label

Feature Type 1 (4)

~~*Tumor*~~	~~*TumorSize*~~	*TumorLocation*	~~*NamingNumber*~~
HCC	*Liver*	~~*OtherSize*~~	*OtherDisease*

Feature Type 2 (2)

~~*Tumor & TumorSize*~~	*TumorLocation & TumorSize*	~~*Special & TumorSize*~~
TumorLocation & Tumor	~~*Special & Tumor*~~	

Feature Type 3 (4)

Tumor & TumorSize	~~*TumorLocation & TumorSize*~~	*Special & TumorSize*
TumorLocation & Tumor	*Special & Tumor*	

Training Set		Testing Set	
Sensitivity	Specificity	Sensitivity	Specificity
91%	**97%**	**80%**	**95%**

Figure 19 Sensitivity and specificity of using all 10 features.

Chapter 4 Results and Discussions

4.1 Results

Figure 20 shows we obtain 1271 feature patterns which are the best results from the validation system in trying parts of the combination of the 18 features. We select 11 features that have higher occurrence in the group of better results and try all feature patterns of the combination of the 11 features. Then, we obtain 83 better results which are shown in Figure 21.

The feature patterns can affect the performance of the validation system to predict the correctness of extracted results. Therefore, the validation system tries different combinations of feature to get better performance. If the validation system uses all the 18 features to predict the correctness of extracted results, it gets 90% sensitivity and 98% specificity on predicting the training set; 75% sensitivity and 96% specificity on predicting the testing set. If the validation system uses the 11 features shown in Table 3 to predict the correctness of extracted results, it gets 91% sensitivity and 97% specificity on predicting the training set; 76% sensitivity and 96% specificity on predicting the testing set. The other comparisons of the best performance on predicting the training set and the testing set between using the different combination of the 18 features and the 11 features are shown in Table 4. Currently, the best sensitivity to recognize the wrong extracted results in the testing set is 80%.

The sensitivity was used for presenting the predicative performance of the extracted results with low confidence score. The specificity was used for presenting the predicative performance of the extracted results with high confidence score.

The sensitivity = number of true positives / (number of true positives + number of false negatives). The specificity used for presenting the predicative performance of the extracted results with high confidence score. The specificity = number of true negatives / (number of true negatives + number of false positives). The extracted results are divided into training set and testing set.

In Figure 20 and Figure 21, we find 4 features to be the key features to the validation system on predicting the correctness of extracted results.

1. The occurrence of LiverConcept
2. The occurrence of OtherDiseaseConcept
3. The distance between TumorConcept and TumorSizeConcept
4. The distance between TumorLocationConcept and TumorConceot

Training Set		Testing Set		
Sensitivity	Specificity	Sensitivity	Specificity	Amount of results
>=90%	>=90%	>=75%	>=85%	1271

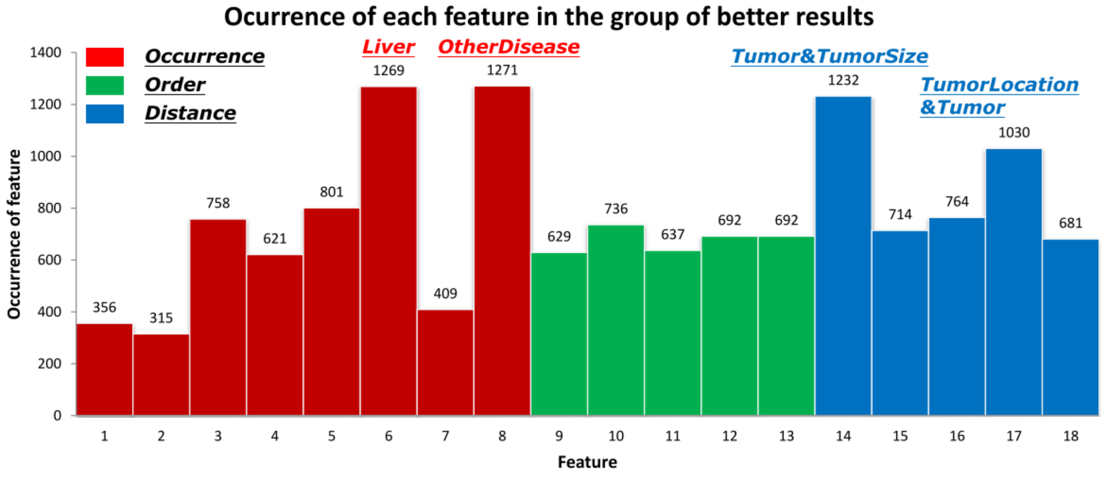
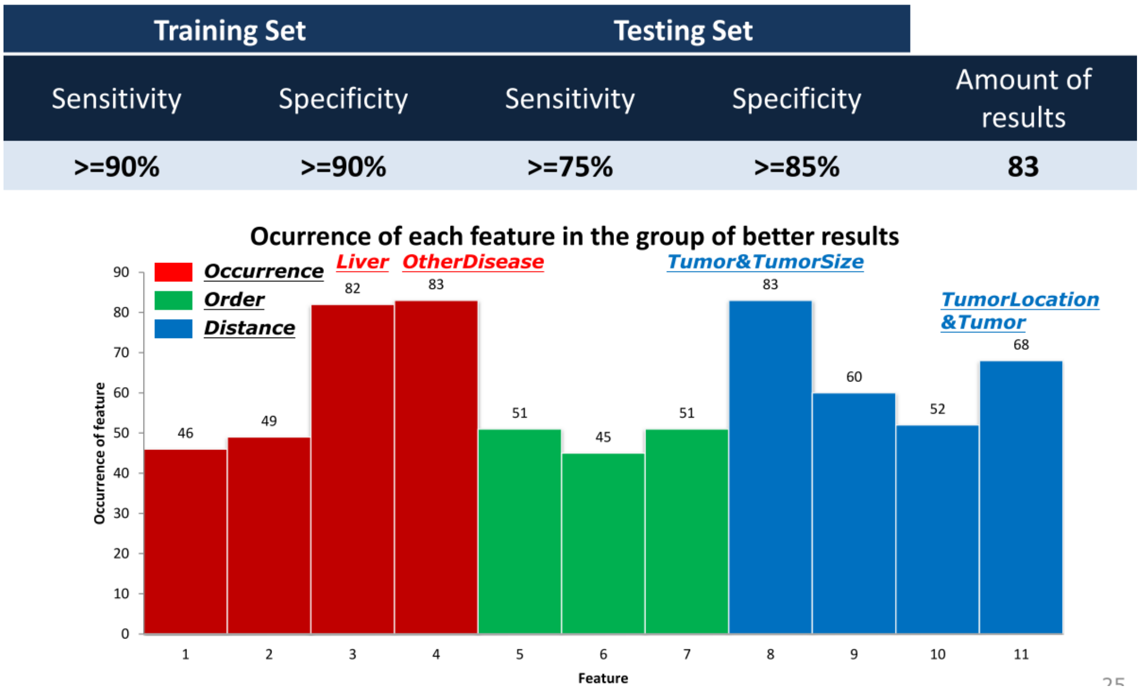


Figure 20 Occurrence of 18 features in the group of better results.

Table 3 The 11 features used by the validation system.

<table>
<tr><th colspan="12">Feature type 1: The occurrences of concept in the source sentence (4)</th></tr>
<tr><td colspan="3"><em>TumorLocation</em></td><td colspan="3"><em>Liver</em></td><td colspan="3"><em>NamingNumber</em></td><td colspan="3"><em>OtherDisease</em></td></tr>
<tr><th colspan="12">Feature type 2: The order between two concepts in the source sentence (3)</th></tr>
<tr><td colspan="4"><em>TumorLocation&TumorSize</em></td><td colspan="4"><em>TumorLocation & Tumor</em></td><td colspan="4"><em>Special & Tumor</em></td></tr>
<tr><th colspan="12">Feature type 3: The distance between two concepts in the source sentence (4)</th></tr>
<tr><td colspan="4"><em>Tumor & TumorSize</em></td><td colspan="4"><em>TumorLocation&TumorSize</em></td><td colspan="4"><em>Special & TumorSize</em></td></tr>
<tr><td colspan="4"><em>TumorLocation & Tumor</em></td><td colspan="4"></td><td colspan="4"></td></tr>
</table>



Training Set		Testing Set		
Sensitivity	Specificity	Sensitivity	Specificity	Amount of results
>=90%	**>=90%**	**>=75%**	**>=85%**	**83**

Figure 21 Occurrence of 11 features in the group of better results.

Table 4 Result Comparisons of using 18 features and 11 features

Description	Total Feature	Training Set		Testing Set		Feature Number
-	-	Sensitivity	Specificity	Sensitivity	Specificity	-
Using total features	**18**	*90%*	*98%*	*75%*	*96%*	*18*
	11	*91%*	*97%*	*76%*	*96%*	*11*
Best training set	**18**	*95%*	*98%*	*75%*	*94%*	*13*
	11	*93%*	*98%*	*75%*	*95%*	*8*
Best testing set	**18**	*91%*	*97%*	*80%*	*95%*	*10*
	11	*90%*	*98%*	*78%*	*95%*	*10*

4.2 Discussions

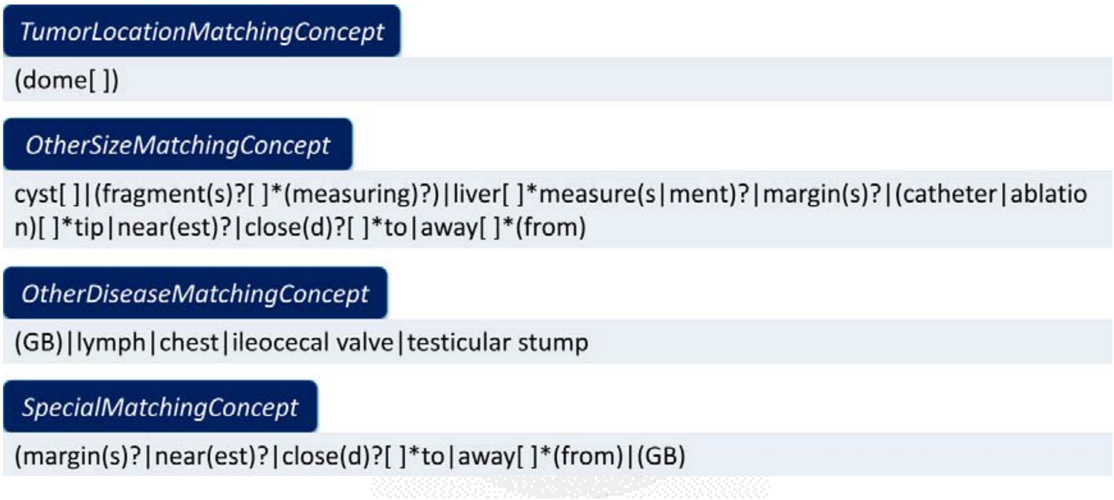
In this research, we find two factors that can affect the performance of the validation system on predicting the correctness of extracted results.

1. The modification of regular expression.
2. The combination of features.

The performance of the IE system differs from the performance of the validation system. The performance of the IE system is determined by whether the IE system has extracted the target information correctly. The performance of the validation system is determined by whether the validation system predicts the correctness of extracted information correctly. The performance of the IE system can affect the performance of the validation system.

In the study, the liver cancer patients' textual medical records are processed by the IE system to produce the extracted results. There are 9 concepts in the IE system and each concept has its own regular expression. Regular expression is the core of the IE system. They help the IE system to recognize the information hidden in the textual medical reports. The modification of regular expression can affect the performance of the IE system to extract information correctly. If the IE system has a better performance on extracting target information, the amount of wrong extracted results would decrease and it becomes more difficult for the validation system to find the wrong extracted results. Figure 22 shows the regular expression added in the human-labeling process. Compared to original version of regular expression, the wrong extracted results decrease in the IE system with updated version regular expression. For detailed information see Table 5.

The validation system uses the features of extracted results in training set, testing set and the selected feature pattern to predict the confidence score of extracted results. The features can affect the performance of the validation system. But the more features input into the validation system cannot promise we can get a better result. The combination of features can exhibit the difference between uncertain extracted results, wrong extracted results and correct extracted results and let the validation system recognize them. The validation system will try different combination of features to get the feature pattern which has better capability to predict the correctness of extracted results.



TumorLocationMatchingConcept

```
(dome[ ])
```

OtherSizeMatchingConcept

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

OtherDiseaseMatchingConcept

```
(GB)|lymph|chest|ileocecal valve|testicular stump
```

SpecialMatchingConcept

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

Figure 22 The Regular expression added in the manual reviewing process.

Table 5 The Effect of modifications of regular expression.

Original Version of Regular Expression			
Uncertain Extracted Results	**280**	*Wrong Extracted Results*	**140**
Extended Version of Regular Expression			
Uncertain Extracted Results	**280**	*Wrong Extracted Results*	**88**

Chapter 5 Conclusions and Future Works

5.1 Conclusions

Textual medical records about patients with liver cancer contain lots of medical related information. But these records are unstructured, domain-specific and the volumes of these records are too enormous to retrieve the information efficiently and correctly. Therefore, we are to provide a combination of the IE system and validation system to collect and review these extracted results from medical records.

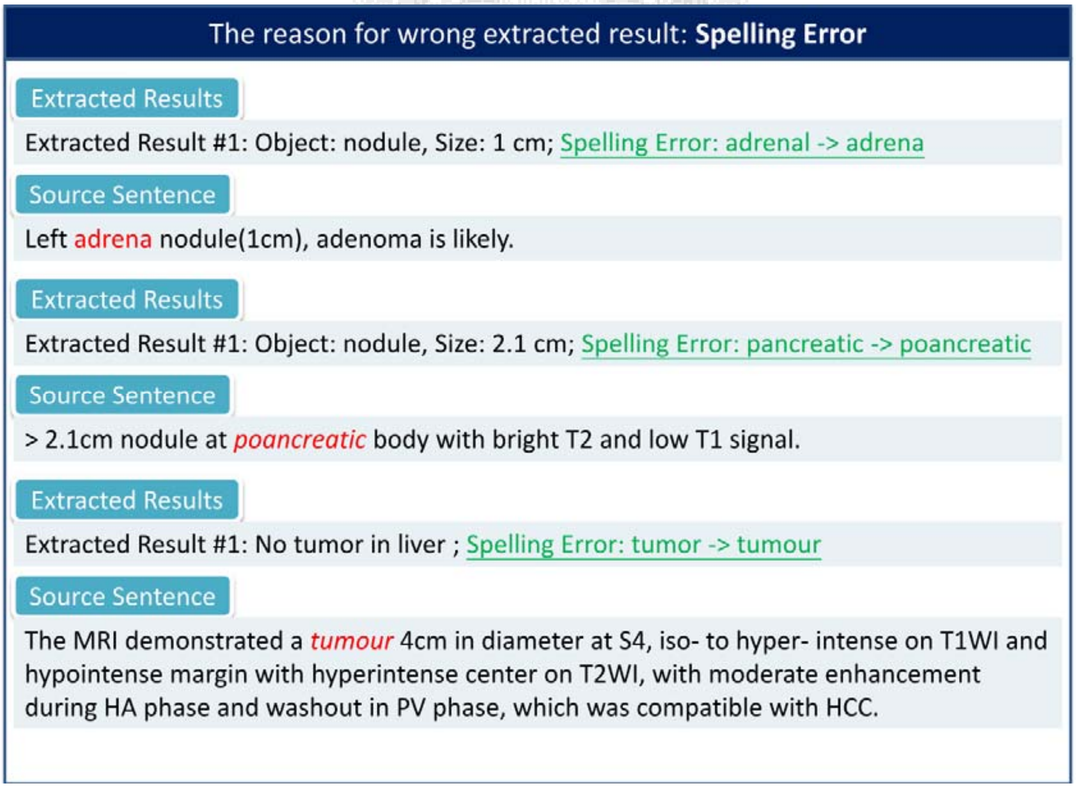
The data quality of extracted results is important to further medical application. But manual reviewing process is very inefficient. In this research, we propose the validation system to overcome this bottleneck. In the validation system, we will give a confidence score to each extracted result. The confidence score represents the status of each extracted result. If the extracted result was labeled with a lower confidence score, it would remind reviewers to check this extracted result. If the validation system has a higher accuracy for predicting the confidence score on each extracted result, it has higher probability to recognize the uncertain-extracted information or wrong-extracted information and can led the reviewers to check and correct them. The combination of the IE system and the validation system can provide a reliable way to access information hidden in the textual medical records for further application.

5.2 Future Works

The procedures from the IE system producing the extracted results and used the validation system to predict their correctness, there are 2 shortcomings should be improved in the future.

1. The spelling error in the textual medical records.
2. The other organs or other diseases information in the textual medical records.

First, the IE system cannot recognize the information spelled incorrectly and led to the consequence of wrong extracted results. This type of wrong extracted results cannot be correctly recognized by the validation system. We should propose a solution to identify and extract information correctly. Second, the information not defined in the regular expression cannot be recognized by the IE system. Therefore, the tumor information corresponding to other organs or other diseases was extracted incorrectly by the IE system. This type of wrong extracted results is also not recognized by the validation system. In order to let the IE system recognize other organs or other diseases information, we should add anatomy and disease terminology to the ontology.



The reason for wrong extracted result: **Spelling Error**

Extracted Results

Extracted Result #1: Object: nodule, Size: 1 cm; Spelling Error: adrenal -> adrena

Source Sentence

Left adrena nodule(1cm), adenoma is likely.

Extracted Results

Extracted Result #1: Object: nodule, Size: 2.1 cm; Spelling Error: pancreatic -> poancreatic

Source Sentence

> 2.1cm nodule at *poancreatic* body with bright T2 and low T1 signal.

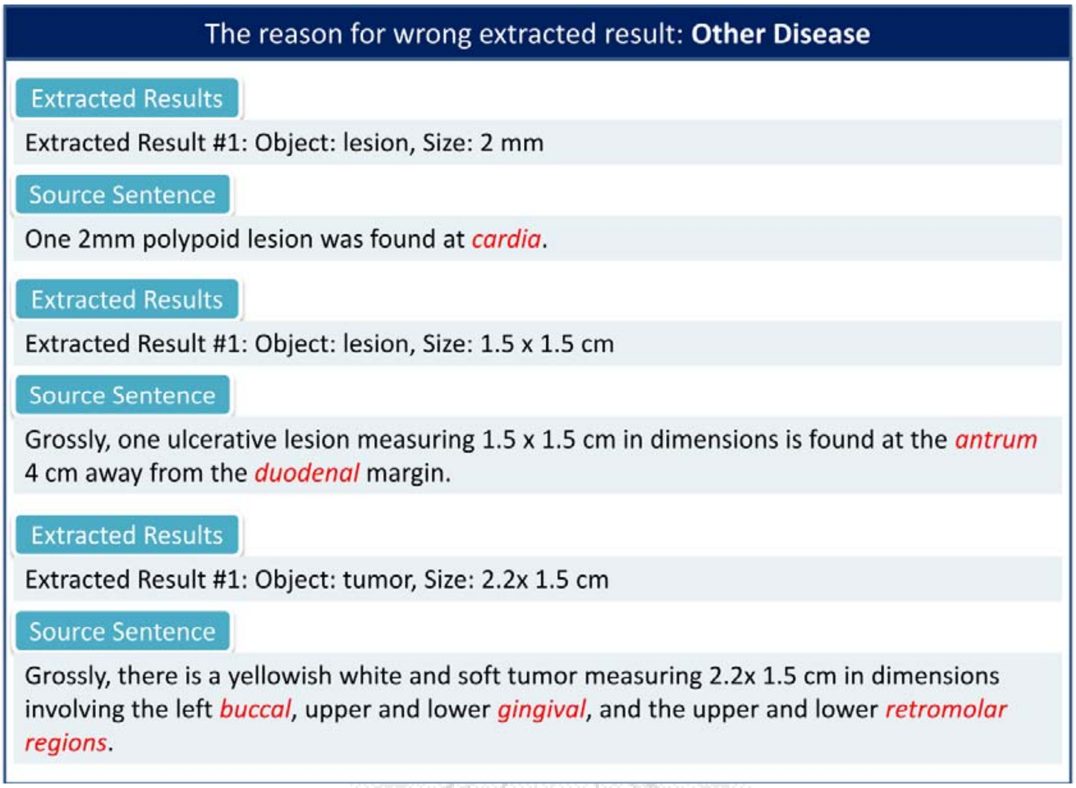
Extracted Results

Extracted Result #1: No tumor in liver ; Spelling Error: tumor -> tumour

Source Sentence

The MRI demonstrated a *tumour* 4cm in diameter at S4, iso- to hyper- intense on T1WI and hypointense margin with hyperintense center on T2WI, with moderate enhancement during HA phase and washout in PV phase, which was compatible with HCC.

Figure 23 Example the spelling error to cause the wrong extracted results.



The reason for wrong extracted result: **Other Disease**

Extracted Results

Extracted Result #1: Object: lesion, Size: 2 mm

Source Sentence

One 2mm polypoid lesion was found at *cardia*.

Extracted Results

Extracted Result #1: Object: lesion, Size: 1.5 x 1.5 cm

Source Sentence

Grossly, one ulcerative lesion measuring 1.5 x 1.5 cm in dimensions is found at the *antrum* 4 cm away from the *duodenal* margin.

Extracted Results

Extracted Result #1: Object: tumor, Size: 2.2x 1.5 cm

Source Sentence

Grossly, there is a yellowish white and soft tumor measuring 2.2x 1.5 cm in dimensions involving the left *buccal*, upper and lower *gingival*, and the upper and lower *retromolar regions*.

Figure 24 Example of the other disease to cause the wrong extracted results.

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. Hripcsak, G., et al., *Use of Natural Language Processing to Translate Clinical Information from a Database of 889,921 Chest Radiographic Reports1.* Radiology, 2002. **224**(1): p. 157-163.
4. Chapman, W.W. and K.B. Cohen, *Current issues in biomedical text mining and natural language processing.* Journal of Biomedical Informatics, 2009. **42**(5): p. 757-759.
5. 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.
6. 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.
7. Keith E. Stuart, M. and M. Melissa Conrad Stöppler. Available from: http://www.medicinenet.com/liver_cancer/article.htm.
8. 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.
9. Available from: http://www.faqs.org/health/topics/75/Liver-cancer.html.
10. Available from: http://cancerhelp.cancerresearchuk.org/type/liver-cancer/.
11. Wimalasuriya, D. and D. Dou, *Ontology-based information extraction: An introduction and a survey of current approaches.* Journal of Information Science, 2010. **36**(3): p. 306-323.
12. Mykowiecka, A. and M. Marciniak, *Domain model for medical information extraction-the lightmedont ontology*, M. Marciniak and A. Mykowiecka, Editors. 2009. p. 333-357.
13. Mykowiecka, A., M. Marciniak, and A. Kupść, *Rule-based information extraction from patients' clinical data.* Journal of Biomedical Informatics, 2009. **42**(5): p. 923-936.
14. Nassif, H., et al. *Information Extraction for Clinical Data Mining: A Mammography Case Study*. in *Data Mining Workshops, 2009. ICDMW '09. IEEE International Conference on*. 2009.
15. Boytcheva, S., et al., *Integrating Patient-Related Entities Using Hospital Information System Data and Automatic Analysis of Free Text Availability, Reliability and Security for Business, Enterprise and Health Information Systems*, A. Tjoa, et al., Editors. 2011, Springer Berlin / Heidelberg. p. 89-101.
16. Goyvaerts, J. 23 October 2011; Available from: http://www.regular-expressions.info/.

17. Spasic, I., et al., *Text mining and ontologies in biomedicine: Making sense of raw text.* Briefings in Bioinformatics, 2005. **6**(3): p. 239-251.

18. McGuinness, D.L. and F.v. Harmelen. Available from: http://www.w3.org/TR/owl-features/.

19. ; Protégé was developed by Stanford Center for Biomedical Informatics Research at the Stanford University School of Medicine.]. Available from: http://protege.stanford.edu/.

20. Chang, C.-C. and C.-J. Lin, *LIBSVM: A library for support vector machines.* ACM Trans. Intell. Syst. Technol., 2011. **2**(3): p. 1-27.