ONTOLOGY BASED TEXT MINING IN TURKISH RADIOLOGY REPORTS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

ONUR DENİZ

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

JANUARY 2012

Approval of the thesis:

ONTOLOGY BASED TEXT MINING IN TURKISH RADIOLOGY REPORTS

submitted by **ONUR DENİZ** in partial fulfillment of the requirements for the degree of **Master of Science in Computer Engineering Department, Middle East Technical University** by,

Prof. Dr. Canan Özgen
Dean, Graduate School of **Natural and Applied Sciences** ________________

Prof. Dr. Adnan Yazıcı
Head of Department, **Computer Engineering** ________________

Prof. Dr. Göktürk Üçoluk
Supervisor, **Computer Engineering Department, METU** ________________

Dr. Meltem Turhan Yöndem
Co-supervisor, **Faculty of Eng. and Natural Sciences, Sabanci University** ________________

Examining Committee Members:

Prof. Dr. İsmail Hakkı Toroslu
Computer Engineering, METU ________________

Prof. Dr. Göktürk Üçoluk
Computer Engineering, METU ________________

Dr. Meltem Turhan Yöndem
Faculty of Engineering and Natural Sciences, Sabanci University ________________

Asst. Prof. Dr. Pınar Şenkul
Computer Engineering, METU ________________

Dr. Onur Tolga Şehitoğlu
Computer Engineering, METU ________________

Date: ________________

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: ONUR DENİZ

Signature :

ABSTRACT

ONTOLOGY BASED TEXT MINING IN TURKISH RADIOLOGY REPORTS

Deniz, Onur
M.S., Department of Computer Engineering
Supervisor : Prof. Dr. Göktürk Üçoluk
Co-Supervisor : Dr. Meltem Turhan Yöndem

January 2012, 96 pages

Vast amount of radiology reports are produced in hospitals. Being in free text format and having errors due to rapid production, it continuously gets more complicated for radiologists and physicians to reach meaningful information. Though application of ontologies into biomedical text mining has gained increasing interest in recent years, less work has been offered for ontology based retrieval tasks in Turkish language. In this work, an information extraction and retrieval system based on SNOMED-CT ontology has been proposed for Turkish radiology reports. Main purpose of this work is to utilize semantic relations in ontology to improve precision and recall rates of search results in domain. Practical problems encountered such as spelling errors, segmentation and tokenization of unstructured medical reports has also been addressed during the work.

Keywords: information retrieval, text mining, ontology, information extraction, description logics

ÖZ

TÜRKÇE RADYOLOJİ RAPORLARINDA ONTOLOJİ TABANLI METİN
MAĞDENCİLİĞİ

Deniz, Onur
Yüksek Lisans, Bilgisayar Mühendisliği Bölümü
Tez Yöneticisi : Prof. Dr. Göktürk Üçoluk
Ortak Tez Yöneticisi : Dr. Meltem Turhan Yöndem

Aralık 2012, 96 sayfa

Hastanelerde yüksek miktarda radyoloji raporu yazılmaktadır. Serbest şekilde yazılmaları ve hızlı üretilmelerinden kaynaklı hatalar bulundurmaları, radyologların ve doktorların raporlardaki anlamlı bilgiye ulaşmalarını zorlaştırmaktadır. Biyomedikal metin işleme alanında varlık bilgisinin uygulanması son yıllarda ilgi çekiyor olsa da, Türkçe radyoloji raporlarından varlık bilgisine dayalı bilgi edinmeye dair çok fazla çalışma yapılmamıştır. Bu çalışmada SNOMED-CT ontolojisi kullanılarak Türkçe raporlardan bilgi çıkarımı ve bilgi edinimine dair metodlar önerilmiştir. Çalışmanın asıl amacı varlık bilgisinin anlamsal ilişkilerinden faydalanarak raporlar üzerinde yapılan arama sonuçlarındaki başarıyı arttırmaktır. Çalışma süresince karşılaşılan medikal alana has yazım hataları, kuralsız oluşturulmuş raporların kısımlarına, kelime veya ölçüm gibi parçalarına ayrılması gibi uygulamada karşılaşılan problemlere de çözümler aranmıştır.

Anahtar Kelimeler: bilgi edinimi, metin işleme, varlık bilgisi, bilgi çıkarımı, tanım mantıkları

To my family...

ACKNOWLEDGMENTS

I would like to thank my advisors, Dr. Meltem Turhan and Dr. Göktürk Üçoluk for their guidance and generous patience through this study.

I also want to thank my family for their unconditional love and support.

I would like to thank my friends Anıl Sınacı, Can Eroğul, Çelebi Kocair, Erdal Sivri, Gökdeniz Karadağ, Hande Çelikkanat, Selma Süloğlu and Sefa Kılıç for their friendship and great support during my study.

I would also like to thank International Health Terminology Standards Development Organization (IHTSDO) and U.S. National Library of Medicine (NLM) for the academic licenses of UMLS and SNOMED-CT.

TABLE OF CONTENTS

CHAPTERS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

CHAPTER 1

INTRODUCTION

Producing hundreds of medical recordings on daily basis, a vast amount of medical narratives are added to information system of a hospital continuously. For example, over 30000 medical reports are generated and included into medical information system of Hacettepe University Hospital approximately within 3 months.

Unfortunately, the generation process of reporting takes place on standard office applications or embedded editors of medical information systems which is lacking of sophisticated textual processing units. In fact, these applications apparently do not possess simple spell correction utilities.

Reporting types produced in medical information systems include but not limited to clinical notes, patient histories, radiology reports after/before reports of surgical operations etc. Among these medical reporting types, this work focused on radiology reports due to practical reasons.

Reports are mainly produced by dictation method: While inspecting the image, radiology specialist dictates what he sees on the image to a tape. Later, the dictation is translated to textual form by a secretary or an assistant quickly. The rush in this process unfortunately results in spelling errors in documents. Furthermore, the reports are written in free text manner without any exact template.

In respect to medical content, radiology reports show a dense nature. Each sentence in reports include some semantic medical entity, even if the sentence mentions a normality finding. Besides simple information contents which is represented by a terminological term or phrase, sentences may include complex information units as well. These complex information entities are expressed with simple ones and some grammatical structures.

The grammatical structures used in reports constitute a narrower subset of language and are easier to process grammatically[32]. This property of language used in medical narratives stems from the nature of domain : An implicit jargon has been developed among medical specialists. This fact can be observed obviously in N-gram analysis of radiology reports.

Having rich medical information content, radiology reports are useful resources for radiologists or other physician to use in their own academic works. Unfortunately, with rapid production of reports, it gets harder to find just useful ones in vast amount of reports. Searching capabilities of information systems in use, on the other hand, cannot fulfill the relevant request.

Search operations in medical information systems are currently carried out by classical information retrieval models, which are designed for general retrieval problems. However, content in radiology reports shows specific characteristics that a classic model cannot deal with. For example, classical models decreases significance of common words by weighting schema's or probabilistic methods. But, common words among the corpus, such as "karaciğer *(liver)*" or "lezyon *(lesion)*" are significant terms depending on the context of reports. Another example of insufficiency in classical models is synonyms: Medical experts frequently use synonyms for medical terms, which results in lower recall values for retrieval tasks.

In this work, we addressed such problems in order to fulfill expectations of radiology experts from a search engine developed for radiology reports. In order to deal with semantic problems of data such as relevancy in retrieval of synonyms and related concepts, we utilized a medical ontology.

Because of common spelling errors in medical content of data, our work also includes a medical-specific adjustment to common spell correction algorithms which exploits pronunciation similarities of medical terms in English and Turkish. Experiments showed significant improvement in precision values of corrected words.

Different types of tokens are used in medical reports such as dates, scientific numerical units etc. These numerical tokens shows a relative complex structure which cannot be tokenized by standard lexical analyzers. A recursive regular expression based tokenizer has been implemented in order to address lexical properties of radiology reports.

We proposed an ontology based retrieval model in order to fulfill information needs of radiologists: Instead of indexing whole report documents in a bag of words approach, we indexed semantic concepts and ontologically expanded (i.e. related concepts) concepts of sentences according to their polarity values in corresponding sentences. We incrementally tested our approach in a subset of radiology documents.

CHAPTER 2

BACKGROUND INFORMATION

2.1 Information Retrieval

Information retrieval (IR) is the process of representing, storing and accessing to information items providing easy access to information in which the user is interested [5]. Various information items can be addressed by retrieval processes : searching for documents, for information within documents, for meta-data about documents, across the World Wide Web.

Basically there are two types of user information needs in retrieval systems: Browsing and Retrieval [5]. When user has a general idea of the topic he is interested although he is not completely aware of what he is looking for, he is simply 'browsing' the documents instead of searching. Carrying out a literature survey about a topic is an example to this behaviour. In the 'retrieval' task on the other hand, the user usually knows his information need and formulates a regarding query to his information need. Asking the google about list of world records in swimming is an example to retrieval task. Although modern retrieval systems might attempt to combine these tasks, combination of retrieval and browsing is not yet a well defined approach.

Another categorization in retrieval systems is about domain. A vertical search engine focuses on a specific domain with the advantages of; increased precision due to limited scope, utilizing domain knowledge such as taxonomies, ontologies and supporting specific unique user tasks. Horizontal search engines, on the other hand, are general search engines (web search engines) used in daily life.

Documents in IR systems are frequently represented through a set of index terms or keywords which are obtained from document texts using various processes. Addition to this logical view of documents, main processes of retrieval systems are also common: Preprocessing of documents so that the documents are represented by normalized index terms, indexing document with an inverted file structure, acquiring the information need from user through an interface and formulating the query along the index structure, retrieving the relevant documents and ranking the documents in the degree of relevantness.

Before going into details on processes of information retrieval, a formal definition is given on IR model from [5].

Definition *An information retrieval model is a quadruple* $[\boldsymbol{D}, \boldsymbol{Q}, F, R(q_i, q_j)]$ *where*

1. $\boldsymbol{D}$ *is a set composed of logical views (or representations) for the documents in the collection*

2. $\boldsymbol{Q}$ *is a set composed of logical views (or representations) for the user information needs. Such representations are called queries.*

3. F *is a framework for modeling document representations, queries, and their relationships.*

4. $R(q_i, q_j)$ *is a ranking function which associates a real number with a query* $q_i \in \boldsymbol{Q}$ *and a document representation* $d_j \in \boldsymbol{D}$*. Such ranking defines an ordering among the documents with regard to the query* q_i*.*

2.1.1 Processes

Having the goal of searching and retrieving information to user, basic components of an information retrieval system barely changes even if different IR models applied since first appearance of an IR system.

Main processes of IR systems is mainly listed in literature as:

- Text operations such as lexical analysis, elimination of stop-words, stemming, spelling correction, selection of index terms and so on.

- Indexing documents for fast retrieval with inverted index, suffix trees and suffix arrays.

- Searching through documents.
- Ranking results with user needs and domain specifications.

Main processes are studied in [5, 20] in detail. Some of these related to our work are discussed shortly.

2.1.1.1 Preprocessing

Generally IR systems are fed on raw data such as free text documents, web sources in HTML. A logical transformation of raw data is therefore required resulting in representations of documents or queries. The logical representation of documents or queries is necessary due to further operations of IR systems, which are indexing, searching and ranking.

A key point of preprocessing in IR systems is that resulting representations of documents and queries should have same formalism in order to successfully search queries among documents.

Lexical Analysis

Process of converting a sequence of characters into a sequence of tokens is known as lexical analysis. The purpose of using lexical analysis in IR systems is to identify and classify tokens in documents and queries so that index terms are distinguished. Then index terms are directed to index or search operations within system.

Although tokenizing character streams just using space characters seems legit at first glance, a detailed lexical analysis is recommended in applications where tokenization and classification of tokens is crucial for domain. Even if the application is intended for general solutions like horizontal search engines, special types of characters or tokens such as digits, hyphens, case of letters and punctuation marks, should be considered carefully. Examples of these cases include but not limited to dates (numerical dates, textual dates -04 March 1985-), abbreviations.

In domain specific applications, on the other hand, lexical analysis becomes more of an issue depending on purpose of IR system on domain. In domains in which scientific measurements are significant a detailed lexical analysis has to be carried out in order to cover measurements in data.

Elimination of Stop-words

Articles, prepositions, conjunctions, adverbs and any other common words which are too frequent among documents are referred as *stop-words*. Such words are removed from candidates of index terms before any indexing or searching process.

Addition to not being good discriminators over documents, elimination of stop-words provides also a reduction in size of index structure.

Morphological Stemming

Users generally initiate queries composed of words each of which constitutes a specific variant of corresponding word. In other words, the users do not keep in mind that the words they are searching for may exist in documents with different variations. In order to overcome such cases, words are substituted with their stems through indexing or searching.

Choice of applying stemming in retrieval tasks is mostly language dependent. In languages involving significant number of suffixed words such as English and Turkish, stemming may improve the search results. In contrast, stemming in languages with little varied words such as Chinese is not effective.

Spelling correction

Spelling errors are common among the documents and queries in IR systems especially when documents are generated under loose conditions. In order to prevent inconsistency in result that stems from spelling errors, appropriate spelling correction solutions should be implemented.

The basic approach used in spelling correction is to find the correct spelling of a word in spelling dictionary by comparing words using a string distance algorithm.

Edit distance (Levenshtein distance) and Jaro-Winkler distance algorithms are the two of most commonly used distance metrics.

The former defines the distance between two strings by the minimum number of edits needed to transform one string to the other. Standard edit distance algorithm includes insertion, deletion or substitution operations of a single character. Distance between strings are calculated by a dynamic algorithm in which each operation applied to the string contributes to the total penalty. In the end, minimum penalty required to transform one string into the other is known as the distance between the two.

The latter, namely Jaro-Winkler distance, is defined as a variant of Jaro distance metric, which measures distance between two strings as weighted sum of matching characters and half number of transpositions. This metric is then increased with a scaled value of length of initial matching characters. This metric is designed and intended to compare person names.

2.1.1.2 Indexing and Searching

Indexing is basically building data structures over documents to provide faster search operations. Inverted files, suffix arrays and signature files are main indexing techniques widely used in literature.

An inverted file (or inverted index) is a mechanism that stores index terms and their *occurrences*. *Occurrences* of a word is a list of documents in which the word exists. Positions of terms can also be included in inverted file structure to provide proximity search.

Searching query terms on inverted index generally takes three steps [5]:

- In **Vocabulary search** step, the query is splitted into single words and words are searched among inverted indices.
- In **retrieval of occurrences** step, the list of all occurrences of all the words in query is retrieved
- In **Manipulation of occurrences** step, occurrences are processed to solve queries such as phrase queries, proximity queries, boolean queries.

2.1.1.3 Ranking

Initiating the query and retrieving documents from indices, each resulting document in this phase is given a score on basis how the document and the query are related. Then list of results is sorted using these scores so that most relevant documents stay on top of the list. The scoring mechanism is dependent on IR model where the relatedness between query and document is mostly measured with a similarity function.

2.1.2 Information Retrieval Models

IR system models are categorized into three with respect to their mathematical basis:

- In set-theoretic models, Standard Boolean, Extended Boolean and Fuzzy Set models, documents and queries are represented as set of index terms. Similarities between documents and queries are calculated by set operations.
- Algebraic models, Vector Space Model, Generalized Vector Space Model, Latent Semantic Index Model, represent documents and queries as vectors, matrices or tuples.
- Finally, examples of Probabilistic models used in IR systems include Probabilistic relevance model, Inference Network, Belief Network. Probabilistic models use probabilistic theorems like Baye's theorem to compute similarities between documents and queries.

Among these models, Standard Boolean Model, Vector Space Model and Probabilistic Model based on Baye's theorem are known as classic models of IR. Standard Boolean Model and Vector Space Model will be briefly explained after some basic definitions are presented.

Before jumping into definitions of these models, however, some basic definitions are to be presented.

Representing documents with index terms using either a set structure or a vector, assignments of numerical *weights* to terms in documents is required since each term in document has its own different contribution. In other words, different terms in documents have different importance in representation.

Definition *Let k_i be an index term, d_j be a document, and $w_{i,j} \geq 0$ be a weight associated with pair (k_i, d_j).*

Definition *Let t be the number of index terms in the system.* k_i *be a generic index term.* $K = \{k_1, ..., k_t\}$ *is the set of all index terms.* $w_{i,j} > 0$ *is the weight of index term* k_i *of a document* d_j. *If an index term does not appear in the document, then* $w_{i,j} = 0$. *Weights of index terms in a document* d_j *is represented by the vector* $\vec{d_j} = (w_{1,j}, w_{2,j}, ..., w_{t,j})$. *Let* g_i *be a function that returns the weight of index term* k_i *in any t-dimensional vector (i.e.* $g_i(\vec{d_j}) = w_{i,j}$*).*

In classic methods, index terms are assumed to be mutually independent and weights of terms are assigned accordingly. Obviously assuming mutual independence between terms is a simplification because terms in documents are mostly correlated. However, it significantly simplifies the indexing, searching and ranking processes in terms of computational power.

2.1.2.1 Standard Boolean Model

This model is based on set theory and Boolean algebra. Weights of terms in documents are assigned in boolean values according to appearance of term. Queries are boolean expressions which are composed of terms and three boolean connectives: *not, and, or*. Similarity between a query and a document in this model is also represented by a boolean menu. A document is either *relevant* to a query with scoring of binary value *1*, or *irrelevant* with binary value *0*.

Definition *For all index terms in all documents, weights of documents are all binary. i.e.* $w_{i,j} \in \{0, 1\} \quad s.t. \quad \forall_{k_i} \in K, \quad \forall_{d_j} \in \boldsymbol{D}$. *Query q is conventional boolean expression. Let* $\vec{q}_{dnf}$ *be the disjunctive normal form of query q. Let* $\vec{q}_{cc}$ *be any of the conjunctive components of* $\vec{q}_{dnf}$. *The similarity of a document* d_j *to the query q is defined as:*

$$sim(d_j, q) = \begin{cases} 1 & if \quad \exists_{\vec{q}_{cc}} \quad | \quad (\vec{q}_{cc} \in \vec{q}_{dnf}) \wedge (\forall_{k_i}, g_i(\vec{d_j}) = g_i(\vec{q}_{cc})) \\ 0 & otherwise \end{cases}$$

Despite its simplicity, this model does not enable partial matching. Another disadvantage is that since its similarity function scores over binary values, ranking of results is not possible.

2.1.2.2 Vector Space Models (VSM)

Documents and queries in VSM is represented as vectors of weighted terms. These term weights are used to compute similarity scores between documents and queries.

Definition *$w_{i,j}$ associated with a pair (k_i, d_j) is positive and non-binary. Terms in query are also weighted. Let $w_{i,q}$ be the weight of index term k_i, where $w_{i,q} \geq 0$. The query vector $\vec{q}$ is defined as $\vec{q} = \{w_{1,q}, w_{2,q}, ..., w_{t,q}\}$ where t is the total number of index terms in the system. A document d_j is still represented by $\vec{d_j} = \{w_{1,q}, w_{2,q}, ..., w_{t,q}\}$.*

Similarity value between two vectors, one for document and the other one for query, is calculated with *cosine of the angle* between vectors:

$$\begin{aligned} sim(d_j, q) &= \frac{\vec{d_j} \bullet \vec{q}}{|\vec{d_j}| \times |\vec{q}|} \\ &= \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{t} w_{i,q}^2}} \end{aligned}$$

Using *cosine of the angle* between vectors, similarity value varies from 0 to +1 enabling partial matches. But to compute rankings weights of index terms are to be specified.

In classical VSM method, weights of index terms in documents are assigned according to their frequencies. A frequent term in a document has a higher weight factor. The frequency of a term in the document is usually called as *tf factor*. On the other hand, terms which are frequent among all documents should not favored in weightings since those terms are not discriminative among the collection. An *inverse document frequency* or *idf factor* is defined to balance such terms so that weight of terms is normalized.

Definition *Let N be the total number of documents in the system and n_i be the number of documents in which the index term k_i appears. Let $freq_{i,j}$ be the raw frequency of term k_i in the document d_j (i.e. the number of occurrences of the term k_i in document d_j). Then the normalized frequency $f_{i,j}$ of term k_i in document d_j is given by*

$$f_{i,j} = \frac{freq_{i,j}}{max_l \quad freq_{l,j}}$$

where maximum frequency of term k_l is computed over all terms in document d_j. If term k_i does not exist in the document d_j then $f_{i,j} = 0$. Let idf_i, inverse document frequency for k_i, be given by

$$idf_i = f_{i,j} \times \log \frac{N}{n_i}$$

One of the advantages of VSM over boolean model is to support partial matches due to scalar values of similarity scoring. The other advantage is providing satisfactory ranking results with *tf-idf* schemes for general search applications.

2.1.3 Evaluation

Several methods can be used for evaluation of a retrieval system. In order to evaluate an IR model, a collection of documents and a set of query should be prepared for ground-truth of evaluation. Each document should be classified as relevant or irrelevant according to each query.

In this section, we will shortly introduce evaluation metrics that we used for this work.

2.1.3.1 Precision

Precision is the fraction of related and retrieved documents to the retrieved documents.

$$\text{Precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

2.1.3.2 Recall

Recall is the fraction of relevant and retrieved documents to the relevant documents.

$$\text{Recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

2.1.3.3 Average Precision and Mean Average Precision

Precision and recall measures the success of the system not considering the order of retrieved documents. However, in a "successful" IR system, the relevant documents expected to be ordered on top of the listings.

Average precision and mean average precision metrics fill this gap in IR evaluations. Average precision metrics gives the performance of a system for a single information need namely for a single query.

$$AvP = \frac{\sum_{k=1}^{n}(P(k) \times rel(k))}{|\text{relevant documents}|}$$

where $rel(k)$ where is a binary function having value 1 if the document k is relevant, 0 otherwise. In other words, it is the average precision value obtained for the set of top k documents existing after each relevant document is retrieved.

Mean average precision is than an averaged value of "average precision" values over a set of queries (Q).

$$MAP = \frac{\sum_{q=1}^{Q} AvP(q)}{Q}$$

2.2 Description Logic

This section covers a brief summary of Description Logic's (DLs): basic definitions, syntactic and semantic properties of Description Languages, reasoning problems of DL system and reasoning algorithms.

Definitions and explanations in the following notes are organized from [22, 4, 3]

2.2.1 Introduction

Description Logic's (DL) are a family of knowledge representation languages used to formally represent knowledge in an application domain. As the name "Description Logic's" implies, this representation is constructed by defining concepts of the domain (its terminology) and describing properties of individuals/objects in the domain (the world description) in terms of other concepts [4]

Unlike from their predecessors such as semantic networks and frames, DLs are equipped with a formal, logic-based semantics so that reasoning capabilities enable to infer implicit knowledge that is not explicitly stated in knowledge base [3]. [3, 22] provides an overview of historical development of DLs from semantic networks.

Being decidable fragments of first-order predicate logic, it is more efficient and practical to use DLs in order to represent domain knowledge. And yet, it is possible to define a translation between a DL and first-order predicate logic that preservers the semantics. As well as this relationship between DL and first-order predicate logic, relationships between DL and other formalisms such as semantic networks, frame systems, conceptual graphs and modal logics are surveyed in [29] in detail.

2.2.2 Basic Definitions

Description logic defines a knowledge base with two components: the *TBox* and the *ABox*. TBox, *terminological part* of knowledge base, provides vocabulary of the application domain by concept descriptions in which complex concepts and roles are described with atomic ones. ABox, on the other hand, is the *assertional part* of knowledge base which introduces named individuals and relationships of these individuals with concepts and roles.

Concepts and *roles* constitute together the vocabulary of application domain. *Concept names* (i.e., *atomic concepts*) and *role names* (i.e., *atomic roles*) are used to define concept descriptions with concept constructors. Concepts are simply sets of individuals, grouped in a semantic manner of the application domain. Roles are the relationships between individuals.

Along with the knowledge base, TBox and ABox, a DL system also offers *reasoning* services, which includes *satisfiability checking* and *subsumption testing* for terminology, *consistency* *checking* and *instance checking* for assertions. Satisfiability checking is to determine whether a description is satisfiable, that is description is not contradictory to the rest of the TBox. Another reasoning task for TBox is subsumption testing which determines whether a description is more general than another one. For ABox, on the other hand, consistency checking is to determine whether the set of assertions, ABox, is consistent with respect to the TBox. Given a query about relationships between concepts, roles and individuals, retrieving individuals satisfying the query is instance checking problem. Refer to [4] for details of these reasoning problems and more explanations.

2.2.3 A basic description language : $\mathcal{AL}$

Using *concept names* and *role names*, complex description of concepts are formed according to *concept constructors* and *role constructors*. $\mathcal{AL}$, attributive language, is introduced as a minimal language with limited capabilities [4, 30]. Starting with $\mathcal{AL}$, constructors such as concept negation (syblomized as $\mathcal{C}$), concept union ($\mathcal{C}$), number restrictions ($\mathcal{N}$) supplement additional expressivity to the languages. Meanwhile, names of description languages vary on these constructors that they provide. For example, $\mathcal{ALC}$ stands for *A*ttributive *L*anguage with *C*omplenets, and obtained from $\mathcal{AL}$ by adding complement constructor [3]. For details on naming conventions of description languages refer to [4, 2]

2.2.4 Syntax and Semantics of $\mathcal{AL}$

Formal definitions of syntax and semantics of constructors for the language $\mathcal{AL}$ is adopted from syntax and semantics for $\mathcal{ALC}$ given in [3].

> **Definition** ($\mathcal{AL}$ syntax) *Let N_C be a set of* concept names *and N_R be a set of* role names. *The set of* $\mathcal{AL}$-concept descriptions *is the smallest set such that*
>
> 1. $\top, \bot$, *every concept name $A \in N_C$ is an AL-concept description,*
> 2. *if C and D are $\mathcal{AL}$ -concept descriptions and $r \in N_R$, then $C \sqcap D$, $\forall R.C$, and $\exists R.C$ are $\mathcal{AL}$ -concept descriptions.*
>
> **Definition** ($\mathcal{AL}$ semantics). *An interpretation $\mathcal{I}=(\Delta^{\mathcal{I}}, \cdot^{\mathcal{I}})$ consists of a non-empty set $\Delta^{\mathcal{I}}$, called the domain of $\mathcal{I}$, and a function $\cdot^{\mathcal{I}}$ that maps every $\mathcal{AL}$-concept to a subset of $\Delta^{\mathcal{I}}$, and every role name to a subset of $\Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}}$ such that, for all $\mathcal{AL}$-concepts C, D and all role names R,*
>
> - $\top^{\mathcal{I}} = \Delta^{\mathcal{I}}, \bot^{\mathcal{I}} = \emptyset$,

- $(C \sqcap D)^{\mathcal{I}} = C^{\mathcal{I}} \cap D^{\mathcal{I}}$,
- $(\exists R.C)^{\mathcal{I}} = \{x \in \Delta^{\mathcal{I}} |$ *There is some* $y \in \Delta^{\mathcal{I}}$ *with* $\langle x, y\rangle \in R^{\mathcal{I}}$ *and* $y \in C^{\mathcal{I}}\}$,
- $(\forall R.C)^{\mathcal{I}} = \{x \in \Delta^{\mathcal{I}} |$ *For all* $y \in \Delta^{\mathcal{I}}$, *if* $\langle x, y\rangle \in R^{\mathcal{I}}$, *then* $y \in C^{\mathcal{I}}\}$.

We say that $C^{\mathcal{I}}(R^{\mathcal{I}})$ *is the extension of the concept C (role name R) in the interpretation* $\mathcal{I}$. *If* $x \in C^{\mathcal{I}}$, *then we say that x is an instance of C in* $\mathcal{I}$.

Definition *A* general concept inclusion *(GCI) is of the form* $C \sqsubseteq D$, *where C, D are* $\mathcal{AL}$*-concepts. A finite set of GCIs is called a TBox. An interpretation* $\mathcal{I}$ *is a model of a GCI* $C \sqsubseteq D$ *if* $C^{\mathcal{I}} \subseteq D^{\mathcal{I}}$; *I is a model of a Tbox* $\mathcal{T}$ *if it is a model of every GCI in* $\mathcal{T}$. *We use* $C \equiv D$ *as an abbreviation for the symmetrical pair of GCIs* $C \sqsubseteq D$ *and* $D \sqsubseteq C$.

Definition *An* assertional axiom *is of the form* $x : C$ *or* $(x, y) : R$, *where C is an* $\mathcal{AL} -$ *concept, R is a role name, and x and y are individual names. A finite set of assertional axioms is called an* ABox. *An interpretation* $\mathcal{I}$ *is a model of an assertional axiom* $x : C$ *if* $x^{\mathcal{I}} \in C^{\mathcal{I}}$, *and* $\mathcal{I}$ *is a model of an assertional axiom* $(x, y) : R$ *if* $\langle x^{\mathcal{I}}, y^{\mathcal{I}}\rangle \in R^{\mathcal{I}}$; $\mathcal{I}$ *is a model of an Abox* $\mathcal{A}$ *if it is a model of every axiom in* $\mathcal{A}$.

Definition *A* knowledge base *(KB) is a pair* $(\mathcal{T}, \mathcal{A})$ *if* $\mathcal{I}$ *is a model of* $\mathcal{T}$ *and* $\mathcal{I}$ *is a model of* $\mathcal{A}$.

$\mathcal{I} \models \mathcal{K}$ *(resp.* $\mathcal{I} \models \mathcal{T}, \mathcal{I} \models \mathcal{A}, \mathcal{I} \models a$*)* denotes that $\mathcal{I}$ is a model of a KB $\mathcal{K}$ (resp., TBox $\mathcal{T}$, ABox $\mathcal{A}$, axiom a).

2.2.5 Inferences

Besides describing concepts and individuals formally, with a proper knowledge representation system it should be possible to discover implicit facts about concepts and individuals from explicitly stated axioms. This feature is enabled by reasoning over TBox and/or ABox in Description Logics systems.

Examples of such inferences include instance checking (i.e. checking for an individual whether being a member of a concept), relation checking (i.e. checking whether a relation holds between concepts and/or individuals), subsumption (i.e. checking whether a concept (or an individual) is subclass of an another), consistency checking (i.e. checking consistency of given TBox or ABox) etc.

Formal definitions of inference problems with respect to a KB consisting of a TBox and an ABox is given in [3] as:

Definition *Given a KB* $\mathcal{K} = (\mathcal{T}, \mathcal{A})$, *where* $\mathcal{T}$ *is a TBox and* $\mathcal{A}$ *is an ABox,* $\mathcal{K}$ *is called* consistent *if it has a model. A concept C is called* satisfiable *with respect to* $\mathcal{K}$ *if there is a model* $\mathcal{I}$ *of* $\mathcal{K}$ *with* $C^{\mathcal{I}} \neq \emptyset$. *Such an interpretation is called* *a* model of C *with respect to* $\mathcal{K}$. *The concept* D subsumes *the concept* C *with respect to* $\mathcal{K}$ *(written* $\mathcal{K} \models C \sqsubseteq D$*) if* $C^{\mathcal{I}} \subseteq D^{\mathcal{I}}$ *holds for all models* $\mathcal{I}$ *of* $\mathcal{K}$. *Two concepts* C, D *are* equivalent *with respect to* $\mathcal{K}$ *(written* $\mathcal{K} \models C \equiv D$*) if they subsume each other with respect to* $\mathcal{K}$. *An individual* a *is an* instance of a concept C with respect to $\mathcal{K}$ *(written* $\mathcal{K} \models a : C$*) if* $a^{\mathcal{I}} \in C^{\mathcal{I}}$ *holds for all models of* $\mathcal{I}$ *of* $\mathcal{K}$. *A pair of individuals* (a, b) *is an* instance of a role name r with respect to $\mathcal{K}$ *(written* $\mathcal{K} \models (a, b) : r$*) if* $\langle a^{\mathcal{I}}, b^{\mathcal{I}} \rangle \in r^{\mathcal{I}}$ *holds for all models* $\mathcal{I}$ *of* $\mathcal{K}$.

Interested readers may refer to [4] for details on these reasoning problems, reasoning algorithms and more importantly reduction processes between these inference tasks.

CHAPTER 3

RELATED WORK

Developments in the field of knowledge representation has mostly focused on ontology based representation systems. With advancements in supportive tools, libraries and its theoretical background, utilization of ontology has been common in information extraction and semantic annotation systems as well [1, 25, 12].

Two primary approaches on semantic annotation systems are pattern-based and machine learning-based. Although machine learning algorithms mostly outperform pattern-based techniques, some rule based systems can perform as well as a machine learning system[27].

An automated disambiguation process of large-scale semantic-tagged content is proposed in [11]. First, labels given in taxonomy is spotted with exact string match. A representative sample of label-in-context data, small amount of manually and large amount of automatically generated meta-data, is then used to calculate distribution of terms for each node of taxonomy. With this distribution a taxonomy based algorithm is utilized in order to disambiguate the rest of the data. Similarity between a node-in-taxonomy and the candidate context is calculated using classic tf-idf scheme in cosine measure since it outperforms the Bayes probability measure. Although 82% accuracy is not such an improvement according to similar approaches, disambiguation in such a large corpora (264 million pages) is a major contribution.

Especially, automatic annotation and extraction of medical entities in documents has gained popularity because of its practical contributions to the field of medicine [11, 31, 6].

An ontology-based information extraction system for Turkish radiological reports is proposed in [32]. Extraction of named entities and relations between entities is accomplished by a rule-based system. In this supervised approach, template-rules are defined in terms of ontology entities and syntactic regular expressions by a domain expert beforehand. Besides forming the information model of extractions and describing relations between entities, ontology is also used for resolution of ambiguities in sentences. Moreover, a context mechanism is also used to discover missing entities in the sentences. One of the major contributions of this work is being the first Turkish information extraction system in medical. Precision (98%) and recall (93%) evaluations shows that it is also one of the most successful information retrieval systems in literature to our knowledge.

On the other hand, utilization of ontologies designed for specific systems instead of a common one brings interoperability and maintenance issues to the extraction problem of information management in medicine. Therefore, widely accepted knowledge representation systems such as SNOMED-CT has to be imported into medical information systems.

However, adaptation of such big ontologies into medical information systems brings scalability problems to be solved. One solution to such problems is to provide browsing tools to end-users: A hybrid categorization and browsing tool for SNOMED encoding system which is a combination of VSM and a regular expression pattern matcher is proposed in [28]. Regular expression pattern matcher part of the system refines the ranking of vector space part of the system. Although the results do not seem to improve the state of the art, the proposed method proves its reliability considering the larger data used in experiments.

Although SNOMED-CT is the most comprehensive medical terminology developed in DL formalism, it cannot be surely claimed to have a compliance with sound ontological principles [10, 7, 9]. Quantitative and qualitative analysis showed lack of organization and incompleteness of SNOMED-CT hierarchy based on the facts concepts having large number of children, absence of difference in the description between children and parents, using the IS-A relation with different meanings (IS-A overloading)[7].

Practices of ontology and semantic representation systems is not limited with only information extraction and annotation systems. There is solid contribution of semantic technologies in information retrieval systems as well. Different semantic approaches appears in literature: Besides the integration of semantic similarity and relatedness measures between terms in documents using graph structures of ontologies such as WordNet, UMLS into classical IR models [24, 15], embedding ontology structures directly in retrieval processes has also gained popularity lately [14, 21].

[33] proposes an information retrieval model which adapts the classic vector space model exploiting ontology-based inferences. The system basically uses inferencing mechanisms in order to expand KB beforehand. Instead of indexing terms in classical VSM, they used ontology instances and classes as indexing items in their model. Low-scale experimentation results presented in [33] surpasses classical VSM when their KB is mature enough. In order to annotate documents they used a semi-automatic annotation process with two basic heuristics: longest match of terms and taxonomic disambiguation.

An adaptation of the vector space model for ontology-based information retrieval [8] is an extended work of [33] . Proposed method is tested with large scale experiments and normalization in ranking aggregation is changed as proposed in [13].

A simple ontology based retrieval model is described in [23] where documents and queries are represented with sets of concepts. Similarity between query and documents is simply adjusted with sets of concepts in both query and document. Even such a simple semantic enhancement of similarity measure gives better precision-recall curve compared to classic models of VSM and LSI.

Unlike most of the existing semantic or ontology-based proposals to IR which uses bag of concepts approach, [26] assumes documents or queries as concept-descriptors in DL. In this approach a document (or query) is nothing but a concept described by annotated concepts within the document (or query) using the additional role 'indexed_by'. Then retrieval process inherently becomes a subsumption testing between document and query. Obtained results of the proposed approach is claimed to be promising even in the absence of any experimental result. Using DL descriptors in retrieval process provides the advantage of being precise in retrieved documents. The main disadvantage of this method is, however, for each given query subsumption testing should be repeated across the whole knowledge base and this repetition is obviously makes the approach unpractical for real-world applications.

CHAPTER 4

METHODOLOGY

4.1 Data and Lexicon Description

4.1.1 Data-set

In this work, a broad data-set of radiology reports provided by Department of Radiology in University of Hacettepe is used. Dataset includes 32305 reports with many different examination types, and many different titles. Examination type denotes which method of examination type is used for corresponding part of body. Although titles are seem to include examination types, reports having different examination types can have same titles as well. For example, reports in examination types “BT alt abdomen (kontrastlı) *(Computerized tomography of lower abdomen with contrast)*” and “BT alt abdomen (kontratsız) *(Computerized tomograhpy of lower abdomen without contrast)*” may have the same title “ÜST VE ALT ABDOMEN BT *(Computerized tomography of upper and lower abdomen*)9”.

Hundreds of new radiology reports are generated using office programs in hospitals on daily basis. Either doctors type reports themselves viewing examination images or medical secretaries/assistants transfer dictated reports into text. In either ways reports are created in hurry of free text format. This means no standardization through generation process of reports.

Reports are physically and semantically consisted of different sections which covers various phases of examination. Most commonly used sections in reports are *Clinical Information, Technique, Findings, Results*. Since there is no standardization in construction process, different tags can be used for same sections or different sections can be joined. Different section separators that we encountered in dataset are *clinic, Technique, findings, operation,* *suggestions, comment, result, endication, request of examination, projection protocol, clinical information, findings and operation, Technique and findings* (*klinik, teknik, bulgular, işlem, öneriler, yorum, sonuç, endikasyon, tetkiki isteyen, görüntüleme protokolü, klinik bilgi, bulgular ve işlem, teknik ve bulgular*).

Clinical information section of reports includes the relevant history of patients' regarding the examination: external reason for the examination like accidents, any suspicious findings that may to point a disorder etc.

The radiology procedure applied in examination is stated in Technique section in detail: quantity of medical substances used in procedure, exposition strategy of medical substances to patients etc.

Findings is the section in which the doctors describes observations on images. Information in this section is represented as full sentences. The information in this section includes but not limited to normalities and abnormalities in patients' examination, predictions of disorders, suggestions to clinician who requested radiologic examination, comparisons with older examinations. These and other information items are usually stated in similar orders, especially physical normalities and abnormalities.

Important information items given in *findigs* section is summarized in *result* section. Although scope of ímportanth́ere is upto the examining doctor's preferences and the special nature of the case, mainly similar information items are summarized in this section. Physical format of this section is either in regular sentences or term phrases without verbal expressions.

Without any standardization on section separating, we have accepted that a report is consisted of *title, examination type, clinical information, textual content, result*. *Textual content* refers to sections *Technique, findings, operation, suggestions, comment, endication, projection protocol, findings and operation, Technique and findings*. By joining these different sections into one, we did not expect to loose any information since we claim semantics of sections lies within the words.

The language used in reports shows significant similarity due to implicit jargon developed in radiology domain. N-gram analysis we applied on 4689 reports (highlighted results in Appendix-A) shows that similar sentence structures are used to express findings. In fact, it is possible to come across even exactly same sentences. Unfortunately a formal definition of a well-defined subset of language cannot be structured unless all of the textual corpus is processed. Even if all of the corpus is used to define the language used in reports, there is still possibility to encounter new data during run-time because of loose property of natural languages. However, we can make use of this unintentionally developed jargon for extraction of information.

4.1.2 Terminology and Lexicon

A list of medical stems, a Turkish-English medical dictionary, a list of Turkish words excluding misspellings in medicine and a list of English words in medicine is used in this work as terminologies and lexicon.

Words that cannot be analyzed with Zemberek Morphological tool from 4688 reports has been listed. Such a word is either medical term that does not exist in Zemberek's own lexicon or a misspelling, or a complete new medical term. A list of 3499 medical stems is formed manually by domain experts using these words that Zemberek library could not analyze.

These stems of medical terms have been queried to an online dictionary. Resulting web-site has been processed using a wrapper so that other terms in resulting page as well as queried term have been extracted. For misspelled words, suggestions of online dictionary have also been queried. Extracted Turkish-English entries are added to our dictionary base if they are labeled with ḿedicalín resulting page. 17240 terminological entries extracted from the dictionary web-site have been joined with medical terms that we obtained from terminological services of Tepe-Technology and a Turkish-English medical dictionary of over 70000 entries have been compiled.

A list of unique Turkish words has been formed using Turkish words in the dictionary, list of stems and morphologically valid words from data-set. This list is used as a base for correct-spelled Turkish terms. Although this list is said to be composed of correct-spelled Turkish terms, different spellings for same words is common due to medical domain. Detailed examination of this subject is covered in Section 4.2.3 spelling correction section.

A list of unique English words gathered from dictionary and terminological concept descriptions of SNOMED-CT has been formed as a base for correct-spelled English terms.

4.2 Preprocessing in Medical Textual Data

Although the language used in radiology reports shows significant similarity among the corpus due to practical reasons:the nature of reports brings many structural and syntactic errors which needs to be resolved before any extraction and retrieval processes.

A simple analysis on tokens in dataset shows the importance of a detailed preprocessing: Over 30000 reports, there are 4326078 tokens, of which 41211 are unique. Over these unique tokens, only 17360 are composed of solely alpha characters (i.e. `[A-Za-z]`). The rest of tokens, which is more than half of the tokens in dataset, includes any number of non-alpha characters (i.e. numeric `[0-9]`, and punctuation). 11283 of these 23851 non-alpha-including tokens are relatively easy to process and clean and of the form `<word><punc>(<word><punc>*)`. *(ex. "bronş-karina" , "büyüktür.Bilateral" , "teatoz,hepatomegali,pankreatosteatoz,myoma").* Rest of the non-alpha-including tokens (12568 of 23851) are unfortunately includes many types of tokens including numeric (numeric, scientific-unit, percentage, area, date etc.).

Analysis shows that text mining for medical documents requires a detailed and comprehensive preprocessing in order to avoid information loss. Preprocessing of radiology reports withing the scope of this thesis is designed and implemented in a pipe-lined manner: After sectioning the report into its parts, lexical analysis is performed in order to normalize and categorize tokens, and then sentence boundaries is detected using regular expressions. Stemming is essential for templates used in information extraction and spell-correction is significant since reports include typing errors and difference in terminology ecole of doctors. Translating terms from Turkish to English is essential since reports are in Turkish despite English SNOMED-CT.

4.2.1 Segmentation, Tokenization and Sentence Splitting

Despite the similarities in partition of reports, since reports are generated via unstructured environments (office programs), markers of report-parts vary with the user who generate the report. Case differences, number of space characters after/before tokens and different markers for same parts are observed along the dataset. We used regular expressions given in Appendix-B.2 in order to cope with these differences. Segmentation of reports to partitions are accomplished with regular expression equivalents of partition tags. Reports are segmented with regex-tag into corresponding partitions.

In order to successfully process the real medical data, a detailed tokenization process is needed since medical reports includes many non-alpha including tokens. Numerical information in reports constitute the majority of these non-alpha including tokens. Although some of these numerical extractions are negligible such as time measurements *(10-14. gün)*, length and area measurements *(17x12 mm, 20 m/sn2)*; numerical anatomic identifiers are considered in order to maximize recall in information extraction *(7-8-9. kosta, L1-2 düzeyi, 4. ventrikül)*.

A detailed lexical analysis was carried out on data set using UNIX command tools and regular expressions. Different types of tokens covering non-alpha tokens are determined. Starting from the smallest one, we developed regular expressions to match and extract different types of tokens 'substitutingly'. In other words after extracting regular expressions of basic types, more complex expressions are automatically generated in run-time. For example, in order to express area token we used the regular expression:

```
'numeric\s*(unit){0,1}(\s*[xX]\s*numeric\s*(unit){0,1})+'
```

where `numeric` and `unit` are substitutions for other regular expressions which may include other regular expressions as well.

A configurable tokenization and token labeling system is developed. Types of tokens, derived or manually developed regular expressions for token types and the order of implication of expressions are defined in a configuration file. Derived regular expressions are the expressions that substitutes other expressions. Given the configuration file and report to be processed, tokens in different types are extracted and labeled correspondingly. Extracted token types, derived or manually given regular expressions for these token types, and implication order of expression used in this work is given in Appendix-C

After tokenization and labelling, normalization of tokens such as clearing successive and redundant punctuation and spaces, separating words which are accidentally written combined with punctuation *(ex. aittir.Sol)* or uniting accidentally separated with punctuation *(ex. büy.ük)* are performed using regular expressions and word lexicon.

Another reason for developing a detailed tokenization application is to clear the text from noise characters which may disrupt sentence splitting process. Sentence splitting is performed with a single regular expression given in Appendix-B.1. Boundary of a sentence in radiology reports is either a sentence-ending punctuation followed by a upper case letter, or a Turkish `-dir` suffix followed by an ending punctuation (except the medical term “kontur”).

4.2.2 Stemming

We used Zemberek, an open-source morphological analyzer for Turkish. We have added 3499 terminological roots to the present lexicon of Zemberek in order to be able to morphologically analyze medical terms either.

In order to add medical roots to the lexicon of morphological analyzer, we made an assumption on medical terms: Since adjectives are not to get any inflectional suffixes in Turkish and medical terms are not to be verbs, we assumed that we do not loose any information if we accepted all medical terms as noun’s.

It is a common fact that using roots of terms increases efficiency of information retrieval tasks [5]. Although we did not make use of inflectional suffixes, derivational suffixes and roots are utilized to generate stems of words. Stem of a word is the base part of the word without any inflectional morphemes. We used stems of medical terms in both extraction and retrieval parts of our methodology since in both parts we heavily utilized information retrieval tasks.

Furthermore, we used Zemberek library in approximate detection of medical terms in dataset: Given a word from dataset, if the word is successfully analyzed with Zemberek tool, it is highly possible that the word is non-medical. (The possibility in this statement comes from that some medical terms are already embedded in Zemberek lexicon such as *alveol, apandisit, ülser, safra, akciğer*. In order to exclude this type of medical terms already exist in Zemberek lexicon, we manually constructed a list with such terms.) If the word cannot be analyzed with standard-Zemberek, but can be analyzed with Zemberek using lexicon enhanced with medical roots, the word is definitely a medical term. In other case, where the word cannot be analyzed even with Zemberek using medical-enhanced lexicon, if the word does not exist in unique Turkish words-list either, the word is either a misspelled word or a new term. We used this empirical method in analyzing spelling errors (Table 4.1, Table 4.2) and a slightly different version in spelling correction with pronunciation expansion. (Section 4.2.3.2)

4.2.3 Spelling correction

As we stated earlier, hundreds of radiology reports are constructed on a daily basis. Unfortunately, generating lots of free-text reports in limited time using regular office tools (such as Notepad or Word) without any spell-checker developed for domain results in spelling errors, mostly in word level.

Table 4.1 shows results of a morphological and terminological analysis on words of two separate datasets.

First words in datasets are analyzed with morphological analysis tool, Zemberek, with its default lexicon. A successful analysis indicates the word being a correctly spelled word. Default lexicon of Zemberek library does not include a comprehensive list of medical terminology as it is developed for general uses. Not covering a wide medical terminology, a morphological success using Zemberek with default lexicon therefore implies that this successfully analyzed word is highly probably a non-medical term.

If the morphological process with default lexicon of tool fails, then Zemberek with enhanced lexicon is used to determine a successful analysis. A success in this step shows that the word is a correctly spelled medical term, since enhancement in lexicon is carried out with medical roots.

With a fail in analysis with enhanced lexicon, the word is finally queried through Valid Turkish Words list. As we stated earlier, this word list is constructed with Turkish words in dictionaries (based on a trivial assumption on correctness of spellings in dictionaries), with correctly spelled words in dataset which we detected using morphological tool with enhanced lexicon. A successful match in this query denotes the word having medical property since most non-medical words are already filtered in first step. Not being successfully analyzed in previous step, namely with enhanced lexicon, results from only a limited list of medical roots is included in enhancement step. Therefore, the medical property of the word determined in this step comes from the dictionaries that we included in Valid Turkish Words list.

Finally, if the word cannot be successfully analyzed using Zemberek tool with any lexicon and also cannot be matched with a Valid Turkish Word in the list, the leftover word is treated as a misspelled one or a new medical term that does not exist in any of the lexicons or word lists we provided. We assume that existence of new medical words in leftovers is negligible because of limited domain of radiology. Therefore leftovers are classified as misspelled words.

Table 4.1: Morphological and terminological analysis on two separate datasets.

		30000set		4000set		
	sequenced process	Number	Ratio	Number	Ratio	category
1	default lexicon	2381755	0.72	390785	0.69	non-medical
2	enhanced lexicon	693229	0.21	124034	0.22	medical
3	valid turkish	122382	0.04	27736	0.05	medical
4	other	106148	0.03	22667	0.04	misspelled
	TOTAL	3303514	100	565222	100	

Examining results of this analysis, spelling error ratio appears to be nearly 3%. However this simple logic does not takes the difference between the number of non-medical words and medical words into account. It is obvious that significant information in medical texts mostly lies within medical words instead of non-medical words. Since nearly 70% of words are non-medical in dataset, this 3% ratio of misspelled words to whole dataset is not confident unless the ratio of medical words within misspelled words is also nearly 27%. Therefore, we need to calculate ratio of misspelled medical terms.

In order to calculate the ratio of medical and non-medical misspelled words, we picked 4 sets of 100 successive words in both datasets. We manually inspected medical and non-medical words in these sets. Table 4.2 gives the ratio of medical terms in misspelled words as approximately 92%.

Table 4.2: Categories of misspelled words.

	30000set		4000set	
set	medical	non-medical	medical	non-medical
1	89	11	97	3
2	89	11	83	17
3	93	7	94	6
4	89	11	97	3
AVERAGE	90	10	92.75	7.25

Generalizing 90-92.75% of medical words in misspelled words, numbers of medical words in misspelled words are estimated as 110143.8-25725.14 respectively. Table 4.3 show the calculations of ratio for correctness/misspellings for medical terms. Ratio of spelling errors within medical words is calculated as approximately 12-14% of which constitutes to a significant proportion.

Table 4.3: Spelling error ratios in medical words. Number of misspelled words are estimated using error rates found in Table 4.2

	30000set		4000set	
spellings	number	ratio	number	ratio
correct	815611	0.88	151770	0.85
misspelled(*)	110143.8	0.12	25725.14	0.14
TOTAL	925754.8	100	177495.14	100

It is also trivial to infer that spelling errors in data is mostly within medical words considering 3% of spelling error rate in general and 12-14% error rate in medical terms. Because of importance of medical terms regarding information quality, these spelling errors within medical words is not to be easily negligible. Therefore we need to handle spelling errors delicately.

In order to work out spelling errors' problem, we need to inspect error types in medical reports. Spelling errors in our medical data can be divide into two sets:

- Typing based errors
- Pronunciation based errors

Typing based errors: Spelling errors stemmed from typing are just typos such as switching successive letters, dropping a letter, adding a letter, mistyping a letter, repeating a phoneme etc.

Most common solution of this type of error is comparing misspelled word against a known list of correctly spelled words. Key point of this solution is to utilize the most suitable comparison metric between words in order to map misspelled words to correctly spelled ones. In our method, we experimented two popular string distances, Levenshtein and Jaro-Winkler

Table 4.4: Typing based spelling error examples, and correct versions

misspelled	correct
anevrizmasnın	anevrizmasının
değerlendiirlen	değerlendirilen
yerlişimlidir	yerleşimlidir
yuuşak	yumuşak
yuymuşak	yumuşak

distances, and chosen Levenshtein distance because of its success in evaluation.

Pronunciation based errors: Second type of spelling errors in Turkish medical reports are due to translational differences in medical terms. Most medical terms in Turkish are originated from English and Latin. Translations of these terms in Turkish are represented mostly with their pronunciations. For example translation of acromioclavicular term in Turkish is written and pronounced with "akromiyoklavikuler". Unfortunately pronunciational representation of medical terms in Turkish are not globally unique. In other words, for the same medical term in English, many pronunciational derivations of the same term in Turkish exists. The English term acromioclavicular is represented with terms "akromiyoklavikuler, akromioklavikuler, akromiyoklavikular, akromioklavikular,akromioklaviküler". This derivation problem is not only limited with practical behaviours of doctors or domain experts but also different dictionaries also have different derivations of same terms. In fact, it would be more convenient to name these errors as pronunciational derivations.

Soundex algorithm and its variants are seem to be an appropriate solution to this pronunciation based spelling errors as these algorithms are already designed for similar problems. Most suitable variant of Soundex algorithm for our problem is New York State Identification and Intelligence System Phonetic Code, shortly named as is NYSIIS. This phonetic algorithm simply changes the string according to its rules defined. Unfortunately, it does not seems possible to design and implement an accurate rule system for medical words. In other words, in medical words, pronuncational derivation rules are not as sharp as rules in NYSIIS coding. Therefore, we implemented a variation of this algorithm which applies its derivation rules with all possible variations. We call this Pronunciational Expansion.

Table 4.5: Pronunciation based spelling derivation examples, and English translations.

translation	pronunciational derivation
acromioclavicular	akromiyoklavikuler akromioklavikuler akromiyoklavikular akromioklavikular akromioklaviküler
angiography	anjiografi anjiyografi
diaphysial	diafizyal diyafizyal diafizyel diyafizyel
intramedullary	intramedüler intramedullar intramedüller

4.2.3.1 Pronunciational Expansion of Medical Words

Our inspections on errors in medical reports shows us using pronunciational derivations of medical words is common. Although most of the spelling errors can be handled with classic spell checking methods such as matching the nearest word from dictionary using edit distance, we observed that major part of errors cannot be handled because string distances between pronunciational derivations of medical words and original words remain out of confident threshold.

Aside from spelling correction problem of misspelled words, pronunciational expansion of medical words is also useful in text based information retrieval for Turkish medical data in which a common terminological naming convention has not been formed yet. For example, results of a textual query such as “akromioklavikuler eklemde” is meaningful only when other variants of the term are also queried.

Similar to NYSIIS coding, we implemented a method which generates derivations of a medical term using pronunciational variations. NYSIIS algorithm applies its rules with conditions resulting a single unique output for an input. Unfortunately, in our problem, it is not possible to define conditional rules for variations in medical terms since the application of pronuciational conversion rules are seem to be random. In other words it is unclear whether to apply a rule. Moreover, we need to expand the given word so that we can capture all possible derivations of the word. For these reasons, our method generates all possible derivations of a medical word by applying all possible combinations of matching variation rules.

Applying a variation rule to a word simply changes the matching substring of the word to the other substring of the rule. In other words a variation rule is defined by two substrings: one for domain and one for range. Since the expansion method generates pronunciational variant of the word, variation rules can run in both directions. If the ‘domain’ substring matches in the words, matching region is replaced by ‘range’ substring and vice versa. Variation rules that we used in our method are in Table 4.6

Table 4.6: Bi-directional variation rules used in expansion

domain	range
e	a i
eal	yal yel
f	ff
l	ll
my	miy
ny	niy
ia	iya iye
ie	iye
io	yo iyo
o	a ö
s	ss z
u	ü
yo	iyo

Since some rules have common characters/substring in their definition, overlapping is possible through extraction matched rule for a word. For such cases, non-overlapping sets of matched rules are detected. Since it is not possible to guess which rules to apply, all possible combinations for non-overlapping sets of rules are calculated and candidate words are generated. Then a successful query on Turkish valid words with the candidate denotes that the pronunciational derivation of the word is actually used in any of the valid terminologies in our lexicon.

However, generating derivations of longer words in which there are many variation rules match, this process takes unreasonable time for real-time applications because of its computational complexity and, more importantly, the queries to valid Turkish words list even with an inverted index.

In order to optimize this handicap, we apply an empirical limit to the number of rules possible for a word. This means that the method generates not all possible combinations but the combinations of rules with a maximum number of empirical limit. We believe that a limit to number of rules applied to a word does not change the result since it is not rational to generate a word with so many rules applied.

In Table 4.7 we have short examples for pronunciational expansion. Although example misspelled words in the table are already solvable by standard levensthein distance spell checker, we give those examples to be clear on pronunciational expansion. In table only bold generations are valid through our lexicon and all the other generations are meaningless words. As we explained before, validness or correctness of pronunciatonal generations are detected with queries to valid Turkish word list.

4.2.3.2 Spell Correction with Pronunciational Expansion

Decision on Spell Correction: Before attempting spell correction to a given word, it is obviously essential to decide on whether the given word is already correctly spelled or misspelled. Since nature of spelling correction is correcting something, application of spell correction to an already correctly spelled word will result an error in correction without nowhere.

Therefore a decision process has to be taken in run-time specifying whether a given word requires spell correction or not. Our correction method made this decision by a query to valid Turkish words list. A successful query denotes that given word does not require correction and vice versa.

Decision on Pronunciationally Expandibility: We already categorized spelling errors in medical reports into two types, namely typing based and pronunciation based errors. Spell

Table 4.7: All pronunciational expansions of relatively short misspelled medical words.

source	generations
diabet	deabat, deabet, deabit, deebat, deebet, deebit, deobat, deobet, deobit, deyabat, deyabet, deyabit, diabat, diabet, diabit, diebat, diebet, diebit, diobat, diobet, diobit, diyabat, **diyabet**, diyabit
lesyon	lasion, lassion, lassyan, lassyon, lassyön, lasyan, lasyon, lasyön, lazion, lazyan, lazyon, lazyön, lesion, lession, lessyan, lessyon, lessyön, lesyan, lesyon, lesyön, lezion, lezyan, **lezyon**, lezyön, lision, lission, lissyan, lissyon, lissyön, lisyan, lisyon, lisyön, lizion, lizyan, lizyon, lizyön, llasion, llassion, llassyan, llassyon, llassyön, llasyan, llasyon, llasyön, llazion, llazyan, llazyon, llazyön, llesion, llession, llessyan, llessyon, llessyön, llesyan, llesyon, llesyön, llezion, llezyan, llezyon, llezyön, llision, llission, llissyan, llissyon, llissyön, llisyan, llisyon, llisyön, llizion, llizyan, llizyon, llizyön
reverzibl	ravarsebl, ravarsebll, ravarsibl, ravarsibll, ravarzebl, ravarzebll, ravarzibl, ravarzibll, raversebl, raversebll, raversibl, raversibll, raverzebl, raverzebll, raverzibl, raverzibll, ravirsebl, ravirsebll, ravirsibl, ravirsibll, ravirzebl, ravirzebll, ravirzibl, ravirzibll, revarsebl, revarsebll, revarsibl, revarsibll, revarzebl, revarzebll, revarzibl, revarzibll, reversebl, reversebll, **reversibl**, reversibll, reverzebl, reverzebll, reverzibl, reverzibll, revirsebl, revirsebll, revirsibl, revirsibll, revirzebl, revirzebll, revirzibl, revirzibll, rivarsebl, rivarsebll, rivarsibl, rivarsibll, rivarzebl, rivarzebll, rivarzibl, rivarzibll, riversebl, riversebll, riversibl, riversibll, riverzebl, riverzebll, riverzibl, riverzibll, rivirsebl, rivirsebll, rivirsibl, rivirsibll, rivirzebl, rivirzebll, rivirzibl, rivirzibll

correction using Levenshtein Distance is an widely accepted solution to the former. And we proposed a pronunciational expansion solution for the latter one. However, although we can easily differentiate two types of errors, an automated spelling correction method needs to separate errors from themselves since an unnecessary pronunciational expansion process of a non-medical word will consume computational resources needlessly and also will increase error rate in spelling correction procedure.

Therefore we defined a function that decides whether a given word is an ‘pronunciationally expandable term’ or otherwise. Similar to detection of a word being a medical term or not, in order to decide whether a given word is ‘pronunciationally expandable term’ we used Zemberek library and medical stems that we mention in Section 4.1.2. For medical terms that are already listed in Zemberek’s own lexicon, we manually extracted such terms into an external list such as “alveol, anjiyografi, apandisit, lezyon, patoloji, tüberküloz”. Remind that medical words such as “akciğer, ülser” and “safra” are still medical words but not ‘pronunciationally expandable’ since they are natural Turkish words. To sum up, if a word cannot be analyzed by Zemberek with its default lexicon and can be analyzed by Zemberek with medical enhanced lexicon, then the word is an ‘pronunciationally expandable term’. If not, but if the word is in manually extracted medical terms list which are not natural Turkish ones, then again the word is said to be an ‘pronunciationally expandable term’.

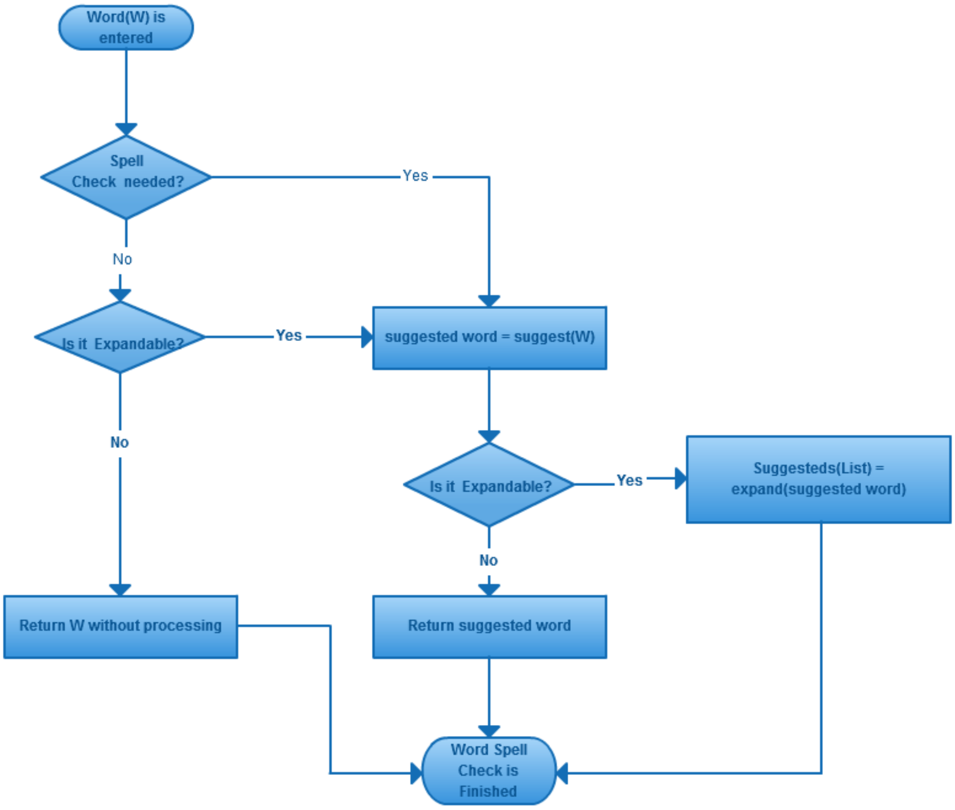
Word(W) is entered
Spell Check needed?
Yes
No
Is it Expandable?
Yes
suggested word = suggest(W)
No
Is it Expandable?
Yes
Suggesteds(List) = expand(suggested word)
No
Return W without processing
Return suggested word
Word Spell Check is Finished

Figure 4.1: Flow diagram for spell-correction

As given in Figure 4.1, if a given word is required to have spelling correction, the nearest word in lexicon according to given string distance metric is selected as a suggestion. Since various terminological resources in medical domain includes pronunciational variants of same medical terms, a pronunciational expansion should still be initiated despite correctness in spelling of given word. In other words the decision on pronunciationally expandibility should be taken independent from the decision on spelling correction. In either way, whether a given word needs spelling correction or not, pronunciational expansions of given word (in case of correctness) or the suggested word (in case of spelling correction) is calculated based on decision on pronunciationally expandibility.

4.3 Information Extraction in Turkish Radiology Reports using SNOMED-CT

In this section we explain our method that extracts information from Turkish textual reports as SNOMED-CT concept format. Being comprehensive and internationally accepted, we used SNOMED-CT terminology in order to represent information in reports. Our short-introductory survey about SNOMED-CT is included in Section 7.1.5. Detailed documentations [19, 17, 18, 16]. about the ontology is also available for further examinations.

As previously described radiology reports are rich information sources due to their professional purposes. Each sentence in reports contains information about the examination that is to artificially represented. Comprehensive coverage of SNOMED-CT contains many concepts used in radiology as in any other medicine sub-domain.

We believe words or phrases and their semantic meanings have the most important role in representation of information on such domains like medicine where the textual content shows a dense nature in terms of conceptual identities. Therefore we focused on medical words and phrases in our data, and map these textual formalisms into ontological concepts that SNOMED-CT provides. In other words we utilized medical words and phrases in radiology reports to capture ontological concepts in order to extract the information within sentences to some extent.

We are aware that we loose deeply formulated information within sentences with this shallow method in which only words and phrases used. By deeply formulated information, we refer to composite information which is formed with simple concepts and syntactic features of sentences. Since we mainly focused in ontological improvement in retrieval process, we implemented our extraction method so that it does not cover deep semantic formations of information for now.

Using words and phrases, however, will not make our extraction process simple as reports are in Turkish free-text format whereas description of ontology concepts is in English. Mapping free-text representations of information ontology into machine-readable concepts of an ontology is already an open problem even in English corpuses. In addition to working in a difficult domain, medicine, working with Turkish free-text data is another challenge that we dealt in our development.

Luckily, information and representation of information in medical textual data shows a significant similarity because of its universal semantics. In our extraction method we basically made use of this property of medical semantics. As we stated before, using medical words and phrases in sentences, we believe that it is possible to capture medical information to some extent. Furthermore, as a result of having universal semantics, information in medical domain is conceptually represented in a parallel way most of the time. In other words, semantics remains while the language changes.

Putting these assumptions together, our overall method is as follows: Verbal polarity for sentences are captured with morphological analysis on verbal phrases of sentences. Turkish medical words and phrases in corpus are detected by elimination of non-medical terms. Then, medical words and phrases, are translated into English using our medical dictionary that we explained in Section 4.1.2. Translated terms then queried in ontology descriptions for matching concepts. Finally, a refinement procedure using subsumption reasoning is used to select the most specific concepts that are used to represent semantics in sentence.

After explaining our information model that we used to represent information in reports, we will explain our method thoroughly: polarity detection,non-medical term elimination, dictionary based translation and finally concept finding and selection.

4.3.1 Information Model

Concepts of SNOMED-CT ontology is the main information entity that we used in our representation. Each report consists of subsections that we previously explained in Section 4.1.1. A subsection is divided into sentences, which we assume as semantically complete information tuples.

A sentence is represented with the polarity of sentence and a list of concepts, of each defining a semantic concept stated in sentence.

Polarity of a sentence defines whether the given conceptual entities in representation do exist or not. A sentence has either a positive or a negative polarity stating the existence of corresponding concepts in sentence. Polarity value of a sentence is detected by investigating its verbal phrase.

Concept list of a sentence, on the other hand, covers the medical information as SNOMED-CT concepts. As explained in Section 7.1.5, ontology of SNOMED-CT is organized in hierarchies. In our information model, only concepts categorized in particular hierarchies are used: Clinical finding, procedure, observable entity, body structure and qualifier value. Concepts in other hierarchies are disregarded because of their irrelevancies.

Table 4.8 shows examples for polarities of sentences and possible concepts lists extracted from sentences.

Table 4.8: Information model examples.

Sentence	Polarity	Concepts
Karaciğerde lezyon görülmemiştir. *(No lesion in liver has been detected)*	Negative	Liver structure; Lesion
Sağ böbrek üst polde 2x3mm'lik kortikal kist izlenmiştir. *(2x3 mm cortical cyst has been detected in upper pole of right kidmey)*	Positive	Upper pole, right kidney; Cortical cystic disease

For compound and complex sentences which have more than one verbal phrases, clauses are splitted from their verbal phrases and each clause of the sentence is assumed as an independent sentence having its own polarity value and concept list.

Finding and splitting of verbal phrases are also accomplished with our regular expression tokenizer.

4.3.2 Polarity Detection

Two types of polarities exist in sentences: Positive and Negative. A *positive* polarity value indicates the existence of corresponding semantic concept in sentence, whereas *negative* value shows the absence of concept.

Polarity values are recognized by simply looking at verbs of sentences: A negative marker on verbal phrases, `-ME/-MA`, or the verb `‘yoktur’`, `‘değildir’` means a *negative* polarity. Verbs that do not include negators gives the *positive* polarity to sentence.

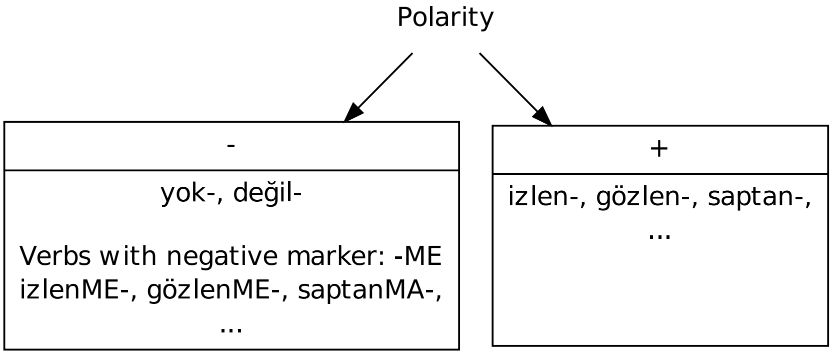
Polarity
-
yok-, değil-
Verbs with negative marker: -ME
izlenME-, gözlenME-, saptanMA-,
...
+
izlen-, gözlen-, saptan-,
...

Figure 4.2: Polarity values

4.3.3 Non-medical term elimination

Since we used only medical terms phrases in reports to extract conceptual information from reports, non-medical words and phrases are ignored during translation and extraction processes.

Two reasons for removing non-medical terms before translation and extraction phrases are to provide simplicity and to avoid possible errors in translation and extraction.

In order to find and remove non-medical phrases within sentences regular expression tokenizer is used. Unfortunately, regular expressions used to find these phrases are formed manually and this process is a tedious task. On the other hand, expressive power of regular expressions provides the tokenizer to find and label a wide range of phrases.

4.3.4 Dictionary Based Translation

After elimination of non-medical and verbal phrases, the sentence is now just a bag of words. In this phase of extraction process, Turkish medical phrases and words are translated to English using dictionary look-ups into our Turkish-English medical dictionary.

Successive look-ups into dictionary is required for translating words or phrases: In order to translate a single word, two look-ups has to be performed. One for the original morphological form of the word, and the other one for the stem of the word.

After an unsuccessful search with original form of the word, the look-up is repeated in case of the failure comes from inflectional suffixes. On the other hand, searching with the root of the word can result in ambiguous translation for the words having same roots with different derivational formation. For an example, the words "büyümektedir, büyümüş" and "büyük" in sentences; "karaciğer parankimi büyümektedir *(parenkima of liver is growing)*", "büyümüş lenf nodları vardır *(enlarged lymph nodes is detected)*", "sağ böbrek normalden büyük olarak gözlenmiştir *(right liver is observed larger than normal)*" have the same morphological root, büyü-, whereas the meanings and therefore translations for all are different: *growing, enlarged, larger*.

Since translating phrases or words requires successive look-ups into dictionary with over 70000 tuples, we implemented an inverted index on our dictionary with Lucene. In fact, two separate indices is constructed for both morphological forms of the words: original and stem.

In case of a failure in dictionary translation, pronunciation based translation is tried. By pronunciation based translation, we mean to convert a Turkish medical term into English by applying pronunciational rules which are similar to the Table 4.6. Pronunciational translation seems legit in case of a failure in dictionary based translation. Reasons for a failure in dictionary are either misspelling in words or a new term that dictionary does not include. Solution of to former is discussed in Section 4.2.3.2. For the latter, however, pronunciational translation solves the problem to some extent.

Application of pronuncational translation successes only if one of the translated candidate words exist in our valid English word list.

4.3.5 Concept Finding and Selection

To our knowledge an association between SNOMED-CT concepts and Turkish medical descriptions is absent in literature. Therefore, a translation is required in order to map medical terms in Turkish radiology reports into SNOMED-CT concepts. As stated in [17] each SNOMED-CT concept has terminological descriptions written in natural language.

After translation of medical words in sentences into English, now the problem is to find corresponding concept in SNOMED-CT. Since we have concept descriptions and English translations in hand, it seems legit to think that it would not be difficult to find correct concept in ontology. However, descriptions in SNOMED-CT may not be unique for concepts. Even if a unique concept-description mapping exist, since we are searching all of the English translations in sentence, number of searching results will be more than required.

In order to deal with ambiguity in text based search results, we applied intuitive refinements using SNOMED-CT ontology: Eliminating and filtering the concepts using relations of ontology.

First, concepts under the unrelated hierarchies such as Organism, Staging and scales, Physical force, Physical object etc. are eliminated from the results because of irrelevancies of those concepts to radiology reports.

Second, redundant concepts are filtered with reasoning.If one of the two concept in results is subclass of the other, subsumer concept is redundant since representation of subsumed concept already inherits properties of subsumer. Therefore, we applied subsumption reasoning among the results in order to find and filter subsumer concepts. In other words, most specific concepts in the results are selected to represent information in sentence.

For example, translation of “sağ böbrek üst polde krotikal kist” yields the terms *adrenal cortical cyst dextro kidney kidneys pole right safe superior supra upper*.

Unique concepts obtained from query of translations among descriptions of SNOMED-CT is given in Table 4.9.

Elimination of concepts under unrelated hierarchies or concepts gives us the Table 4.10, and visualized on sub-graphs of SNOMED-CT in Figures 4.3, Figure 4.4, Figure 4.5, Figure 4.6 with nodes having both descriptions and identifiers. Note that Nodes without identifiers are placed for clearance. Leaf nodes in subgraphs are most specific concepts and are selected for representation.

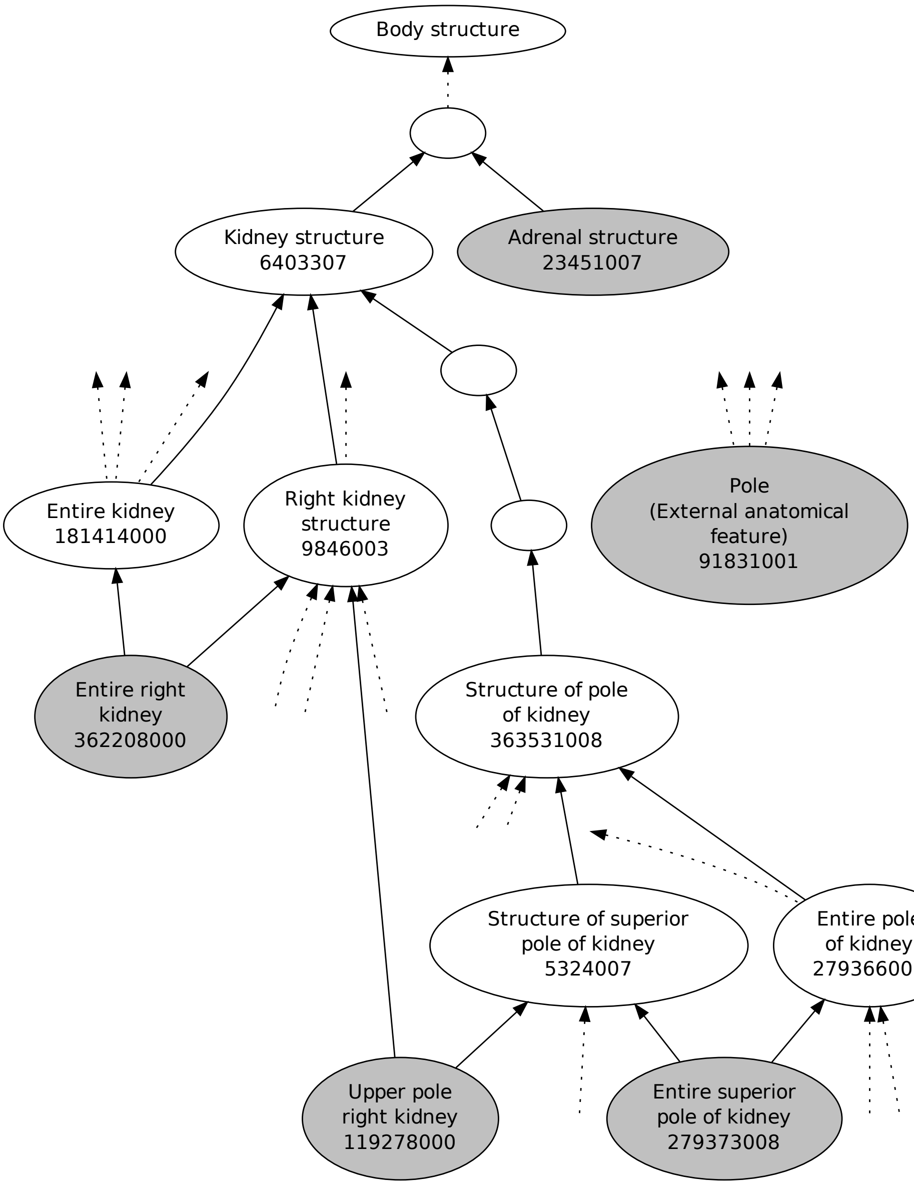
Body structure
Kidney structure 6403307
Adrenal structure 23451007
Entire kidney 181414000
Right kidney structure 9846003
Pole (External anatomical feature) 91831001
Entire right kidney 362208000
Structure of pole of kidney 363531008
Structure of superior pole of kidney 5324007
Entire pole of kidney 279366003
Upper pole right kidney 119278000
Entire superior pole of kidney 279373008

Figure 4.3: Subgraph of SNOMED-CT visualizing concepts under *Body Structure* hierarchy

Table 4.9: Unique concepts from SNOMED-CT with *adrenal cortical cyst dextro kidney kidneys pole right safe superior supra upper* searched in descriptions

Concept-ID	Description
5324007	Structure of superior pole of kidney (body structure)
9846003	Right kidney structure (body structure)
12494005	Cyst (morphologic abnormality)
23451007	Adrenal structure (body structure)
24028007	Right (qualifier value)
30171000	Adrenal disease
43014004	Superior (modifier) (qualifier value)
64033007	Kidney structure (body structure)
77945009	Cyst of kidney
90708001	Kidney disease (disorder)
91818006	Pole (external anatomical feature) (body structure)
103552005	Cyst
119278000	Upper pole, right kidney (body structure)
133846001	Cyst (morphologic abnormality)
156973002	Cystic kidney disease
181414000	Entire kidney (body structure)
236439005	Cystic disease of kidney (disorder)
237784000	Adrenal cyst (disorder)
253881008	Cortical cystic disease (disorder)
261183002	Upper (qualifier value)
264217000	Superior (qualifier value)
264515009	Cyst
264746006	To the right (qualifier value)
266625006	Cyst - kidney
268332003	Cystic kidney disease
279366003	Entire pole of kidney (body structure)
279373008	Entire superior pole of kidney (body structure)
352730000	Supra- (qualifier value)
362208000	Entire right kidney (body structure)
363531008	Structure of pole of kidney (body structure)
367643001	Cyst (morphologic abnormality)
441457006	Cyst (disorder)

Table 4.10: After elimination of unrelated concepts

Concept-ID	Description
5324007	Structure of superior pole of kidney (body structure)
9846003	Right kidney structure (body structure)
23451007	Adrenal structure (body structure)
24028007	Right (qualifier value)
30171000	Adrenal disease
64033007	Kidney structure (body structure)
77945009	Cyst of kidney
90708001	Kidney disease (disorder)
91818006	Pole (external anatomical feature) (body structure)
119278000	Upper pole, right kidney (body structure)
181414000	Entire kidney (body structure)
236439005	Cystic disease of kidney (disorder)
237784000	Adrenal cyst (disorder)
253881008	Cortical cystic disease (disorder)
261183002	Upper (qualifier value)
264217000	Superior (qualifier value)
279366003	Entire pole of kidney (body structure)
279373008	Entire superior pole of kidney (body structure)
362208000	Entire right kidney (body structure)
363531008	Structure of pole of kidney (body structure)
367643001	Cyst (morphologic abnormality)
441457006	Cyst (disorder)

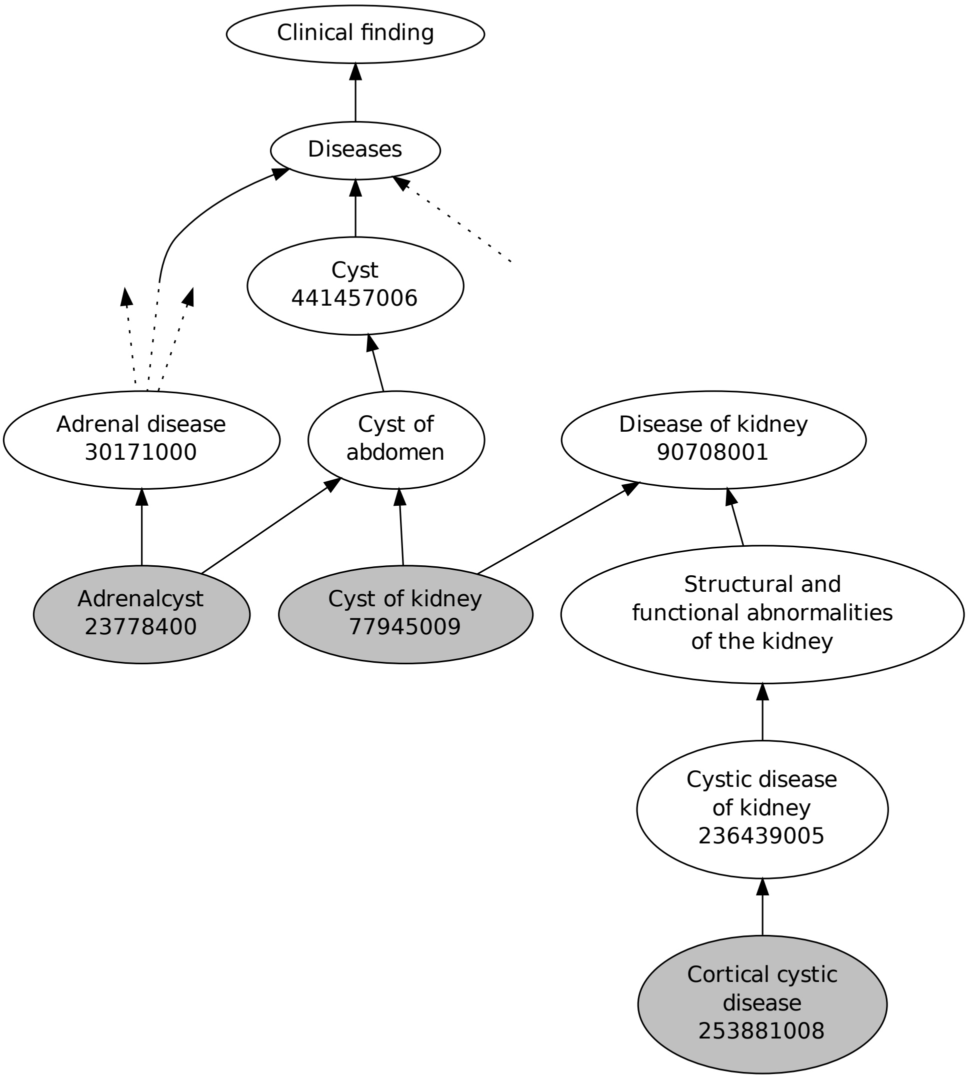

Figure 4.4: Subgraph of SNOMED-CT visualizing concepts under *Disorder* hierarchy

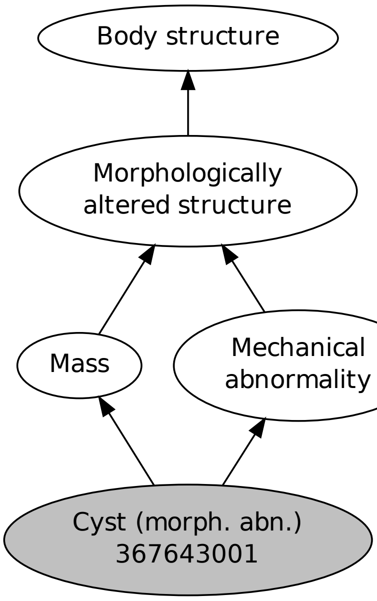
Body structure
Morphologically altered structure
Mass
Mechanical abnormality
Cyst (morph. abn.) 367643001

Figure 4.5: Subgraph of SNOMED-CT visualizing concepts under *Morphological abnormality* hierarchy

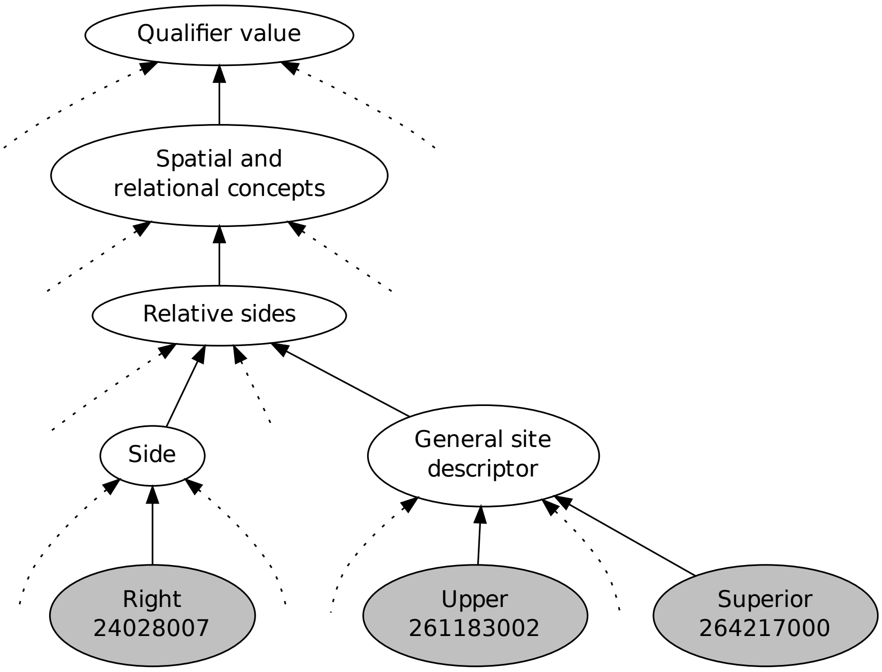

Figure 4.6: Subgraph of SNOMED-CT visualizing concepts under *Qualifier value* hierarchy

Table 4.11: Most specific concepts selected for representation

Concept-ID	Description
23451007	Adrenal structure (body structure)
24028007	Right (qualifier value)
77945009	Cyst of kidney
91818006	Pole (external anatomical feature) (body structure)
119278000	Upper pole, right kidney (body structure)
237784000	Adrenal cyst (disorder)
253881008	Cortical cystic disease (disorder)
261183002	Upper (qualifier value)
264217000	Superior (qualifier value)
279373008	Entire superior pole of kidney (body structure)
362208000	Entire right kidney (body structure)
367643001	Cyst (morphologic abnormality)

4.4 Information Retrieval using Ontology

Although classic IR models such as VSM offer good results on general retrieval tasks, it is not rational to expect same quality in domain specific systems. Thus, retrieval systems designed for special tasks should include further adjustments in order to achieve satisfactory results. Such kind of improvements include but not limited to domain specific operations, sophisticated ranking algorithms with user feedback, complex document representation templates etc.

Our interviews with radiology specialists made us to realize that an expert retrieval system on medical domain should exploit semantic relations between medical terms. As an ontology we chose SNOMED-CT because of its wide medical coverage, its international reputation and finally its theoretical background on a commonly accepted formalism such as Description Logic.

Throughout our analysis and implementation, we noticed structural specifications of radiology reports and how to exploit of these features for a successful retrieval task.

In this section we explain our original method which make use of domain information, a medical ontology, and specific structure of radiology reports in order to increase user satisfaction from retrieval tasks.

Since the method includes improvements on different layers of an standard retrieval system, it will be presented in an incremental way. In other words starting with Lucene implementation as a base model, in each subsection a different adjustment of the proposed method will be explained.

4.4.1 Vector Space Model and Boolean Model : Lucene

Throughout our work, we used Lucene tool as our indexing and searching engine. In this level of implementation, reports are not through in any process. In other words, textual content from reports are given to Lucene engine in a raw format.

Indexing, searching and ranking operations are handled with its standard implementation. Being a combination of VSM and Standard Boolean Model, it provides a variety of query types such as fuzzy query, proximity query and phrase query along with boolean operations.

We choose this model as a base standard for comparison since it is the model used in IR tasks of Turkish medical applications to our knowledge.

4.4.1.1 Document Representation

Document representation in this model is list of terms just as in standard VSM.

4.4.2 Turkish Medical Spell correction and Stemmer extension

In this phase, preprocessing in word level is added to IR model.

As we explained in Section 4.2.3, Turkish medical textual data includes significant amount of spelling errors in medical terms. Correcting misspelled words in documents and queries should improve retrieval results accordingly.

Furthermore, being an agglutinative language, Turkish words in medical reports show a different distribution of morphological formation. In order to catch relate index terms in query and document with different morphological forms in terms of inflectional suffixes, stemming operation is required both in indexing phase and searching phase. For example, a free-text query of “karaciğer lezyonu *(lesion of liver)*” should return the document in which the sentence of “Karaciğerde lezyon görülmüştür *(Lesion is detected in liver)*” occurs. Without a proper stemming operation on both query and the document, retrieval process will naturally fail.

As a result, improvements of this model are spell correction for Turkish medical words and stemming on words.

4.4.2.1 Document Representation

Document representation in this model is list of terms just as in standard VSM. But terms are preprocessed so that spelling errors are corrected and the words are index with their form of stem.

4.4.3 Sentence based Representation and Polarity extension

Because of the procedural examinations, sentences in reports have both negative and positive polarity values. In other words, even if the finding of a body structure is natural, the normality of corresponding feature is still expressed by radiologist. For example, “Karaciğerde lezyon izlenmemiştir. *(No lesion is detected in liver.)*”. However, these kind of sentences with negative polarity in which normality is expressed are still indexed in classic IR models with default document representation of list of terms. With a query of “karaciğerde lezyon *(liver in lesion)*”, which is asking for positive polarity, the results with negative polarity are obviously irrelevant.

In order to deal with such kind of scenarios, document representation is updated so that positive and negative polarity values are handled differently.

4.4.3.1 Document Representation

Instead of list of terms, documents in this IR model are represented by a list of sentences. Each sentence has two fields, one for textual content and the other for polarity value.

Polarity value of a sentence is detected as given in Section 4.3.2. And textual content of the sentence is just index terms of sentence.

4.4.4 Concept Expansion with Relations

With this final modification in our IR model, we aimed to exploit semantic relations between medical concepts.

First advantage of using an semantic knowledge representation in retrieval task is to relate synonym terms to the same concept in ontology. For example "yağlanma *(fattening)*" and "steatosis *(medical term for fattening)*" are synonyms used in reports for the term concept of fatty degeneration. Another example is that "calculus" and "stone" terms in English are synonyms for the same concept in SNOMED-CT. Despite the semantic uniqueness of concepts in ontology, there is still a need for accurate mapping from textual representations into concept descriptions.

Another advantage of ontology utilization in information systems is exploiting inter-conceptual relations. This usage of semantic structure enables information systems to infer implicit knowledge in data. For example, Figure 4.7 presents a sub-graph of IS-A relations from SNOMED-CT. The sub-graph shows hierarchical existence of some of the morphological abnormalities. Calculus and Cyst are Mechanical abnormality. Cyst, Nodule, Soft mass and Proliferative mass are subclasses of Mass. Similarly, Proliferative mass is subclass of Proliferation too. Lesion is the most general morphological abnormality among these concepts.

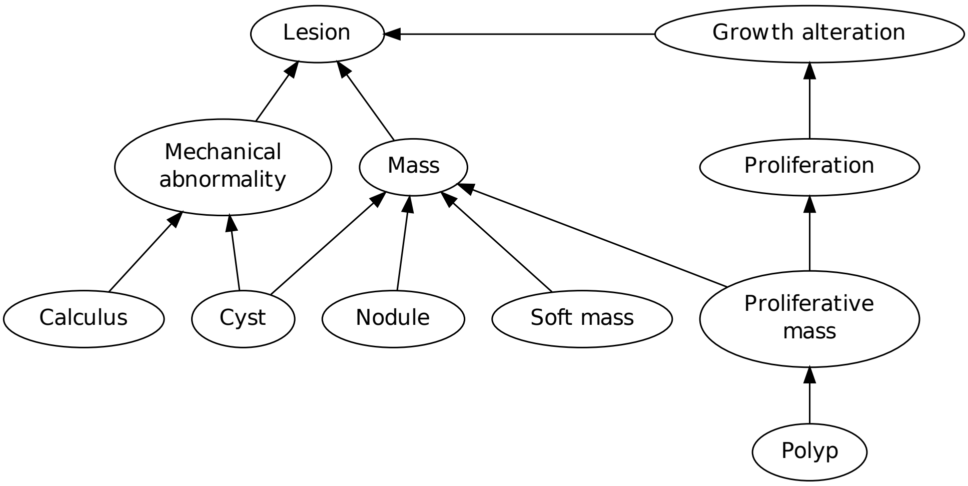

Figure 4.7: Morphological Abnormality Subgraph

According to this taxonomic classification of those medical concepts, a medical specialist querying for “karaciğer lezyonu *(lesion of liver)*” may also be interested with results for sub-classes of Lesion, which are all the other concepts in the sub-graph. In fact, results of a query of “karaciğerde kist *(Cyst in liver)*” should not be limited to concept of Cyst only since the doctor can be interested with similar concepts too.

Generally this task of inferring implicit information from data is accomplished with a reasoner over the ontology. Unfortunately, being computationally expensive, reasoning over the whole corpus online is an un-practical solution for real world applications.

Therefore, we propose a semantic indexing algorithm which partially preserves the structure of the ontology embedded in index structure. The idea of this technique is to expand the conceptual terms in documents or queries with related concepts which fill the range field of defining relationships.

4.4.4.1 Document Representation

Representation of documents in this phase are parallel to information model explained in Section 4.3.1: Each radiology report is represented by a set of sentences each of which has a polarity value and an expanded list of SNOMED-CT concepts.

4.4.4.2 Expansion method

Either extracted automatically with method explained in Section 4.3 or manually, concepts in sentences and queries are expanded in a recursive manner with its related concepts. By related concepts, we mean concepts in range property of defining relationships.

Expansion of a concept should continue with higher levels. For example, in order to enable retrieval of a document with concept of “Polyp” using a query with concepts of “Lesion”, concepts in document has to be expanded in such a way that concept of “Lesion” should be included as well. Nevertheless, there is not any obvious way of cutting expansion procedure at a level with superficial properties of concepts such as inbound or outbound relations. Thus, expansion of a concept continues until the root concept has been reached in our implementation.

Expanding concepts upto root concept without any scoring, on the other hand, will result in inaccuracy of retrieved results. Significance of a concept expanded should decrease in manner of its distance to original concept. In other words, the more distant candidate concept is, the less significant it should be.

We provide this alternation in significance of concepts with a variable we called "Expansion factor". Harmonic series are used to determine expansion factor of concepts depending on their distance from the starting concept:

Definition Let $\boldsymbol{P}$ be all paths from concept C, to the root concept R. Let L be the shortest distance from C to R. Expansion factor, e_C is given by L. Let $p \in P$ is any path from C to R. Expansion factor of any concept $c_i \in p$ is then calculated by $(L + 1)/(n_i + 1)$ where n_i is shortest distance from C to c_i

We will give an example on calculation of expansion factors for documents having concepts from sub-graph in Figure 4.7. For simplicity, we will this subgraph as a full ontology, and "Lesion" as the root of ontology. Assume example collection has 6 documents(sentences) with concepts given below:

$$D_1 = \{polyp\} \quad D_2 = \{polyp\} \quad D_3 = \{cyst\} \quad D_4 = \{cyst\} \quad D_5 = \{nodule\} \quad D_6 = \{calculus\}$$

Table 4.12: All paths from concepts in documents to root

Document	Paths to root
D_1, D_2	{polyp, proliferative mass, mass, lesion}
	{polyp, proliferative mass, proliferation, growth alteration, lesion}
D_3, D_4	{cyst, mechanical abnormality, lesion}
	{cyst, mass, lesion}
D_5	{nodule, mass, lesion}
D_6	{calculus, mechanical abnormality, lesion}

Expanding concepts, documents will be indexed with their expanded concept lists (as in Table 4.14). While preserving main concepts significance with harmonically highest expansion factor, related concepts are also included to document as index terms. Thus, expanded documents now will match with queries including related concepts.

Table 4.13: Minimum distances of concepts along paths

Document	$n_i + 1$ for c_i in p
D_1, D_2	polyp=1, proliferative mass=2, mass=3, proliferation=3, growth alteration=4, lesion=4
D_3, D_4	cyst=1, mechanical abnormality=2, mass=2, lesion=3
D_5	nodule=1, mass=2, lesion=3
D_6	calculus=1, mechanical abnormality=2, lesion=3

Table 4.14: Expansion of concepts with e_i

Document	Expanded list of concepts
D_1, D_2	$\{4 \times polyp,\ 2 \times proliferative\ mass,\ mass,\ proliferation,$ $growth\ alteration,\ lesion\}$
D_3, D_4	$\{3 \times cyst,\ 2 \times mechanical\ abnormality,\ 2 \times mass,\ lesion\}$
D_5	$\{3 \times nodule,\ 2 \times mass,\ lesion\}$
D_6	$\{3 \times calculus,\ 2 \times mechanical\ abnormality,\ lesion\}$

For example, given the query $Q = \{nodule\}$, same expansion procedure will apply to the query and expanded query will be $Q_{exp} = \{3 \times nodule,\ 2 \times mass,\ lesion\}$.

When the query initiated, D_5 will be the first result because of its equality. After that, as expected the documents $D_3 and D_4$ will share the second results because of their common parent concept $mass$ with concept $nodule$.

Since $tf - idf$ weighting schema disfavour the term common in all (or many) documents, the concept of $lesion$ in this example (which we assumed as root concept) has no effect in scoring because of 0 value in idf score. In other words, $tf - idf$ schema conserves the fact that the more general expanded concept in document is, the less significant it becomes in scoring. That means, general concepts in ontology will have less importance in scoring because they occur in many documents and further they have less expansion factor because of its distance to the starting concept.

For simplicity, the example we illustrated here is based on single-concept sentences. In case of multiple concepts, the procedure will apply with no difference.

CHAPTER 5

EXPERIMENTS

5.1 Spelling Correction with Pronunciational Expansion

5.1.1 Test Data

A list of misspelled words and correct version is constructed by medical experts from 4689 medical reports. This list includes 4314 misspelled-corrected word mappings. 274 entries of this list is eliminated since they include punctuation marks mostly meaning regarding spelling error involves multiple words, which is out of scope for a word spell corrector we assumed.

5.1.2 Evaluation

Evaluation of spelling correction module has been performed on two common string distances, namely Levenshtein Distance and Jaro Winkler Distance. Our proposed post-processing methodology, Pronunciational Expansion is applied to output of spelling correction methods of these distance metrics. In other words, evaluations include success of common correction algorithms and then additional success of proposed post-processing.

Abbreviations and naming convention used throughout the evaluation are descried in tables below.

For example, in the first experiment in Table 5.5 LR is applied as data exclusion function. 148 entries of data set ,of which 147 are derivable, are excluded in this experiment. Therefore 844 non-derivable and 3048 derivable entries used in experiment. Contain comparison function is used for result comparison. Pronunciational expansions of suggested words that

Table 5.1: Spelling Correction methods used in evaluation.

Symbol/Name	Explanation
SD	is default spelling correction using common string distance metrics, Levenshtein and Jaro Winkler distances.
PE	is abbreviation for Pronunciational Expansion method.

Table 5.2: Input/Output of correction methods

Symbol/Name	Explanation
Left	represents misspelled word in testing data.
Right/Ground	are used for correct versions of regarding misspelled words.
Suggested	is the word that an SD suggests without post-processing.
Suggested List	is the list of pronunciational derivations for the Suggested.

default spelling correction algorithm suggests are compared to ground words with contain comparison. Default spelling algorithm, which is Levenshtein Distance based spell checking, suggests corrected-words with success of 77.6% in non-derivable words, 64.9% in derivable words and 67.9% in all included data. Pronunciational expansion post-processing gives an additional success of 16.3% to derivable words and therefore a 12.7% to all words. Since PE is applied to derivable words during post-processing, additional success of PE in non-derivable is not expected.

Success results of experiments in which CC and SWG are used as result comparison are nearly the same, whereas success of experiments use EC are much different. Similar success results of CC-SWC and different results of EC is a natural effect of Turkish suffixes. Since we did not apply morphological processing in spell correction, different inflectional derivations of same words are evaluated as mismatch. Although comparisons with CC or SWC catch some of these inflectional derivations in where one of the words is directly inflectional derivation of the other such as *anjiografiler, anjiografilerinde*, between words having same root but completely different inflectional suffixes such as *anjiografiyle, anjiyografiler* CC or SWC will also evaluate to a mismatch. We assume that evaluations using SWC will be more realistic. However, we still included results with CC and EC in order to show inflectional nature of Turkish words, even in spelling correction, has a major effect in results.

Table 5.3: Data separation

Symbol/Name	Explanation
excluded	data is the subset of test data which is ignored according to exclusion metric.
included	data is the subset of test data used in regarding experiment. Included and excluded sets are complementary.
non/non-derivable	words are natural Turkish words in included set, to which pronunciational expansion is not applied.
derivable	words are terminological terms which are not natural Turkish. Post-processing of pronunciatonal expansion is applied to such words.

We excluded testing data in which Left and Right are equal in all of the experiments as these entries are not to be misspelled and correct version. Such entries are included in testing data by medical experts because of case differences between misspelled and correct words. Inclusion of such test data in experiments would definitely increase the results of spelling correction but then the results would be unrealistic since spell correction is based on a list of correct words.

We observed in our experiments that suggested word of a default correction is equal to misspelled word, namely left in our naming convention. Default correction algorithm simply looks for nearest suggested word for a misspelled in list of correctly spelled words. Since our list of correct words has been constructed using various medical dictionaries, different spellings of same medical words are included in this list. However, medical experts who composed testing data of spelling correction used only one of these spellings. As a result such kind of entries in data, which are LSE in our naming convention, are seem to be misspelled entries in evaluation. Although they are not to be assumed as misspelled, different spelling conventions in medical words are handled as if they were misspelled. Exclusion of such kind of data in experiments resulted in increase of success of default correction algorithm and overall algorithm whereas success of PE decreases since most of the data excluded are derivable medical terms which are to solved by PE. SCN is similar to LSE because of its lexicon based decision on whether the given word needs a spelling correction or not. Results of SCN is almost the same as LSE. Success rates of PE in LR experiments is significantly higher than LSE and SCN experiments. This difference in results comes from PE ability of handling different spelling corrections in medical words.

Table 5.4: Functions used in evaluation.

Symbol/Name	Explanation
PEE	denotes the evaluation function used for pronunciational expansion results. It compares given Right/Ground and Suggested List for a match.
PRC	compares expansions given in suggested list with ground Ground using corresponding RC.
PWG	compares expansions of expansion given in suggested list with Ground using corresponding RC.
EX	denotes exclusion metric used in experiments.
LR	excludes data when Left is Equal to Right.
LSE	excludes data when LR or Left is Suggested to Right.
SCN	excludes data when LR or LSE or spell check is not needed.
RC	is abbreviation for result comparison. An RC function compares given Left and Right/Ground words for a match.
CC	gives a match if any of the given strings contains the other
SWC	gives a match if any of the given strings starts with the other
EC	gives a match if any of the given strings is exactly equal to the other

To sum up, experiments show us that PE post-processing applied to default spelling correction algorithms provide significant additional success in overall system, because of different spelling conventions in medical words. Although classic correction methods using Levenshtein and Jaro Winkler distances have acceptable success rate in non-derivable words which are also natural Turkish words, a different approach in derivable words is needed. Our PE post-processing handles such kind of spelling conventions in medical words and contribute to success of total correction system with a success rate between 4.6% - 12.7%.

Table 5.5: Spell correction results using Levenshtein Distance and Pronunciational expansion.

				number of filtered data				success results over data								
	methods used			excluded		included		non-derivable			derivable			all		
exp	PEE	EX	RC	non	der	non	der	SD	PE	total	SD	PE	total	SD	PE	total
1	PRC	LR	CC	1	147	844	3048	77.6	0.0	77.6	64.9	16.3	81.2	67.6	12.7	80.4
2	PRC	LR	SWC	1	147	844	3048	77.3	0.0	77.3	64.8	16.3	81.1	67.5	12.7	80.3
3	PRC	LR	EC	1	147	844	3048	71.6	0.0	71.6	58.6	14.4	73.1	61.5	11.3	72.8
4	PRC	LSE	CC	4	767	841	2428	77.8	0.0	77.8	77.4	8.1	85.6	77.5	6.0	83.6
5	PRC	LSE	SWC	4	767	841	2428	77.6	0.0	77.6	77.3	8.1	85.5	77.4	6.0	83.4
6	PRC	LSE	EC	4	767	841	2428	71.9	0.0	71.9	73.6	6.2	79.9	73.2	4.6	77.8
7	PRC	SCN	CC	4	769	841	2426	77.8	0.0	77.8	77.4	8.1	85.6	77.5	6.0	83.6
8	PRC	SCN	SWC	4	769	841	2426	77.6	0.0	77.6	77.3	8.1	85.5	77.4	6.0	83.5
9	PRC	SCN	EC	4	769	841	2426	71.9	0.0	71.9	73.7	6.2	79.9	73.2	4.6	77.8
10	PWG	LR	CC	1	147	844	3048	77.6	1.1	78.7	64.9	16.7	81.6	67.6	13.3	81.0
11	PWG	LR	SWC	1	147	844	3048	77.3	0.8	78.1	64.8	16.6	81.4	67.5	13.2	80.7
12	PWG	LR	EC	1	147	844	3048	71.6	0.4	72.1	58.6	14.5	73.2	61.5	11.5	73.0
13	PWG	LSE	CC	4	767	841	2428	77.8	1.1	79.0	77.4	8.5	85.9	77.5	6.6	84.2
14	PWG	LSE	SWC	4	767	841	2428	77.6	0.8	78.4	77.3	8.4	85.7	77.4	6.4	83.9
15	PWG	LSE	EC	4	767	841	2428	71.9	0.4	72.4	73.6	6.2	79.9	73.2	4.7	78.0
16	PWG	SCN	CC	4	769	841	2426	77.8	1.1	79.0	77.4	8.5	86.0	77.5	6.6	84.2
17	PWG	SCN	SWC	4	769	841	2426	77.6	0.8	78.4	77.3	8.4	85.8	77.4	6.4	83.9
18	PWG	SCN	EC	4	769	841	2426	71.9	0.4	72.4	73.7	6.2	79.9	73.2	4.7	78.0

Table 5.6: Spell correction results using Jaro Winkler Distance and Pronunciational expansion.

	methods used			number of filtered data				success results over data								
				excluded		included		non-derivable			derivable			all		
exp	PEE	EX	RC	non	der	non	der	SD	PE	total	SD	PE	total	SD	PE	total
19	PRC	LR	CC	1	147	830	3062	72.4	0.0	72.4	57.4	18.7	76.1	60.6	14.7	75.3
20	PRC	LR	SWC	1	147	830	3062	72.0	0.0	72.0	57.4	18.7	76.1	60.5	14.7	75.2
21	PRC	LR	EC	1	147	830	3062	64.9	0.0	64.9	50.0	14.0	64.1	53.2	11.0	64.3
22	PRC	LSE	CC	4	767	827	2442	72.6	0.0	72.6	68.0	11.2	79.2	69.2	8.3	77.6
23	PRC	LSE	SWC	4	767	827	2442	72.3	0.0	72.3	67.9	11.2	79.1	69.0	8.3	77.4
24	PRC	LSE	EC	4	767	827	2442	65.1	0.0	65.1	62.8	5.7	68.5	63.4	4.3	67.7
25	PRC	SCN	CC	4	769	827	2440	72.6	0.0	72.6	68.0	11.2	79.3	69.2	8.3	77.6
26	PRC	SCN	SWC	4	769	827	2440	72.3	0.0	72.3	67.9	11.2	79.2	69.0	8.3	77.4
27	PRC	SCN	EC	4	769	827	2440	65.1	0.0	65.1	62.8	5.7	68.6	63.4	4.3	67.7
28	PWG	LR	CC	1	147	830	3062	72.4	2.0	74.4	57.4	19.4	76.9	60.6	15.7	76.4
29	PWG	LR	SWC	1	147	830	3062	72.0	1.5	73.6	57.4	19.3	76.8	60.5	15.5	76.1
30	PWG	LR	EC	1	147	830	3062	64.9	0.6	65.5	50.0	14.1	64.2	53.2	11.2	64.5
31	PWG	LSE	CC	4	767	827	2442	72.6	2.0	74.7	68.0	12.0	80.0	69.2	9.5	78.7
32	PWG	LSE	SWC	4	767	827	2442	72.3	1.5	73.8	67.9	11.9	79.8	69.0	9.2	78.3
33	PWG	LSE	EC	4	767	827	2442	65.1	0.6	65.7	62.8	5.8	68.6	63.4	4.4	67.9
34	PWG	SCN	CC	4	769	827	2440	72.6	2.0	74.7	68.0	12.0	80.1	69.2	9.5	78.7
35	PWG	SCN	SWC	4	769	827	2440	72.3	1.5	73.8	67.9	11.9	79.9	69.0	9.3	78.3
36	PWG	SCN	EC	4	769	827	2440	65.1	0.6	65.7	62.8	5.8	68.6	63.4	4.4	67.9

5.2 Information Retrieval in Turkish Radiology Reports

5.2.1 Test Data

Evaluation of information retrieval has been conducted with 40 radiology reports and 9 textual queries. Queries are intended to emphasize the effect of ontology on retrieval performance. Therefore, related morphological abnormality concepts are selected for queries.

Last query, on the other hand, are selected in order to test the performance of ontology based algorithm on specific concepts.

Table 5.7: Queries used for evaluation

Number	Text
1	karaciğerde lezyon
2	karaciğerde kist
3	karaciğerde kitle
4	böbrekte kist
5	böbrekte kortikal kist
6	böbrekte kitle
7	böbrekte taş
8	sağ böbrek üst polde kortikal kist

5.2.2 Evaluation

Each queries are associated with related reports manually for ground-truth. Evaluation of retrieval methods have been performed with precision-recall values, average precision values at recall cut-offs and mean average precision value for each method over all queries.

Since we have evaluated the four algorithms with a limited number of reports, recall values for all three incrementing methods are 100% as expected.

Looking at evaluation results, as we expected spell correction and more importantly stemming of terms have increased standard methods performance noticeably.

With the addition of document representation in sentence and polarity values, precision values and more importantly average precision values have significantly improved.

However, ontology increment did not satisfy our expectations even if promising result have been observed in queries with more general concepts.

Table 5.8: Precision-Recall values of queries with each method

	standard		spellstem		sentpol		ontology	
Query	Pr	Rc	Pr	Rc	Pr	Rc	Pr	Rc
1	0.43	0.75	0.50	1.00	0.65	1.00	0.50	1.00
2	0.24	1.00	0.20	1.00	0.23	1.00	0.20	1.00
3	0.31	1.00	0.28	1.00	0.35	1.00	0.28	1.00
4	0.67	1.00	0.40	1.00	0.52	1.00	0.40	1.00
5	0.54	1.00	0.33	1.00	0.42	1.00	0.33	1.00
6	0.56	0.90	0.50	1.00	0.71	1.00	0.50	1.00
7	0.10	1.00	0.07	1.00	0.11	1.00	0.07	1.00
8	0.03	1.00	0.03	1.00	0.03	1.00	0.03	1.00

Table 5.9: Average precision values for queries

Query	standard	spellstem	sentpol	ontology
1	0.36	0.64	0.75	0.80
2	0.40	0.74	0.97	0.91
3	0.24	0.47	0.60	0.73
4	0.67	0.76	0.90	0.80
5	0.70	0.98	0.99	0.97
6	0.55	0.63	0.87	0.91
7	0.19	0.12	0.92	0.81
8	0.25	0.20	0.50	1.00

Table 5.10: Mean Average Precision values for each method

standard	spellstem	sentpol	ontology
0.42	0.57	0.81	0.87

Table 5.11: Mean Average Precision values excluding for each method excluding query 8

standard	spellstem	sentpol	ontology
0.45	0.61	0.89	0.84

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1 CONCLUSION

In medical centers and hospitals, vast amount of radiology reports are generated on daily basis. Unfortunately, the way these reports produced is error prone. Physical structures added to medical information systems vary from user to user. Because of rapid generation of reports, spelling errors are also common. Furthermore, these medical reports include many types of tokens which cannot be analyzed using standard tokenization systems.

Solution to these problems mostly related with preprocessing of textual data of medical reports is crucial in order to increase performance of any text mining tasks.

Most of spelling errors in radiology reports are observed in medical terms. Implementation of standard edit-distance based spell-checking solves the problem to some extent. Inspecting spelling errors in radiology reports, we noticed that a significant portion of errors stem from pronunciational differences of English medical terms in Turkish. We have exploited this property of errors and improved precision at spelling correction: Our method basically tries all possible pronunciational variation of the word and check if new state exists in valid words. Results of our experiences showed 5-15% improvements in correction.

In order to correctly tokenize medical sentences, we have implemented a regular expression based tokenizer. Recursively defined regular expressions in a configuration file are used to label various token types.

Having rich semantic content, radiology reports are useful resources for radiologists and other medicine specialists. Unfortunately, current applications in medical information systems are based on classical IR models, which do not satisfy requirements of experts. Main objective of this work is to address problem of information needs of radiologists.

An ontology based information retrieval model has been tested incrementally. The base model is a combination of VSM and standard boolean model. Documents in this model are traditionally represented: By bag of words. An increment to this model is to index documents after stemming and spell correction tasks. This increment improves the average precision values comparing the standard model as expected. The second increment to the model is to represent the documents as a list of sentences in which polarity values are extracted. Only positive polarity valued sentences are indexed because of information need: Doctors usually search for positive findings such as "karaciğerde lezyon görülmüştür *Lesion in liver is detected.*" instead of negative findings. This increment contributed a significant improvement according to two other models as we expected. Indexing in classical models do not distinguish sentences according to polarity values. However, a negative polarity value is to be excluded form such operation because obviously it should not be a target in retrieval task. Finally, our last increment is expansion of concepts in sentences using a medical ontology. Each sentence is represented by polarity value and a list of SNOMED-CT concepts in this phase. Semantic relations between concepts in SNOMED-CT are used to expand concepts in sentences. With expansion of related concepts in sentences, we aimed to observe higher recall value. Results seem to be promising, however did not justify our predictions.

One possible reason for the unexpected results of ontology based retrieval model could be possible inaccuracy in our concept extraction approach. In order to provide evaluation data to retrieval task, we applied an intuitive method of concept extraction from Turkish textual data: Once Turkish words in sentences are translated to English by dictionary look-up, matching concepts from SNOMED-CT are refined using subsumption reasoning so that the sentence is to represented by only the most specific concepts matched. Since we could not evaluate this method, a problem here can trigger another problem in retrieval task.

6.2 FUTURE WORK

The method we used in concept extraction is intuitively implemented and is not tested. An evaluation is to be performed. However, in order to exact comparison of automatically extracted concepts with ground-truth concepts is another research problem: Evaluation of similarity and relatedness between sets of concepts. Further research is required here to provide a reliable evaluation.

Concept expansion method that we applied depends on some assumptions: a concept is expanded proportional to shortest distances to the original concept. Distance between concepts is measured according to only the number of edges between concepts. Nevertheless, it is not realistic to assume equality of all edges in ontology. For example, the similarity between root concept “SNOMED-CT Concept” and its child concept “Body Structures” should not be same with similarity between “Liver structure” and child concept Structure of ligament of liver”. A special scoring system should be developed for weighting concepts in concept expansion phase if SNOMED-CT will be used again.

We also ignored “negative” polarity sentences in retrieval whereas it is possible for radiologists to query an expression of both positive and negative valued sentences. In our next version, we will check out this assumption to provide user satisfaction.

Developing an ontology for only radiology which differentiates spatio relationships between body structures, takes radiology-specific relations into account will most likely improve the results.

Finally and more importantly, representation of relationships between concepts are ignored in this work for simplicity. A detailed work is required in order to represent inter-conceptual relationships in radiology reports for more accurate representation. To our knowledge DL expressions will represent medical knowledge on reports sufficiently. One of the problems with this approach is how to map natural language medical information into DL axioms. The other problem is how to use DL axioms for retrieval tasks with performance concerns.

CHAPTER 7

IMPLEMENTATION

7.1 Tools

7.1.1 Lucene

Apache Lucene[1] is an open source information retrieval library which provides indexing and searching operations. It uses a combination of VSM and Boolean model for retrieval and scoring. Boolean model reduces the number of matching results and VSM sorts the documents according to their scores. It uses a slight modification of $tf - idf$ weighting schema in order to calculate scores.

Lucene provides a wide-range of facilities in all phases of retrieval process: Stemmers, spell-checkers, tokenizers and analyzers for preprocessing; ability to change weighting schema for ranking.

7.1.2 Protege

Protege[2] is an open source ontology editing, knowledge-base modeling and reasoning framework. It is a plug-in based framework. In this work protege is used in order to visualize and browse SNOMED-CT ontology.

Being an open source project and having active support of developers and users, protege is the most popular and updated ontology editor. Developing plug-ins for protege is relatively

[1] http://lucene.apache.org/

[2] http://protege.stanford.edu/

easy because it uses OSGI[3] framework as a plug-in infrastructure and its generous developer documentation.

7.1.3 ELK Reasoner

ELK[4] is an ontology reasoner developed for OWL-2 and supports OWL-$\mathcal{EL}$ language.

Although it does not support all of the OWL features and all reasoning tasks currently, it provides very fast reasoning for OWL-$\mathcal{EL}$. Classification of SNOMED-CT is accomplished in 15-20 seconds using ELK reasoner whereas other reasoners (Pellet or Fact++) classify the ontology about 30-45 minutes.

7.1.4 Zemberek 2

Zemberek[5] is an open source morphological analyzer and spell-checker for Turkish language. It morphologically analyzes given word upto the words root form in which derivational suffixes also separated.

The library can provide more than one analysis for a word. For such cases it sort analysis using its statistical model. Although this approach gives successful results for a general domain application/problem, in a specific domain due to its sophisticated statistical language model this approach may fail. Thus, an external morphological disambiguation algorithm has to be applied in specific applications. However, in our problem, since most of the words in domain are already in noun form mainly having few suffixes, we disregarded disambiguation for morphological analysis and assumed first result as correct.

7.1.5 SNOMED-CT

7.1.5.1 Definition

SNOMED CT is a comprehensive medical terminology system that provides medical content to express clinical documentation and reporting. It is an ontological structure that provides

[3] http://www.osgi.org/About/Technology

[4] http://code.google.com/p/elk-reasoner/

[5] http://code.google.com/p/zemberek/

a consistent way for coding, retrieval and analyze of clinical data. This semantically organized terminology is basically comprised of concepts, terms and relationships. SNOMED-CT covers a wide range of medical information such as findings, diseases, body structures, organisms, procedures, pharmaceutical products, which are represented as hierarchies in its structure (Figure 7.1).

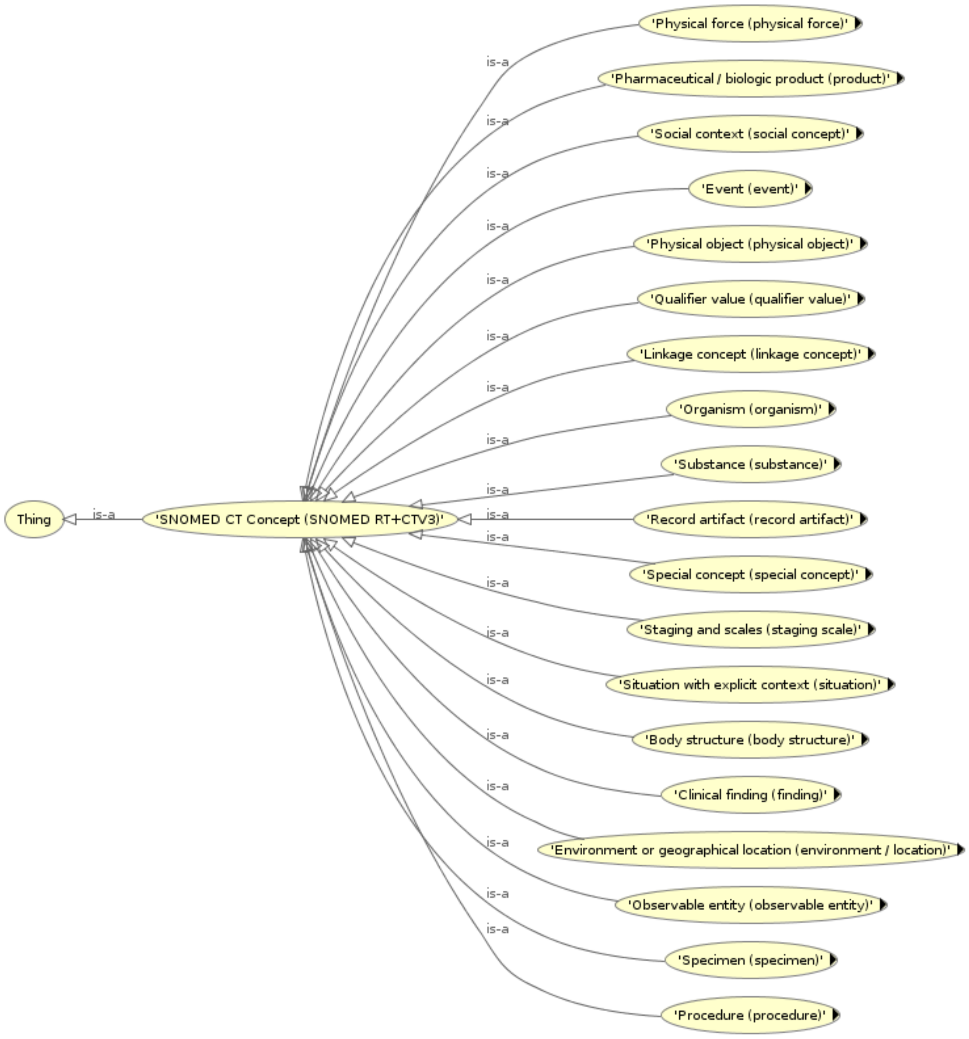
Thing
is-a
'SNOMED CT Concept (SNOMED RT+CTV3)'
'Physical force (physical force)'
'Pharmaceutical / biologic product (product)'
'Social context (social concept)'
'Event (event)'
'Physical object (physical object)'
'Qualifier value (qualifier value)'
'Linkage concept (linkage concept)'
'Organism (organism)'
'Substance (substance)'
'Record artifact (record artifact)'
'Special concept (special concept)'
'Staging and scales (staging scale)'
'Situation with explicit context (situation)'
'Body structure (body structure)'
'Clinical finding (finding)'
'Environment or geographical location (environment / location)'
'Observable entity (observable entity)'
'Specimen (specimen)'
'Procedure (procedure)'

Figure 7.1: Top level hierarchies

7.1.5.2 Physical Structure

Basic components of SNOMED CT are concepts, descriptions and relationships, each of which are stored as database tables in its regular distribution. Refer to [17, 18] for detailed table structures of SNOMED CT distribution and technical recommendations about database implementations.

More than 386000 unique concepts are organized in hierarchies from the most general to the most specific. Concepts being defined in different levels of abstraction allows to record or retrieve clinical data in detail.

Descriptions are string representations of concepts in human-readable way. Descriptions are terms or names assigned to a concept. In other words, descriptions are string representations for concepts in human-readable way. In SNOMED, nearly 1145000 descriptions exist. Most concepts have several descriptions which are called synonyms.

And finally, approximately 1385000 relationships between concepts constitute the semantic structure of ontology. These links among the concepts provide logical definitions or specifications for concepts.

Concepts

Concept is the basic structure of SNOMED terminology which represents a semantically defined medical meaning. Although a concept is a medical meaning by definition, it does not have to cover a single atomic meaning. It may, and generally most of the concepts in SNOMED are, to be a set of semantic meaning.

“Concepts are formally defined in terms of their relationships with other concepts.” [19] That means, coverage of a concept is specified with its relationships. Above these logical definitions, *IS_A* relation , the most distinguishing and commonly used one in SNOMED, contributes to the granularity of concepts at most.Concepts being whether general or specific depends on their position in ontology regarding *IS_A* relations. In other words, the nearer to root, the more coverage concept has.

Being equivalent of ‘concept’ in Description Logic, table of concepts in SNOMED-CT has a field called IsPrimitive. This field specifies whether the concept is primitive (i.e. concept has necessary conditions) or fully defined (i.e. concept has necessary and sufficient conditions).

Descriptions

Concept descriptions are human-readable phrases or names associated with SNOMED CT concepts. Descriptions are usually referred as 'terms' in context of SNOMED literature. Multiple descriptions (of three different types) might have been assigned to a concept, whereas each concept have at least two descriptions of types Fully Specified Name *(FSN)* and Preferred Term.

Each concept has exactly one unique *FSN* that unambiguously represents the concept. A *FSN* of a concept provides a unique phrase for that concept. Similarly, each concept has exactly one Preferred Term, that is the most common phrase used by clinicians. Another type of descriptions, Synonym, captures any other terms that represent the same concept as the *FSN*. Synonyms and Preferred Terms, unlike *FSN*s, are not necessarily to be unique.

Relationships

Relationships are the linking structures between concepts that gives semantic network property of SNOMED-CT. A relationship, a.k.a. attribute, is defined by two related concepts and a relating concept, referred as *RelationshipType*, in terminology.

Attributes in SNOMED-CT are defined with 'domain' and 'range' restrictions. 'Domain' of an attribute is the set of concepts to which the attribute can be applied. 'Range' of an attribute is, similarly, the set of value concepts allowed for the attribute. Domain and range properties for defining attributes is detailly explained in [19].

There are four types of relationships in SNOMED-CT: defining, qualifying, historical and additional.

- **Defining relationships** provide the logical definitions of concepts in terms of other concepts. In addition to subtype relationship, the main relationship that relates concepts to other concepts in *IS_A* hierarchy, SNOMED-CT uses over 50 defining attributes. Examples include *ASSOCIATED MORPHOLOGY, FINDING SITE, FINDING METHOD, LATERALITY, MEASUREMENT METHOD*. Although full list of defining attributes, valid domain - range values and examples is explained detailly in [19, 17], a brief explanation on defining characteristics and semantic properties of these attributes are discussed in the next section.

- **Qualifying relationships**, unlike Defining relationships, are optional attributes that may have one of several values for a particular concept. An expression with a qualifying attribute-value pair represents a more tightly defined concept than the original.

- **Historical relationships** are used to link inactive concepts to other concepts. Examples include *SAME AS, REPLACED BY, MAY BE A, WAS A*.

- **Additional relationships** represent context specific characteristics, which may vary according to place or time. For example: "prescription only medicine" is a context specific characteristic of the concept "ampxycilin 250mg capsule". It is true in the UK but not in some other countries [18].

7.1.5.3 Semantic Properties

"Each concept in SNOMED-CT is logically defined through its defining relationships to other concepts [19]." In other words, logical representation of a concept is established using other concepts and defining relationships such as *IS_A, FINDING_SITE, ASSOCIATED_MORPHOLOGY*.

Logical definition of the concept "Fracture of tarsal bone (disorder)" is given as an example:

Fracture of tarsal bone (disorder)

IS_A *Fracture of foot (disorder)*

FINDING SITE *Bone structure of tarsus (body structure)*

ASSOCIATED MORPHOLOGY *Fracture (morphologic abnormality)*

Compositional grammar of SNOMED-CT enables to build new concepts using concepts defined in ontology. These concepts which are already defined in ontology and related to a concept identifier are referred as *pre-coordinated*. An expression containing two or more concept identifier defining a new concept in ontology is said to be *post-coordinated*.

Definition of a *primitive* concept (or expression) "does not sufficiently express its meaning so that its subtypes can be computably recognized [19]". That means, an expression with a concept identifier implies the logical definitions of the concept, however it is not possible to imply the concept from the logical definitions.

For example, the concept of “Structure of inferior pole of kidney” is a primitive concept defined with an *IS_A* relationship to the concept of “Structure of pole of kidney”. Being primitive states that a “Structure of inferior pole of kidney” concept is always a “Structure of pole of kidney”, but not the other way.

Definition of a *fully defined* or *sufficiently defined* concept on the other hand, is “sufficient to computably recognize (automatically subsume) all its subtypes [19]”. In other words, both the concept and its logical definitions imply the other.

For example, definition of “Fracture of tarsal bone (disorder)” is fully defined which means a concept having relations *IS_A* to “Fracture of foot (disorder)”, *FINDING SITE* to “Bone structure of tarsus (body structure)” and *ASSOCIATED MORPHOLOGY* to “Fracture (morphologic abnormality)” is equivalent to the concept of “Fracture of tarsal bone (disorder)”

REFERENCES

[1] H. Alani, Sanghee Kim, D.E. Millard, M.J. Weal, W. Hall, P.H. Lewis, and N.R. Shadbolt. Automatic ontology-based knowledge extraction from web documents. *Intelligent Systems, IEEE*, 18(1):14–21, jan-feb 2003.

[2] Franz Baader. The description logic handbook. chapter Appendix 1 Description Logic Terminology, pages 495–505. Cambridge University Press, New York, NY, USA, 2003.

[3] Franz Baader, Ian Horrocks, and Ulrike Sattler. Description Logics. In Frank van Harmelen, Vladimir Lifschitz, and Bruce Porter, editors, *Handbook of Knowledge Representation*, chapter 3, pages 135–180. Elsevier, 2008.

[4] Franz Baader and Werner Nutt. The description logic handbook. chapter Basic description logics, pages 43–95. Cambridge University Press, New York, NY, USA, 2003.

[5] Ricardo A. Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1999.

[6] Dalila Bekhouche, Yann Pollet, Bruno Grilheres, and Xavier Denis. Architecture of a medical information extraction system. In Farid Meziane and Elisabeth Métais, editors, *Natural Language Processing and Information Systems*, volume 3136 of *Lecture Notes in Computer Science*, pages 522–531. Springer Berlin / Heidelberg, 2004. 10.1007/978-3-540-27779-8_35.

[7] Olivier Bodenreider, Barry Smith, Anand Kumar, and Anita Burgun. Investigating subsumption in snomed ct: An exploration into large description logic-based biomedical terminologies. *Artif. Intell. Med.*, 39:183–195, March 2007.

[8] P. Castells, M. Fernandez, and D. Vallet. An adaptation of the vector-space model for ontology-based information retrieval. *IEEE Transactions on Knowledge and Data Engineering*, pages 261–272, 2007.

[9] Werner Ceusters, Barry Smith, and Jim Flanagan. Ontology and medical terminology: Why description logics are not enough.

[10] Werner Ceusters, Barry Smith, Anand Kumar, and Christoffel Dhaen. Ontology-based error detection in snomed-ct. *Studies In Health Technology And Informatics*, 107(Pt 1):482–486, 2004.

[11] Stephen Dill, Nadav Eiron, David Gibson, Daniel Gruhl, R. Guha, Anant Jhingran, Tapas Kanungo, Kevin S. Mccurley, Sridhar Rajagopalan, Andrew Tomkins, John A. Tomlin, and Jason Y. Zien. A case for automated large scale semantic annotations. *Journal of Web Semantics*, 1:115–132, 2003.

[12] David W. Embley, Douglas M. Campbell, Randy D. Smith, and Stephen W. Liddle. Ontology-based extraction and structuring of information from data-rich unstructured

documents. In *Proceedings of the seventh international conference on Information and knowledge management*, CIKM ’98, pages 52–59, New York, NY, USA, 1998. ACM.

[13] M. Fernández, D. Vallet, and P. Castells. Probabilistic score normalization for rank aggregation. *Advances in Information Retrieval*, pages 553–556, 2006.

[14] R. Guha, Rob McCool, and Eric Miller. Semantic search. In *Proceedings of the 12th international conference on World Wide Web*, WWW ’03, pages 700–709, New York, NY, USA, 2003. ACM.

[15] Angelos Hliaoutakis, Giannis Varelas, Epimeneidis Voutsakis, Euripides G. M. Petrakis, and Evangelos Milios. Information retrieval by semantic similarity. In *Intern. Journal on Semantic Web and Information Systems (IJSWIS), 3(3):55–73, July/Sept. 2006. Special Issue of Multimedia Semantics*, 2006.

[16] The International Health Terminology Standarts Development Organisation (IHTSDO). Snomed clinical terms - abstract logical models and representational forms. Technical report, January 2009.

[17] The International Health Terminology Standarts Development Organisation (IHTSDO). Snomed clinical terms - technical implementation guide. Technical report, January 2009.

[18] The International Health Terminology Standarts Development Organisation (IHTSDO). Snomed clinical terms - technical reference guide. Technical report, January 2009.

[19] The International Health Terminology Standarts Development Organisation (IHTSDO). Snomed clinical terms - user guide. Technical report, January 2009.

[20] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schtze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[21] Sougata Mukherjea. Information retrieval and knowledge discovery utilising a biomedical semantic web. *Briefings in Bioinformatics*, 6(3):252–262, 2005.

[22] Daniele Nardi and Ronald J. Brachman. The description logic handbook. chapter An introduction to description logics, pages 1–40. Cambridge University Press, New York, NY, USA, 2003.

[23] J. Paralic and I. Kostial. Ontology-based information retrieval. In *Proceedings of the 14th International Conference on Information and Intelligent systems (IIS 2003), Varazdin, Croatia*, pages 23–28. Citeseer, 2003.

[24] Ted Pedersen, Serguei V. S. Pakhomov, Siddharth Patwardhan, and Christopher G. Chute. Measures of semantic similarity and relatedness in the biomedical domain. *J. of Biomedical Informatics*, 40:288–299, June 2007.

[25] Borislav Popov, Atanas Kiryakov, Damyan Ognyanoff, Dimitar Manov, and Angel Kirilov. Kim - a semantic platform for information extraction and retrieval. *Nat. Lang. Eng.*, 10:375–392, September 2004.

[26] S. Radhouani, G. Falquet, and J.P. Chevalletinst. Description logic to model a domain specific information retrieval system. In *Database and Expert Systems Applications*, pages 142–149. Springer, 2008.

[27] Lawrence Reeve and Hyoil Han. Survey of semantic annotation platforms. In *Proceedings of the 2005 ACM symposium on Applied computing*, SAC '05, pages 1634–1638, New York, NY, USA, 2005. ACM.

[28] P. Ruch, Julien Gobeill, Christian Lovis, and Antoine Geissbühler. Automatic medical encoding with snomed categories. *BMC medical informatics and decision making*, 8 Suppl 1:S6, 2008 2008.

[29] Ulrike Sattler, Diego Calvanese, and Ralf Molitor. The description logic handbook. chapter Relationships with other formalisms, pages 137–177. Cambridge University Press, New York, NY, USA, 2003.

[30] Manfred Schmidt-Schauß and Gert Smolka. Attributive concept descriptions with complements. *Artif. Intell.*, 48(1):1–26, 1991.

[31] Parikshit Sondhi, Manish Gupta, ChengXiang Zhai, and Julia Hockenmaier. Shallow information extraction from medical forum data. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 1158–1166, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[32] Ergin Soysal, Ilyas Cicekli, and Nazife Baykal. Design and evaluation of an ontology based information extraction system for radiological reports. *Comput. Biol. Med.*, 40:900–911, November 2010.

[33] D. Vallet, M. Fernández, and P. Castells. An ontology-based information retrieval model. *The Semantic Web: Research and Applications*, pages 103–110, 2005.

Appendix A

N-Gram Analysis on 4689 reports

Table A.1: 2-grams

Phrase	Frequency	Phrase	Frequency
her iki	5007	kontrast madde	618
ile uyumlu	1694	iki akciğerde	605
lezyon izlenmemiştir	1441	venöz sistem	602
normal olup	1419	kortikal sulkuslar	594
fokal lezyon	1209	iki adrenal	591
Kesitlere dahil	1118	Beyin sapı	589
ya da	1085	hepatik venöz	586
dikkati çekmektedir	980	Trakea ve	586
konturu büyüklüğü	929	sistem patenttir	573
dahil kemik	874	safra yollarında	571
ve lateral	847	kuyruk kesimi	567
lezyon saptanmamıştır	829	ve kuyruk	564
ve her	827	normal görünümdedir	562
lenf nodu	775	büyüklüğü parankimi	557
Safra kesesi	773	madde tutulumu	551
kemik yapılarda	767	izlenmiş olup	549
olarak değerlendirilmiştir	738	iki böbreğin	546
ana vasküler	710	kesimi normal	535
lateral ventriküller	706	daha belirgin	530
normal olarak	677	Pankreasın baş	527
bez normaldir	645	büyüklüğü normal	527
ve hepatik	645	arter ve	526
büyüklüğü normaldir	631	korpus ve	521
adrenal bez	621	bazal gangliyonlar	520
Portal ve	621	bulgu saptanmamıştır	518

Table A.2: 3-grams

Phrase	Frequency	Phrase	Frequency
fokal lezyon izlenmemiştir	1011	parankim içerisinde fokal	487
ve her iki	825	konturu büyüklüğü normal	483
dahil kemik yapılarda	712	normal olup parankim	483
Kesitlere dahil kemik	684	olup parankim içerisinde	482
normal olarak değerlendirilmiştir	627	ana vasküler yapılar	481
Portal ve hepatik	620	İntra ve ekstrahepatik	463
ve lateral ventriküller	599	safra yollarında dilatasyon	455
hepatik venöz sistem	586	Mediastinal ana vasküler	454
iki adrenal bez	585	ekstrahepatik safra yollarında	444
venöz sistem patenttir	566	3 ve lateral	425
ve hepatik venöz	563	Her iki akciğerde	421
ve kuyruk kesimi	556	madde tutulumu saptanmamıştır	413
adrenal bez normaldir	549	yollarında dilatasyon izlenmemistir	409
kontrast madde tutulumu	549	her iki hiler	403
Her iki böbreğin	535	anormal kontrast madde	400
kuyruk kesimi normal	528	veya anormal kontrast	399
kesimi normal olarak	526	duvarında kalınlaşma veya	399
Her iki adrenal	520	kesesi normal olup	399
korpus ve kuyruk	515	Safra kesesi normal	399
baş korpus ve	511	kalınlaşma veya anormal	399
ve ekstrahepatik safra	507	normal olup duvarında	399
Pankreasın baş korpus	506	olup duvarında kalınlaşma	399
içerisinde fokal lezyon	499	iki hiler bölgede	395
büyüklüğü normal olup	498	lehine radyolojik bulgu	392
Dalağın konturu büyüklüğü	492	ve ana bronşlar	389

Table A.3: 4-grams

Phrase	Frequency
ve hepatik venöz sistem	563
hepatik venöz sistem patenttir	556
Portal ve hepatik venöz	545
Kesitlere dahil kemik yapılarda	541
iki adrenal bez normaldir	527
ve kuyruk kesimi normal	527
kuyruk kesimi normal olarak	526
kesimi normal olarak değerlendirilmiştir	521
Her iki adrenal bez	516
korpus ve kuyruk kesimi	512
baş korpus ve kuyruk	511
Pankreasın baş korpus ve	506
içerisinde fokal lezyon izlenmemiştir	488
parankim içerisinde fokal lezyon	487
konturu büyüklüğü normal olup	482
Dalağın konturu büyüklüğü normal	482
normal olup parankim içerisinde	482
olup parankim içerisinde fokal	482
büyüklüğü normal olup parankim	479
İntra ve ekstrahepatik safra	462
ve ekstrahepatik safra yollarında	444
ekstrahepatik safra yollarında dilatasyon	436
kontrast madde tutulumu saptanmamıştır	413
safra yollarında dilatasyon izlenmemistir	409
anormal kontrast madde tutulumu	400
veya anormal kontrast madde	399
duvarında kalınlaşma veya anormal	399
kesesi normal olup duvarında	399
kalınlaşma veya anormal kontrast	399
normal olup duvarında kalınlaşma	399
olup duvarında kalınlaşma veya	399
ve her iki hiler	398
Safra kesesi normal olup	395
her iki hiler bölgede	392
Trakea ve ana bronşlar	378
Mediastende ve her iki	369
iki böbreğin konturu büyüklüğü	366
Her iki böbreğin konturu	362
ve ana bronşlar normaldir	357
ve toplayıcı sistemi normaldir	352

Table A.4: 11-grams

Phrase	Frequency
kesesi normal olup duvarında kalınlaşma veya anormal kontrast madde tutulumu saptanmamıştır	397
Safra kesesi normal olup duvarında kalınlaşma veya anormal kontrast madde tutulumu	395
Mediastende ve her iki hiler bölgede kitle ya da lenfadenopati saptanmamıştır	165
bazal gangliyonlar talamuslar kapsula interna ve eksternalar her iki sentrum semiovale	118
Bilateral bazal gangliyonlar talamuslar kapsula interna ve eksternalar her iki sentrum	118
gangliyonlar talamuslar kapsula interna ve eksternalar her iki sentrum semiovale normaldir	117
İntraabdominal veya retroperitoneal koleksiyon veya solid kitle yoktur İntraabdominal serbest sıvı	78
veya retroperitoneal koleksiyon veya solid kitle yoktur İntraabdominal serbest sıvı izlenmemiştir	78
Paraaortik parakaval parailiyak ve mezenterik patolojik boyutta büyümüş lenf nodu saptanmamıştır	63
bazal gangliyonlar kapsüla internalar talamuslar her iki korona radyata ve hemisferik	60
Bilateral bazal gangliyonlar kapsüla internalar talamuslar her iki korona radyata ve	60
gangliyonlar kapsüla internalar talamuslar her iki korona radyata ve hemisferik kortikal	57
internalar talamuslar her iki korona radyata ve hemisferik kortikal sulkuslar normaldir	53
kapsüla internalar talamuslar her iki korona radyata ve hemisferik kortikal sulkuslar	53
Her iki tarafta bazal gangliyonlar talamuslar ve kapsula interna ve eksternalar	40
iki tarafta bazal gangliyonlar talamuslar ve kapsula interna ve eksternalar normaldir	40

Appendix B

Regular Expressions

B.1 Sentence Splitter Regex

```
((?<=(?<!kon)[dt][iİüÜuUıI]r)[ˆA-Za-zöçşğüıÖÇŞĞÜİ]|\.\s*(?=[A-ZÖÇŞĞÜİ]))
```

B.2 Report Section Tags

```
kl[iIıİ]n[[iIıİ]]k\s*:
kl[iIıİ]n[[iIıİ]]k\s+b[iIıİ]lg[iIıİ]\s*:
tekn[iIıİ]k\s*:
tetk[iIıİ]k\s*:
bulgular\s+ve\s+[iIıİ][ŞşSs]lem\s*:
tekn[iIıİ]k\s+ve\s+bulgular\s*:
bulgular\s*:
[iIıİ][ŞşSs]lem\s*:
[öÖOo]ner[iIıİ](ler)?\s*:
radyofarmas[öÖoO]t[iIıİ]k\s*:
g[öÖOo]r[üÜuU]nt[üÜuU]leme\s+protokol[üÜuU]\s*:
[öÖOo]ner[iIıİ]ler\s*:
yorum\s*:
kar[sSşŞ][ıiIİ]la[sSşŞ]t[ıiIİ]rma\s*:
referans\s+de[gğGĞ]erler[iIıİ]\s*:
sonu[cçCÇ]\s*:
tetk[iIıİ]k[iIıİ]\s*[iIıİ]steyen\s*
```

Appendix C

REGEX Tokenizer Configuration File

```
 <regex_token>
#=================================
basicNumeric:
>>[0-9]+([\.,][0-9]+){0,1}
orn; 1.2 ; 23,3 ; 565 ; 1.51

<regex_token>

rangedNumeric:
>[0-9]{1,5}([\.,][0-9]+){0,1}\s*-\s*basicNumeric(\s*-\s*basicNumeric)*
>>[0-9]{1,5}([\.,][0-9]+){0,1}\s*-\s*[0-9]+([\.,][0-9]+){0,1}
(\s*-\s*[0-9]+([\.,][0-9]+){0,1})*
orn; 12-20 ; 4-4,5 ; 3- 4 ; 20 -30 ; 3-6-8

<regex_token>

unit:
>>(dcbead|mikron|msn|joule|cm|mm|sn|ml|lt|ng|atım|dk|cc|cpm|mci|m|hg)
[0-9]{0,1}(/\s*(cm|sn|ml|dk|ci|s|hg)[0-9]{0,1}){0,1}
orn; mm ; cm2 ; hg/s2

<regex_token>
```

```
numeric:
>(\brangedNumeric|basicNumeric)
>>(\b[0-9]{1,5}([\.,][0-9]+){0,1}\s*-\s*[0-9]+([\.,][0-9]+){0,1}
(\s*-\s*[0-9]+([\.,][0-9]+){0,1})*|[0-9]+([\.,][0-9]+){0,1})
```

```
<regex_token>
```

```
area:
>numeric\s*(unit){0,1}(\s*[xX]\s*numeric\s*(unit){0,1})+
>(rangedNumeric|basicNumeric)\s*(unit){0,1}(\s*[xX]\s*(rangedNumeric
|basicNumeric)\s*(unit){0,1})+
>>([0-9]{1,5}([\.,][0-9]+){0,1}\s*-\s*[0-9]+([\.,][0-9]+){0,1}
(\s*-\s*[0-9]+([\.,][0-9]+){0,1})*|[0-9]+([\.,][0-9]+){0,1})\s*
((dcbead|mikron|msn|joule|cm|mm|sn|ml|lt|ng|atım|dk|cc|cpm|mci|m|hg)
[0-9]{0,1}(/\s*(cm|sn|ml|dk|ci|s|hg)[0-9]{0,1}){0,1}){0,1}(\s*[xX]
\s*([0-9]{1,5}([\.,][0-9]+){0,1}\s*-\s*[0-9]+([\.,][0-9]+){0,1}
(\s*-\s*[0-9]+([\.,][0-9]+){0,1})*|[0-9]+([\.,][0-9]+){0,1})\s*
((dcbead|mikron|msn|joule|cm|mm|sn|ml|lt|ng|atım|dk|cc|cpm|mci|m|hg)
[0-9]{0,1}(/\s*(cm|sn|ml|dk|ci|s|hg)[0-9]{0,1}){0,1}){0,1})+
orn; 17x12 mm ; 10x9.5 mm ; 22x15x20 mm ; 3,5 x 2 cm ;
 6-8mmx40 mm ; 4,5mmx37 mm ; 3,5 X 4,5 cm ; 28-20x170 mm ;
  16-16 mmx120 mm ; 25- 16x166 mm
```

```
<regex_token>
```

```
numericUnit:
>numeric\s*unit
>(rangedNumeric|basicNumeric)\s*unit
>>([0-9]{1,5}([\.,][0-9]+){0,1}\s*-\s*[0-9]+([\.,][0-9]+){0,1}(\s*-\s*
[0-9]+([\.,][0-9]+){0,1})*|[0-9]+([\.,][0-9]+){0,1})\s*(dcbead|mikron
|msn|joule|cm|mm|sn|ml|lt|ng|atım|dk|cc|cpm|mci|m|hg)[0-9]{0,1}
(/\s*(cm|sn|ml|dk|ci|s|hg)[0-9]{0,1}){0,1}
orn; 88 mm ; 20 m/sn2 ; 0.073 sn ; 16,5 mm ; 35-45-60 mm
```

```
<regex_token>
numericInci: #(harf kismi dahl degil)
>(numeric)\.(?=\s*[a-zıöçşğü])
>(rangedNumeric|basicNumeric)\.(?=\s*[a-zıöçşğü])
>>([0-9]{1,5}([\.,][0-9]+){0,1}\s*-\s*[0-9]+([\.,][0-9]+){0,1}
(\s*-\s*[0-9]+([\.,][0-9]+){0,1})*|[0-9]+([\.,][0-9]+){0,1})\.
(?=\s*[a-zıöçşğü])
orn; 1. ventrikul ; 5. kosta ; 3-6-8-11. kosta ; 24.saat ;

<regex_token>
#==================================
tarih:
>>([0-9]{1,2}\s*(/|\.|-)\s*){0,1}[0-9]{1,2}\s*(/|\.|-)\s*[0-9]{4}
orn; 23.12.2008 ; 2/2/1111 ; 13/1234

<regex_token>


tarih2:
>>([0-9]{1,2}\s*){0,1}\b((ocak|[sşŞ]ubat|mart|n[ıIiİ]san|may[ıIiİ]s|
haz[ıIiİ]ran|temmuz|a[ğgGĞ]ustos|eyl[üÜuU]l|ek[ıIiİ]m|kas[ıIiİ]m|
aral[ıIiİ]k))\b\s*-{0,1}\s*([0-9]{4})
orn; Ocak 2008 ; mart-2001 ; 6 şubat 2009 ; Ocak-2007

<regex_token>


tarihtotal:
>(tarih2|tarih)
>>(([0-9]{1,2}\s*){0,1}\b((ocak|[sşŞ]ubat|mart|n[ıIiİ]san|may[ıIiİ]s|
haz[ıIiİ]ran|temmuz|a[ğgGĞ]ustos|eyl[üÜuU]l|ek[ıIiİ]m|kas[ıIiİ]m|
aral[ıIiİ]k))\b\s*-{0,1}\s*([0-9]{4})|([0-9]{1,2}\s*(/|\.|-)\s*){0,1}
[0-9]{1,2}\s*(/|\.|-)\s*[0-9]{4})
orn; 05.02.2009 ; 11.2008 ; 14/05/2009 ; Şubat 2008
```

```
<regex_token>

#-------
tarihrange1:tarihtotal-tarihtotal
>(tarihtotal\s*-\s*tarihtotal)
>>((([0-9]{1,2}\s*){0,1}\b((ocak|[sşŞ]ubat|mart|n[ıIiİ]san|may[ıIiİ]s|
haz[ıIiİ]ran|temmuz|a[ğgGĞ]ustos|eyl[üÜuU]l|ek[ıIiİ]m|kas[ıIiİ]m|
aral[ıIiİ]k))\b\s*-{0,1}\s*([0-9]{4})|([0-9]{1,2}\s*(/|\.|-)\s*){0,1}
[0-9]{1,2}\s*(/|\.|-)\s*[0-9]{4})\s*-\s*(([0-9]{1,2}\s*){0,1}\b((ocak|
[sşŞ]ubat|mart|n[ıIiİ]san|may[ıIiİ]s|haz[ıIiİ]ran|temmuz|a[ğgGĞ]ustos|
eyl[üÜuU]l|ek[ıIiİ]m|kas[ıIiİ]m|aral[ıIiİ]k))\b\s*-{0,1}\s*([0-9]{4})|
([0-9]{1,2}\s*(/|\.|-)\s*){0,1}[0-9]{1,2}\s*(/|\.|-)\s*[0-9]{4}))
orn; 26.06.2009-09.06.2009 ; 08.12.2008 -13.11.200 ;

<regex_token>

tarihpartialrange2:26-27.05.2009
>>[0-9]{1,2}\s*-[0-9]{1,2}\s*(/|\.)\s*[0-9]{1,2}\s*(/|\.)\s*[0-9]{4}
orn; 26-27.05.2009 ;

<regex_token>

tarihrange:
>(tarihrange1|tarihpartialrange2)
>>(((([0-9]{1,2}\s*){0,1}\b((ocak|[sşŞ]ubat|mart|n[ıIiİ]san|may[ıIiİ]s
|haz[ıIiİ]ran|temmuz|a[ğgGĞ]ustos|eyl[üÜuU]l|ek[ıIiİ]m|kas[ıIiİ]m|
aral[ıIiİ]k))\b\s*-{0,1}\s*([0-9]{4})|([0-9]{1,2}\s*(/|\.|-)\s*){0,1}
[0-9]{1,2}\s*(/|\.|-)\s*[0-9]{4})\s*-\s*(([0-9]{1,2}\s*){0,1}\b((ocak|
[sşŞ]ubat|mart|n[ıIiİ]san|may[ıIiİ]s|haz[ıIiİ]ran|temmuz|a[ğgGĞ]ustos
|eyl[üÜuU]l|ek[ıIiİ]m|kas[ıIiİ]m|aral[ıIiİ]k))\b\s*-{0,1}\s*([0-9]{4})
|([0-9]{1,2}\s*(/|\.|-)\s*){0,1}[0-9]{1,2}\s*(/|\.|-)\s*[0-9]{4}))|
[0-9]{1,2}\s*-[0-9]{1,2}\s*(/|\.)\s*[0-9]{1,2}\s*(/|\.)\s*[0-9]{4})
```

```
orn; 26-27.05.2008 ; 26.06.2009-09.06.2008 ; 11 Mart 2008-8 Ekim 2008

<regex_token>
#=================================


kesirliNumeric:tarih uygulandiktan sonra uygula
>>[0-9]+/[0-9]+

<regex_token>

kesirlirange:
>>[0-9]+/[0-9]+\s*-\s*[0-9]+/[0-9]+

<regex_token>



upperkisaltma:
>>\b[A-ZÖÇŞİĞÜ]([A-ZÖÇŞİĞÜ]|[-+\.x]|[0-9]|\s)*([A-ZÖÇŞİĞÜ]|
[-+\.x]|[0-9])+
orn; İTP. ; US ; TAH+BSO. ; SMA ; TV US ; BI-RADS 2 ; BIRADS 2 ;

<regex_token>

lowerkisaltma:
>>([a-zıöçşğü][\.-]){2,}
orn; d.m. ; v.p. ; i.v.

<regex_token>

tirnakEk:
>>'\s*[a-zığüöçş]+
orn; 'nın ; 'ye ; 'sının ; ' lık ;
```

```
<regex_token>

percNumeric:
>%\s*numeric

<regex_token>

sequenceNumber:
>>[0-9]{1,3}(\s*(-|,|ve)\s*[0-9]+)*
orn; 2-3-4 ; 4,5,6,8 ; 6 ve 5;

<regex_token>

segmentSeq:
>([Ss]egment\s*sequenceNumber)(\s*\.){0,1}
#>>(([Ss]egment)\s*[0-9]+(\s*(,|-|ve)+\s*[0-9]+)*)
orn; Segment 7-8 ; segment 6 ve 5 ; segment 4-5-8 ;
 segment 2-4-6 ve 8

<regex_token>

capitalSeqNumber:
>[A-Z]sequenceNumber(\s*\.){0,1}
>>[A-Z]([0-9]+(\s*(-|,|ve)\s*[0-9]+)*)(\s*\.){0,1}
orn; T12 ; L5 ; T3-4 ; C2-3-4-5 ; T4,5,6,8. ; L5.

<regex_token>

capitalSeqNumberPhrase: #raporlar/13979.txt
>(capitalSeqNumber(\s*(ve|-|,)+\s*capitalSeqNumber)*)(\s*\.{0,1})
orn; T12-L1 ve T3-4 ; M1, M2 ve M3 ; T5-T7, T10-11. ;

<regex_token>
```

```
bolgeselBirim:
>>((kesit|vertebra|segment|disk|düzey|seri|sinir|ventrik[uü]l|
lokalizasyon|seviye|kosta)[^\s]*)
orn; kesitler ; vertebra ; vertebraların ;

<regex_token>

capitalSeqNumberPhraseBolgeselBirim:
>(capitalSeqNumberPhrase(\s*bolgeselBirim){0,1})
orn; T6-L3 vertebraları ; L5,S1, S2 vertebra ; segment 2-4-6 ve 8

<regex_token>

sequenceNumPhraseBirim:
>sequenceNumber(\s*\.{0,1})(\s*bolgeselBirim)
onr; 56-57-68. kesit ; 2-3-4. kostalarda

<regex_token>

numberedBodyStructure:
>(capitalSeqNumberPhraseBolgeselBirim|segmentSeq
|sequenceNumPhraseBirim)

<regex_token>



#sıra:
<<area,
<<percNumeric,
<<tarihrange, tarihtotal,
<<kesirlirange, kesirliNumeric,
<<numberedBodyStructure,
<<numericUnit, numericInci,
<<upperkisaltma, lowerkisaltma,
<<numeric,
<<tirnakEk
#<<word
```

Appendix D

Information Retrieval Precision-Recall curves

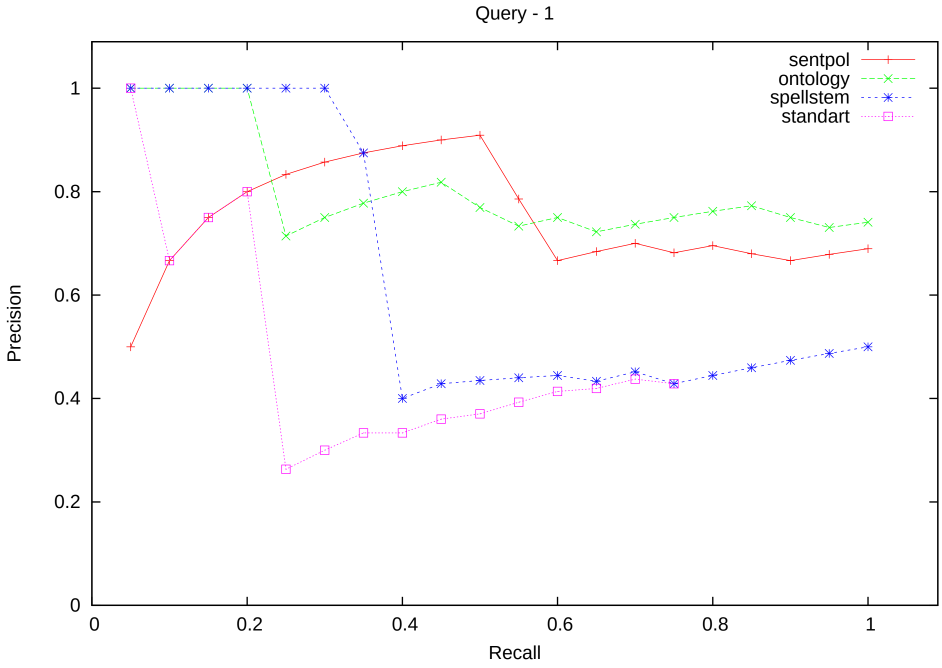

Figure D.1: Query:“karaciğerde lezyon”

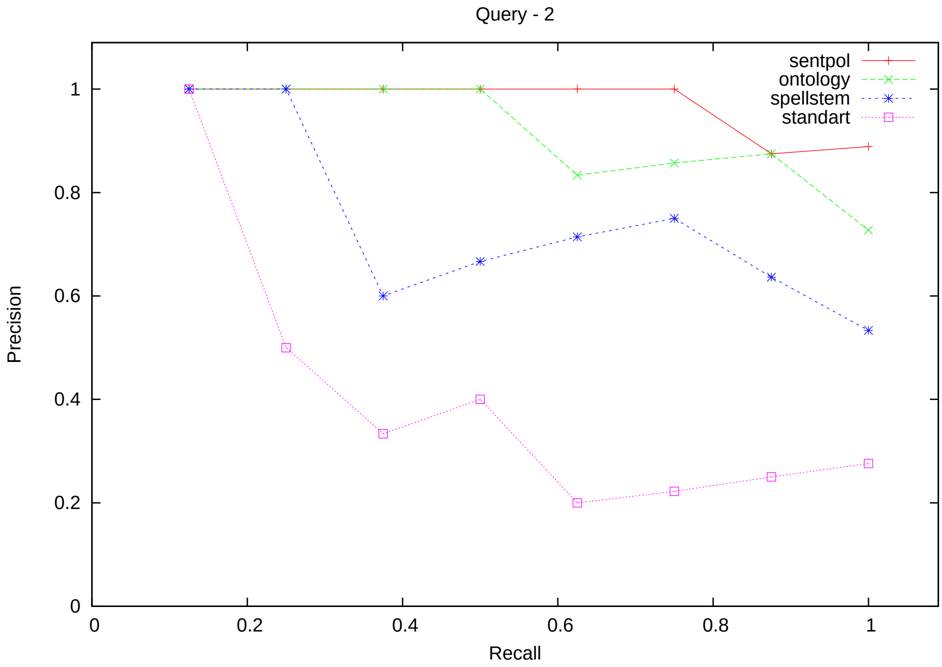

Figure D.2: Query:“karaciğerde kist”

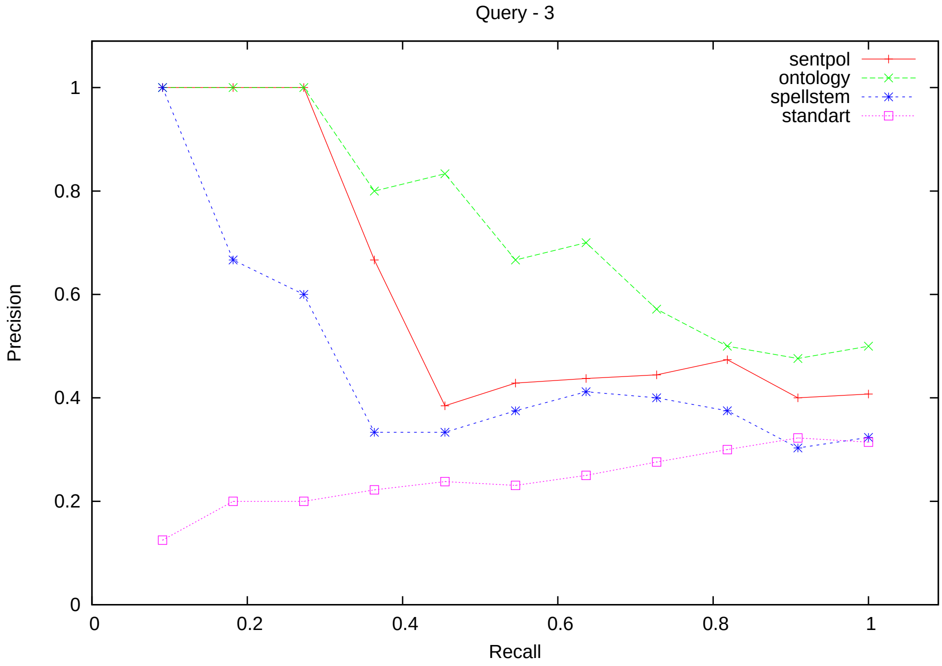

Figure D.3: Query:“karaciğerde kitle”

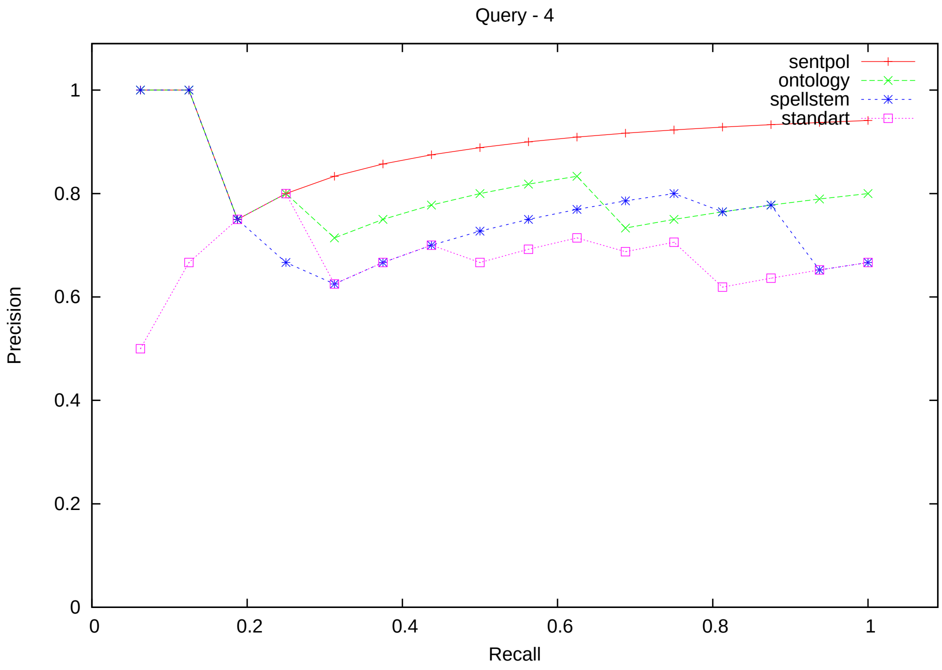

Figure D.4: Query:“böbrekte kist”

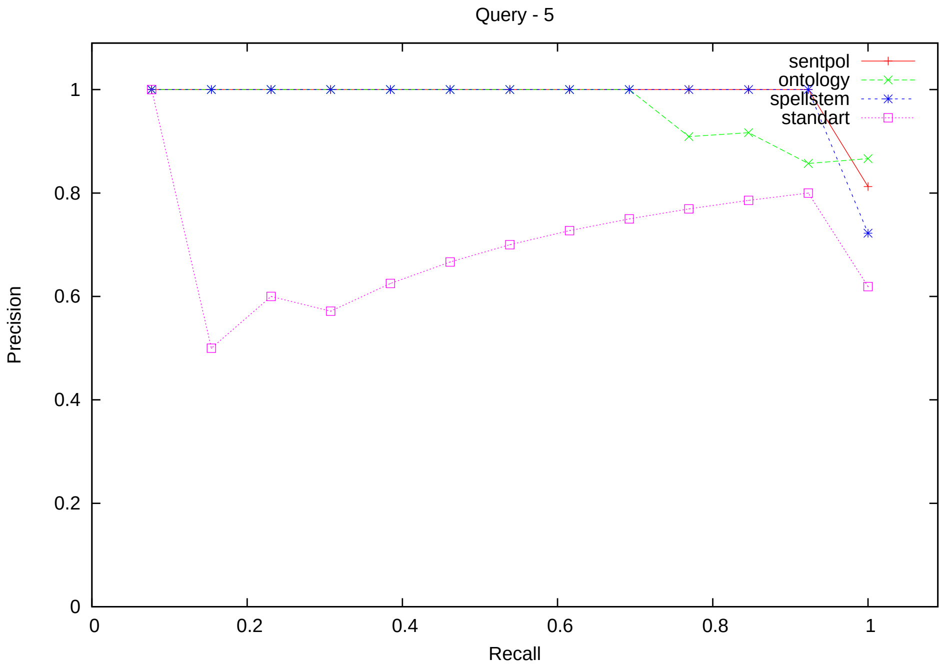

Figure D.5: Query:“böbrekte kortikal kist”

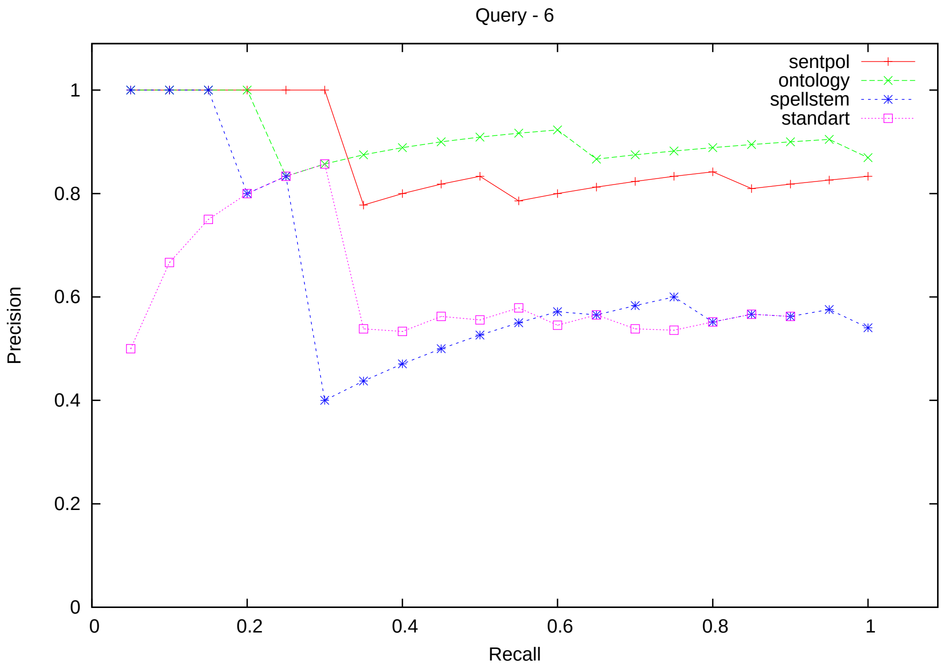

Figure D.6: Query:“böbrekte kitle”

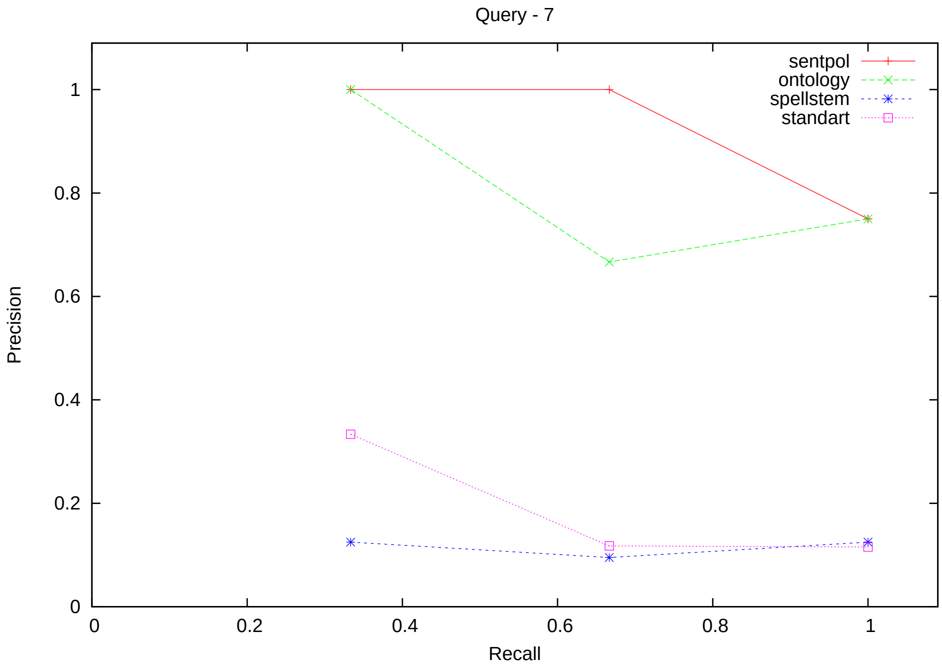

Figure D.7: Query:“böbrekte taş”

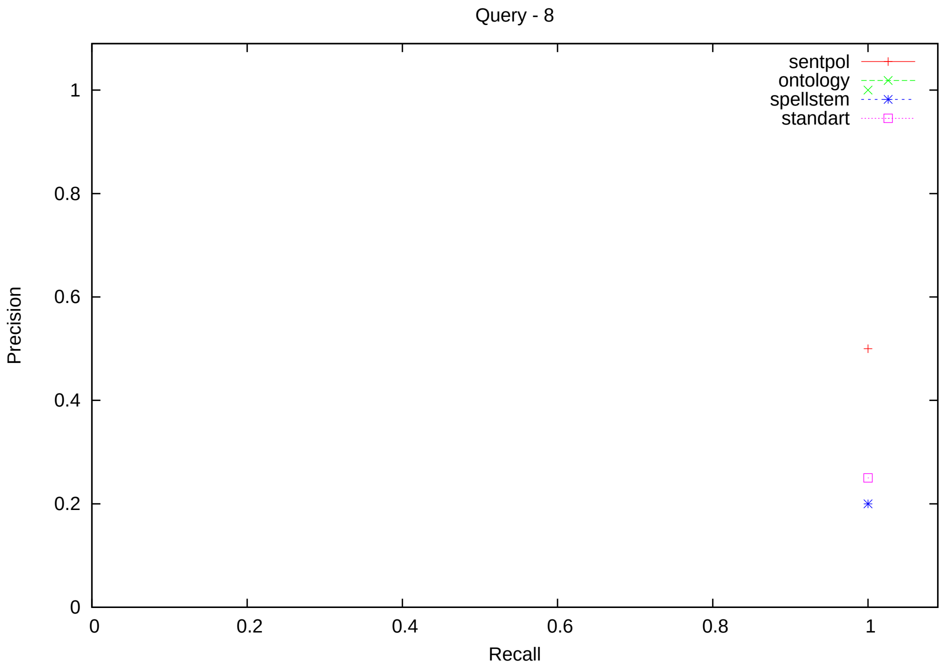

Figure D.8: Query:“sağ böbrek üst polde kortikal kist”

Appendix E

Sample report

ABDOMİNOPELVİK US

Klinik bilgi: Opere meme ca.

Bulgular: Karaciğerin konturu, büyüklüğü ve eko paterni normaldir. Karaciğer parankim ekojenitesi difüz minimal artmıştır. Karaciğer sağ lobda biri 9 mm diğeri 6 mm boyutlarında ölçülen 2 adet hiperekojen lezyon izlenmiş olup öncelikle hemanjiyom lehine değerlendirilmiştir. Hepatik ve portal venöz sistemler normaldir. İntra ve ekstrahepatik safra yolları normaldir. Safra kesesi boyutları ve duvar kalınlığı normal olup, lümen içerisinde taş ya da poliple uyumlu olabilecek görünüm saptanmamıştır.

Dalak konturu, büyüklüğü, parankim ekojenitesi ve eko paterni normaldir. Pankreas parankim kalınlığı, ekojenitesi ve eko paterni normaldir.

Her iki böbreğin konturu, büyüklüğü, lokalizasyonu, parankim kalınlığı, ekojenitesi ve eko paterni normaldir. Sinüs yankıları doğaldır. Toplayıcı sistem normal olarak değerlendirilmiştir. Sağ böbrekte parapelvik yerleşimli 11.5 mm çapında kist izlenmiştir.

Abdominal büyük damarlar, paraaortik, parakaval ve parailyak sahalar normaldir. İntraabdominal kitle, serbest mayi ve lenfadenopati saptanmamıştır.

Mesanenin kontur, kapasitesi ve duvar kalınlığı normaldir. Lümen içerisinde taş, ya da poliple uyumlu görünüm saptanmamıştır.

Uterus ve bilateral overler izlenmemiştir (opere).

<SONUC>
<SONUC>
Sonuç: Minimal hepatosteatoz, karaciğerde hemanjiyom ile uyumlu hiperekojen lezyonlar, sağ böbrekte kist, TAH+BSO.
</SONUC></SONUC>

Dr. Vural

Hacettepe Üniversitesi Hastaneleri Radyoloji Anabilim Dalı’nın radyolojik inceleme raporudur.