Computers in Biology and Medicine 40 (2010) 900–911

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

Design and evaluation of an ontology based information extraction system for radiological reports

Ergin Soysal [a,*], Ilyas Cicekli [b], Nazife Baykal [c]

[a] Ankara University, Dept. of Medical Education and Informatics, Ankara, Turkey
[b] Bilkent University, Dept. of Computer Engineering, Ankara, Turkey
[c] Middle East Technical University, Institute of Informatics, Ankara, Turkey

ARTICLE INFO

Article history:
Received 26 December 2009
Accepted 5 October 2010

Keywords:
Information extraction
Ontology
Natural language processing
Medical language processing
Turkish radiology reports

ABSTRACT

This paper describes an information extraction system that extracts and converts the available information in free text Turkish radiology reports into a structured information model using manually created extraction rules and domain ontology. The ontology provides flexibility in the design of extraction rules, and determines the information model for the extracted semantic information. Although our information extraction system mainly concentrates on abdominal radiology reports, the system can be used in another field of medicine by adapting its ontology and extraction rule set. We achieved very high precision and recall results during the evaluation of the developed system with unseen radiology reports.

1. Introduction

Health information systems and electronic health records are expected to lower costs and improve health care quality through improved access to information [1]. Free unstructured text is still the most common information source in medical records. Many medical disciplines such as radiology, pathology, and nuclear medicine almost completely rely on unstructured free text as the route of dissemination for information. This format is widely used for both storage and exchange of information about an individual patient, and the file of an individual patient usually contains several different free text reports such as clinical notes, patient history, or discharge summaries. Information covered in these reports is a valuable data resource for management, research, or educational purposes. Medical applications such as clinical decision support systems require utilizing this information. Nevertheless, this form of information is not as useful as structured and coded data for decision making nor knowledge discovery related to public health. Although the required information to answer many medical questions is stored electronically, we cannot answer precisely many questions like "What is the rate of non-pathological renal cysts in patients without renal complaints?", "What are the average sizes of left and right kidneys in our population?", and "How is renal parenchymal echogenic structure changing over the time, before a renal cancer is diagnosed?" since the required information is not available computationally.

As more and more text becomes available electronically, there is a growing need for systems that extract information automatically from narrative data. Manual extraction of this information is quite costly and time consuming process. As the text source grows, machine evaluation becomes mandatory to be able to use this huge amount of text. Information extraction (IE) and natural language processing (NLP) techniques are required to extract the useful information from these free texts.

Information extraction which is a sub-discipline of NLP focuses on the identification of the specific facts and relations within unstructured texts, the extraction of the relevant values, and their transformation into standardized codes and/or structured information. An information extraction task takes two inputs, namely a free text document that is the source of information and predefined templates, and fills these templates with suitable information extracted from the given document. The filled templates are the structured representation of the information available in the given document.

IE has become a popular research topic since late eighties by the promotion of Message Understanding Conferences (MUCs) sponsored by Defense Advanced Research Projects Agency (DARPA). The MUCs have a great impact on the research on information extraction. Many new IE problems have been identified, and the algorithms are developed to solve these problems. The MUCs have helped the development of the evaluation metrics that are used in the comparisons of the information extraction systems participated in the competitions.

* Corresponding author. Tel.: +90 312 508 3009.
E-mail address: esoysal@gmail.com (E. Soysal).

0010-4825/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiomed.2010.10.002

A typical information extraction system may have two main subtasks: entity recognition and relation extraction. Entity recognition tries to identify the boundaries of the text segments representing entities in natural language texts. For example, protein name extraction is an entity recognition task that tries to identify text segments representing protein names in medical texts. Relation extraction tries to identify the relations between entities in order to fill predefined templates. For example, the extraction of interaction relations among proteins is a relation extraction task. Both of these tasks use pattern matching techniques in order to extract the required information. The extraction rules that are generally regular expressions are applied to a given document in order to extract entities or relations.

A successful IE system at least relies to some degree on *domain knowledge* and some level of *grammatical information*. All the facts, relations and implicit assumptions of the domain, which are required to identify semantic entities and extract the information within the text properly, must be conveyed to the IE system. The success of a system closely correlates with the coverage of the required domain knowledge which is made available to the system as data sources. The domain knowledge is very complex and covers all of our world knowledge for general natural language texts, and the complexity of the required grammatical information for general natural language texts are complex as the whole grammar of that natural language. On the other hand, medical narratives are relatively easier to process from grammatical point of view because of their nature. Like many other technical subjects, medical texts also use a narrower subset of the language with limited number of information types [2], relatively unambiguous terminology [3] and predictable presentation patterns [4]. In other words, an information extraction system targeting a specific field such as medical texts which use a specific domain knowledge and sublanguage can be more successful than a general information extraction system because of the less ambiguity problem in those texts. Our information extraction system concentrates only on Turkish abdominal radiology reports that have less ambiguity problem, and its required domain knowledge is limited.

There are two basic approaches for information extraction: a supervised methodology, also known as *Knowledge Engineering Approach*, and an unsupervised (or semi-supervised) methodology referred as Automatic Training Approach [5]. In the supervised approach, extraction rules are manually developed by a domain expert or a knowledge engineer in consultation with a domain expert. The system performance is affected by the performance of the knowledge engineer and/or the domain expert. The main disadvantages of these systems are difficulties in the adaptation to another domain, and the requirement of a domain expert for the domain knowledge. On the other hand, it is expected to have a higher performance in comparison to automatic training approach, as a consequence of human intelligence in the construction of the system parameters. The information extraction system described in this paper uses a supervised methodology, and its extraction rules and ontology are developed by a domain expert.

In the unsupervised approach, IE system is trained by means of an annotated training set data using statistical approaches. For example, after manual annotation of entity names, the text can be used to train the system on named entity recognition. During the training period, the system may interact with a user to test whether the extracted data is correct or not, so that it can fix its rules accordingly [5]. One of the major obstacles in IE is the manual adaptation of an IE system to a newer domain since the manual adaptation is a costly process. The manual adaptation requires recreation of rule-sets and templates on the basics of the new domain. The difficulty of the domain knowledge creation for a new domain is another limitation for the performance. As a consequence of these problems, machine learning techniques for information extraction are viable alternatives, and they are discussed as a research topic for information extraction [6].

Traditionally, IE systems do not try a deep semantic analysis of all aspects of a text. They generally use pattern matching techniques such as finite state methods or regular expressions [7]. The ontology is a formal specification of a shared understanding of the domain of interest [8], and it is getting more popular to share knowledge across the systems. In IE systems, it is claimed that the use of a formal ontology as one of the system's resources improves the performance of entity recognition and semantic annotation tasks [9]. There are some published systems that use ontology during the information extraction task [10–14].

Ontologies are getting more and more popular to model knowledge in medical domain. OpenGalen is an initiative to create open source resources, which includes an ontology development environment and a large open source description logic-based ontology for the medical domain [15]. Rosse and Mejino published a reference ontology for functional model of anatomy (FMA) [16]. Another medical ontology, RadLEX, is derived from FMA, and it is extending FMA to cover radiological anatomy [17,18]. A related work RadiO was developed as a prototype application ontology to close the gap between radiology reports and RadLEX [19].

Our IE system uses ontology in both entity recognition and rule extraction. We use the ontology to determine not only the possible attributes, attribute values and entities appearing in the radiology reports, but also missing entities, attributes and attribute values in the sentences of the reports. In other words, we use the ontology to extract the semantic knowledge by disambiguating the sentences. Since our rules that are used in entity recognition and relation extraction contain ontological concepts, they have more expressive power than the rules based on textual items.

In this paper, we present a prototype IE system for Turkish radiology reports. Our system is designed to process all kinds of reports from different types of radiological examinations such as ultrasonography, magnetic resonance imaging, computerized tomography and plain X-Rays, and all of them are referred as radiology reports in this paper. Although the prototype system presented here is designed to handle the radiology reports from different sources, it is tested with abdominal ultrasonography reports. Our IE system converts a complete report into a target relational information model. The Turkish radiological information extraction system (TRIES) uses rules as grammatical knowledge and ontology as both domain knowledge for named entity recognition and semantic analysis. One of the main contributions of this paper is the usage of ontology in information extraction that increases the expressive power of extraction rules and helps to determine missing items in the sentences. Our system is the first information extraction system for Turkish texts. Since Turkish is a morphologically rich language, we use a morphological analyzer and our extraction rules are also based on the morphological features.

The rest of the paper is organized as follows. Section 2 discusses the related work in medical information extraction systems and ontology-based information extraction systems. In Section 3, we present the details of our ontology-based information extraction system. The performance results of our information extraction system are given in Section 4. We give the concluding remarks in Section 5.

2. Related work

After the initial introduction of information extraction approaches, the medical domain has become a popular application field for these systems. Many different research groups have emerged, mainly focusing on indexing reports as a free medical

text search facility, automatic term coding such as diseases or physical findings, and detection of abnormal conditions such as disease findings. Recently, many medical IE extraction systems have been developed using different approaches, and some of them are discussed in this section. The recall and precision values are frequently used in evaluation of performances of information extraction systems. The precision is calculated as the ratio of the relevant findings in all findings of the system, whereas, the recall is the ratio of relevant findings within the total numbers of all expected findings.

Linguistic String Project (LSP) [2,20] is one of the earliest rule based systems aiming to extract data from medical narratives to populate predetermined template slots, aiming to improve search on these texts. The project is based on a subset of natural language so called *sublanguage*. Since medical narratives only use a subset of natural language, LSP aims to recognize the texts in this sublanguage and uses the patterns that are specific to the sublanguage to achieve information extraction without a complete language processing. Additionally, LSP tries to code entities using SNOMED. For their evaluation test set, LSP showed a performance of a recall value of 82.1% and a precision value of 82.5%. Our IE system also uses a sublanguage that covers the sentence structures used in Turkish radiology reports.

Haug describes Special Purpose Radiology Understanding System (SPRUS) for the extraction of coded findings from free-text radiologic reports, and the evaluation results for the prototype system are reported as 87% recall and 95% precision [21]. The system mainly relies on semantic approach rather than syntactic methods. SymText is developed on top of SPRUS, extending its functionality to syntactic analysis of the text with different statistical methods [22,23] SymText is evaluated with the reports of acute pulmonary embolism patients. 92% recall and 88% precision values are achieved for making a diagnosis in chest radiography reports [24]. Our IE system also targets the radiology reports in Turkish.

Medical Language Extraction and Encoding System (MedLEE) [25] has been developed to extract clinical information from clinical texts. Its initial application domain was radiology reports. The system used a controlled vocabulary to code entries. The initial evaluation of this rule based system resulted in 85% recall and 87% precision results. Later Hripcsak evaluated MedLEE [26] to use the coded data for automated decision-support. The system was tested for identification of six medical conditions from radiology reports. Recall and precision were found to be 81% and 98%, respectively.

MENELAS is a multilingual medical language system, primarily focusing on discharge summaries and coding diseases using International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9-CM) [27]. The overall recall and precision results are measured at 48% and 63% on the coding task, and 66% and 77% on the questionnaire task, respectively.

MedSyndikate is developed to extract medical information automatically from findings reports in German language [28]. It uses a semi-automatic tool to acquire the domain knowledge. Its recall and precision values are found to be 93%. Recently, Mykowiecka et al. [29] have developed a rule based IE system for medical narratives in Polish. The system uses a syntactic parser and relies on ontology for named entity recognition. Its recall and precision values are over 80%.

Buitelaar et al. [10] describe an ontology based system named as SOBA, focusing on extraction of sports events from soccer web sites. The system transforms linguistic annotations into an ontology based representation, so that resources crawled from different web sites can be integrated to form a knowledge base.

Textpresso is another ontology based system, mainly aiming to index biomedical papers for better information retrieval from literature [13]. Ontology is used for term tagging and clarifying the underlying semantics – terms and relations among them – for the domain of interest. It has an overall performance of 94.7% and 30.4% for recall and precision in keyword search in full texts, whereas the same values are 44.6% and 52.3% in abstract search.

The IE system described in this paper, named TRIES, extracts the semantic information from Turkish radiology reports. In this sense, a sublanguage of Turkish is targeted in this information extraction task. TRIES is a rule-based system, and its extraction rules are hand-coded by a domain expert. Ontology is widely used in TRIES. The structure of TRIES ontology also determines the structure of its information model. The usage of ontological concepts in the extraction rules increased their flexibility. TRIES ontology is also used in the reference resolution problem in order to determine missing entities and attributes in sentences. To the best of our knowledge, TRIES is the first Turkish medical IE system. TRIES achieved 93% recall, and 98% precision results.

3. Ontology based information extraction

TRIES is an information extraction system aiming to parse free text Turkish radiological reports into computationally usable structured information. The major components of TRIES are given in Fig. 1. All the words in a given report are analyzed by a Turkish morphological analyzer. Each word is converted into a sequence consisting of a root word followed by possible morphemes. Morphological analyzer uses a lexicon, which is the source of lexical information for a set of Turkish root words. The root words of all possible words that can be seen in radiology reports are available in the lexicon, The words in the lexicon are grouped according to their functional properties of words such as verbs, nouns, adjectives, as well as abbreviations (e.g. units—mm, cm, ml, cc, mgr). In case of any failure during morphological analysis of a word in the report, the spell-corrector is invoked in order to fix a possible typing error. The fixed word returns back to morphological analyzer.

After the morphological analysis, a sentence can be seen as a sequence of root words and morphemes. Then, the entity recognition module recognizes some substrings of the sentence as terms, and marks them as a named entity term such as an ontological concept, an attribute, or an attribute value. TRIES ontology is designed at the conceptual level. The verbal representation of each ontological concept is maintained as a terminology attachment to conceptual ontology. These terms are commonly represented by morphological structures to let term analyzer to distinguish the morpheme belongs to term itself and the morphemes related to syntactic structures. In a sentence like

```
Safra kesesinde 3 mm taş izlendi.
Gall bladder 3 mm stone observed
(A stone of 3 mm was observed in gall bladder.)
```

Morphological analysis of *kesesinde* yields *kese*+POSS3SG+LOC (bladder+POSS3SG+LOC). During term analysis, the terminology part of ontology provides the Turkish term "safra kese+POSS3SG" as a representation of *GallBladder* entity. The remaining morphemes are attached to the newly formed term to be processed further during rule extraction such as *GallBlader*+LOC for the above example. So, the morphemes taking place in the formation of a named entity term are merged, and they are treated as a single unit after the entity recognition phase. The remaining morphemes are kept as modifiers. Turkish strings that can be named entity terms are determined with the help of the knowledge stored in the terminology part of TRIES ontology.

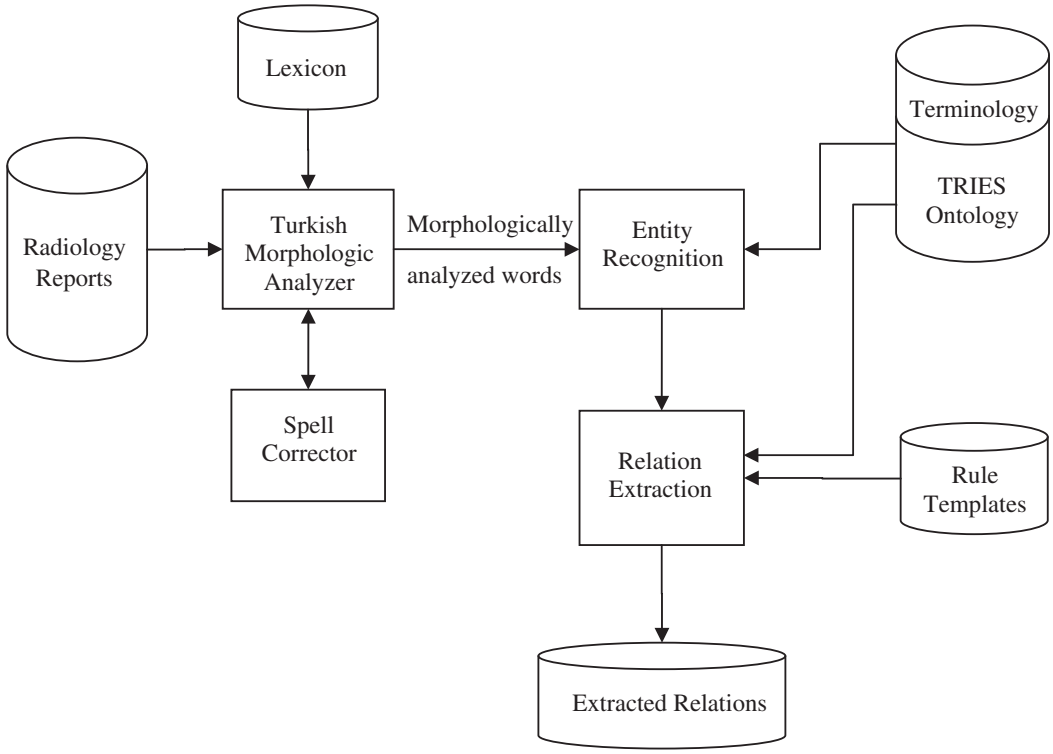

Fig. 1. Components of Turkish radiological information extraction system (TRIES).

Table 1
Application of TRIES to a sample sentence. (POSS3SG: possessive suffix for 3rd singular person, NESS: -ness suffix, COP: copula).

Text
Karaciğer vertikal uzunluğu 14 cm'dir.
The height of liver is 14 cm.
Morphological analysis
Karaciğer vertikal uzun+NESS+POSS3SG 14 cm+COP
Liver vertical tall+NESS+POSS3SG 14 cm+COP
Named entity recognition
[Karaciğer] [vertikal uzun+NESS] +POSS3SG [14 cm] +COP
[entity:Liver] [attribute:height]+POSS3SG [value:NUMERIC: 14 cm] +COP
Relation extraction—rule to be matched, and rule constraints to be satisfied:
〈VisibleStructure O〉 〈O:Attribute A〉 +POSS3SG 〈O:A:Value V〉 +COP
obj_has_attribute(Object, Attribute) – (Liver, height)
obj_attribute_accept_value(Object, Attribute, Value) – (Liver, height, 14 cm)
Extracted relation
Liver.height = 14 cm

In the next step, a sentence is processed by the relation extractor to match against TRIES rule templates, and the semantic information in the sentence is extracted as a set of relations. In the definition of the rule templates, the entity terms appearing in ontology are used in order to have more flexible rules. Rule templates may also utilize morphological elements to capture semantics gained by natural language grammar. So a typical rule template is made up of ontological concept elements and syntactic elements that are bound by regular expression elements.

Table 1 gives the steps of TRIES which is applied to a sample sentence. After morphological analysis, the sample sentence can be seen as a sequence of root words and morphemes. The TRIES entity recognition module recognizes the root word "*karaciğer*" (liver) as the ontological concept "*Liver*", the morpheme sequence "*vertikal uzun +NESS*" (height) as the attribute "*height*", and the sequence "*14 cm*" as an attribute value of "*NUMERIC*" type. After entity recognition, if the sentence matches a rule template and satisfies its rule constraints, a set of relations is extracted from that sentence. In our example, the entity "*Liver*" matches the entity "*VisibleStructure*" in the rule template since "*Liver*" is a sibling of "*VisibleStucture*" according to TRIES ontology. The attribute "*height*" matches the attribute field in the template, and it satisfies the rule constraint since "*height*" is an attribute of "*Liver*" according to our ontology. Similarly, the string "*14 cm*" matches the value field in the template, and it satisfies the rule constraint since the "*height*" attribute of the "*Liver*" entity accepts a numeric value as its attribute value. The relation "*Liver.height = 14 cm*" is extracted from the sentence since the sentence matches the rule template, and all the rule constraints are satisfied.

The infrastructure of TRIES is created from the reports of 756 abdominal ultrasonography (USG) examinations consisting of 11780 sentences. On the average, a report has 15.58 sentences and 107.23 words. Based on the examined reports, TRIES ontology is developed for abdominal USG reports by a domain expert. TRIES ontology consists of 740 items: 135 entities, 70 attributes, 56 topographical descriptions and the remaining items are the general terms that can be attribute values. After a manual processing of these reports, 150 grammatical rule templates based on ontological concepts are identified.

3.1. Turkish morphologic analysis

Turkish language is an agglutinative language, and it has very rich morphological structures. Many grammatical functions are represented by affixes in Turkish [30]. Since English language does not have such complex morphological structures, many NLP systems do not use morphologic analysis. On the other hand, the usage of the morphological analysis in Turkish systems increases their flexibility. In our IE system, recognizing morphemes enables it to handle words much more flexibly [31]. For example, the place of

a single accusative morpheme determines the whole meaning of the sentence in the following sentences:

Doktor, hastayı muayene etti (The doctor examined the patient)
Doktor hasta+ACC muayene et+PAST(Doctor patient+ACC examine+PAST)
Doktoru, hasta muayene etti (Patient examined the doctor)
Doktor+ACC hasta muayene et+PAST (Doctor+ACC patient examine+PAST)

TRIES has a Turkish morphological analyzer that looks like a PC-Kimmo [32] based morphological analyzer. As an initial preparatory step, morphological analyzer tokenizes the sentence into tokens. At this step, words, symbols, numeric expressions and punctuation marks are identified and marked by means of regular expressions. Then the words are taken into the analyzer. The morphological analyzer uses finite state methods (FSM) and its own restricted lexicon that is generated from the ultrasonography reports repository. We explicitly used a restricted lexicon for the morphological analyzer in order to reduce the amount of ambiguity. This analyzer parses a given word into possible morpheme combinations using its own lexicon. The lexicon provides the word roots together with their part of speeches such as noun, adjective, verbs, abbreviations, units, etc. The morphological parser can handle Turkish specific phonological rules such as vowel harmony, consonant softening and consonant doubling, and it uses a PC-Kimmo compatible phonological rules that are compiled by KGEN component of PC-Kimmo. It can also identify the different Turkish specific suffixes and use morphotactic rules in order to determine the morpheme sequence, based on the functional role of the word obtained from the lexicon.

The morphological analyzer is tightly coupled to a spell-corrector, so that it can fix some simple typing errors such as a missing letter, an extra letter, or two transposed letters. This integrated spell-corrector algorithm is developed to overcome typing errors that can break the pattern recognition tasks that are used during entity recognition or relation extraction. This integrated spell-corrector helped to improve the performance of our IE system.

In Turkish, the average number of morphological parses for a given word is 2.5. As a side effect, the morphological analysis introduces ambiguities [33]. The usage of the restricted lexicon in our morphological analyzer reduces the ambiguity problem for our system. Although we have a reduced ambiguity problem, still there are morphologically ambiguous words in our sentences. A separate sentence is created for each of the morphological parse combinations of the words, and they are processed by the other steps of TRIES in order to extract templates.

3.2. Ontology

TRIES ontology is created by examining 756 abdominal ultrasonography reports consisting of 11780 sentences in order to model the abdominal region organs that appear in the reports. The ontology currently contains 135 hierarchical entities with possible 70 attributes. In addition to entities and attributes, the ontology contains the terms that can be possible values for attributes. In TRIES ontology, currently there are 740 terms, and some of them are associated with a set of Turkish strings to indicate their representations in Turkish.

TRIES ontology entities implemented two types of relations. The former one, "Is a" relation creates the skeleton of TRIES ontology (Fig. 2), which is closely correlated to target information model for the extracted information. On the other hand, the next relation type is a family of relationship that helps to create parent–child relationships. The parts of the entities and other owned entities are linked to parent entity by means of a corresponding relationship specialized to for the target entity such as has_lobe, has_cyst or has_mass. By definition, these relationships may require varying instances for that particular entity class (e.g. one to one, or zero to many). This approach simplifies the relationship of ontology and information model, and the semantics of represented information. Furthermore, it plays an important role in the validation process of rule constraints.

TRIES ontology is created using Protégé ontology creation tool (Fig. 3) [34]. Entities inherit particular attributes in an *is_a* hierarchy. Entity–entity relationships other than *is_a*, are maintained by slots. For example, *Kidney* has several attributes inherited from its parent entities, and it also defines its own specific attributes. The *parenchyma* and *cyst* attributes of *Kidney* can be seen as the examples of specialized *part_of* relations. *Kidney* can have a single instance of *Parenchyma* (1 to 1), and it can also have multiple instances of *Cyst* (0 to many). These slots host proper instances of these entities at during rule extraction, satisfying the rule constraint conditions.

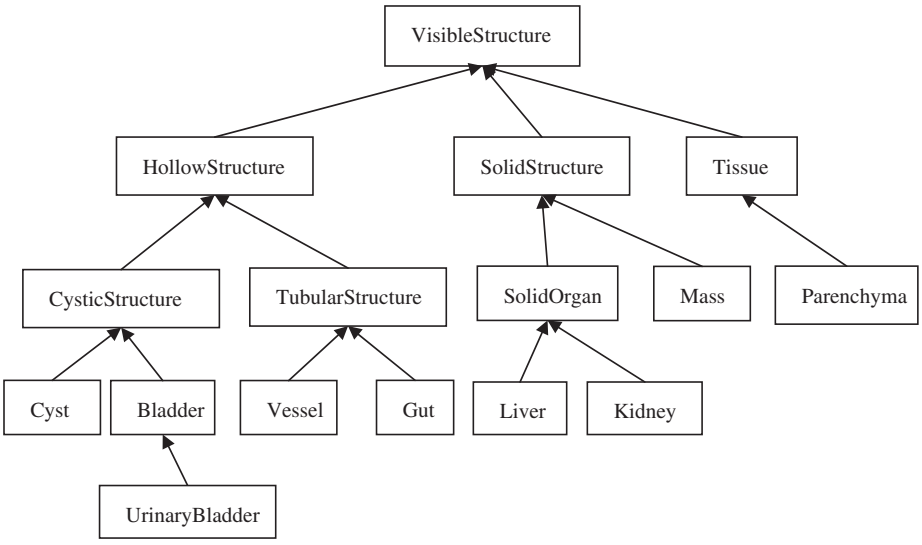

Fig. 2. An excerpt from TRIES ontology that was designed using Protégé: VisibleStructure is the parent for all other entities.

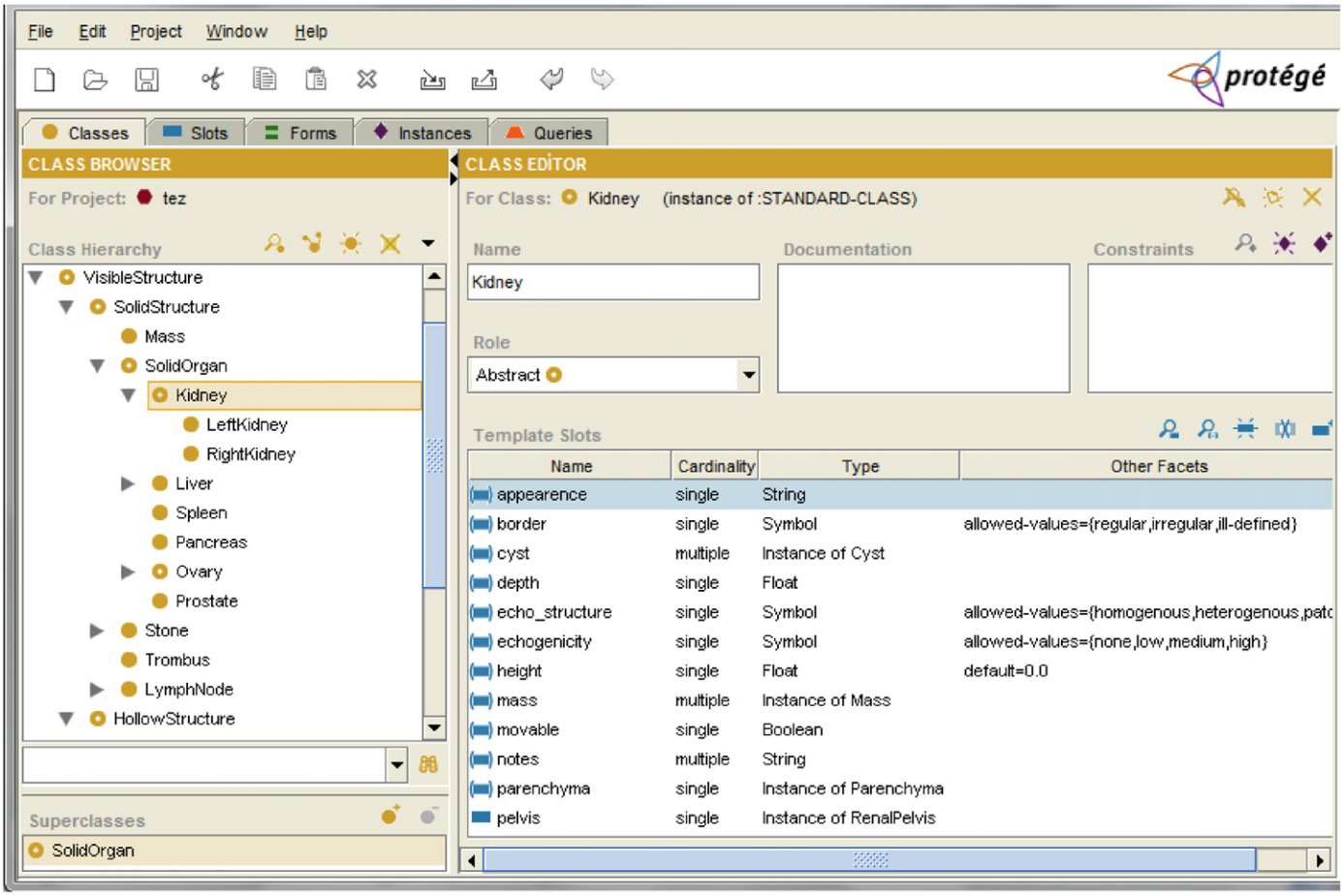

Fig. 3. TRIES ontology was designed by Protégé.

TRIES ontology is a tree, and its root is the entity *VisibleStructure* that is the parent entity of any visible structure on a radiological image. The top portion of TRIES ontology is given in Fig. 2. TRIES entities can be the entities that represent the observable visible structures on radiological images, or abstract entities that are used to create the entity hierarchy. Entities inherit all of the attributes of its parent. Furthermore, they may define their own additional attributes and override inherited attributes. For example, *Liver* entity is derived from *SolidOrgan* entity. So, it inherits all attributes of *SolidOrgan* entity such as *parenchyma*, *mass* and *cyst*. The entity *SolidOrgan* itself is derived from *SolidStructure* entity and *Liver* entity borrows some of its attributes such as *height*, *width* and *depth* from *SolidStructure* entity. *Liver* entity further defines its own attributes such as *bile_ducts* to model the information belonging to intra hepatic bile ducts.

Attributes of entities correspond to information slots in the extracted relations, and they may have strict or loose type checking to allow or disallow the assignment of an attribute value. This means that each attribute is associated with a set of constraints to limit the type of attribute values that it can take. The type of an attribute is one of the constraints, and it may be a simple type such as number, date, enumeration and string. An attribute type may also be some other entity name, or a collection of entity names defined within the ontology. So, the ontology also plays the role of controlled vocabulary for types. For example, if the type of an attribute is the simple type *NUMERIC*, it means that it can only be instantiated with a numeric attribute value. On the other hand, since the *parenchyma* attribute is typed as *Parenchyma* entity in TRIES ontology, the *parenchyma* attribute of *Liver* entity can only be bound to an instance of *Parenchyma* entity with its own instantiated attributes.

When some of the attributes of an entity are associated with values, it is called as an instantiated entity. An instantiated entity may define a non-empty set of relations in the extracted information. Although the instances of some entities can directly appear in the extracted information, the instances of some other entities cannot be directly seen, and their instances must be the attribute values of other instantiated entities. We refer the first group entities as normal entities and the latter as *sub-entities* because their instances can only be attribute values. For example, *Liver* entity is a normal entity, and its instances can directly appear as a set of extracted relations. On the other hand, *Parenchyma* is marked as a sub-entity in TRIES ontology because its instance can be a value of the *parenchyma* attribute of an instantiated *Liver* entity.

TRIES ontology requires to model a collection of items such as the *cyst* attribute of *Liver* entity. An attribute value can be a collection of instantiated instances of sub-entities. For example, the *cyst* attribute of an instantiated *Liver* entity is a collection of instantiated instances of *Cyst* sub-entities. Table 2 gives some attributes of *Liver* entity together with their types and sources.

Entities in TRIES ontology are also categorized as *instantiable* and *abstract* entities depending on whether their instances can be creatable or not. The instances of *instantiable* entities can be creatable, and they are further categorized as *standalone* entities and *sub-entities*. The instances of standalone entities are directly represented as a set of relations in the TRIES information model. On the other hand, the instances of sub-entities can only be attribute

Table 2
Some attributes of Liver class with attribute types and sources.

Attribute	Type	Attribute source
Border	ENUM	VisibleStructure
Height	NUMERIC	SolidStructure
Width	NUMERIC	SolidStructure
Parenchyma	Parenchyma	SolidOrgan
Cyst	Collection	SolidOrgan

values of other instantiable entities. The usage of sub-entities makes it easy to model the relations in the form of

Entity.entity2.attribute2 = value2

where *Entity* is the name of an instantiated standalone entity with the attribute *entity2*. The value of the attribute *entity2* is an instance of a sub-entity *Entity2*, and that instance contains an attribute named as *attribute2* with a value named as *value2*. The approach that we use for sub-entities is similar to the model defined by Archbold and Evans [4].

The instances of *abstract* entities cannot be created. They help to organize TRIES ontology, and their siblings inherit the attributes that are defined for them. Of course, each abstract entity must have at least one *instantiable* entity as its sibling. In fact, all inner nodes in TRIES ontology are abstract entities and all leaves are instantiable entities.

The strings representing abstract entities often appear in radiology reports, and they cause ambiguity. Let us consider the following example:

Safra kesesi normal boyuttadır. (The size of **gall bladder** is normal.)
Kese içinde taş ya da kitle izlenmedi. (Stone or mass is not observed inside the **bladder**.)

The expression "*bladder*" may be used as a shorthand for either "*gall bladder*" or "*urinary bladder*". This ambiguity must be resolved before the semantic information is extracted from these sentences. TRIES handles this ambiguity problem through abstract entities. At the entity recognition level, these terms are recognized as abstract entities. For example, TRIES entity recognition module recognizes the Turkish string "*safra kesesi*" as the entity *GallBladder* which is an instantiable entity, and the string "*kese*" as the entity *Bladder* which is an abstract entity. During the relation extraction, an abstract entity is replaced by one of its proper instantiable offspring entities using the context information. In our example, *Bladder* abstract entity is replaced with *GallBladder* instantiable entity because *GallBlader* is an offspring of *Bladder*, and it appears in the previous sentence.

Another kind of ambiguity that is caused by a string representing an abstract entity is that the string can refer to all instantiable siblings of that abstract entity. In order to solve this problem, the abstract entities whose usages in the reports refer to all of its possible instantiable siblings are marked as *propagable* entities. Although an instance of a propagable abstract entity is not created, any value assigned to the attributes of this entity is propagated to siblings. In other words, the instances of its instantiable siblings are created, and all assigned values are copied into these instances. For example, the abstract entity *Kidney* is *propagable*, and all assigned values are copied into the instances of its instantiable siblings *LeftKidney* and *RightKidney*. When TRIES considers the following sentence, the Turkish string "*böbreklerin*" is recognized as the entity *Kidney* by the entity recognition. All extracted attribute values from this sentence are copied into the instances of *LeftKidney* and *RightKidney* entities, and the following relations are extracted:

Böbreklerin büyüklükleri, şekilleri ve yerleri normaldir.
Kidneys are normal in sizes, shapes and locations.
Extracted relations
LeftKidney.size=normal
LeftKidney.shape=normal
LeftKidney.location=normal
RightKidney.size=normal
RightKidney.shape=normal
RightKidney.location=normal

3.3. Named entity recognition

After all words in a sentence are broken into their morphemes, the sentence is passed to the entity recognizer. The entity recognizer identifies phrases as named entities together with their named entity type. TRIES supports five types of named entities:

Entity—Strings representing ontology entries such as organs and major vessels are recognized as named entities of type *Entity*. In fact, any entity that is not a sub-entity in TRIES ontology is recognized as *Entity*.
Sub-Entity—A string representing an entity that is marked as a sub-entity in TRIES ontology is recognized as a named entity of type *Sub-Entity*.
Attribute—Strings representing the defined attributes in the ontology are recognized as named entities of type *Attribute*.
Value—The possible attribute values are recognized as named entities of type *Value*, and the types of value strings are also determined.
Location—Strings representing topographic locations are recognized as named entities of type *Location*, and they are also used as attribute values.

The strings that are recognized as named entities are packed as a single unit, and replaced with appropriate named entities. The information about all strings that represent named entities is stored in TRIES ontology, and entity recognition module uses this information together with simple regular expressions to determine the named entity strings. Some of the ambiguity introduced at the morphological analysis level is eliminated by the help of this process.

3.4. Information model

One of the main problems for IE systems in medical domain is the proper computational usability of the extracted information. An information model for TRIES is created based on domain expert opinions (Radiologist and Clinician) and guidelines of Turkish Ultrasonography Association. This is a key challenge for the usability of the extracted data for decision making and knowledge discovery. The solution to this problem is achieved by means of domain experts. TRIES ontology is heavily influenced by the target knowledge structures. The complete information model is integrated into the ontology as entities and attributes. So, the ontology also hosts the information model for TRIES. The information extracted from a sentence is populated from the instances of entities of TRIES ontology.

The extracted information is represented as a set of relations. Each relation represents an attribute with its value. Of course, the entity that owns the attribute also appears in the relation.

A relation is in the following form:

$Entity.attribute_1.\ldots..attribute_n.simpleattribute = simplevalue$

where $attribute_1.\ldots..attribute_n$ is optional, *Entity* is an instantiable entity, *simpleattribute* is an attribute whose value cannot be an entity instance and *simplevalue* is its value. If $attribute_1.\ldots..attribute_n$ are present, all of them are attributes whose values can be the entity instances, $attribute_1$ is an attribute of *Entity*, each $attribute_{i+1}$ is an attribute of $attribute_i$, and *simpleattribute* is an attribute of $attribute_n$.

3.5. Relation extraction and rule templates

The set of rule templates is a classical component of an information extraction system. TRIES uses a set of rule templates that are manually extracted by means of a domain expert. Each rule template is combined with a set of constraints to further eliminate ambiguities. Rule templates in our system correspond to grammar rules. These rule templates are also tightly integrated with TRIES ontology. Ontology entities are used in both expressions and constraints of the rule templates. Each rule template may have additional constraints such as "may this object have this attribute?" or "may this attribute of this object have this value?". A rule template is a regular expression that consists of entities from TRIES ontology. For example, the following is a simple rule template:

〈VisibleStructure O〉 〈O:Attribute A〉 +POSS3SG 〈O:A:Value V〉 +COP

This rule template matches sentences that start with a *VisibleStructure* entity O (i.e. any entity in TRIES ontology since *VisibleStructure* is the root of the ontology tree), and continues with an attribute A that can be an attribute of the entity O and the morpheme "*+POSS3SG*". The sentence must finish with an attribute value V that can be taken by the attribute A, and the morpheme "*+COP*". There is also an implicit constraint, and it says that O must be an instantible entity. If this rule matches a sentence, the relation "O.A=V" is extracted.

Some words or punctuations usually denote a set of similar grammatical functions. For example, the comma and the Turkish conjunction word "ve" (and) play similar grammatical roles in Turkish sentences. TRIES rules also support macros, which are used for some sort of shorthand, and expand to full instructions. For example, a list of similar items can be expressed as a macro. A rule template using macros is given in Table 3. The first row gives the defined macros, the second row gives the rule template, the third row gives some sample sentences that can match this rule template, and the last row gives the extracted relations from these sentences. In the third row, the sentences are given together with their forms after the entity recognition (the morphologically analyzed Turkish words are not given for simplicity reasons). This rule template can match a sentence, if and only if the matched entity must accept all the attributes in the list item, and all the attributes in the list item must accept the matched value in the sentence.

3.5.1. Reference resolution

The reference resolution is one of the most important problems in the relation extraction. TRIES uses a context mechanism integrated into the relation extractor in order to solve the reference resolution problem. This context mechanism keeps track of the ontology entities appearing in the sentences in a stack, and tries to estimate the missing (omitted) terms along the sentences using this stack. Whenever the relation extractor faces a missing entity, the context is taken into account in "last in first out" fashion. The extractor tries to estimate the missing entity by referencing ontological properties of entities within the context. In some cases, TRIES ontology is used alone to solve some of the reference resolution problems. The reference resolution is an important utility to further overcome ambiguity.

Table 3
A sample rule and real life sentences matching this rule (LOC: locative suffix, COP: copula). 〈O:A:Value V〉 term will be assigned to the list of given list of attributes to an ontology entity derived from VisibleObject, if entity O possesses these attributes.

Macros:
CONJ={ ",", "and"}
LIST(X)=X [〈CONJ〉 X]*

Rule template:
[〈VisibleStructure O〉]? 〈O:A:Value V〉 LIST(〈O:Attribute A〉) +LOC +COP

Sentences:
Abdominal aorta normal görünümdedir (Abdominal aorta is in a normal appearance)
[AbdominalAorta] [normal] [appearance] +LOC +COP
Böbrekler normal boyuttadır (Kidneys are normal in sizes)
[Kidney] [normal] [size] +LOC +COP
Dalak 10.5 × 2.5 cm boyuttadır (Spleen is 10.5 × 2.5 cm in size)
[Spleen] [10.5 × 2.5 cm] [size] +LOC +COP
Karaciğer normal şekil ve boyutlardadır (Liver is normal in shape and size)
[Liver] [normal] [shape] 〈CONJ〉 [size] +LOC +COP

Extracted relations:
AbdominalAorta.appearance=normal
LeftKidney.size=normal
RightKidney.size=normal
Spleen.size=10.5 × 2.5 cm
Liver.shape=normal
Liver.size=normal

In some cases, the well known entity attributes can be omitted. For example, although the entity *LeftKidney* and the attribute value *smaller_than_normal* are available, the *size* attribute is missing in the following sentence:

```
Sol böbrek normalden küçüktür. (Left kidney is
  smaller than normal.)
[entity:LeftKidney]
  [value:string:smaller_than_normal] +COP
```

Although this sentence is grammatically and semantically a normal sentence, the extracted attribute value must be assigned to the attribute *size* according to the information model, and the relation "*LeftKidney.size=smaller_than_normal*" must be extracted. But this attribute is not present in the sentence, because it is very-well known by a human reader. In order to determine the missing attribute, TRIES ontology is used to find an attribute of *LeftKidney* such that the found attribute accepts *smaller_than_normal* as its value.

In some cases, entities themselves are missing in the sentences. An instantiable entity does not appear in the last two of the following three sentences, and it must be found using the context information.

```
Karaciğer sağ lob vertikal uzunluğu 17 cm'dir.
(Liver right lobe vertical length is 17 cm.)
[entity:Liver] [subentity:RightLobe]
  [attribute:height]+POSS3SG [value:string:17 cm]
  +COP
Parankim ekosu steatozla uyumlu olarak
  artmıştır.(Parenchymal echo is increased in
  accordance with steatosis.)
[subentity:Parenchyma] [attribute:echogenity]
+POSS3SG [value:string:steatosis] +LOC
uyumlu olarak [value:string:increased] +COP
```

```
Kitle içermemektedir. (It does not contain a mass.)
[subentity:Mass] [value:string:not_exist] +COP
```

The instantiable entity *Liver* is mentioned in the first sentence, but it is not mentioned in the next two sentences. Thus, the missing instantiable entity *Liver* in the last two sentences is deduced with the help of the context mechanism. The second sentence contains two attribute values, but it contains only one attribute. This means that one attribute is missing. Since the attribute *echogenity* can get the attribute value *increased* in that sentence, it is associated with that value. In order to find out the missing attribute, a *Parenchyma* attribute that can accept the attribute value *steatosis* is searched among *Parenchyma* attributes using the knowledge available in TRIES ontology. Since the *impression* attribute satisfies this constraint, it is identified as the missing attribute. The third sentence has also a missing attribute. That missing attribute is similarly found, and it is identified as the *appearance* attribute of *Mass* sub-entity. After all reference resolutions are determined, the following relations are extracted from the three sentences given above:

Liver.rightlobe.height = 17 cm
Liver.parenchyma. echogenity = increased
Liver.parenchyma.impression = steatosis
Liver.mass.appearance = not_exist

The resolution problem will be even worse if we append the following sentence to the sentences above:

```
Parenkim homojendir. (Its parencyma is homogeneous.)
[subentity:Parenchyma] [value:string:homogenous]
 +COP
```

In this sentence, there is a sub-entity, namely *Parenchyma*, but there is not any main entity or attribute. The main entity will be found with the help of context information, and the missing attribute will be found with the help of ontology. According to ontology and context information, this sentence must be presented as "*Liver* parenchymal *echogenic structure* is homogenous". In other words, the missing entity is *Liver*, and the missing attribute is *echogenic structure*.

The relation extractor refers to the ontology as a source of domain knowledge for the resolution of some more issues like disparities of verbal expressions and the information model. In the following two sentences, there are such disparities.

```
Barsak duvarlarında aşikar duvar kalınlığı
izlenmedi. (A prominent thickening was not observed
in the intestinal wall.)
[entity:Intestine] [subentity:wall]
 [attribute:thickeness] [value:string:not_exist]
Karaciğer parenkim görünümü homojendir. (Liver
 parenchymal appearance is homogeneous.)
[entity:Liver] [subentity:Parenchyma]
 [attribute:appearance]
 [value:string:homogeneous]
```

In the first sentence, the attribute *thickness* does not accept the attribute value *not_exist*. The acceptable values of the attribute *thickness* are searched in order to determine whether one of them has similar meaning with that value in this context, or not. An acceptable value *normal* for the attribute *thickness* is spotted, and the attribute value *not_exist* is replaced with this new found value. The second sentence has also a similar problem. Here, the attribute value *homogeneous* is not an acceptable value for the attribute *appearance*, and the *Parenchyma* sub-entity does not have the *appearance* attribute. In this case, the attributes of the *Parenchyma* sub-entity are searched to find an attribute that has a similar semantic meaning with the attribute *appearance* in this context, and accepts the attribute value *homogeneous*. Thus, the attribute *echogenic_structure* is identified, and it replaces the attribute *appearance* in the second sentence. After all reference resolutions are resolved, the following relations are extracted from the sentences above:

Intestine.wall.thickness = normal
Liver.parenchyma. echogenic_structure = homogeneous

Sometimes, entities or attributes are expressed as if owned by other entities. In the following sentence, although *diverticulum* attribute belongs to the *wall* sub-entity of urinary bladder, it is referred as an attribute of bladder itself:

```
Mesanede 2 cm çaplı divertikül izlenmiştir. (In
urinary bladder, a diverticulum in 2 cm diameter was
observed.)
[entity:UrinaryBladder] +LOC
 [value:numeric:2 cm]çaplı
 [attribute:diverticulum] izlenmiştir
```

It looks like the sentence contains all the required named entities. The relation extractor can determine that there is a missing sub-entity attribute by observing that *UrinaryBladder* cannot have the attribute *diverticulum* but its sub-entity *Wall* can have it. With the help of the ontology, the relation extractor can model the information in this sentence as the following relation:

UrinaryBladder.wall.diverticulum = 2 cm

Since the extracted data may be required in different formats for different purposes, some attributes may require multiple entries for a single value. For example, *size* is a common attribute frequently used for entities derived from *SolidStructure* either with qualitative values such as "decreased", "slightly increased", etc. or quantitative values at one to three dimensions such as 10.5×2.5 *cm*. These multidimensional values represent length, width and depth for the given entity. *SolidStructure* also have separate attributes for *length*, *width* and *depth*. For the consistency of extracted data, this multidimensional *size* must be separated into corresponding dimension attributes. TRIES completes this by an optional post-processing. Although this obviously results in redundancy of data, this is a required step for data consistency.

As a rule based system, semantics are fixed by the rules in TRIES. The negative meanings in Turkish are expressed using negation morpheme attached to verbs. The rule templates containing the negation morpheme are used to recognize negative information in clinical reports. For example

Karaciğer kitle içermemektedir (Liver does not contain a mass)
Liver mass içer+NEG+PRESENT+COP ("içer" means *contain* in English)

The negation morpheme attached to the verb "içer" indicates the negative information. This negative information is represented with "not_exist" attribute value, and the extracted information from this sentence will be as follows:

Liver.mass = not_exist

4. Evaluation

For the performance evaluation of TRIES, 100 radiology reports are randomly selected as unseen data. On the average, each report is composed of 14.34 sentences and 105.43 words. The configuration of the system was frozen prior to analyzing the test set. A human domain expert is considered as the gold standard, and the domain expert extracted the relations from these 100 reports. Then, the relations extracted by TRIES are compared against the relations extracted by the domain expert. Table 4 summarizes how the extracted relations are classified. A relation that is extracted by both the domain expert and TRIES is classified as TP (true positive), and a relation that is extracted by TRIES but not extracted by the domain expert is classified as FP (false positive). A relation that is extracted by the domain expert but not TRIES is categorized as FN (false negative).

For the evaluation of IE systems, recall and precision values are frequently used [35]. The *recall* of an information extraction system can be defined as the ratio of the number of relevant findings returned to the total number of findings that are present. The *precision* is the ratio of the number of relevant findings returned to the total numbers of all findings returned. The recall and precision can be formulated in terms of TP, FP and FN as follows:

$$Recall = \frac{TP}{TP+FN} \quad Precision = \frac{TP}{TP+FP}$$

Table 5 gives the evaluation results of TRIES. For the evaluation set, the average number of extracted relations for each report is 51.7. For all extracted relations, the overall recall value is 93% and the precision value is 98%. This means that only 2% of the extracted relations are incorrect, and only 7% of the available information is not extracted.

In addition to the general performance of TRIES, its performances in specific cases are also measured and they are given in the rows 2–5 of Table 5. The average number of relations extracted from the sentences containing non-propagable abstract entities is 0.9 per report. In this group of extracted relations, a recall of 92% and precision of 98% have been achieved. Although some sentences contain both an attribute and an attribute value, the appearing value may not be the proper value for the attribute. In those sentences, the attribute value is assigned to another attribute that is found with the help ontology (e.g. parenchymal appearance is homogeneous; appearance mapped to echo structure). For those sentences, the average number of extracted relations is 2.5 per report, the recall and precision values are 91% and 97%, respectively. The average number of extracted relations from the sentences containing missing entities or attributes is 8.1 per report. For this group of sentences, the recall and precision values are 92% and 98%, respectively. Finally, for the group of sentences where attribute values are given by means of a general parent class (e.g. Kidneys are normal in size, instead of declaring left and right kidneys separately), the average number of extracted relations are 21.6 per report, and recall value is 93% and precision value is 98%. These numbers indicate that the performances of our system in special cases are very similar to its overall performance.

SpellCorrector has a prominent contribution to the success of information extraction. Many typing errors that might break the patterns are automatically fixed at the rate of 91% of all misspelled words. The detected errors contain only one error belonging to one of the following cases: a missing letter (25%), an extra letter (39%—frequently doubling of the same letter), a wrong letter (17%—including Turkish letter) and finally two adjacent letters interchanged (9%).

Table 4
Evaluation table (DE: domain expert, TP: true positive, FP: false positive, FN: false negative, TN: true negative).

	Extracted by DE	
	Yes	No
Extracted by TRIES	**TP**	**FP**
Yes	**TP**	**FP**
No	**FN**	**TN**

5. Conclusion

In this paper, we introduced an information extraction system TRIES that uses an ontology as the domain knowledge for Turkish radiological reports. The ontology is main source of domain knowledge in TRIES. It is referenced by the term analyzer in the named entity recognition phase, by the relation extractor in pattern matching, and by the target information model. TRIES uses domain ontology to incorporate the knowledge of relevant concepts and their semantic relations into the system. TRIES uses hand-coded rule templates as grammatical expressions. The extracted semantic knowledge is constrained by the rule templates, the rule constraints and the ontological relations used within the rule templates. The usage of ontology concepts provides flexibility in the design of rule templates. The structure of TRIES ontology also determines the information model that describes the structure of the extracted semantic information.

TRIES ontology is used in not only the design of the relation extractor, but also the resolution of the ambiguities caused by the missing terms in the sentences. Some of the missing terms are determined by the constraints implied by TRIES ontology. A context mechanism that holds the history of referred entities is also used to figure out the missing terms.

The use of ontology is an important tool for the adaptation of the system to another domain. TRIES ontology is relatively a small ontology designed to model the concepts appearing on abdominal ultrasonography reports. A future work may concern a statistical formation of a bigger ontology to model all the concepts appearing on different radiology reports.

Table 5
Average numbers of attributes per report, recall and precision values.

	n per report	***Recall* (%)**	***Precision* (%)**
Total extracted relations	51.7	93	98
Relations extracted from sentences containing non-propagable abstract entities	0.9	92	98
Relations extracted from sentences containing attribute value mapped to another attribute	2.5	91	97
Relations extracted from sentences containing missing entity or attribute	8.1	92	98
Relations extracted from sentences containing propagable abstract entities	21.6	93	98

Since English morphological structure is not too complex, the morphological analysis is overlooked in most of the IE systems designed for English texts. On the other hand, the morphological analysis is an important part of TRIES, since Turkish has a rich morphological structure. The morphological analysis in TRIES increases the flexibility of entity recognition and relation extraction.

TRIES achieved 93% recall and 98% precision results in the performance evaluations. The scores are very high when compared with other IE systems. The reason for these high scores can be the usage of effective hand-coded rules and ontology in the information extraction.

Information that is extracted by TRIES can be utilized by various applications such as research tools, text summarization, information visualization or report validation during report entry. But further applications of TRIES should regard that TRIES extract explicitly expressed information in reports and not the implied ones.

Ontology is the most important component of TRIES. It is not a general purpose ontology, besides, specifically developed regarding the knowledge requirements of an information extraction system and how the entities are described in reports with a point of angle of domain experts. In near future it will continue to expand to include other body parts, and enriching relationship types, entity and attribute set based on.

6. Summary

Free texts are still the main source of information in medical domain, and they are widely used for both storage and exchange of information. Nevertheless, this form of information is not as useful as structured and coded data for decision making nor knowledge discovery related to public health because of computational inaccessibility of the information in unstructured reports. Since the access to the information in free texts requires extensive efforts, information extraction systems can reduce this inaccessibility problem by converting unstructured data into structured data.

There are two basic approaches for information extraction: a supervised methodology and an unsupervised. In the supervised approach, extraction rules are manually developed by a domain expert or a knowledge engineer in consultation with a domain expert. In the unsupervised approach, IE system is trained by means of an annotated training set data using statistical approaches. The usage of effective hand-coded rules is still one of the best approaches in order to get a medical information extraction system with high precision and recall values. For this reason, we preferred to use hand-coded extraction rules in our information extraction system.

This paper describes an information extraction system that is designed to process free text Turkish radiology reports in order to extract and convert the available information into a structured information model. The system uses natural language processing techniques together with domain ontology in order to transform verbal descriptions into a target information model so that they can be used for computational purposes. The developed domain ontology is effectively used in entity recognition and relation extraction phases of the information extraction task. The ontology provides the flexibility in the design of extraction rules, and the structure of the ontology also determines the information model that describes the structure of the extracted semantic information. In addition, some of the missing terms in the sentences are identified with the help of the ontology. One of the main contributions of this paper is the usage of ontology in information extraction that increases the expressive power of extraction rules and helps to determine missing items in the sentences. Our system is the first information extraction system for Turkish texts. Since Turkish is a morphologically rich language, we use a morphological analyzer and our extraction rules are also based on the morphological features.

Our information extraction system extracts the structured information from Turkish radiology reports using manually created rules and domain ontology. Although our prototype information extraction system mainly concentrates on abdominal radiology reports, the system can be used in another field of medicine by adapting its ontology and its extraction rule set. We achieved very high precision and recall results during the evaluation of the developed system with unseen radiology reports.

Conflict of interest statement

Authors declare that they do not have any conflict of interest with any people or organization.

Acknowledgment

We would like to thank to Prof. Dr. Serdar Akyar, Prof. Dr. Mustafa Özmen and Prof. Dr. Utku Şenol for their support and their allowance to access to radiology reports during our research.

References

[1] J.M. Corrigan, M.S. Donaldson, L.T. Kohn, S.K. Maguire, K.C. Pike, Crossing the Quality Chasm: A New Health System for the 21st Century, National Academy Press, Washington, DC, 2001.
[2] N. Sager, C. Friedman, M.S. Lyman, Medical Language Processing: Computer Management of Narrative Data, Addison-Wesley Longman Publishing Co., Boston, MA, 1987.
[3] A.M. Rassinoux, J.C. Wagner, C. Lovis, R.H. Baud, A. Rector, J.R. Scherrer, Analysis of medical texts based on a sound medical model, in: Proceedings of the Annual Symposium on Computer Application in Medical Care, 1995, p. 27.
[4] A.A. Archbold, D.A. Evans, On the Topical Structure of Medical Charts, in: Proceedings of the 13th Annual Symposium on Computer Applications in Medical Care, IEEE Press, Washington, DC, 1989, pp. 543–547.
[5] D.E. Appelt, Introduction to information extraction, AI Communications 12 (3) (1999) 161–172.
[6] J. Turmo, A. Ageno, N. Català, Adaptive information extraction, ACM Computing Surveys (CSUR) 38 (2) (2006) 4.
[7] D.A. Evans, N.D. Brownlow, W.R. Hersh, E.M. Campbell, Automating concept identification in the electronic medical record: an experiment in extracting dosage information, in: Proceedings of the American Medical Informatics Association Annual Fall Symposium, 1996, pp. 388–392.
[8] M. Uschold, M. Gruninger, Ontologies: principles, methods and applications, Knowledge Engineering Review. 11 (2) (1996) 93–136.
[9] K. Bontcheva, H. Cunningham, A. Kiryakov, V. Tablan, Semantic annotation and human language technology, in: Semantic Web Technologies: Trends and Research in Ontology Based Systems, John Wiley & Sons, Chichester, West Sussex, UK, 2006, pp. 29–50.
[10] P. Buitelaar, P. Cimiano, S. Racioppa, M. Siegel, Ontology-based information extraction with soba, in: Proceedings of the International Conference on Language Resources and Evaluation, Genoa, Italy, 2006, pp. 2321–2324.
[11] D.W. Embley, D.M. Campbell, R.D. Smith, S.W. Liddle, Ontology-based extraction and structuring of information from data-rich unstructured documents, in: Proceedings of the 17th International Conference on Information and Knowledge Management, ACM Press, New York, NY, USA, 1998, pp. 52–59.
[12] A. Maedche, G. Neumann, S. Staab, G. Saarbruecken, Bootstrapping an ontology-based information extraction system, in: Studies in Fuzziness and Soft Computing, Intelligent Exploration of the Web, Springer, New York, NY, 2002, pp. 345–360.
[13] H. Müller, E.E. Kenny, P.W. Sternberg, Textpresso: an ontology-based information retrieval and extraction system for biological literature, PLoS Biology 2 (11) (2004) 309.
[14] A. Todirascu, L. Romary, D. Bekhouche, Vulcain—an ontology-based information extraction system, in: Proceedings of Natural Language Processing and Information Systems. Springer, Stockholm, Sweden, 2002, pp. 64–75.
[15] A.L. Rector, J.E. Rogers, P.E. Zanstra, E. van der Haring, OpenGALEN: open source medical terminology and tools, in: Proceedings of AMIA Annual Symposium 2003, 2003, pp. 982–982.
[16] C. Rosse, J.L. Mejino, A reference ontology for biomedical informatics: the Foundational Model of Anatomy, Journal of Biomedical Informatics 36 (2003).
[17] J.L.M. Jr, D.L. Rubin, J.F. Brinkley, FMA-RadLex: an application ontology of radiological anatomy derived from the Foundational Model of Anatomy

reference ontology, in: AMIA Annual Symposium Proceedings 2008, 2008, pp. 465–469.
[18] D. Rubin, Creating and curating a terminology for radiology: ontology modeling and analysis, Journal of Digital Imaging 21 (2008).
[19] D. Marwede, M. Fielding, T. Kahn, RadiO: a prototype application ontology for radiology reporting tasks, in: AMIA Annual Symposium Proceedings, 2007, p. 513.
[20] N. Sager, M. Lyman, N.T. Nhan, L.J. Tick, Medical language processing: applications to patient data representation and automatic encoding, Methods of Information in Medicine 34 (1–2) (1995) 140–146.
[21] P.J. Haug, D.L. Ranum, P.R. Frederick, Computerized extraction of coded findings from free-text radiologic reports, Radiology 174 (2) (1990) 543–548.
[22] P.J. Haug, S. Koehler, L.M. Lau, P. Wang, R. Rocha, S.M. Huff, A natural language understanding system combining syntactic and semantic techniques, in: Proceedings of the Annual Symposium on Computer Application in Medical Care, 1994, pp. 247–251.
[23] P.J. Haug, S. Koehler, L.M. Lau, P. Wang, R. Rocha, S.M. Huff, Experience with a mixed semantic/syntactic parser, in: Proceedings of the Annual Symposium on Computer Application in Medical Care, 1995, p. 284.
[24] D.F. Worsley, A. Alavi, J.M. Aronchick, J.T. Chen, R.H. Greenspan, C.E. Ravin, Chest radiographic findings in patients with acute pulmonary embolism: observations from the PIOPED study, Radiology 189 (1) (1993) 133–136.
[25] C. Friedman, P. Alderson, J. Austin, J. Cimino, S. Johnson, A general natural-language text processor for clinical radiology, Journal of American Medical Informatics Association 1 (2) (1994) 161–174.
[26] G. Hripcsak, C. Friedman, P.O. Alderson, W. DuMouchel, S.B. Johnson, P.D. Clayton, Unlocking clinical data from narrative reports: a study of natural language processing, Annals of Internal Medicine 122 (9) (1995) 681–688.
[27] P. Zweigenbaum, B. Bachimont, J. Bouaud, J. Charlet, J.F. Boisvieux, A multilingual architecture for building a normalised conceptual representation from medical language, in: Proceedings of the Annual Symposium on Computer Applications in Medical Care, 1995, pp. 357–361.
[28] U. Hahn, M. Romacker, S. Schulz, MEDSYNDIKATE a natural language system for the extraction of medical information from findings reports, International Journal of Medical Informatics 67 (1–3) (2002) 63–74.
[29] A. Mykowiecka, M. Marciniak, A. Kupsc, Rule-based information extraction from patients' clinical data, Journal of Biomedical Informatics 42 (5) (2009) 923–936.
[30] K. Oflazer, Two-level description of Turkish morphology, Literary and Linguistic Computing 9 (2) (1994) 137–148.
[31] C. Friedman, S.B. Johnson, Natural language and text processing in biomedicine, in: Biomedical Informatics: Computer Applications in Health Care and Biomedicine, 3rd ed., Springer, 2006, pp. 312–343.
[32] E.L. Antworth, PC-KIMMO: A Two-Level Processor for Morphological Analysis. Occasional Publications in Academic Computing, Summer Institute of Linguistics, Dallas, TX, 1990.
[33] M. Temizsoy, I. Cicekli, An ontology-based approach to parsing turkish sentences, in: Machine Translation and the Information Soup, Lecture Notes in Computer Science, Springer, Berlin, 1998, pp. 124–135.
[34] W. Grosso, H. Eriksson, R. Fergerson, J. Gennari, S. Tu, M. Musen, Knowledge modelling at the millennium—the design and evolution of Protege2000, in: Proceedings of the 12th Knowledge Acquisition, Modelling, and Management (KAW'99), 1999.
[35] G. Hripcsak, G.J. Kuperman, C. Friedman, D.F. Heitjan, A reliability study for evaluating information extraction from radiology reports, Journal of American Medical Informatics Association 6 (2) (1999) 143–150.