Journal of Biomedical Informatics 42 (2009) 937–949

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Automatically extracting cancer disease characteristics from pathology reports into a Disease Knowledge Representation Model

Anni Coden [a,*], Guergana Savova [b], Igor Sominsky [a], Michael Tanenblatt [a], James Masanz [b], Karin Schuler [b], James Cooper [a], Wei Guan [d,1], Piet C. de Groen [b,c]

[a] IBM T.J. Watson Research Center, 19 Skyline Drive, Hawthorne, NY 10532, USA
[b] Division of Biomedical Informatics, Mayo Clinic College of Medicine, Rochester, MN, USA
[c] Division of Gastroenterology and Hepatology, Mayo Clinic College of Medicine, Rochester, MN, USA
[d] Georgia Institute of Technology, Atlanta, GA, USA

ARTICLE INFO

Article history:
Received 20 May 2008
Available online 27 December 2008

Keywords:
Cancer Disease Knowledge Representation Model
Analysis system
Natural language processing
Concept formation
Information retrieval
Medical records

ABSTRACT

We introduce an extensible and modifiable knowledge representation model to represent cancer disease characteristics in a comparable and consistent fashion. We describe a system, MedTAS/P which automatically instantiates the knowledge representation model from free-text pathology reports. MedTAS/P is based on an open-source framework and its components use natural language processing principles, machine learning and rules to discover and populate elements of the model. To validate the model and measure the accuracy of MedTAS/P, we developed a gold-standard corpus of manually annotated colon cancer pathology reports. MedTAS/P achieves *F*1-scores of 0.97–1.0 for instantiating classes in the knowledge representation model such as histologies or anatomical sites, and *F*1-scores of 0.82–0.93 for primary tumors or lymph nodes, which require the extractions of relations. An *F*1-score of 0.65 is reported for metastatic tumors, a lower score predominantly due to a very small number of instances in the training and test sets.

1. Introduction

Recently, cancer has become the number one cause of death in the United States, surpassing cardiovascular diseases. According to the American Cancer Society approximately 1.4 million new cancer cases occurred in 2007 and 1500 people died each day from cancer [1] in the United States. In order to perform research, to improve standards of care and to evaluate cancer treatment outcomes easy—and ideally, in an automated fashion—access to a variety of data sources is required. The knowledge contained in unstructured textual documents (e.g., pathology reports, clinical notes), is critical to achieve all these goals. For instance, clinical research requires cohort identifications which follow precisely defined patient- and disease-related inclusion and exclusion parameters. In most medical institutions such information is scattered among structured and unstructured data sources.

In the past few years, guided by a National Cancer Institute initiative caBIG [2], a new emphasis on creating a network of cancer communities has emerged. The necessity of consistent and comparable data is one of the key elements of a cancer network. The cancer common ontologic representation environment (caCORE) [3] and its associated software development kit provide an infrastructure to achieve "interoperability across the systems it develops or sponsors." A central component is a UML coded "domain information model." The caTIES—[4] effort within this initiative focused on extracting "coded information from free-text surgical pathology reports" with the goal of enabling and facilitating cancer research. Coded information refers here to some concepts (i.e., named entities, or textual mentions that belong to the same semantic class) specified within the NCI Metathesaurus [5].

The National Program of Cancer Registries (NPCR), administered by the Centers for Disease Control (CDC) in 1992, collects "population-based cancer incidence data" [6]. Some of the information is culled manually from surgical pathology reports to create comparable data. Automating this process would result not only in substantial cost savings but also be a step forward in the creation of consistent reports at the national level.

From these examples it is apparent that the medical community is striving towards a structured cancer representation. At the same time, the community currently relies on information stored mostly in unstructured free-text pathology reports. To bridge the gap between free-text and structured representation, an automatic and highly accurate mapping of free-text reports onto a structured representation is required.

* Corresponding author. Fax: +1 718 796 2445.
E-mail address: anni@us.ibm.com (A. Coden).
[1] Work done during internship at IBM T.J. Watson Research Center, Hawthorne, New York.

1532-0464/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2008.12.005

Natural language processing (NLP) systems exist that can retrieve named entities such as histology and anatomical site, and may be able to provide links between them in case of a relationship. However an NLP system able to map free-text pathology reports into a model that incorporates all the nuances of grading and staging cancer as is present in textual reports does not exist. Such system would have great clinical value: it could dramatically increase sub-classification of patients with cancer for practice management (how many patients with which grade/stage do we serve), research (how many patients with which grade/stage are available for protocol enrollment), quality control (how many patients with ... did we see and what where their outcome). In addition, it would allow physicians to continue to practice using their current descriptive language in transcribed dictations without a requirement to enter structured data in a complex, time consuming computer-based system.

We hypothesize that an NLP model can be built for pathology reports of colon cancer, using the latest NLP techniques, with retrieval recall and precision of at least 90%. Indeed, this is the main goal of this work.

In this paper we describe such a structured cancer representation which we call the *Cancer Disease Knowledge Representation Model* (CDKRM). It is an extension and generalization of the caTIES model, adding relations between cancer characteristics and defining additional concepts. We describe our system, MedTAS/P, which automatically determines the mapping between free-text pathology reports and the CDKRM. In addition, we report on the performance of MedTAS/P, measured on a set of colon cancer pathology reports.

This paper is organized as follows: Section 3 covers the methodology of our approach. In particular, the *Cancer Disease Knowledge Representation Model* is described in Section 3.1 and Section 3.2 discusses the terminologies used and Section 3.3 provides details about the corpora and the manual annotation process. In Section 3.4 we describe the development and test sets and conclude the section with a description of the evaluation procedures and metrics in Section 3.5. The Medical Text Analysis System MedTAS/P is the focus of Section 4. Section 4.1 provides details about the algorithms used to automatically populate the CDKRM and Section 4.2 explains the codification of the unstructured information. Section 5 presents the evaluation results of automatically populating the CDKRM from a set of free-text pathology reports. We discuss our work in Section 6 and conclude in Section 7.

2. Related work

The gold-standard diagnosis for cancer in general is obtained by analyzing tissue samples under a microscope. A pathology report captures the interpretation of a pathologist after careful inspection of such samples. The reports are generally in a semi-structured or unstructured text format and contain the detailed information about the presence or absence of cancer and—when present—its associated characteristics, such as tumor grade, the status of lymph nodes and presence or absence of metastases. It should be emphasized that the reports also capture the absence of cancer characteristics and that such information is as important as cancer-related information.

Recognizing the need to "ensure complete and consistent retrieval and transmission of cancer cases" [7], the College of American Pathologist (CAP) launched a project to create templates to "report results of surgical specimen examination" in a standardized form. A study showed that up to "96% of cancer diagnoses originate in the surgical pathology laboratory" [6]. Such templates, also referred to as synoptic reports or checklists, specify a set of values for key cancer characteristics and their attributes, e.g., histology or anatomical site. Several major medical institutions are starting to adopt synoptic reporting. However, mapping of previously written free-text pathology reports into synoptic reports with sufficient accuracy has so far eluded the community, and may be a contributing factor to the slow adoption of synoptic reporting. Some medical institutions are developing proprietary variations of the CAP checklists, and creating new ones for non-invasive cancers, which presents challenges to data sharing and comparative studies. Barriers to adoption of synoptic reporting in the community are currently being studied under the guidance of the US Department of Health and Human Services, with recommendations expected in 2009.

The goal of information extraction (IE) is to extract structured and semantically well defined concepts from unstructured data sources to facilitate access and retrieval of information. In the clinical domain, information extraction has the potential to help investigators rapidly answer questions such as: How many patients were diagnosed in 2004 with primary colon cancer? What percentage of these patients also had metastatic tumors in the liver? For which patients with invasive breast cancer does the tissue bank have more than four tissue blocks in storage and what are their block identifiers? Two excellent overviews on the state-of-the art of IE are presented in [8] and [9] including extensive references to work in the biomedical and clinical domains. In the pathology sub-domain, publications report on the automatic extraction of named entities such as histology or anatomical site, but not on extraction of *concepts* such as *primary tumor or metastatic tumor.* The relations between the named entities (for instance histology and anatomical site) and the higher-level concepts (for instance primary or metastatic tumor), have to be discovered from the pathology report.

There are multiple approaches to building information extraction systems. In general, such systems have natural language processing (NLP) components, such as tokenizers, part-of-speech taggers and parsers. Other components may be based on machine-learning techniques. Recently, two separate frameworks for building information extraction systems were developed and made available as open-source components—one is the Generalized Architecture for Text Engineering (GATE) [10], the other is the Unstructured Information Management Architecture (UIMA) [11]. The caTIES system described earlier is based on the GATE framework, whereas MedTAS [12]—the Medical Text Analysis System developed by IBM—is UIMA-based. It is noteworthy that components developed within the GATE framework can also be used within UIMA.

The caTIES system extracts several types of named entities (NE) such as histology, anatomical site, size and grade. The NE extraction is based on the NCI Metathesaurus [5] in conjunction with rule-based filters. Negation detection is an integral part of the system. The public UML model of caTIES does not seem to reflect relations between concepts (e.g., tumor grade and histology). Relations between the extracted named entities are not discovered, hence concepts like "primary tumor" or "metastatic tumor" and their associated characteristics are not found. Unfortunately, to the best of our knowledge, no accuracy results have yet been published.

MedLEE, another clinical natural language processing IE system, extracts domain knowledge from a variety of unstructured reports, such as discharge summaries, radiology reports and pathology reports [13]. MedLEE seems to focus on extracting named entities and not on the relations between them. For instance, a pathology report can describe multiple tissue samples, each having different cancer characteristics, e.g., primary tumor vs. metastatic tumor, MedLEE does not distinguish which characteristic (e.g., histology) describes which tissue sample. A substantial number of pathology reports in several major medical institutions describe multiple tissue samples. There seem to be no accuracy results published for

extracting information from unstructured pathology reports using MedLEE. We conjecture that this is partially due to the lack of a gold standard in the clinical community against which IE systems can be evaluated [14].

In this paper we describe a *Cancer Disease Knowledge Representation Model* for capturing cancer and its disease progression. In addition, we illustrate a novel method and system MedTAS/P (Medical Text Analysis System/Pathology) for automatic conversion of unstructured pathology reports into a structured and codified knowledge source according to a subpart of this model. It is part of a clinical NLP-based system as described in [15]. For instance, MedTAS/P automatically determines whether a free-text pathology report describes a "grade 3 primary tumor with histology of adenocarcinoma in the ascending colon." A user can then retrieve pathology reports of patients with "grade 3 primary tumors with histology of adenocarcinoma in the ascending colon" via a model-based user interface.

3. Methodology

3.1. The Cancer Disease Knowledge Representation Model

In this section, we describe our extensible knowledge representation model for storing cancer characteristics and their relations, including temporal information and inference (Fig. 1). We refer to this model as the *Cancer Disease Knowledge Representation Model* (CDKRM). Each node in the model is referred to as a *class*. Each class can have multiple attributes which can be filled with individual values of a given type, e.g., strings, integers or other classes. Subsequent figures describe some of the classes in more detail. We propose to use the CDKRM as the formalism to record a patient's disease state, track disease progression and draw inferences on outcome in conjunction with available structured information.

Classes whose attributes are only values are referred to as leaf classes. Our model has five leaf classes which describe cancer characteristics: anatomical site, histology, grade value, dimension and stage and three other leaf classes: document type and tumor block and tissue bank. Classes whose attributes are either values or other classes are referred to as container classes. Each leaf class can be thought of as a named entity with associated specific attributes. Fig. 2 shows details for the *anatomical site* class.

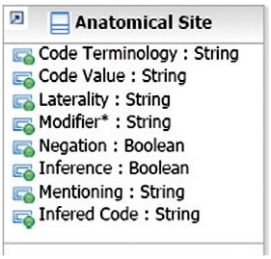

Fig. 2. Anatomical site.

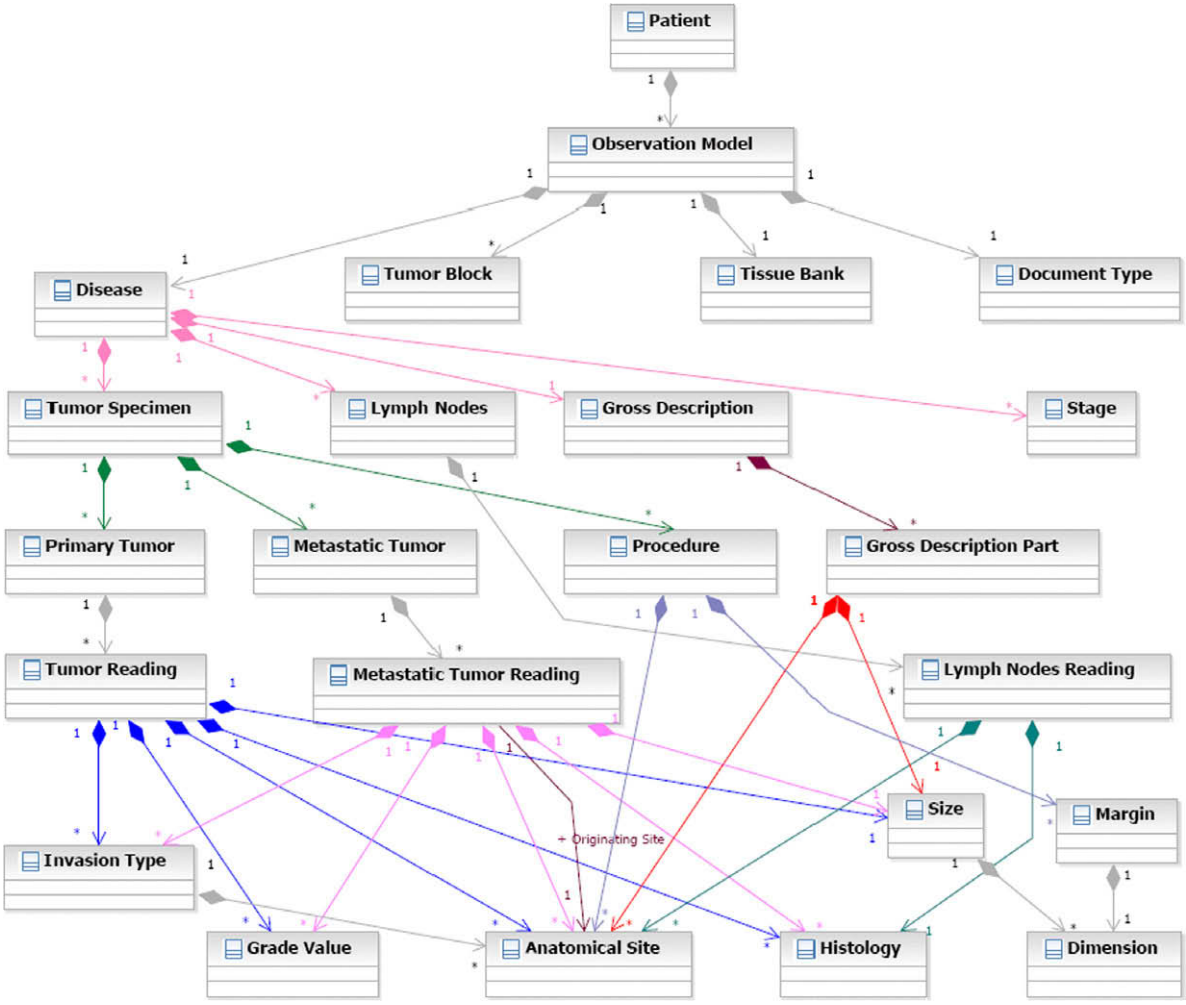

Fig. 1. Cancer Disease Knowledge Representation Model.

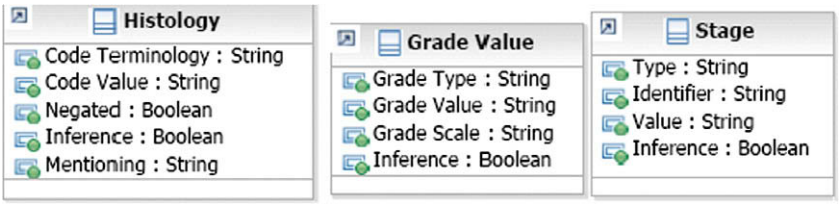

Fig. 3. Histology, grade value and stage classes.

Fig. 4. Dimension and size classes.

Here, anatomical site attributes specify the *code terminology* and *code value* associated with the attribute *mentioning* whose value is extracted from the text. Other attributes are laterality, negation and modifiers. An asterisk next to an attribute label indicates that multiple instances of an attribute can be specified. In addition, the anatomical site leaf class—like many other classes in the model—has attributes to specify whether a particular instance of a class contains inferred values. For instance, the text may refer to *lymph nodes—code value LN*—but from the context one could infer that *mesenteric lymph nodes—code value MLN*—were described. In this case, an instance of the anatomical site class would have the string LN in the *code value* attribute, the string MLN in the *inferred code* attribute and the *inference* attribute set to true.

Fig. 3 shows three other leaf classes capturing cancer disease characteristics. Histology and grade value are attributes of several container classes, such as primary and metastatic tumor. The stage of a cancer is either mentioned explicitly within a pathology reports or can be derived from other information from the primary tumor, the lymph node status and the occurrence or absence of metastases.

Instances of the *dimension* class can describe a measurement in a single dimension, such as linear extent or a weight (Fig. 4). The container class *size* has multiple attributes, each of which can be filled by a *dimension* class.

The leaf classes shown in Fig. 5 capture information important to evaluating the data extracted from unstructured textual sources and relating it to other data in a medical institutional system. The *document type* class is used to specify the type—e.g., clinical note, pathology report, treatment report—of the data source and the physician who signed for the data. The *tumor block* and *tissue bank* classes capture specific identifiers about the location the tissue

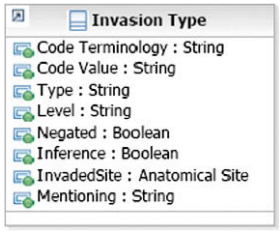

Fig. 7. Invasion type class.

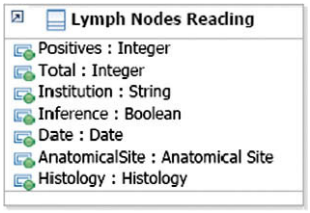

Fig. 8. Lymph node reading.

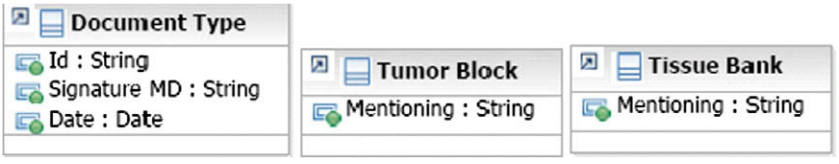

Fig. 5. Leaf classes.

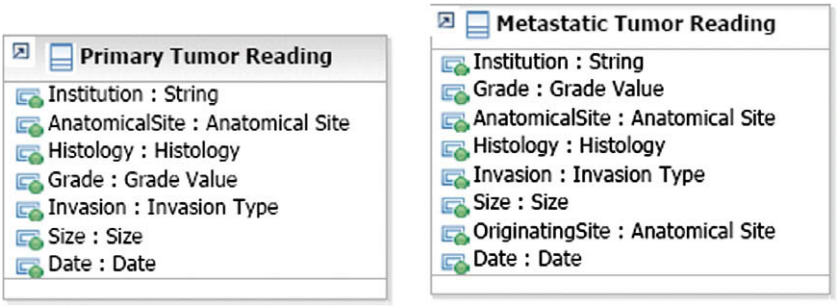

Fig. 6. Primary and metastatic tumor reading classes.

Fig. 9. Gross description classes.

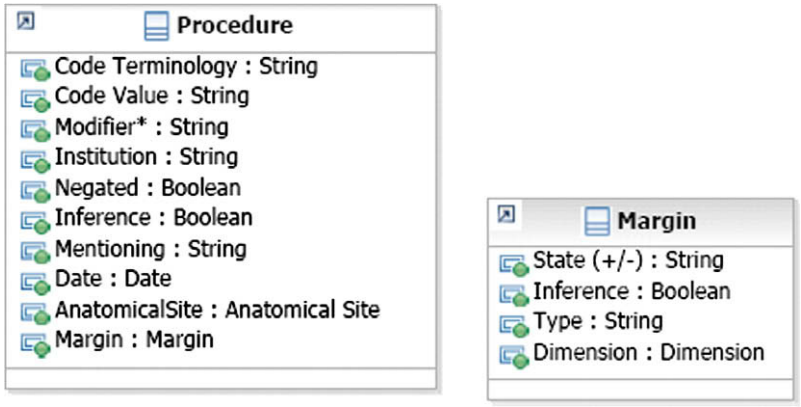

Fig. 10. Procedure and margin classes.

samples within the institution as specified within a pathology report.

The *primary* and *metastatic tumor reading* classes depicted in Fig. 6 are examples of container classes in the model. A tumor reading class contains the following attributes: *histology*, *anatomical site*, *size* and *invasion type* (invasion type class is shown in Fig. 7). In addition, the institution where the analysis on the tissue sample was performed and its date are attributes of a tumor reading class. The *metastatic tumor reading* class specifies two anatomical sites: originating and metastatic.

Fig. 1 shows that a tumor class (primary or metastatic) can contain multiple instances of tumor reading classes, capturing the notion of multiple interpretations of the same tissue sample. For instance, two doctors in the same or different institutions can reach different conclusions about the type and severity (e.g., histology, grade) of the disease based on the same tissue sample. Different interpretations are not that common in pathology reports, and based on some preliminary observations seem to be rather common in clinical notes.

Fig. 8 describes the *lymph node reading* class. Noteworthy attributes are the number of *positive* nodes and *total* number of lymph nodes excised. Similarly to the tumor classes, a *lymph nodes* class can contain multiple lymph nodes reading classes.

The *gross description* classes are shown in Fig. 9. The *gross description part* classes describe each excised tissue sample, whereas the institution where the procedure was performed and the date are associated with the *gross description* class.

The *procedure* class has a number of identifying attributes. One significant attribute that has not been seen in other classes is *margin* (Fig. 10). Multiple margins can be specified, with each margin specification including an attribute indicating its type (e.g., anatomical, surgical).

Over the course of time a patient can have multiple disease episodes; each episode is captured in an observation model, which can have time stamps or sequence numbers associated with it. In general, a single pathology report does not reflect multiple episodes; however a single clinical note often describes the patient's disease progression.

The CDKRM is easily extended by adding additional concepts and relations. Such models, instantiated from textual sources have multiple use cases: examples include identification of cohorts of patients who have similar disease progression or summarization of disease progression of a single patient from multiple reports.

3.2. Terminology

In our project, we used the International Classification of Diseases—Oncology version 3 (ICD-O) [16] as the underlying terminology for anatomical sites and histology. In particular, the value of the mention attribute within these classes was mapped to this terminology. ICD-O is the lingua franca of pathologists and is in widespread use within tumor registries. The ICD-O terminology encapsulates two features which are exploited for co-reference resolution and for instantiating classes within the CDKRM. The first feature is the specification of the "is-part-of" relation between anatomical sites. For instance, ICD-O specifies that the *hepatic flexure of the colon* is part of the *colon*. The ICD-O code values for anatomical sites denote the "is-part-of" relation. Anatomical sites are codified as C*x.y* where all anatomical sites having the same "is-part-of" relation to another anatomical site have the same *x*. The *y* value 9 denotes the most generic description of the group of anatomical sites sharing the same *x* and will be referred to as the *generic anatomical site*. For instance, C18 refers to the colon, C18.3 to the hepatic flexure of the colon and C18.9 to the colon, NOS (*n*ot otherwise specified). The second feature applies to histologies, where a "has-behavior" relation is specified. This behavior code within ICD-O codifies this relation. The code value for histology is written as M*x*/*y*, where *y* defines the behavior relation for a particular diagnosis. For instance, the behavior code "6" indicates that the diagnosis refers to a metastatic tumor.

Similarly to other terminologies, ICD-O defines a "canonical" or "normal" form for the concepts it contains. For instance, "adeno-

carcinoma, NOS" can be considered to be the canonical form for the synonyms "adenocarcinoma" and "adenocarcinoma NOS." Data access and retrieval can be simplified by mapping all synonyms onto their canonical form. Synonyms are defined as textual strings which always have the same meaning (i.e., terminology code) as their canonical form.

3.3. Corpus and its annotations

The biomedical community has yet to develop gold-standard training and test beds of annotated clinical pathology reports which can be used as a shared standard for evaluating automatic knowledge extraction systems. Therefore, we developed not only a detailed CDKRM, but also a detailed manually annotated corpus for training and testing MedTAS/P which is the system that automatically populates the CDKRM from free-text pathology report. This section describes the construction of our gold-standard data set and the manual annotation process.

3.3.1. Corpus description

The training, validation and evaluation of the work reported in this paper is based on a corpus of 302 pathology reports of patients who had an assigned billing (ICD-9 CM [17]) code for a diagnosis of colon cancer. The first step in creating the corpus of pathology reports was creating a list of all patients who had a diagnosis of colon cancer in 2004 (year was picked randomly). The initial pool of patients was created by retrieving patients with one of the following colon cancer codes: ICD-9 CM = {153.0, 153.1, 153.2, 153.3, 153.4, 153.7, 153.7, 154.0, 154.1}. The extraction was done through a query accessing the billing database.

The query results consisted of patient ID's, their corresponding ICD-9 CM diagnoses and dates, specimen dates, final diagnoses and gross description. From that list, patients were selected randomly; pathology reports were manually reviewed, discarding reports that did not contain information about a malignancy or about the colon. When 302 unique reports had been found, the selection process was stopped. The 302 reports represented 222 patients. From each of the 302 reports, the final diagnosis section and gross description section were converted to plain text (i.e., formatting was removed) and combined programmatically into one report, with the insertion of section headings.

All personal information from all notes sent from the Mayo Clinic NLP team to the IBM NLP team was manually removed (de-identification process) following the Health Insurance Portability and Accountability Act (HIPAA) [18] requirements for Protected Health Information (PHI). The safe harbor methodology was used; however the dates were not de-identified. The data was exchanged only between the two institutions.

3.3.2. Manual corpus annotation

Four domain experts trained as medical retrieval specialists manually annotated the pathology corpus. The *Cancer Disease Knowledge Representation Model* was implemented within Knowtator [19], a Protégé [20] plug in. The domain experts then manually filled in the attributes and relations in the classes of the CDKRM with information from the pathology reports using the Knowtator tool.

The manual annotation process proceeded as follows: six sets containing 50–51 documents each were created and two domain experts were assigned the task of independently annotating each set. By dividing the documents into six sets, we ensured that each of the four annotators would work with each of the others on a set. After the independent annotations were finished, each pair of annotators consented over their divergences. Subsequently, the manual annotations were *adjusted* to correct for errors and omissions by the annotator teams and clinicians. Thus we have two sets of manual annotations: a set of annotations based on the consented agreement of domain experts ("raw") and a set of annotations where errors were corrected by the entire team and clinicians ("adjusted"). The final step was to annotate co-referenced mentions. Words and/or phrases are co-referenced if they have the same meaning within the context of the document. A precise definition is provided later in this section.

During the annotation process we developed a detailed annotation guidelines document. Of special interest are the interpretations of linguistic concepts such as nominal ellipsis and co-reference. One type of ellipsis is conjunctive phrases—e.g., "invasive and ductal carcinoma." Such linguistic structures (ADJECTIVE-A CONJUNCTION ADJECTIVE-B NOUN) are interpreted as "invasive carcinoma and/or ductal carcinoma" (ADJECTIVE-A NOUN CONJUNCTION ADJECTIVE-B NOUN). Another type of ellipsis is specified by a comma between two nouns—e.g., "colon, rectum" (NOUN-A COMMA NOUN-B). Such structures are interpreted as NOUN-A, NOUN-B and NOUN-A NOUN-B. In our example, the phrase "colon, rectum" would be interpreted as referring to the "colon," "rectum" and the "colon rectum."

Co-reference was also annotated. We define co-reference for two of the classes in the CDKRM, *anatomical site* and *histology*. For co-reference, only the mentioning, code terminology and code value attributes are considered. The laterality attribute is also taken into account for anatomical sites. Two classes which are the same with respect to these attributes are co-referenced. In addition, classes can be co-referenced based on the underlying ICD-O terminology. In particular, for each group of anatomical sites, ICD-O specifies a *generic anatomical site* as defined in Section 3.2. Each anatomical site is co-referenced to any instances of its *generic anatomical site*. Histologies are also co-referenced to their *generic histology*. Unfortunately, ICD-O does not address the notion of *generic histology*. In practice, there is a set of terms such as "tumor" and "carcinoma" which are used in such a fashion. More precisely, within ICD-O, histologies with the following codes are referred to as generic: m8000/*x* (neoplasm), m8001/*x* (tumor), m8010/2 (carcinoma in situ), m8010/3 (carcinoma) and m8010/6 (metastatic carcinoma).

3.3.3. Inter-annotator agreement

The inter-annotator agreement (IAA) for the pathology notes was calculated as a two-way agreement between human annotators for six sets of 50–51 notes each on the pre-consensus data. In Table 1 below we summarize the inter-annotator agreement for some classes in the CDKRM in each of the six sets.

IAA was calculated simply as a percentage agreement given by the total number of annotations where the human annotators agree over the total number of annotations. Since in our task the expected agreement can be considered zero due to the large number of possible values, the κ coefficient [21], [22] and accuracy yield similar results. It is argued in [23], for many NLP tasks (named entity tasks including) the κ results will approximate *F*-score and accuracy. The expected probability term $E(\text{Accuracy})$ approaches zero, and κ is just:

$$\kappa = \frac{\text{Accuracy} - E(\text{Accuracy})}{1 - E(\text{Accuracy})}$$

and κ approaches Accuracy as $E(\text{Accuracy})$ approaches 0.

In general, IAA results show strong inter-annotator agreements (>85%). Classes which have no span associated with them are counted as not matching by Knowtator, the tool used to compute IAA. Hence, some values, such as positive lymph nodes had very little agreement because in many documents, there was no explicit mention of the number and therefore no text was spanned. The IAA results show that our manually annotated corpus is of good quality and that the task is suited for automation.

Table 1
Inter-annotator agreement.

	Set 1 (%)	Set 2 (%)	Set 3 (%)	Set 4 (%)	Set 5 (%)	Set 6 (%)	Average (%)
Anatomical site	96.0	95.5	96.6	95.5	96.8	93.8	95.7
Histology	100.0	97.0	98.6	98.6	95.9	96.4	97.8
Grade scale	100.0	100.0	100.0	100.0	100.0	98.5	99.7
Grade value	100.0	99.1	100.0	100.0	100.0	100.0	99.9
Excised nodes	98.0	90.9	87.0	96.7	100.0	92.9	94.2
Positive nodes	49.1	21.8	17.9	30.0	35.3	22.2	29.4
Dimension extent	100.0	99.3	99.8	99.4	99.8	99.8	99.7
Dimension unit	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Date	96.1	89.1	100.0	100.0	94.7	98.7	96.4

3.4. Development and evaluation sets

For the purposes of development and evaluation, we divided the 302 annotated documents into three datasets, two sets (set 1 and set 2) were used to train our knowledge extraction engine, and one set was used as an evaluation set.

Table 2 shows the number of annotations for some classes. The numbers labeled *raw* refer to the manual annotations done by the domain experts which were used to compute inter-annotator agreement. The automatic population of the CDKRM is evaluated against the "raw" and the "adjusted" corpora. A pathology report describes the excised tissue at a single point in time hence the observation model contains only a single disease model. In our corpus, we observed a one-to-one correspondence between reading classes and their containers (e.g., primary tumor reading class and tumor class). There are only few instances of metastatic tumor in our set, hence the results on metastatic tumor should be taken as preliminary results.

3.5. Evaluation procedure and metrics

We evaluated the automatic population by MedTAS/P of the following classes:

1. Anatomical site
2. Histology
3. Grade value
4. Dimension
5. Date
6. Gross description
7. Primary tumor
8. Metastatic tumor
9. Lymph node

We computed standard metrics for our evaluation: precision, recall and *F*1-score and exact confidence interval [24].

1. Precision = P = TP/PP
2. Recall = R = TP/AP
3. $F1\text{-score} = (2*P*R)/(P + R)$

where,

TP—true positives (correctly recognized)
AP—actual positives (total number of entities in the test set)
PP—predicted positives (total predictions of that entity made by the system).

The metrics precision and recall were computed by comparing the automatically extracted classes against the gold standard. Two classes are deemed to be equivalent, if and only if all their individual attributes are equivalent and if they have the same type and number of relations. Two attributes are equivalent if and only if their values are the same. A value can be either one or more classes (e.g., an anatomical site), a string/integer value (e.g., an ICD-O code) or a mention in the text composed of the string in the text and its offsets. Hence the equivalence relation is recursive. An attribute may also be filled by co-reference objects for which the following equivalence definition applies: let A and B be slots filled by co-reference objects. B is considered to be equivalent to A if and only if each co-reference object in slot A has at least one member object which is also an object in slot B. Furthermore, each object in B has to belong to a co-reference object in A. Let us assume that slot A was filled with a manual annotation process, whereas B was filled with an automatic process. Then our definition of equivalence captures that, even if a single object was included in the automatic annotation but not manually annotated, the entire automatically extracted class would be deemed incorrect.

Table 3 shows which attributes are considered in the evaluation for each class.

The metrics and evaluation procedures just described were implemented using a tool we developed called common feature extractor (CFE) [25] which is fully integrated within the UIMA framework.

Table 2
Number of annotations.

Number of instances	Raw				Adjusted			
	Set 1	Set 2	Set 3	Total	Set 1	Set 2	Set 3	Total
Reports	101	100	101	302	101	100	101	302
Anatomical site	735	814	767	2316	743	812	761	2316
Histology	325	326	325	976	327	324	325	976
Grade value	118	124	118	360	118	124	118	360
Dimension	414	528	471	1413	412	524	471	1407
Date	85	60	88	233	85	60	88	233
Gross description part	120	156	132	408	120	156	132	408
Primary tumor	112	117	119	348	114	117	119	350
Metastatic tumor	10	18	19	47	8	19	19	46
Lymph node	58	59	59	176	57	59	59	175

Table 3
Attributes considered for evaluation.

Class	Attribute	Attribute	Attribute	Attribute	Attribute
Histology	Mention	Terminology code			
Anatomical site	Mention	Terminology code			
Grade	Value	Scale	Type		
Dimension	Extent	Unit			
Date	Day	Month	Year		
Gross description part	Anatomical site	Size			
Primary tumor	Anatomical site	Histology	Size	Grade	
Metastatic tumor	Anatomical site	Originating anatomical site	Histology	Size	Grade
Lymph node	Anatomical site	Histology	No. of positive nodes	Total no. of nodes excised	

4. Medical Text Analysis System/Pathology

4.1. Automatic population of the Cancer Disease Knowledge Representation Model

Our text analysis system (MedTAS) is tailored to the medical domain, the pathology version (/P) containing additional components for extracting cancer-specific characteristics from unstructured text. It is based on natural language processing (NLP) principles, and contains both rule-based and machine-learning based components and runs within the UIMA framework. An application within such a framework consists of a set of programs (annotators), each having a configuration file in XML format. The execution sequence, or pipeline, of annotators is also described in a configuration file. Configuration files can be modified with any text (XML) editor. In addition, MedTAS/P provides a mechanism to ingest, process and use external resources, such as terminologies and ontologies. The rest of this section will describe the MedTAS/P system in more detail.

Our pipeline can be broken into several components:

1. Ingestion—a component which extracts implicit meaning from the structure of a document
2. General natural language processing—components for tokenization, sentence discovery, part-of-speech tagging and shallow parsing
3. ConceptFinding—a component which determines concepts based on specified terminology and determines negation
4. Cancer-specific annotators for grade, stage, size, margin, date, tumor blocks
5. RelationFinder—a component which populates the CDKRM and resolves co-references

4.1.1. Ingestion and tokenization

In general terms, annotators mark up an unstructured textual document, inserting ‘‘annotations” that can be associated with a particular piece of text or which can be container objects for other annotations. A subsequent annotator can read and process all previously created annotations. The document ingestion annotator converts embedded tags of an input document (if present) into annotations and simultaneously adds derived information, such as the number of sections and subsections, header information and correlations between disjoint pieces of text describing the same tissue specimen. The derived information is based both on textual labels and visual (formatting) cues within the document.

The pipeline can use any tokenizer. For optimal performance, all textual resource used within the pipeline, such as terminologies and ontologies, are expected to be tokenized in the same fashion as the documents that are analyzed.

4.1.2. Sentence annotations

There are a variety of approaches and implementations to determine sentence boundaries—both rule-based and statistical. Within the pipeline, a domain specific sentence annotator (DSSP) is executed to adjust previously determined sentence annotations to take into account the structure of medical documents and their implied meaning. Examples of potential issues for a general sentence detector are list and header processing, parenthesis processing and non-standard use of punctuation symbols. We experimented both with a rule-based and a statistical sentence annotator. The rule-based sentence annotator in conjunction with DSSP led to improvements in the automatic population of classes within the CDKRM. The improvements in $F1$-score ranged from 0% to 2.2% for leaf classes, and from no improvement to a 10.4% improvement for most of the container classes with the exception of a single container class where the $F1$-score decreased by 0.8%. For this reason, the results reported later in this paper are based on using a rule-based sentence detector.

4.1.3. Part-of-speech tagger and shallow parser

We use a statistical part-of-speech (POS) tagger, which takes a model as input. We experimented with a model based on the Penn Treebank corpus [26], and a model based on the Penn Treebank corpus augmented with a manually annotated corpus of clinical notes developed at the Mayo Clinic. Differences in performance on clinical notes have been reported in [27], demonstrating that accuracy on a particular domain is improved, if a general English training corpus is augmented with a small domain specific training corpus. We came to a similar conclusion in the context of pathology reports. For instance, terms like ‘‘left, descending, ascending” seem to always be used as adjectives within the context of pathology reports in contrast to clinical notes and general English language use. We added a small dictionary of terms and their respective part-of-speech tags, as applicable within pathology reports to MedTAS/P, which is used to overwrite the part-of-speech tags determined by the statistical POS tagger.

In several of the algorithms the context plays an integral part. Context is defined here as the range of text within a document used to determine the semantic meaning of a word or phrase. To determine context, we experimented with a noun phrase chunker and with a shallow parser that provided more precise information, important in the next steps of the pipeline. The shallow parser is based on finite-state transducer technology and a set of cascading grammars [28] originally developed for parsing general English. One of the grammars was modified to extend the pattern for noun phrases: parenthesized numerical expressions are considered part of a noun phrase. The shallow parser identifies (1) noun phrases, (2) noun phrase lists and (3) prepositional noun phrases, amongst other constructs. We define the concept of a *generalized noun phrase* as a noun phrase which belongs to any of these three types and we define a noun phrase hierarchy in the order of (1–3).

4.1.4. Concept Identification

ConceptFinding is one of the most critical components in our system. It maps textual mentions to terminology concepts to create codified information. We built two separate annotators, (1) *conceptMapper* which creates candidate matches between concept structures based on a terminology and unstructured text and (2) *conceptFilter*, a set of rule-based annotators that filter out matches depending on the desired application as exemplified later in this section. The concept structures themselves are specified in an XML format,

```
<token canonical="adenocarcinoma, nos" SemClass="Diagnosis"
AttributeType="ICDO" AttributeValue="m8140/3" POS="NN">
    <variant base="adenocarcinoma, nos" POS="NN"/>
    <variant base="adenocarcinoma, n.o.s." POS="NN"/>
    <variant base="adenocarcinoma" POS="NN"/>
</token>
```

Such structures allow for inclusion of canonical forms (first line in above specification) and their synonyms. Each synonym can have multiple features such as associated semantic type, terminology and part-of-speech tag. The set of these features and how they map to UIMA annotations is specified within the configuration file.

The mapping of textual mentions to concepts depends on the final intended application. For example, the snippet "hepatic flexure of the colon" can be codified using ICD-O as "hepatic" (C22.0), "colon" (C18.9) and "hepatic flexure of the colon" (C18.3). Note that the ICD-O entry for C18.3 actually reads as "hepatic flexure of colon" (note the absence of "the" before "colon"), which demonstrates why exact string matching is not sufficient to codify unstructured text. *ConceptMapper* finds all possible mappings between the terminology and the free-text (C22.0, C18.9, C18.3). *ConceptFilter* marks the ones to be ignored due to a potential term subsumption (C22.0, C18.9). Of course, while using the longest match heuristic would avoid the need for such filtering in this simplified case, there are more complex examples where that is not sufficient.

ConceptMapper has multiple user-defined parameters specified in the associated XML configuration file:

1. Tokenizer to be used for text and resources
2. Case matching: on/off
3. Stemming: a stemmer can be specified
4. Word order independent lookup: on/off
5. Tokens to be skipped (user-specified stop words, semantic classes)
6. Context within which matching is executed (e.g., sentence, paragraph)
7. Matching algorithms
   a. Start looking for a match at every token
   b. Start looking for a match starting after the last match

Parameter 4 allows for the words within a phrase to be mentioned in the unstructured text in a different order than in the specified resources. For instance, the terminology may specify "ascending colon" as an anatomical site, but the mention in the text may read as "colon, ascending." If parameter 4 is set to "on," the mention "colon, ascending" would be correctly identified as an anatomical site. Parameters 5 and 6 allow the user to specify which tokens should be considered for lookup. Parameter 5 allows for skipping terms as candidates for lookup if they appear in a stop word list or belong to a user-specified specified semantic class. For instance, "invasive duct carcinoma" is codified as the histology m8500/3 within ICD-O. Suppose that a pathology report reads as "invasive well differentiated duct carcinoma" and that the phrase "well differentiated" was already identified as having the semantic meaning of "grade." Parameter 5 allows for skipping of terms considered for lookup if they have a specified semantic class (e.g., "grade") resulting in the correct identification of the histology in the pathology report. In parameter 6 the user can specify the context within which tokens for matching should be considered—for instance, within noun phrases, sentences or sentence fragments. It is noteworthy, that *conceptMapper* preserves the structure/relation of the underlying terminology or ontology.

The "hepatic flexure of the colon" example described earlier showed the necessity for some filtering which is executed in *conceptFilter*. For example, one such filtering rule is the *subsumption rule*, which specifies whether contained annotations, e.g., "hepatic" should be exposed or not. Other filters mark annotations based on particular values of one of their attributes and another removes duplicate and identical annotations. One filter discovers subsumed generic anatomical sites or histologies and marks them. *ConceptFilter* also handles nominal ellipsis. Consider for example the phrase "Colon, ascending and transverse." One interpretation is that it refers to two sites (ascending colon and transverse colon), another interpretation is that the physician described three sites, colon, ascending colon and transverse colon, as denoted by the use of the comma. Depending on user-defined configuration parameters, *conceptFilter* will apply the appropriate one of these interpretations. We evaluated *conceptMapper* and c onceptFilter with a multitude of parameter settings and the results will be reported in a separate paper. Within the evaluation section we describe the parameters which proved to be optimal for our particular task.

4.1.5. Pattern recognition and negation

MedTAS/P also contains a general regular expression annotator which in conjunction with a terminology discovers textual mentions describing dimensions and sizes, dates, number of excised and positive lymph nodes and stage. MedTAS/P has tools for building machine-learning models, and annotators to discover concepts based on such models. Instances of the grade value class are populated based on pattern matching.

The negation detector is a generalized algorithm as reported in [29]. Negation trigger words (e.g., "no," "ruled out") are specified in a user modifiable dictionary. The trigger words become the anchors around which negated phrases are discovered within a user-specified window. Here we report on results assuming that the window is a sentence. Within the predefined window, the algorithm examines generalized noun phrases starting to the right of the negation keyword. If none is found, then it continues examining phrases to the left. When a generalized noun phrase is found, all semantic entities are marked as negated.

4.1.6. Relation identification

The next step in the pipeline—performed by the Relationfinder—is to discover the relationship between the appropriate leaf classes (e.g., histology, anatomical sites, size, grade) to populate container classes such as the *primary* and *metastatic tumor* classes, the *lymph node class* and the *gross description part* class. The relations between the classes are "contains" and "is-part-of." There is a common methodology in filling container classes, coupled with certain class-specific rules as outlined in the next paragraphs. We will first outline the common methodology applied to instantiate the primary and metastatic tumor classes, the lymph node class and the gross description part class and then provide more specific details and examples. At this point in the processing, the pipeline has already identified the leaf classes of anatomical site, histology, grade value, stage and dimension and the container class size.

The first step is to determine which section of a document should be considered for instantiating a container class. For instance, in general *gross description part* classes are populated only from the gross description section of a pathology report. The tumor and lymph node classes take information from the *final diagnosis* section. The relation between class and document section is specified in a user configurable file. Second, certain classes are categorized according to multiple criteria (e.g., primary vs. metastatic vs. benign, tumor size vs. margin size). Third, we determine which mentions refer to each other (i.e., are co-referring). Fourth, we determine which instances of classes (e.g., histology or anatomical site) should be considered candidates for populating each of the container classes (e.g., primary tumor class or lymph node class). In the final fifth step container classes are merged or split according to class-specific rules.

In step two described above, some classes are categorized. One categorization labels classes as positive or negated with respect to a particular class. A class whose negation attribute is set to true by the negation detector is negated with respect to all classes. An example of a negated histology is the phrase "tumor free," where tumor is a histology that our negation algorithm (previously described) marked as absent. A class which is negated with respect to a particular class is referred to as *excluded* with respect to a class. Exclusion states that an instance of a class can be part of only a single container class. For instance, an anatomical site mentioned as part of an invasion class is excluded from consideration for filling a tumor class. Anatomical sites are categorized into originating sites, lymph nodes, invasion sites and other sites. Histology classes are categorized as metastatic and non-metastatic. Size mentions are categorized as tumor sizes and other sizes. Our categorization algorithm is based on a set of trigger phrases (specific to a particular categorization) and the noun phrase hierarchy previously described. For each class instance to be categorized, the algorithm checks whether an appropriate trigger word is co-occurrent with the mention attribute of the class. Co-occurrence is defined based on the noun hierarchy, which means that noun phrases, noun phrase lists and prepositional noun phrases are inspected in turn. In addition, ICD-O codes are used for categorization of histologies and anatomical sites.

Although pronominal anaphora resolution is not required for analysis of pathology reports, co-reference resolution is critical in populating the CDKRM. In Section 3.3.2 co-reference was defined for anatomical sites and histologies. The methodology for discovering co-referenced generic histology classes is similar to pronoun resolution [30]. For each histology H, examine each *generic histology* GH mentioned after H and which is categorized equivalently to H. Only generic histologies {GH} occurring between H and a subsequent equally categorized histology H1 are considered. The set of generic histologies {GH} is co-referenced with H. The following examples may clarify the algorithm some more. Let HM1 and HM2 be two metastatic histologies, HN1 a non-metastatic histology and GHM1, GHM2 and GHN1 be generic metastatic and non-metastatic histologies occurring in the following sequence:

HM1 GHM1 GHM2 HN1 GHM3 GHN1 HM2.

Here GHM1 GHM2 and GHM3 will be co-referenced with HM1 and GHN1 with HN1 respectively.

Besides generic rules for populating class instances, we used class-specific rules as well. For example, the *gross description part* class may contain one or more anatomical sites and a size. Processing of the document starts with an initial anatomical site within the gross description section and continues with all the other anatomical sites within the same context (i.e., hierarchy of noun phrases) until either a size is found, or the hierarchy is exhausted. At this point, the size expression is parsed to determine whether a single size or a range of sizes was specified. In the latter case, two *gross description part classes* are instantiated, both having the same anatomical sites but different sizes. This is a class-specific implementation of the fifth step as described earlier.

The *primary* and *metastatic tumor classes* are populated simultaneously by a single TumorModelAnnotator. The assumption is that tumor classes are populated with information within a user-defined portion of the document—the tumor context TC. The algorithm iterates through multiple steps. First it identifies all non-negated histologies within TC. Second, for all identified histologies, it examines the noun phrase containing the histology for all occurrences of any of these three classes: anatomical sites, grade values and sizes and associates them with the histology. It is noteworthy that each of these classes can be associated with only a single histology and hence, once an association is found, it is removed from further consideration. Third, for histologies missing one or more of these associations, step two is repeated, but for noun phrase lists instead of noun phrases. Fourth, step two is repeated for any histology missing any associations within the context of a sentence. Ultimately, tumor classes which have co-referenced histologies are merged into a single instance. Classes which refer to the same exact anatomical site(s), grade value(s) and sizes and differ only in the histologies are merged as well. An artifact of pathology reports is that anatomical sites are at times implied to be the same as the sites mentioned in the gross description. To account for this, for any non-negated histology that has no anatomical site associations, we extend the context to the gross description part of the document. For the particular pathology reports used for this study, as is the norm, the first sentence of the tumor context TC is considered to be the gross description. The final step is to instantiate the tumor classes based on the categorization of the histology and anatomical sites. It is important to consider all histologies (including benign ones) and all anatomical sites in the process to identify associations correctly, but neither benign histologies, nor histologies with an associated anatomical site that is a lymph node, are considered for primary or metastatic tumor classes. The category of the remaining histologies (metastatic or non-metastatic) determines which type of tumor class is instantiated.

Lymph nodes classes have attributes that are anatomical sites, histologies and lymph node expressions (LNE). In particular, only anatomical sites which have been categorized as lymph nodes (AS-L) are considered. LNE describe either the general state (positive/negative) of the lymph nodes or provide more detail in terms of number of positive lymph nodes and excised nodes (from which the state is deduced). For each AS-L, the algorithm for instantiating lymph node classes determines the histologies and LNE co-occurring with the anatomical site (AS-L) in the same sentence. If they are not found, the context is expanded to sentences within the same section. A set of rule-based filters is applied to derive the correct associations, taking the categorization of histologies and anatomical sites into positive and negative classes into account.

To populate the *gross description part* class, we introduced two new syntactic structures—the ParenthesesSeparatedNoun-phrase (PSN) and the ParenthesesPhrase (PPH). A sequence of a noun phrase, a parenthesis, a noun phrase followed by another parentheses is called a PSN. Any expression enclosed with matching parenthesis is a PPH. We define a hierarchy of syntactic constructs consisting of the following levels: noun phrase, PSN, *generalized noun phrase* and PPH. The algorithm for populating a gross description part examines the syntactic hierarchy in order, with noun phrases being at level 0 and PPH being at level 4. If anatomical site(s) and size expression(s) co-occur in the same syntactic structure, one or more gross description part classes are instantiated. The number of instantiated classes depends on the type and number of size expressions found. If an anatomical site AS occurs without a size expression within a syntactic structure, a set of rules determines whether the AS should be associated with an already existing gross description part class or a new class be instantiated.

The rules depend on the lexical ordering of the anatomical site and size mentions.

4.2. Text to codified concept mapping

There are two goals in the process of mapping an unstructured pathology report onto a CDKRM—first, to identify each class and its attributes (e.g., primary tumor, lymph nodes), and second to codify each of the attributes in the classes for easy access and comparison. In this section we describe the codification of the CDKRM. Although, specific codification schemata will be discussed here, the underlying principles can be applied to a variety of terminologies and ontologies within the medical domain. There are two basic approaches to map entities to a standard terminology. One approach is to discover the entities first (e.g., using a machine-learning algorithm) and then to map them to a terminology. The other approach is to appropriately augment a standard terminology, for example, with plural forms and synonyms, and in a subsequent step create the links between the text and the augmented terminology. Our project took the second approach.

Domain experts suggested ICD-O as the underlying terminology for histology and anatomical sites for pathology reports. Although ICD-O specifies a number of synonyms, our experience revealed that additional synonyms are used within pathology reports. We devised a semi-automated rule-based system to augment the terminology by:

1. Creating plurals
2. Creating synonyms by:
   - 2.1. using variations in punctuation
   - 2.2. removing parenthesized expressions from a mention
   - 2.3. removing stop words, punctuations, medical stop words (e.g., NOS)
   - 2.4. using some common abbreviations
   - 2.5. using adjectival forms of a word as specified in the specialist lexicon of the UMLS.

For multi-term entries, these rules are applied to each of the terms in the entry.

Besides histology and anatomical site, codifications for other entities within the CDKRM need also to be specified. Day, month and year as part of a date expression are codified as integers in the expected form. There are many cancer grade schemata, and MedTAS/P recognizes most of them. MedTAS/P supports the notion of a range of values for a grade specification. Suppose a grading system specifies possible values being integers between 1 and 4, and a pathologist chooses to assign a grade 2–3 to a particular diagnosis. In this case, the range is detected and the lower and the upper bound are both stored separately.

For lymph nodes, the total number of excised nodes is recorded as is the number of positive ones. If the number of positive notes is stated without specification of the total number of excised nodes, the total number of excised nodes is mapped to "99999," a number bigger than the total number of lymph nodes in a human body. If only the total number of negative nodes is reported, it is assumed to be the same as the number of excised nodes, as this is the only medically consistent conclusion from the given information.

5. Results

In this section we compare MedTAS/P annotations with gold-standard annotations of 101 pathology reports (set 3 from Table 2). The comparisons are between MedTAS/P and the manual annotations (reported as "raw" results) as well as corrected manual annotations (reported as "adjusted"). Results are reported in Table 4.

Table 5 shows the exact 95% confidence intervals for the precision and recall metrics.

The width of the confidence intervals is quite small (0.01–0.05) for the adjusted precision metrics for the leaf classes and slightly increased for the container classes. The relative big confidence interval (0.44–0.53 depending on the metrics) for metastatic tumor reflects the sparseness of the available data.

The errors in the manual annotation ranged from mechanical errors (e.g., picking the wrong text span, specifying the wrong ICD-O code), to more difficult ones (e.g., determining the originating site for a metastatic tumor). These annotation errors were discussed with domain experts to reach a consensus opinion regarding a correct, final manually annotated corpus. The adjusted precision, recall and *F*1-score reported in Table 4 are performance numbers measured against an manually revised annotated corpus.

Without the ability to identify leaf classes (e.g., anatomical site and histology) with high precision, it would be impossible to populate container classes (e.g., primary and metastatic tumor) with precision that is acceptable for clinical research purposes. One of the central components to identify mentions in leaf classes is *conceptMapper*. Detailed experimentation resulted in the following parameters for *conceptMapper*.

Table 5
95% exact confidence interval.

	P (raw)	*R* (raw)	*P* (adj)	*R* (adj)
Anatomical site	0.94–0.97	0.94–0.97	0.96–0.98	0.96–0.98
Histology	0.93–0.98	0.93–0.98	0.97–1.00	0.96–1.00
Dimension	0.99–1.00	0.99–1.00	0.99–1.00	0.99–1.00
Date	0.96–1.00	0.96–1.00	0.96–1.00	0.96–1.00
Grade	0.87–0.97	0.92–0.99	0.95–1.00	0.93–0.99
Gross description part	0.79–0.92	0.78–0.90	0.84–0.98	0.83–0.94
Lymph node	0.73–0.93	0.71–0.92	0.84–0.98	0.84–0.98
Primary tumor	0.70–0.85	0.72–0.87	0.72–0.87	0.76–0.90
Metastatic tumor	0.18–0.71	0.13–0.57	0.45–0.92	0.34–0.80

Table 4
Results for automatically populating the CDKRM with MedTAS/P.

	Precision (raw)	Recall (raw)	*F*1-score (raw)	Precision (adjusted)	Recall (adjusted)	*F*1-score (adjusted)
Anatomical site	0.96	0.95	0.96	0.97	0.98	0.97
Histology	0.96	0.98	0.97	0.99	0.98	0.99
Grade value	0.93	0.97	0.95	0.99	0.97	0.98
Dimension	1.00	1.00	1.00	1.00	1.00	1.00
Date	1.00	1.00	1.00	1.00	1.00	1.00
Gross description part	0.86	0.85	0.85	0.91	0.89	0.90
Primary tumor	0.78	0.82	0.80	0.80	0.84	0.82
Metastatic tumor	0.43	0.32	0.36	0.73	0.58	0.65
Lymph node	0.84	0.83	0.84	0.93	0.93	0.93

1. Case matching: off
2. Stemming: off
3. Word order independent lookup: on
4. Tokens to be skipped: stop words: of, in, with and
5. Context within which matching is executed: sentence
6. Matching algorithms: Start looking for a match at every token.

All filters of *conceptFilter* were applied. Additionally, we considered two versions of the dictionaries for both histology and anatomical sites. The first version, which we call the *base* dictionary, was created as described in Section 3.2. Subsequently, we generated an *augmented* dictionary, wherein we supplemented the base dictionary with all of the alternative forms listed in the National Library of Medicine's SPECIALIST Lexicon [31]. This lexicon contains variant forms for many biomedical terms. This augmentation process involved taking each of the base dictionary entries and generating all possible variants using the SPECIALIST Lexicon data. For anatomical sites, the use of the augmented dictionary resulted in an increased recall but decreased precision which translated into a measurable drop in *F*1-score. On the other hand, no such drop was observed for histology, as both precision and recall were either unchanged or improved by using the augmented dictionary over the base dictionary. Therefore, for the current version of MedTAS/P, we use the augmented dictionary for histology, and the base dictionary for the anatomical site discovery.

The majority of the errors in MedTAS/P in filling leaf classes came from erroneous adjectival attachments and from missing synonyms within the terminology. For instance, from the snippet "colon, right hemicolectomy" MedTAS/P extracted the anatomical site "right colon." This is correctly marked as an error against the gold standard where the term "right" is interpreted as a modifier to the procedure hemicolectomy. Another source of errors is caused by the incorrect interpretations of punctuations such as commas and parentheses, which were used inconsistently within the corpus. For instance, from the snippet "colon (hepatic flexure)" MedTAS/P extracted a single anatomical site "hepatic flexure of the colon" which was deemed incorrect. The gold-standard specified two anatomical sites, colon and hepatic flexure. It is noteworthy, that there was only a single instance of word sense error—the word "head" was used in the context of head of a polyp and not as a body part in its own right. Errors in filling the grade value class were a result of grade specifications not present in the training corpus.

Errors in the leaf classes are a major source of errors in filling the container classes. Other errors are due to incorrect categorization of sizes and anatomical sites. A small number of failed co-referring class discoveries cause the wrong filling of primary tumor classes. Errors arise in merging candidate primary tumor classes where the resulting primary tumor specifies correct and incorrect values for some of its slots. The relatively low precision and recall numbers for filling the metastatic tumor classes are mostly due to a very low number of training and testing instances (see counts in Table 2). For example, one of the lymph node classes in the test set encountered a previously unseen way of expressing lymph node counts that was not represented in the training data.

6. Discussion

The CDKRM we developed is able to capture valuable cancer-related information contained within unstructured pathology reports. In the work reported in this paper we focused on automatically populating a critical subset of the knowledge representation model and showed that it can be implemented with high precision.

We envision this knowledge representation model to be enhanced over time. In particular, it seems that an "uncertainty" attribute would be beneficial to capture hedging which can be found in pathology reports and even more frequently in clinical notes. An example of such hedging is the phrase "mass consistent with lymph node."

The process of manually annotating a corpus proved to be difficult and led to the creation of detailed guidelines. Capturing phrases containing non-repeating head nouns and processing punctuation consistently proved to be especially challenging. Do colons or semi-colons always denote a sentence break? How should parentheses be interpreted? Does the term within parentheses stand on its own right, or is it a modifier to a previously stated term? For example how should one interpret the phrase "colon (hepatic flexure)"—as a single anatomical site "hepatic flexure of the colon" or as two sites "colon" and "hepatic flexure"?

MedTAS/P had several challenges to overcome. Document quality, represented by misspelled words, formatting inconsistencies and errors related to size description and dates along with ambiguous section headings have implications for all natural language processing tools in general. Although it was previously shown that adaptation of basic NLP tools to the medical domain is critical, it was not *a priori* apparent that adaptation to a medical sub-domain would be necessary. This, however, was the case as pathology reports have their own conventions and style which has to be taken into account. In addition, MedTAS/P has to handle language which often does not adhere to common rules of grammar.

We previously alluded to challenges for some of the components within MedTAS/P. One such component is the part-of-speech tagger. Certain out-of-vocabulary words are mislabeled—for instance "nodes" was labeled as a verb instead of a noun in the context of "lymph nodes." A wrong part-of-speech tag has ripple effects through the whole NLP pipeline, as it will often cause an erroneous shallow parse which in turn will cause the wrong determination of context for a certain term or concept. The grammars for general English for the shallow parser were modified for pathology reports. However, additional modifications may be beneficial, in particular to include the use of punctuation within syntactic constructs.

Categorization of classes was done with rules, based on syntactic structures. We experimented with machine-learning based categorization; however the training sets proved too small for satisfactory results.

Although the focus was on malignant cancers, both malignant and benign masses had to be identified to correctly understand the relations between disease characteristics. The inclusion of benign histologies in our underlying terminology was very helpful. It remains an open question whether a more detailed terminology for the anatomy, such as the foundational model of anatomy [32] would improve the results even further. Preliminary experience suggests that it is necessary for modeling invasion.

7. Conclusions

In this paper, we describe a *Cancer Disease Knowledge Representation Model* (CDKRM) and an information extraction system MedTAS/P to automatically populate pertinent parts of the model from unstructured free-text pathology reports. The model was validated against pathology reports describing tissue specimens of patients with colon cancer. Based on this CDKRM, detailed manual annotation guidelines were created and a corpus of pathology reports was then manually annotated by four coding domain experts. Algorithms for automatically populating the CDKRM from free-text pathology reports were developed. The precision and recall of these algorithms were evaluated against the gold-standard annotations: *F*1-scores ranged from 0.9 to 1.0 for most tasks. The *F*1-scores for populating the primary tumor class was 0.82 and

the metastatic tumor class was 0.65. The results for the metastatic tumor models are significantly lower than for all other models. This is mainly due to two factors: (1) there are relatively few metastatic tumor models in the set of reports, so each incorrectly built model has a significant negative impact and (2) metastatic tumor models contain one more leaf class than primary tumors. Since the requirement for model equivalence is that all members must match against the manually annotated gold-standard corpus, this additional leaf class greatly decreases the chance for concordance and thus increases the chance for disagreement.

The *Cancer Disease Knowledge Representation Model* can be expanded as more attributes and relations become desirable. The MedTAS/P system is modular: new cancer-specific components can easily be added to accommodate a modified knowledge representation model. In addition, cancer-specific components can be replaced by non-cancer-related modules to allow for adaptation of MedTAS to an ever changing and increasing number of knowledge representation models within and outside the space of medicine.

Acknowledgments

Our deepest thanks are extended to Dr. Marc Rosenblum and Dr. Victor Reuter from Memorial Sloan-Kettering Cancer Center for jointly developing the initial Cancer Disease Knowledge Model. We are also very grateful to Philip Ogren for his work on developing the Knowtator model and guiding the initial gold-standard development. We are indebted to Debra Albrecht, Barbara Abbott, Pauline Funk and Donna Ihrke, who manually tagged the data, and for their meticulous work that consisted of manual annotation of the pathology reports.

References

[1] Jemal A, Siegel R, Ward E, Murray T, Xu J, Thun MJ. Cancer statistics 2007. CA Cancer J Clin 2007;57(1):43–66.
[2] Fenstermacher D, Street C, McSherry T, Nayak V, Overby C, Feldman M. The cancer biomedical informatics grid (caBig™). Conf Proc IEEE Eng Med Biol Soc 2005;1:743–6.
[3] Phillops J, Chilukuri R, Fragoso G, Warzel D, Covitz P. The caCORE Software Development Kit: Streamlining construction of interoperable biomedical information services. BMC Med Inform Decis Mak 2006;6:2.
[4] <http://caties.cabig.upmc.edu/overview.html>.
[5] Fraqoso G, de Coronado S, Haber M, Hartel F, Wright L. Overview and utilization of the NCI thesaurus. Comp Funct Genomics 2004;5(8):648–54.
[6] <http://www.cdc.gov/Cancer/npcr/datarelease.htm>.
[7] Van Berkum MM. SNOMED CT encoded cancer protocols. In: Amia Annual Symposium Proceedings; 2003. p. 1039.
[8] Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. IMIA Yearbook of Medical Informatics; 2008.
[9] Ananiadou S, McNaught J, editors. Text mining for biology and biomedicine. Artech House; 2006.
[10] Cunningham H, Maynard D, Bontcheva K, Tablan V. GATE: a framework and graphical development environment for robust NLP tools and applications. In: Proceedings of the 40th anniversary meeting of the association for computational linguistics (ACL'02), Philadelphia, July; 2002.
[11] Unstructured Information Management. IBM Syst J 2004;43(3) and <http://incubator.apache.org/uima/>.
[12] <http://domino.research.ibm.com/comm/research_projects.nsf/pages/medicalinformatics.index.html>.
[13] Friedman C, Johnson SB, Starren J. Architectural requirements for multipurpose natural language processor in the clinical environment. In: Proceedings of the annual symposium on computer applications in medical care; 1995. p. 347–51.
[14] Xu H, Friedman C. Facilitating research in pathology using natural language processing. AMIA Annu Symp Proc 2003: 1057.
[15] Coden AR, Savova GK, Buntrock JD, Sominsky IL, Ogren PV, Chute CG, et al. Text analysis integration into a medical information retrieval system: challenges related to word sense disambiguation. In: MedInfo 2007, Brisbane, Australia; 2007.
[16] Fritz A et al., editors. International classification of diseases for oncology. 3rd ed. World Health Organization.
[17] <http://www.cdc.gov/nchs/about/otheract/icd9/abticd9.htm>.
[18] <http://www.hipaa.org>.
[19] Ogren, PV. Knowtator: a plug-in for creating training and evaluation data sets for biomedical natural language systems. In: Proceedings of the ninth international Protégé conference; 2006. p. 73–6.
[20] Noy NF, Fergerson R, Musen MA. The knowledge model of Protégé-2000: combining interoperability and flexibility. In: Second international conference on knowledge engineering and knowledge management, Juan-les-Pins; 2000.
[21] Carletta J. Assessing agreement on classification tasks: the kappa statistics. Comput Linguist 1996;22(2):249–54.
[22] Poesio M, Vieira R. A corpus-based investigation of definite description use. Comput Linguist 1998;24(2):183–216.
[23] Hripcsak G, Rothschild AS. Agreement, the f-measure, and reliability in information retrieval. J Am Med Inform Assoc 2005;12(3):296–8.
[24] Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrica 1934;26:404–13.
[25] Sominsky I, Coden A, Tanenblatt M. CFE—a system for testing, evaluation and machine learning of UIMA based applications. In: LREC; 2008.
[26] Marcus M, Kim G, Marcinkiewicz MA, Macintyre R, Bies A, Ferguson M, et al. The Penn Treebank: annotating predicate argument structure. ARPA Human Language Technology Workshop; 1994.
[27] Coden A, Pakhamov SV, Ando RK, Duffy P, Chute CG. Domain-specific language models and lexicons for tagging. J Biomed Inform 2005;38:422–30.
[28] Boguraev B. Towards finite-state analysis of lexical cohesion. In: Proceedings of the third international conference on finite state methods for NLP, Liege, Belgium; 2000.
[29] Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform 2001;34:301–10.
[30] Kennedy C, Boguraev B: Anaphora for everyone: pronominal anaphora resolution without a parser. In: COLING 1996. p. 113–8.
[31] McCray AT, Aronson AR, Browne AC, Rindflesh TC, Razi A, Srinivasann S. UMLS knowledge for biomedical language processing. Bull Med Libr Assoc 1993;81(2):184–94.
[32] Cook DL, Mejino JL, Rosse C. The foundational model of anatomy: a template for the symbolic representation of multi-scale physiological functions. Conf Proc IEEE Eng Med Biol Soc 2004;7:5415–8.