Journal of Biomedical Informatics 64 (2016) 179–191

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Tumor reference resolution and characteristic extraction in radiology reports for liver cancer stage prediction

Wen-wai Yim [a], Sharon W. Kwan [b], Meliha Yetisgen [a,c,*]

[a] Biomedical and Health Informatics, University of Washington, United States
[b] Radiology, University of Washington, United States
[c] Linguistics, University of Washington, United States

ARTICLE INFO

Article history:
Received 27 May 2016
Revised 9 August 2016
Accepted 7 October 2016
Available online 8 October 2016

Keywords:
Natural language processing
Information extraction
Reference resolution
Radiology report
Cancer stages
Liver cancer

ABSTRACT

Background: Anaphoric references occur ubiquitously in clinical narrative text. However, the problem, still very much an open challenge, is typically less aggressively focused on in clinical text domain applications. Furthermore, existing research on reference resolution is often conducted disjointly from real-world motivating tasks.

Objective: In this paper, we present our machine-learning system that automatically performs reference resolution and a rule-based system to extract tumor characteristics, with component-based and end-to-end evaluations. Specifically, our goal was to build an algorithm that takes in tumor templates and outputs tumor characteristic, e.g. tumor number and largest tumor sizes, necessary for identifying patient liver cancer stage phenotypes.

Results: Our reference resolution system reached a modest performance of 0.66 F1 for the averaged MUC, B-cubed, and CEAF scores for coreference resolution and 0.43 F1 for particularization relations. However, even this modest performance was helpful to increase the automatic tumor characteristics annotation substantially over no reference resolution.

Conclusion: Experiments revealed the benefit of reference resolution even for relatively simple tumor characteristics variables such as largest tumor size. However we found that different overall variables had different tolerances to reference resolution upstream errors, highlighting the need to characterize systems by end-to-end evaluations.

1. Introduction

Reference resolution is the task of identifying expressions in text that refer to the same real-world entities. In natural discourse, humans readily employ reference resolution to thread together discrete pieces of information, creating a cohesive picture of discussed entities for both disseminating and processing information. For example, consider the following excerpt from a radiology report:

The three mentions of the *hypervascular lesion* appear in separate sentences, yet the reader will naturally group them as one real world entity.

Solving reference resolution is imperative to unearthing the complex web of information trapped in clinical narrative text. Unfortunately, the state-of-the-art in reference resolution in the general domain is still limited; capabilities are even more modest in the clinical domain, in which there is a relative scarcity of annotated corpora. Furthermore, in the clinical domain, there are still well-known unsolved text processing problems such as ill-formed, ungrammatical, telegraphic, semi-structured, abbreviation-ridden narratives. This paper focuses specifically on reference resolution for tumor references found in radiology reports.

More precisely, our work targets *coreference* and *particularization* forms of anaphoric references, where *coreference* refers to instances in which two template heads are equivalent, as in Fig. 1. We define a *particularization* relationship as a parent-child reference. For example, in Fig. 2 **Lesions**$^{(1)}$ is a reference that represents a set of items, particularized by **Lesion**$^{(2)}$, **Lesion**$^{(3)}$, and **Lesion**$^{(4)}$.

Additionally, we are interested in the real-world task of automatically categorizing patients into liver cancer staging phenotypes, for which reference resolution is a critical intermediate

* Corresponding author at: Biomedical and Health Informatics, Department of Biomedical Informatics and Medical Education, University of Washington, Box 358047, Seattle, WA 98195, United States.
E-mail address: melihay@uw.edu (M. Yetisgen).

http://dx.doi.org/10.1016/j.jbi.2016.10.005
1532-0464/

Within hepatic segment II/III there is a 14 x 9 mm **hypervascular lesion**$^{(1)}$ is isotense to liver parenchyma on portal venous phase ...

This **lesion**$^{(1)}$ is suspicious for hepatocellular carcinoma.

.....

Impression:

1. **Hypervascular lesion**$^{(1)}$ in hepatic segment II/III with imaging features suspicious for hepatocellular carcinoma.

Fig. 1. Radiology report excerpt.

Lesions$^{(1)}$ consistent with HCC given their enhancement characteristics:

1. **Lesion**$^{(2)}$ in segment 8 measuring 1.2 x 1.3 cm previously measuring 1.3 x 1.7 cm
2. **Lesion**$^{(3)}$ in segment 4A measuring 1.2 x 0.9 cm not clearly seen on the previous study.
3. **Lesion**$^{(4)}$ on the border between segment 4A and 8 measuring 2.9 x 3.5 cm previously measured 2.5 x 2.6 cm

Fig. 2. Example of one reference and its particularizations.

step. To this end, we are motivated to identify three summative tumor characteristic variables important for staging: (1) largest size of a malignant tumor, (2) tumor counts, and (3) whether 50% of the liver organ is invaded by tumors. These are relevant for three liver cancer staging algorithms: AJCC (American Joint Committee on Cancer), BCLC (Barcelona Cancer of the Liver Clinic), and CLIP (Cancer of the Liver Italian Program). Since these variables require aggregate knowledge of tumor-related attributes, an end-to-end evaluation, incorporating reference resolution, using these staging variables would provide a worthy perspective.

In this paper, we (a) detail our annotation for tumor reference resolution and tumor characteristics, (b) present our machine-learning reference resolution and rule-based tumor characteristics annotator approach for our tasks, and (c) report our reference resolution results, as well as an end-to-end result for the final tumor characteristics extraction.

2. Related work

Reference resolution is an active area of research in the natural language processing domain. General english NLP focus on reference resolution has primarily been on newswire text, with several notable information events such as the Message Understanding Conference (MUC) [1] and the Automatic Content Extraction (ACE) program [2]. The OntoNotes project includes coreference annotations across three languages (English, Chinese, and Arabic) for various text [3]. Similar to our goals, one previous work attempts to classify event, subevents, etc. using a pairwise logistic regression classifier [4].

In the biomedical domain, the BioNLP 2011 Shared Task featured anaphoric coreference of biomedical entities, e.g. biological entities, processes, and gene expressions [5].

In the clinical domain, annotation of a variety of concept types, e.g. person, tests, problems, for coreference, has been the focus of the 2011 i2b2/VA Cincinnati challenge [6]. Some difference between our task and that of the 2011 i2b2/VA Cincinnati challenge are the following: (a) we target very few specific mentions (tumor references instead of large classes such as person, test, or problems) and (b) our annotation is based on smaller noun phrase chunks. For example, the i2b2 challenge puts references between long noun phrases which includes descriptors such as: "'a left facial mass", "a right parietal hyper dense and heterogeneously enhancing mass", "an endobronchial tumor of the right upper lobe bronchus", "a 5 mm linear, focal area of enhancement in the left central semiovale". In contrast, our references are between shorter phrases, e.g. "hypervascular lesion" or "tumor". Similar to our task, the Ontology Development and Information Extraction (ODIE) part of the corpus has been annotated with anaphoric references, with identity, set/subset, and part/whole relations [7].

Related works on reference resolution relevant to tumors or clinical findings have been the subject of several works. Coden et al. [8] identified coreferences in pathology reports using a rule-based system. Son et al. [9] classified coreferrent tumor templates between documents with a MUC score of 0.72 precision and 0.63 recall. Sevenster et al. [10] paired numerical finding measurements between documents.

Actual reference resolution tasks vary widely in scope. For example, nouns, pronouns, and noun phrases are common; however, coreference for nested noun phrases or nested named entities, (e.g. "America" in "Bank of America"), relative pronouns, and gerunds may not be annotated in a corpus [11]. Here our references are between the template heads of tumor templates. Our corpus does not include pronominal cases and nested references.

3. Methods

3.1. Dataset

Our dataset is a set of 101 abdomen radiology reports drawn from 160 hepatocellular carcinoma (HCC) patients, annotated for 6 important entities, e.g. *tumor reference* and *measurement* entities, and 7 relations, e.g. *hasSize*. Several examples of the template annotations are shown in Fig. 3.

Entities and entity attributes are described in Table 1, and relations described in Table 2. The total numbers of entities, relations, and strict templates for the were 3211, 2283 and 1006, respectively. The corpus is described in a previous work [12]. The number of relaxed templates which encode, *isNegated*, *hadMeasurement*, and *hasTumorEvid* relations as attributes, and re-attach nested relations to the highest head entity, is 999 (the number drop is due to tumor evidence and negation singletons

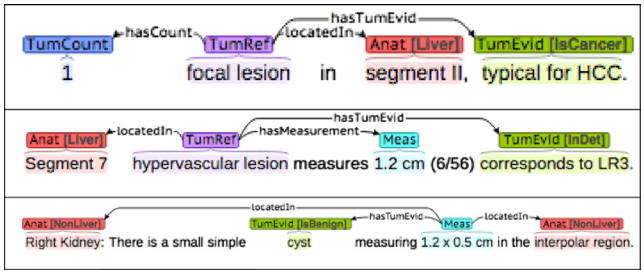

Fig. 3. Three canonical template annotation examples. The last one is a case in which the template head is measurement entity.

Table 1
Entity description.

Label	Description	Freq.
Anatomy	Anatomic locations, e.g. "liver" with attributes (Liver, NonLiver)	1043
Measurement	Measurements findings, e.g. "2.4 cm"	489
Negation	Negation cue, e.g. "no"	73
Tumor count	Number of possible tumors, e.g. "2", "multiple"	174
Tumorhood evidence	Evidence indicating identity, type, or diagnostic information of a tumor referring expression, e.g. "cyst", "suspicious for HCC", with attributes (isCancer, isBenign, inDeterminate)	630
Tumor reference	References to possible tumors, e.g. "lesion"	802
ALL		3211

Table 2
Relation description.

Label	Description	Freq.
hadMeasurement	Past tense indication relation between tumor reference or measurement to another measurement	32
hasCount	Relates a tumor reference its corresponding tumor count	171
hasMeasurement	Relates a tumor reference its corresponding measurement	334
hasTumEvid	For a tumor reference or measurement, marks corresponding evidence to tumorhood evidence	656
isNegated	Relates a tumor reference to a negation cue	75
locatedIn	Identifies anatomy entity where a tumor reference or measurement is found	955
refersTo	Relates a measurement to an anatomy entity, indicating a measurement (rather than a *locatedIn*)	60
ALL		2283

Table 3
Relaxed template frequencies.

Label	Description	Freq.
AnatomyMeas	Events with refersTo relations	53
Negation	Events with isNegated relations	75
OtherSingleton	Events with a single entity, which are not TumorSingleton events	70
TumorSingleton	Events with a single entity, in which the entity is a tumor reference or measurement	157
Tumor	Events not part of the previous event types	644
ALL		999

being removed). The breakdown of relaxed templates by category is detailed in Table 3.

3.2. Annotation for tumor reference resolution and tumor characteristics

Annotation for both reference resolution and tumor characteristics was performed on all 101 reports. 20 reports were used to measure inter-annotator agreement between a medical student and a biomedical informatics graduate student. The rest of the corpus was single-annotated by the biomedical informatics student.

3.2.1. Reference resolution annotation

Our reference resolution annotation are based on two types of relations for tumor-related templates (*TumorSingleton* and *Tumor* templates) heads:

- **coreference** – equivalence relations between mentions, e.g. Fig. 1
- **particularization** – a directed relation in which the first argument represents a set of tumor reference(s) that contains the second argument tumor reference(s), e.g. Fig. 2

Pronominal cases, e.g. "it", "they", and "these" are unmarked.

For our annotation software, we used brat [13], a web-based, annotation software for our reference resolution annotation. Since the number of coreference and particularization relations would visually render the annotations to be highly cluttered, we augmented the software to output text information regarding the clusters and particularizations annotated whenever the user selected a "show references" button, as shown in Fig. 4.

We measured inter-annotator agreement for coreference in terms of MUC [14], B-cubed [15], and CEAF [16] for tumor-related template heads. The agreements were at 0.956, 0.969 and 0.916 F1, respectively. A more detailed description of these metrics and our exact evaluation is described in Section 3.4. For annotator 2, there were 20 clusters (no singletons), 149 clusters (with singletons), with the average size of 2.7 entities per cluster. The cluster-normalized F1 measure for particularization relations was at 0.837.

Some ambiguities did occur between coreference and particularization, which accounted for some of the disparity in inter-annotator agreement. Mainly, as given in the examples of Fig. 5, some mentions are singular but may be equivalent to the plural form of another mention.

The final corpus has 210 clusters (no singletons), 479 cluster (with singletons), with an average of 2.60 mentions per cluster. Inferred particularization relations amounted to 573. The average and median number of sentences between the closest pairwise mentions in the same cluster were 10 and 6 sentences respectively. The large difference between mean and median suggests the existence of some very long-distance coreference relations. The mean proportion of mentions that were exact matches in a cluster was 37%, 43% if normalized for capitalizations. The average proportion of mentions found in the Findings and Impression section per cluster respectively, were 57% and 38%. The proportion of particularization relations that connect mentions in different sections was 47%.

3.2.2. Tumor characteristics annotation

Our tumor characteristics annotation included a spreadsheet that referenced each document name and (1) the number of tumor counts by type (benign, indeterminate, unknown, and malignant), (2) the largest size for malignant tumors, and (3) whether or not more than 50% of the liver is invaded. We decided to mark inequalities, as at times the documents do not in fact give a clear number. Meanwhile, we also collected information regarding the various tumor counts for each of the Findings and Impression sections, as well as the entire document. A sample of this is shown in Fig. 6.

Because the measurement for (3) is not readily quantifiable given the information in reports, we used a series of expert-created guidelines to determine the criteria for (3), as outlined in the Fig. 7.

The inter-annotator agreement is shown in Table 4. The explanation of our evaluation for inter annotator agreement is the same evaluation as those in our system. This is described more thoroughly in Section 3.4.3.

Tumor characteristics annotation were subject to various gray areas. For example for tumor counts, at times there were many ambiguous statements regarding the numbers. One example of this is in the case of conjunctions, several examples of which are shown in Fig. 8.

The first statement can imply either one 5–6-mm focus in segment 6/7 and one in segment 5, or one 5–6-mm touching segments 6/7 and segment 5; or multiple 5–6-mm foci in the areas of segment 6/7 and 5. Similarly, "enhancing area," in the latter

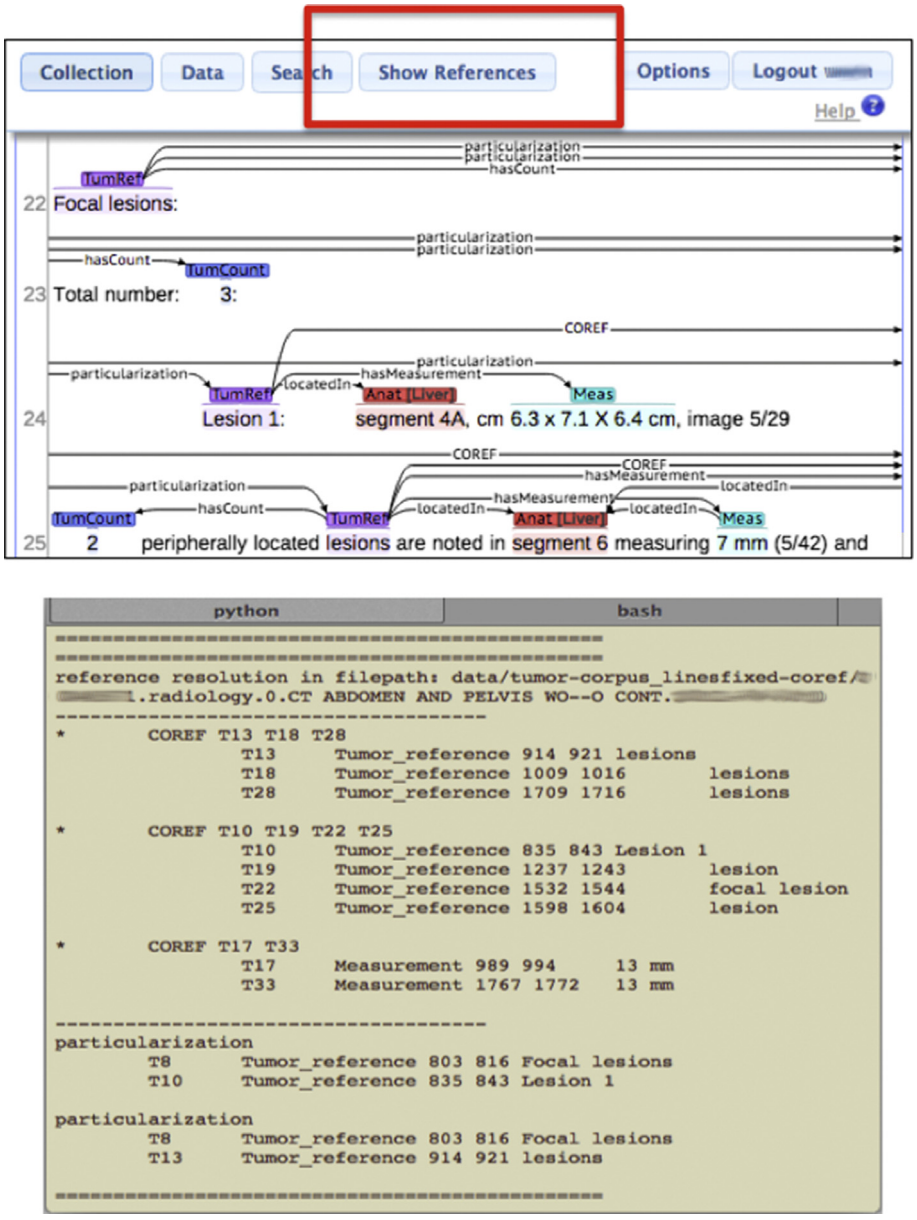

Fig. 4. Brat annotation with augmentations.

Lesion 3: Multiple satellite **lesions** for example, segment 4 measures 2 cm
This **region** is heterogeneously hyper intense with numerous **regions** of focal washout
Mild increase in size of segment 6 hyper vascular **focus** from 1.2 to 2.2 cm with central **area** of nodular hyper vascular focus

Fig. 5. Examples of coreference relations that can be mistaken as particularizations.

statement, may be one large area inside segment 6 and 7 (which are adjacent) or separate areas in 6 and 7. Furthermore, to gather the most accurate number bounds for tumor counts, it was at times necessary to add multiple inequalities, e.g. if there are multiple (but unspecified or only partially specified) lesions in separate areas, which added to the cognitive load.

Annotating the largest size was the least controversial, though this too has some ambiguity. For example the same lesion may have two different measurements in a single report. For example in the Findings section, the largest size might be "2.5 cm" but the same lesion is later referred to as "2.4 cm" in the Impression section. In another example, one measurement mentioned may be specific, e.g. "$6.3 \times 6.1 \times 9.8$ cm", and but later rounded, e.g. "$6 \times 6 \times 10$ cm". Moreover, the amount of text and lengths of the documents, including many possible repetitions, could make it difficult to locate the best representable sizes.

The >50% variable was at times still unclear, even with the guideline. Analyzing Fig. 9 as an example, it is obvious that there are multiple tumors in both the right lobe and in the left lobe; however only 3 segments are specifically mentioned. It is therefore not clear if the unmentioned numerous tumors may be all over the liver or only in those specific parts.

The distribution for the full corpus of the tumor characteristics annotation, binned along critical threshholds, is shown in Table 5.

	A	B	C	D	E	F	G	H
1	FileName	Section	Malignant	Indet	Benign	Unk	>50%	largestmalignant
2	[redacted].radiology.0.CT A	Findings				3		
3		Impression	1		2			
4		Whole	1		2		NO	6.3 x 7.1 x 6.4 cm
5	[redacted].radiology.1.MRI	Findings	1		5			
6		Impression	1		>1			
7		Whole	1		5		NO	6.3 x 7.1 x 6.1 cm

Fig. 6. Tumor characteristics annotation

Tumor extension for malignant tumors are considered over 50% if ANY of the following conditions are met:

1. Tumor ≥ 10 cm
2. >4 segments involved
3. "majority" of "right lobe" involved
4. All right lobe segments involved
5. Entire left lobe plus some right lobe involved
6. Some description to suggest much of liver involved, e.g. massive very extensive

Fig. 7. Logic for >50% of liver invaded.

Table 4
Tumor characteristics inter-annotator agreement.

Label	TP	F1	F1 (relaxed)
Benign	17	0.85	0.95
Indet	18	0.90	0.95
Malignant	17	0.85	0.95
Unk	20	1.00	1.00
LargestSize	17	0.85	0.95
>50%	20	1.00	1.00

5-6-mm **segment 6/7 and 5** hyper vascular **foci**

A small enhancing **area** seen along the lateral aspect of **segment 6, segment 7, and segment 4b/5**

Fig. 8. Conjunction ambiguities.

Focal lesions:

Multifocal HCC with hypervascular washout

In the **right hepatic lobe**, the largest is in **segment 5/6**.

It measures 4.5 x 4.4 cm image 21/7

In the **left hepatic lobe**, the largest is in **segment 4a**.

It measures 3.6 x 3.5 cm

Impression:

Multifocal HCC with malignant vascular invasion of the right portal vein and IVC

Fig. 9. Ambiguity in tumor invasion area.

Table 5
Tumor characteristics annotation distributions, binned according to crucial staging values. The value of "[0, 1, 2–3, >3]" was for a case in which the full number of lesions was given, but it was unclear how many were malignant, resulting in an unknown lesion inequality after subtraction < 5.

Annotation categories		
Tumor counts	Number	Freq.
Benign	0	69
	1	8
	2–3	10
	>3	8
	[2–3, >3]	6
Indet	0	62
	1	19
	2–3	9
	>3	4
	[2–3, >3]	7
Malig	0	3
	1	54
	2–3	25
	>3	13
	[2–3, >3]	6
Unk	0	89
	1	5
	2–3	2
	>3	2
	[2–3, >3]	2
	[0, 1, 2–3, >3]	1
Largest size	Size (cm)	Freq.
	[0,3)	43
	[3,5]	26
	(5–10)	17
	[10,)	10
	n/a	5
>50%	Label	Freq.
	n/a	4
	no	83
	yes	14

3.3. System

Our system consists of two distinct components: the **reference resolution classifier** and the **tumor characteristics annotator**. The reference resolution classifier takes structured tumor templates from a document and categorize which templates are equivalent or in a part-of relation. The tumor characteristics annotator receives grouped tumor templates and outputs (1) the number of tumor for each malignancy category, (2) the largest size malignant tumor, and (3) whether 50% of the liver is taken up by malignant tumors. The two components are further described in Sections 4 and 5.

3.4. Evaluation metrics

Evaluation for our various information extraction goals were measured using F1:

$$F1 = \frac{2PR}{P+R} \quad (1)$$

where $P = precision$ and $R = recall$ used different definitions of instances, precision, and recall depending on each task. In the next sections, we detail the specific instance and F1 definitions for each evaluation task.

3.4.1. Coreference evaluation

For coreference evaluation, we used the F1 metrics for: MUC [14], B-cubed metric [15], and CEAF [16]. We also used the unweighted average of F1 between the three metrics as a separate measure.

3.4.2. Relation evaluation

A relation is a labeled directed connection between two mentions. A correct relation requires the correct label (in this case *particularization*) and the correct identification of the first mention to the second mention. Here, $P = \frac{TP}{TP+FP}$ (precision), $R = \frac{TP}{TP+FN}$ (recall), TP is true positives, FP is false positives, and FN is false negatives.

3.4.3. Tumor characteristics evaluation

Tumor characteristics evaluation was based on the correct label for each document and tumor characteristic variable: (1) tumor counts for benign, indeterminate, malignant, and unknown and (2) largest size for malignant tumors, and (3) whether >50% of liver is invaded. Although we also labeled tumor counts for specific sections in a document (Findings and Impression) we only evaluate values for the entire document in this work.

We also introduced a relaxed match motivated by our specific extraction needs for liver cancer staging for AJCC, BCLC, and CLIP liver cancer algorithms. Based on staging criteria, there are only certain critical thresholds that affect the score. For example, given malignant tumor measurements all under 3 cm, it does not make a difference if our algorithm cannot distinguish between 2 or 3 tumors, or if it cannot distinguish between 5 and 10 tumors; however, if the system cannot distinguish between a single tumor and multiple tumors, the cancer stage is changed drastically. The case is the same for certain sizes. Thus, our relaxed match measures based on the bins discretized from the critical values of our staging algorithms. The bin thresholds are the same as those summarizing our tumor characteristics annotation distribution in Table 5.

4. Reference resolution classifier

4.1. Approach

Our reference resolution classifier consists of a greedy algorithm which visits each template in the order of appearance in each document, and classifies the head of a template as EQUIV, SUBSETOF, SUPERSETOF, and NONE for each available cluster. If the template is EQUIV to one or more clusters, the template is added to the clusters and merged. For all other choices, the template forms a new cluster.

Fig. 10 depicts the choice of a new potential cluster being being classified with one of the relation labels for each available existing cluster. At each round, classifications of the current template with existing clusters are done independently. Relations between clusters are updated at each round. When cycles emerge, all relevant clusters are merged. If there is a conflict due to a NONE classification and another label, the other labels take precedence. Classifications were trained using LibSVM and MALLET, for a support vector machine with a linear kernel with default settings. Feature values were scaled by the difference between the minimum and maximum values. All features (described in Section 4.2) were used for the classification.

After the assignment of EQUIV, SUBSETOF, SUPERSETOF, and NONE for each cluster in a document, the relations were translated back into **coreference** (EQUIV) or **particularization** (SUBSETOF or SUPERSETOF were converted back into a directed relation labeled as a particularization) relations for evaluation.

4.2. Reference resolution features

We detail several types of features shown in Table 6. Classes of these features are described in the following section.

4.2.1. Normalized anatomic location features

If anatomical entities are detected for a template, they are normalized to an anatomic concept. Based on this concept, we designed features based on anatomic hierarchy, e.g. "segment VIII" is contained in "liver". The processing and normalization of anatomic entities is further described in Section 6. Normalization was based on Unified Medical Language System (UMLS) [17] concept names. Relevant related features are **containedIn**, **containerOf**, and **sameLocations**.

4.2.2. Positional features

Whether or not a template appears at the top or near the bottom of the template will affect how many options it will be clustered to and the necessary thresholds for cluster similarity in order to be successfuly classified. We included several features related to the position of a template over all templates in a document. For example, **nthTemplate** gives both the absolute number and the ratio of the template position normalized to the number of all the templates.

4.2.3. Relative features

Relative features identify differences between candidate clusters. For example, **onlySameMal** is in the case of when a candidate cluster is the only one of the candidate clusters which has the same malignancy status. A similar feature exists for same measurement.

4.2.4. Static features

Static features includes a variety of features, such as the **section** of the template, **n-grams** in the sentence, and number of measurements (**numOfMeas**). These features remain the same regardless of what candidate cluster a template head reference is being classified with.

4.2.5. Similarity features

Similarity features (**simvecfeats** and **sim**) are measured from the current template head to be classified to an existing candidate cluster. The similarity with the entire cluster is measured by taking the maximum of each similarity dimension among all the templates in the existing candidate clusters.

Similarity features include the sentence similarity features, tumor reference similarity, as well as similarity between template attributes. For example, tumor reference similarity, measurement similarities, anatomy similarities, and anatomy similarities. The total of all similarity features combine to form a similarity vector of 9 dimensions. Each dimension is described in the Table 7.

5. Tumor characteristics annotator

The tumor characteristics annotator takes tumor templates and reference resolution information as input and delivers the three types of variables for our liver cancer staging (size, number, and whether >50% of liver is invaded), using a series of heuristic rule-based algorithms, as output. The various system components parts are shown in Fig. 11. First the templates are updated to a new malignancy status depending on their coreference and particularization relations to other templates, next the templates are sent

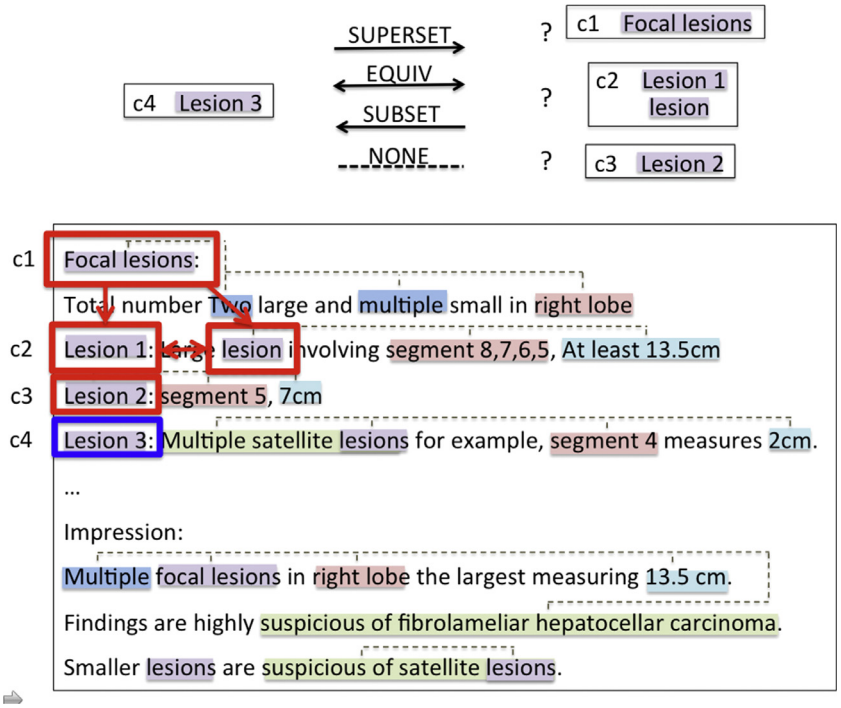

Fig. 10. Reference resolution set up.

Table 6
Reference resolution features.

Feature name	Feature types	Description
closestTempDist	Other	The distance of the closest template in a candidate cluster to the current template
containedIn	Anatomic	If any of the anatomies in the current template are contained in the anatomy in the candidate cluster
containerOf	Anatomic	If any of the anatomies in the candidate cluster are contained in the current template
header	Static	If the sentence of the template looks like a section header
isSuperset	Other	If the candidate cluster is already a superset of another cluster
malignancy	Static	Malignancy status of template
malignancyOfCandCluster	static	Malignancy status of the cluster
nextBestSim	Relative, similarity	L-2 norm of the next best similarity vector
ngrams	Static	1-, 2-, and 3- grams (using lemma) for sentences of template and a candidate cluster
ngramsMatching	Other	Matching 1-, 2-, and 3- grams (using raw words) for sentences of template and a candidate cluster
nthTemplate	Positional, static	The number template in the document
numOfCand	Other	The number of candidate clusters
numOfMeas	Static	The number of measurements
numOfTempInCluster	Other	The number of templates in the candidate cluster
onlySameMal	Relative	The only candidate cluster with matching malignancy as template
onlySameMeas	Relative	The only candidate cluster with matching measurement malignancy as template
sameOrgan	Anatomic	If the organ in the sentence matches organ in a cluster
sameLocations	Anatomic	The matching locations of all
section	Static	Section of the template
sim	Similarity	The L2-norm of similarity vector
simvecfeats	Similarity	This feature extends from the similarity vector features so that each individual similarity vector dimensions are each considered their own feature
summaryOf	Static	If tumor reference is preceded with "the", "this", "these"
totalNumOfTemp	Static	Total number of templates in the document
totalNumOfImpTemp	Static	Total number of templates in the Impressions section
UMLS	Other	Matching UMLS concept between the template and the cluster
Underheading	Other	If there is a sentence belonging in the cluster that looks like a header of the current template

through several various pipelines depending on the chosen variable. In the following sections, we describe several of the non-obvious components in the pipeline: the module for updating malignancy status, the module for classifying whether >50% of liver is invaded, and the module for consolidating referenced tumors.

5.1. Updating malignancy status

The malignancy statuses for related tumor templates are updated in the following way. The malignancy status for coreferrent templates are updated to the most critical case. Thus, anything coreferrent to a malignant tumor template is also malignant; if the

Table 7
Similarity features description.

Target	Description
Sentence similarity	Jaccard proximity for sentence, word-tokenized
Tumor reference similarity	Jarowinkler string proximity
Number of measurements	Difference between number of measurement entities divided by the larger number of measurements
Tumor count similarity	Difference in tumor count divided by the larger tumor count
Matching measurement1	The number of matching measurements divided by the total number of measurements in template 1 (Measurements considered matching if within 0.1 cm)
Matching measurement2	The number of matching measurements divided by the total number of measurements in template 2
Anatomy1	Sum of pairwise jarowinkler proximity for all anatomy entity combinations between template 1 and 2, divided over the number of anatomy entities in template 1
Anatomy2	Sum of pairwise jarowinkler proximity for all anatomy entity combinations between template 1 and 2, divided over the number of anatomy entities in template 2
Malignancy1	The number of matching malignancy status (combined malignancy status' get broken up, e.g. "INDET-BENIGN" becomes "INDET" and "BENIGN"), divided by the total number of malignancy status for template 1
Malignancy2	The number of matching malignancy status, divided by the total number of malignancy status for template 2

most critical status is indeterminate then all templates are updated to indeterminate. In regards to particularizations (superset/subset relations), we take a top-to-bottom approach. The status of the superset is transferred down to the templates in the subset. After this top-down-transfer, the inter-cluster malignancy status is updated once more. Extension of this updating algorithm continuously is left for future work.

5.2. Invasion of >50% of liver logic

The logic for deciding whether or not >50% of the liver is invaded, as shown in Fig. 12. The algorithm is based on the expert guidelines introduced in Fig. 7.

Concepts such as "right lobe", "left lobe", and "liver" are based on the anatomy normalizations from the anatomy normalization module, described in Section 6.

5.3. Reference consolidator

The reference consolidator is responsible for updating templates to the most current set of information and removing extraneous other templates. The premise is to be able to refine all the given information to a few representative templates. For example, if a reference in "Several liver lesions, suspicious for HCC" has the particularizations of "Lesion 1: segment 8, 3.0 cm" and "Lesion 2: segment 5: 2.1 × 1.1 cm" then the template associated with the first passage will be (1) updated with measurements of "3.0 cm" and "2.1 × 1.1 cm", (2) updated with anatomies of "segment 8" and "segment 5", and (3) updated to have a number of "2" for tumor count. Furthermore, if the particularization templates match the malignancy status of its superset template then those are deleted. The final result should yield a set of tumor templates with updated count, measurement, anatomy, and malignancy attributes that can be easily summed to determine the number of each type of tumors found in the radiology report.

Our exact algorithm includes heuristics for deciding for unambiguous cases, for example:

- If the tumor count is set to 3 what happens if there are more than 3 measurements?
- If the tumor count is not reliably determinable, how should it be decided based on the number of associated measurements?

Both coreference and particularization relations are used in the decisions.

6. Anatomy normalizer module

Even with properly marked anatomy entities, concept normalization requires both conjunction normalization as well as concept disambiguation. For example, "segment 2, 4A/B, and 5" must be normalized to "segment 2", "segment 4A", "segment 4B", and "segment 5". Furthermore, "left lobe" may refer to "lung" or "liver".

Anatomy named entity clauses are normalized to discrete concepts by first determining the organ dictionary to use, using the organ context of the sentence. Afterwards, text-spans adjusted to account for missed endings for system entities, e.g. "**segments VIII and V**/IVb", and terms are conjunction-normalized. Finally, concepts are matched based on the lowest score of summing together the matching edit distance with any leftover substrings.

In the following sections, we describe our rule-based algorithms for how to map sentences to an organ context and how to normalize for conjunctions; as well as our automatic creation of organ-specific hierarchal dictionaries using the Foundation Model of Human Anatomy (FMA) ontology [18].

6.1. Mapping sentences to organ scope

In order to differentiate between ambiguous anatomic locations, e.g. "left lobe", the organ context for a sentence must be understood. However, this information is not always available within a sentence, requiring external information. An example of this is shown in Fig. 13.

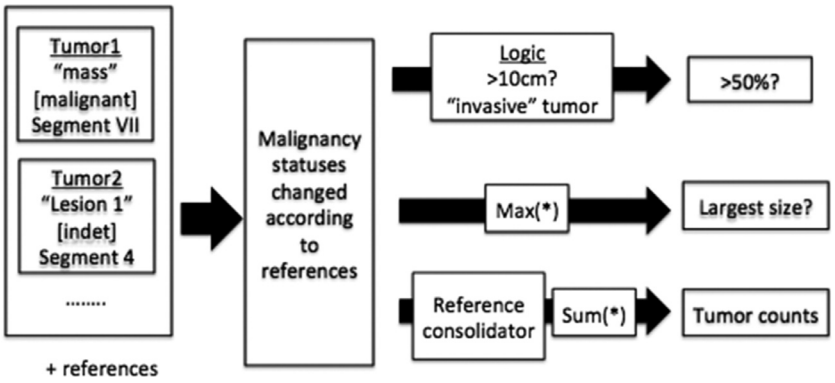

Fig. 11. Tumor characteristics annotator.

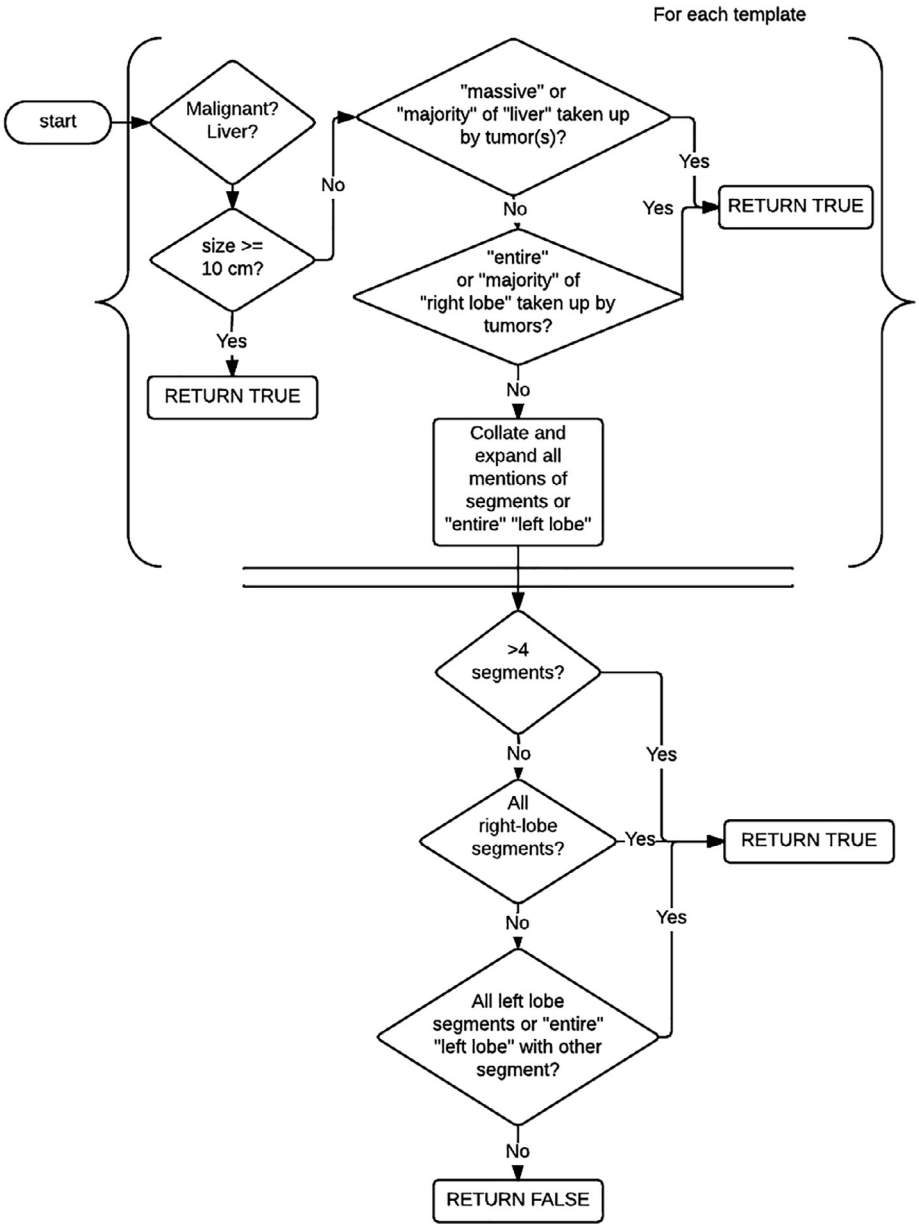

Fig. 12. Algorithm for >50% liver is invaded.

Findings:

Lungs bases: There is calcification of the coronary arteries.

There is a new 1.3 x 0.9 cm a sub pleural nodule in the right base.

No pleural effusion.

Abdomen:

Liver: Nodular cirrhotic liver.

Fig. 13. Different parts of the report have anatomical context not necessarily immediately available in the same sentence or not explicitly clear. In the third sentence, "right base" can be inferred to be part of the lungs by the reference to "Lungs bases" in the previous sentence or the mention of "pleural" in the same sentence.

Our algorithm is detailed as follows. From starting at the beginning of a document to the end, each sentence, previously tagged with UMLS concepts using MetaMap [19], was categorized as related to one or more organ concepts, if these two conditions were met: (1) an anatomic location semantic type was found and (2) the corresponding matched string was matched to the organ dictionary. The list of semantic type abbreviations included in the anatomic location list are: **anst**, **bdsy**, **blor**, **bpoc**, **bsoj**, and **tisu**. The dictionary of organ-related UMLS concept identifiers was created by recursively identifying "is-a" relations starting from the top (non-inclusive) concepts listed in Fig. 14.

Our algorithm additionally assigns organ context by matching to organ-related adjectives, e.g. "hepatic" refers to the liver. The mapping from a organ-related adjective to an organ was created by taking pertainyms from WordNet [20] which point to a MetaMap-matched organ. Examples of the resulting dictionary is shown in Table 8. If no match occurs, the previous line's organ is set for the current line. At the start of each section, the assigned organ is reset to a default state. In our case, the default organ concept was set to the liver.

A manual review of 5 randomly drawn documents (215 sentences) revealed a precision of 94% for this procedure.

"Organ with caveated organ parts" (C0927231)

"Organ with organ parts" (C0927230)

"Nonparenchymatous organ" (C0935295)

"Lobular organ" (C0927223)

"Corticomedullary organ" (C0927224)

"Homogeneous organ" (C0927225)

Fig. 14. Starting organ concept identifiers.

Table 8
Organ adjectives identified using WordNet pertainyms. As bones are considered organs in the FMA, adjective forms of specific bones were also captured (tibial).

Organ	Adjective forms
Kidney	Nephritic, renal, adrenal
Liver	Hepatic
Lung	Pulmonic, lung-like, pulmonary, pneumogastric, pneumonic, cardiopulmonary, intrapulmonary
Prostate	Prostatic, prostate
Spleen	Lienal, splenetic, splenic
Tibia	Tibial

6.2. Normalizing for conjunctions

Conjunctions were normalized by first finding the longest match from organ-specific dictionaries. The automatic creation of these hierarchal organ-specific dictionaries is detailed in Section 6.3. The overlap of the longest match was then intersected with the highest node of the sentence dependency tree. The longest match was determined by finding the terms with the lowest edit distance. Starting from this match, the center-most word is popped off. Then, each unused word from the anatomy entity is paired with the match, ignoring terms such as "and", "or", "/", "-" and ".". The construction of the pairings for "segments 4A and 4B, and 2 and 6", are shown in Fig. 15. The same algorithm was designed to also be used for cases such as "Tumor thrombus within **main, right and proximal left portal veins**".

Our aim here was to provide a way to capture both types of conjunction problems that we encounter for our anatomy entities, such as the right-branching conjunctions of "segments x, x, and x" as well as the left-branching conjunctions of "x, x and x portal veins" in the least assuming way possible. Generalization of this heuristic for other cases is left for future investigation.

6.3. Organ-specific hierarchal dictionary creation

Portions of each organ's hierarchal constituent structures were extracted starting from the organ concept identifiers listed in the previous section. The concepts were collected by recursively following relations: **has_regional_part**, **has_constitutional_part**, and **has_attributed_part**.

Synonym dictionaries for each concept was augmented by adding synonyms in which roman numerals were replaced with numbers (1–12), e.g. "segment II" would be duplicated with variant "segment 2". Synonyms that required mentions of the specific

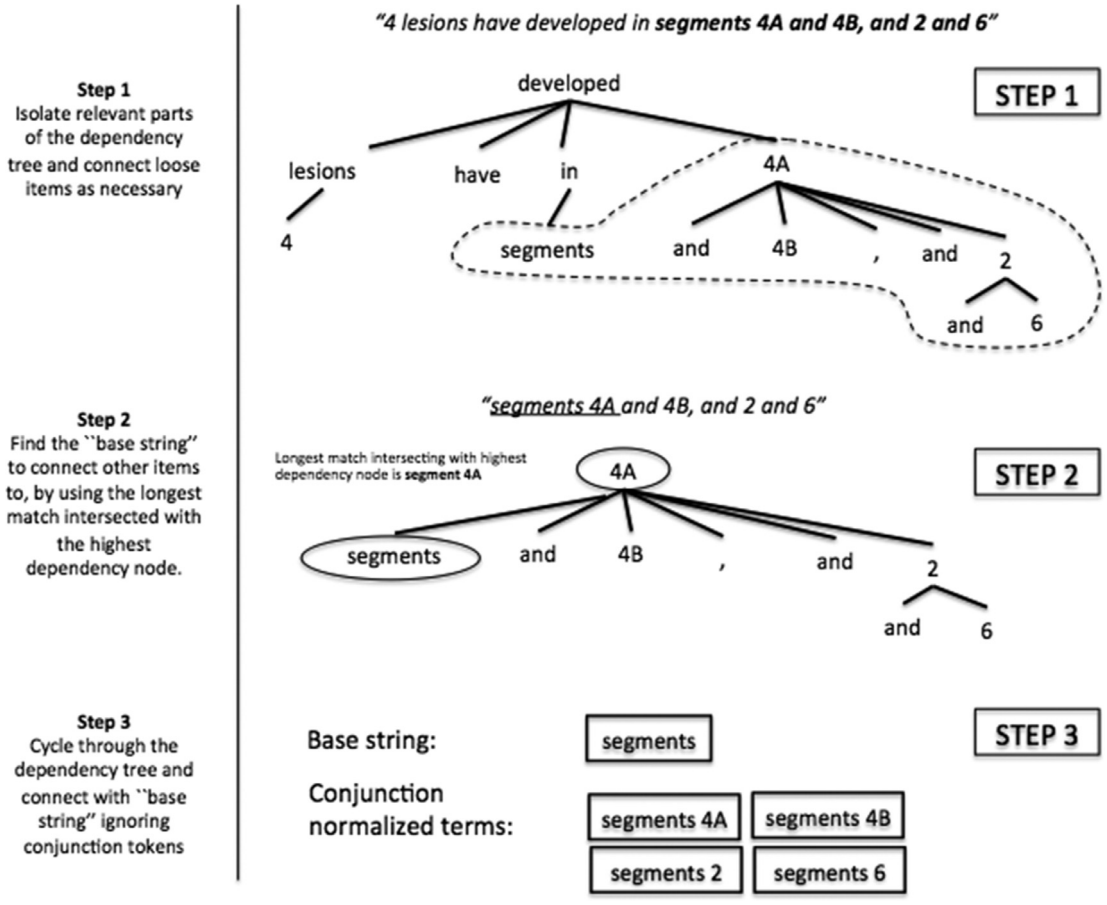

Fig. 15. Conjunction normalization process. **Step 1**: Isolate relevant parts of the dependency tree and connect loose items as necessary. **Step 2**: Find the "base string" to connect other items to, by using the longest match intersected with the highest dependency node. **Step 3**: Cycle through the dependency tree and connect with "base string" ignoring conjunction tokens.

organ, e.g. "right lobe of the liver", were also duplicated with the removal the organ mention to allow better matching, e.g. "right lobe". The following regular expressions were used to identify portions of synonyms to be augmented: "of [organ]$", "⌈organ⌉", "⌈organ-adjective⌉". These regular expressions were created after studying the naming conventions of the FMA. As a caveat, this may not generalizable to all ontologies and is subject to changes of the FMA terminology.

7. Results

Table 9 show coreference and particularization classifications results, using gold standard tumor reference and templates, with a simple baseline of **ngrams** and **ngrams-matching** features compared with our system with the full set of features. Though there was little improvement for particularizations, the coreference performance increased sizably.

In order to quantify how well our tumor characteristics annotator works, we experimented with using no reference resolution, using gold standard reference resolution with gold standard templates, and, finally, using system reference resolution with gold standard templates. The results are shown in Table 10. Given system reference resolution annotations, the tumor characteristics significantly dropped, however the performance remained high for the >50% variable, and dropped less drastically for the largest size variable, compared to those for the tumor count variables.

We were also interested in knowing how the two components affect end-to-end system performances. That is, given **system produced templates**, what is our tumor characteristics annotation results? The comparison results are shown in Table 11. From these results we see the >50% variable remains high, suggesting that it is a variable that is more robust to changes in reference resolution errors as well as template extraction problems. The tumor count variables for all types of malignancies are shown once again to drop substantially. However, this makes sense as even with perfect gold reference resolution, our annotation logic would not get past 0.80 exact F1; furthermore, calculating numbers of tumors requires very exact reference resolution information, making the tolerance for errors very much lower. The largest size variable was the least affected using both the system references and system templates; this accounts to the ability of the metric to absorb errors (it uses a maximum function).

Finally, we experimented with our system to only process certain sections of the document, e.g. Findings only, Impression only, or both (default). We present our results of doing so for our three important variables in Table 12, with different combinations of gold and system reference resolution and template annotations.

While the tumor count variable for malignant tumors was classified more accurately when using only the Impression section, the other two variables benefitted from receiving information across both the Findings and Impression section. Interestingly, the largest size variable performance was much lower for the Impression section compared to the Findings section, which reinforces our observations in which we have found that, in general, detailed information were kept in the Findings section, whereas summary information in the Impression section.

8. Error analysis and discussion

8.1. Tumor reference resolution classification

Analyzing the misclassification of relations, we found that of the 358 FP particularizations, 350 were represented in the gold standard except with the opposites direction (supersetof/subsetof switch) and 27 corresponded to equivalent relations in the gold standard (they may overlap with the supersetof/subsetof switch since they are not mutually exclusive). Similarly, for the 253 FN, 217 were reversed and 34 were related to equivalent relations in the system.

There are many areas for improvement with this classification. Firstly, the greedy merge approach for all coreference and particularization loops is simplistic. An algorithm that resolves this issue

Table 9
Reference resolution results. (P = precision, R = recall, F1 = F1-score).

Evaluation	N-grams + Ngrams matching			All Features		
	P	R	F1	P	R	F1
			Coreference classification			
MUC	0.49	0.35	0.41	0.63	0.54	0.58
B-cubed	0.82	0.72	0.77	0.84	0.78	0.81
CEAF	0.61	0.36	0.45	0.68	0.52	0.59
$\frac{MUC+B3+CEAF}{3}$			0.54			0.66
			Particularization relation			
Particularization	0.51	0.39	0.44	0.42	0.44	0.43

Table 10
Tumor characteristics annotation results (gold standard templates).

	No ref. res.		Gold ref. res.		System ref. res.	
	TP	F1	TP	F1	TP	F1
>50%	94	0.93	95	0.94	95	0.94
#benign	72	0.71	80	0.79	76	0.75
#indet	67	0.66	78	0.77	76	0.75
#malig	14	0.14	70	0.69	56	0.55
#unk	12	0.12	60	0.59	52	0.51
Largest size	80	0.79	94	0.93	87	0.86

Table 11
Tumor characteristics annotation results (system templates).

	No ref. res.		System ref. res.		System ref. res. (relaxed)	
	TP	F1	TP	F1	TP	F1
>50%	89	0.90	90	0.90	90	0.90
#benign	67	0.67	66	0.66	68	0.68
#indet	64	0.64	68	0.68	72	0.72
#malig	18	0.18	50	0.50	62	0.62
#unk	2	0.02	34	0.34	34	0.34
Largest size	63	0.63	77	0.77	79	0.79

Table 12
Tumor characteristics annotation results restricted by section measured in accuracy (gold-templates, gold references/gold-templates, system references/system-templates, system references). The bolded values represent the highest performing sections.

	Section		
	Findings	Impression	Both
>50%	0.81/ 0.80/0.69	0.89/0.89/0.78	**0.94/0.93/0.90**
#malig	0.67/0.56/0.40	**0.69/0.61/0.56**	0.69/0.50/0.50
Largest size	0.76/0.70/0.57	0.43/0.39/0.37	**0.93/0.86/0.77**

by ranking probabilities of each individual relation may provide more nuance at handling conflicts and would presumably lessen the chances of large chain reaction merges. In the general english domain, there are constraints such as pronoun agreement ("John" and "he" vs "her") that are used for coreference systems. We did not implement any such constraints, partly because of our small corpus. However, one idea in this vein could be prohibiting same cluster membership for different "named lesions", e.g. "Lesion 1" and "Lesion 2". Our classification for each template to all candidate clusters were done individually, though perhaps joint classification could yield better results. Finally, our system aggregated clusters from top to bottom in a greedy fashion, allowing the possibility of cascading errors.

8.2. Tumor characteristics annotation

Analysis of the tumor characteristics annotator using gold standard templates and reference resolution annotations revealed some interesting phenomenon.

While the tumor count errors were partly due to our system not producing inequalities (which is required in the gold standard under strict evaluation), it was also due to the heuristic rules of changing malignancy status (only if coreferrent or top-down) and in merging. Furthermore, while particularization hierarchies may go down several levels, we limited our number, measurement, and anatomy update rules to a scope of 3 levels.

In the case of >50% invasion of the liver, there were only a handful of mistakes. One false positive was due to a possible typo in the report (listed as 24 cm in Findings but 24 mm in the Impression), one false negative in which no template was attached to a malignancy evidence finding (it was outside the Findings/Impression section), and one case which was labeled "n/a" due to no size or anatomy in the report at all. The remaining cases included one false negative in which "both segments" was not converted to mean segments 1–8 in the liver and a false positive in which "majority" was not meant to modify "liver" in the sentence. The performance for this was quite high regardless of reference resolution for two reasons. There was a skew in population towards <50% invasion of the liver, which was the default. Secondly, for positive cases, only some documents required reference resolution. Of those that required reference resolution, it was not necessary to be as precise as for calculating tumor counts. For example, a lesion greater than 10 cm may only be known to be malignant through a reference in another sentence. As long as the measurement can be labeled as malignant through either a coreference or particularization relation (regardless which one is true), then the overall <50% invasion of the liver would be easily determined.

For the largest size variable, two errors were due to no malignancy evidence attached to templates, four errors were due to either differences in reported measurements (mistakes or simply precision differences, e.g. 2.5 cm vs. 2.4 cm). Finally, one error was due to malignancy status not being updated in a down-up fashion. This variable improved even with modest reference resolution performance because it only required reference resolution for better assignment of malignancy status; therefore as in the previous case, it requires less precise reference resolution classification. Afterwards, if other measurements with their malignancy statuses are correctly identified, the largest size could be calculated easily.

9. Conclusions and future work

In this work, we present our annotation as well as our system design for tumor reference resolution and tumor characteristics annotation. Although our reference resolution and tumor count results are modest, importantly, our experiments demonstrated several cases in which improvements in reference resolution led to improvements in downstream clinical tasks.

Finding the number of tumors proved to be the most difficult variable, as it requires very precise reference annotations. Meanwhile the other variables, >50% invasion of the liver and largest size, were more tolerant to errors.

Some limitations to this work is that our dataset is small and from a single institution, and annotations were mostly single-annotated. Finally, the corpus annotations in this work are specific for tumors and not generalizable towards general medical concepts.

In future work, we will incorporate our system into our overall patient liver cancer staging system.

Acknowledgments

We thank Tyler Denman for his help with annotation. This project was partially funded by the National Institutes of Health, National Center for Advancing Translational Sciences (KL2 TR000421) and the UW Institute of Translational Health Sciences (UL1TR000423).

References

[1] R. Grishman, B. Sundheim, Message understanding conference-6: a brief history, COLING, vol. 96, 1996, pp. 466–471.
[2] G.R. Doddington, A. Mitchell, M.A. Przybocki, L.A. Ramshaw, S. Strassel, R.M. Weischedel, The automatic content extraction (ace) program-tasks, data, and evaluation, LREC, vol. 2, 2004, p. 1.
[3] OntoNotes Release 5.0 – Linguistic Data Consortium, <https://catalog.ldc.upenn.edu/LDC2013T19>.

[4] J. Araki, Z. Liu, E. Hovy, T. Mitamura, Detecting subevent structure for event coreference resolution, <http://citeseerx.ist.psu.edu/viewdoc/citations;jsessionid=AC6C5BDE654DDC3D6C1940135817B1A7?doi=10.1.1.650.8871>.

[5] J.-D. Kim, N. Nguyen, Y. Wang, J. Tsujii, T. Takagi, A. Yonezawa, The genia event and protein coreference tasks of the BioNLP shared task 2011, BMC Bioinform. 13 (11) (2012) 1–12, http://dx.doi.org/10.1186/1471-2105-13-S11-S1. http://dx.doi.org/10.1186/1471-2105-13-S11-S1.

[6] O. Uzuner, A. Bodnari, S. Shen, T. Forbush, J. Pestian, B.R. South, Evaluating the state of the art in coreference resolution for electronic medical records, J. Am. Med. Inform. Assoc.: JAMIA 19 (5) (2012) 786–791, http://dx.doi.org/10.1136/amiajnl-2011-000784. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3422835/.

[7] W.W. Chapman, G.K. Savova, J. Zheng, M. Tharp, R. Crowley, Anaphoric reference in clinical reports: characteristics of an annotated corpus, J. Biomed. Inform. 45 (3) (2012) 507–521, http://dx.doi.org/10.1016/j.jbi.2012.01.010.

[8] A. Coden, G. Savova, I. Sominsky, M. Tanenblatt, J. Masanz, K. Schuler, J. Cooper, W. Guan, P.C. de Groen, Automatically extracting cancer disease characteristics from pathology reports into a disease knowledge representation model, J. Biomed. Inform. 42 (5) (2009) 937–949, http://dx.doi.org/10.1016/j.jbi.2008.12.005.

[9] R.Y. Son, R.K. Taira, H. Kangarloo, Inter-document coreference resolution of abnormal findings in radiology documents, Stud. Health Technol. Inform. 107 (Pt 2) (2004) 1388–1392.

[10] M. Sevenster, J. Bozeman, A. Cowhy, W. Trost, A natural language processing pipeline for pairing measurements uniquely across free-text CT reports, J. Biomed. Informat. 53 (2015) 36–48, http://dx.doi.org/10.1016/j.jbi.2014.08.015.

[11] V. Stoyanov, N. Gilbert, C. Cardie, E. Riloff, Conundrums in noun phrase coreference resolution: making sense of the state-of-the-art, Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: ACL '09, vol. 2, Association for Computational Linguistics, Stroudsburg, PA, USA, 2009, pp. 656–664. http://dl.acm.org/citation.cfm?id=1690219.1690238.

[12] W.-w. Yim, T. Denman, S. Kwan, M. Yetisgen, Tumor information extraction in radiology reports for hepatocellular carcinoma patients, in: Proceedings of AMIA 2016 Joint Summits on Translational Science, San Francisco, USA, 2016.

[13] P. Stenetorp, S. Pyysalo, G. Topi, T. Ohta, S. Ananiadou, J. Tsujii, BRAT: a web-based tool for NLP-assisted text annotation, in: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, EACL '12, Association for Computational Linguistics, Stroudsburg, PA, USA, 2012, pp. 102–107. http://dl.acm.org/citation.cfm?id=2380921.2380942.

[14] M. Vilain, J. Burger, J. Aberdeen, D. Connolly, L. Hirschman, A model-theoretic coreference scoring scheme, in: Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6–8, 1995, 1995, pp. 45–52. http://www.aclweb.org/anthology/M95-1005.

[15] A. Bagga, B. Baldwin, Algorithms for scoring coreference chains, in: In The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference, 1998, pp. 563–566.

[16] X. Luo, On coreference resolution performance metrics, in: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, HLT '05, Association for Computational Linguistics, Stroudsburg, PA, USA, 2005, pp. 25–32, http://dx.doi.org/10.3115/1220575.1220579. http://dx.doi.org/10.3115/1220575.1220579.

[17] O. Bodenreider, The unified medical language system (UMLS): integrating biomedical terminology, Nucl. Acids Res. 32 (Database issue) (2004), http://dx.doi.org/10.1093/nar/gkh061. pp. D267-270.

[18] C. Rosse, J.L.V.M. Jr, The foundational model of anatomy ontology, in: A.B.B. MSc, D.D. BSc, R.B. BSc (Eds.), Anatomy Ontologies for Bioinformatics, Computational Biology, vol. 6, Springer, London, 2008, pp. 59–117, doi: 10.1007/978-1-84628-885-2_4. <http://link.springer.com/chapter/10.1007/978-1-84628-885-2_4>.

[19] A.R. Aronson, Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program, in: Proceedings/AMIA... Annual Symposium. AMIA Symposium, 2001, pp. 17–21.

[20] L. Bentivogli, P. Forner, B. Magnini, E. Pianta, Revising the Wordnet domains hierarchy: semantics, coverage and balancing, in: Proceedings of the Workshop on Multilingual Linguistic Ressources, MLR '04, Association for Computational Linguistics, Stroudsburg, PA, USA, 2004, pp. 101–108. http://dl.acm.org/citation.cfm?id=1706238.1706254.