UCLA
UCLA Electronic Theses and Dissertations

Title

Towards a Discourse-Level Natural Language Processing Algorithm: Characterizing Tumor Existence, Change of Existence, and its Progression from Unstructured Radiology Reports

Permalink

https://escholarship.org/uc/item/3t0067df

Author

Huang, Ruiqi

Publication Date

2021

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA

Los Angeles

Towards a Discourse-Level Natural Language Processing Algorithm: Characterizing Tumor Existence, Change of Existence, and its Progression from Unstructured Radiology Reports

A thesis submitted in partial satisfaction of the
requirements for the degree of Master of Science
in Bioengineering

by

Ruiqi Huang

2021

© Copyright by

Ruiqi Huang

2021

ABSTRACT OF THE THESIS

Towards a Discourse-Level Natural Language Processing Algorithm:
Characterizing Tumor Existence, Change of Existence, and its Progression
from Unstructured Radiology Reports

by

Ruiqi Huang

Master of Science in Bioengineering
University of California, Los Angeles, 2021
Professor Ricky Kiyotaka Taira, Chair

Cancer has been the second leading cause of death in the US[1]. To provide care for cancer patients and retrospectively study this disease, clinicians and researchers need to manually analyze patient-level medical history to determine whether *a tumor exists*, *has the state of existence changed*, and *does the change implicate disease progression*. With the growing adoption of the electronic health records (EHRs), it is now possible to access these data and automate the discourse-level analysis on unstructured clinical texts using natural language processing (NLP) techniques.

This thesis focuses on developing, training, and evaluating a transformer-based text classification algorithm that will capture contexts from unstructured radiology reports and output the **discourse-level analysis** on the tumor status and its progression through three conceptual frames: *existence of a tumor*, *change of existence*, and *significance of change*. This

is the first clinical NLP work that conceptualize these representations using a wide range of systemic inferences, including contexts from presuppositions. The model shows promising results and can be extended to improve on casual reasoning, logical reasoning, numerical reasoning, and temporal reasoning in the future.

The thesis of Ruiqi Huang is approved.

Corey Wells Arnold

William F. Speier

Dino Di Carlo

Ricky Kiyotaka Taira, Committee Chair

University of California, Los Angeles

2021

To Jewel Zhu and my family!

Thank you for your endless supports.

Contents

Contents **v**

List of Figures **vii**

List of Tables **viii**

1 Motivation **1**

1.1 Representation of Tumor Status and its Progression 4

1.2 Automated Summarization of Tumor Status and its Progression 8

1.3 Thesis Outline 10

2 Related Works **11**

2.1 Existence 11

2.2 Change of Existence and Significance of Change 15

2.3 This Work 16

3 Conceptualizing Tumor Existence from Presuppositions in Medical Texts 19

3.1 Introduction 19

3.2 Learning from Presuppositions Projected by Stable Change-of-state Predicates 24

3.3 Learning from Presuppositions Projected by Directional Change-of-state Predicates 30

4 Materials and Methods **42**

4.1 Materials: Mass Dataset 42

4.2 Methodology 43

5 Results and Discussion **50**

5.1 Results 50

5.2 Discussion 54

6 Conclusion **64**

Bibliography **66**

List of Figures

4.1 Sample output from the pipeline 45

4.2 Overview of our methodology for characterizing *existence*, *change of existence*, and *significance of change for unstructured radiology reports* 46

4.3 BERTbased Text Classification for 1 sub-task within the dataset 48

5.1 Confusion Matrices for *Prior Existence*, *Current Existence*, *Collection-level COS*, *Instance-level COS*, and *Significance of Change* Tasks 53

5.2 Distribution of Mismatch Instances in Test Set Based on the Types of Existence Pairs 56

5.3 Sources of Errors in Incorrect Existence Pairs (N = 162) from Test Set 57

5.4 Types of Errors in Test Set for the *Instance-Level COS* Task (N = 1181) 60

5.5 Types of Errors in Test Set for the *Collection-Level COS* Task (N = 1181) . . . 61

List of Tables

1.1 Examples of mapping from *'starting existence'* and *'ending existence state'* to *'states of change of existence'* 4

2.1 Compare and contrast existing works with respect to their knowledge representation 12

4.1 Corpus Statistics: text characteristics and dataset size 43

4.2 Corpus Statistics: number of class instances in the five conceptual tasks 44

5.1 Precision, recall, and F_1 scores on the test set for the five conceptual frames . . 51

5.2 Difficult Cases in Change of Existence Tasks 62

Acknowledgments

First, I would like to express my sincere gratitude to all of my thesis committees. This thesis would not have been possible without you all.

Words cannot express my gratitude and appreciation towards my graduate advisor, Prof. Ricky Taira who is also my mentor. From day 1, he challenged me to find a more clinically significant project, and I am very glad I did. His knowledge, patience, and supports have guided me through the hardest times and shaped me into a better researcher. He always encourages me to think about the bigger picture and then "just keep laying the bricks!" This is the most rewarding experience I have had at UCLA (besides being a teaching assistant).

Besides my advisor, I would also like to thank the rest of my thesis committees: Prof. Corey Arnold and Prof. William Speier. It is my honor to have them as my committees and participate in the projects they have previously led. The foundations later motivated me to apply these deep learning methods into the works presented in this thesis.

Finally, I owe special thanks to Dr. Ethan Poole from the UCLA Linguistics Department for confirming the linguistics phenomena observed in few important clinical cases, and, most importantly, kindly sharing his notes and insights on various topics in *Semantics* and *Pragmatics*. And, thank you Tianran Zhang for being my collaborator in the annotation process and also in other project(s).

Chapter 1

Motivation

One motivation central to medicine is the establishment of what evidence are observed and/or inferred in a medical examination. Knowing whether *a finding exists*, *has the state of existence changed*, and *does the change implicate disease progression* are critical components that physicians needed to assess. Interpreted by physicians, radiologists and medical professionals, the collected evidence (i.e., findings) and the analysis (i.e., disease progression) become pieces of knowledge summarizing the overall health condition of a patient at that particular visit. With the adoption of electronic health records (EHRs) in the United States (US), EHRs pave the way for an effective communication among physicians and staff, but also enable the sharing of clinical data (containing expertise knowledge on the patient's condition) to stimulate scientific research[2].

However, raw clinical data (e.g., text) do not exist in a generally useful form. The bulk of the clinical information still remains in an unstructured or semi-structured representation, making these knowledge inaccessible for quick retrieval and further analysis. Transforming unstructured **texts** into structured **knowledge** is an important step to unlock the potential for a variety of downstream applications to improve clinical care and expedite research[2, 3, 4].

Particularly in oncology, there are critical needs for summarizing cancer-related information[5, 6]. At a case level, physicians suffer greatly from information overload[7, 8]. Physicians have to review massive amounts of medical history to establish *what exists* and *what has changed* in order to quickly grasp the health condition of their patients. At the cohort level, translational researchers find it challenging to navigate through medical history and identify potential candidates eligible for clinical trials. At the population level, cancer registrars need to manually compile patient-level cancer-related information into hospital-level information, and report them to a higher chain of command in a timely manner[9]. These reported information allow policy makers to analyze trends (e.g., forecast the number of cancer cases), estimate the burden of illness (e.g., costs associated with the episode of care), and strategically allocate limited resources to the areas with high medical needs[10].

Cancer has been the leading cause of death globally and the second in the US[1]. The US alone has diagnosed 1,708,921 cancer incidences during 2018[1]. Though cancer mortality has dropped recently[11], approximately one third of the cancer survivors are still at risk of recurrence[6]. There is a continuous stream of incoming data documenting the observations collected for both cancer survivors and new cancer patients. In order to provide and improve the care for these patients, different parties including medical personnel need to manually perform information extraction and interpret the data (i.e., radiology report). As the number of cases grow, the manual review process can become even more labor intensive. To keep up with the increasing cancer cases and demands of data usage [6, 12], this project aims to focus on characterizing tumor status and its progression.

Summarizing the existence of a tumor can be difficult. As a progressive disease, the journey of cancerous cells have started long before its diagnosis. They first begin as a small group of cells proliferating abnormally, which eventually develop into a tumor[13]. The tumor dynamics vary from patient to patient. It may regress or remain stable, but also metastasize to colonize new body sites. As clues for tumor trajectory, medical care often view the data

as part of a time continuum and monitor the change in such data to assess whether it *has* or *will* become cancerous. Change in existence (e.g., resolved, recurred, still existing, new tumor) and trend in the data values for the tumor characteristics (e.g., size trend) are clues for its regression, stability, or deterioration. Thus, the presence and progression of a tumor between visits are especially important pieces of intermediate knowledge in need for abstraction. The core functionality of such abstraction should answer three questions from an unstructured report: *whether a tumor exists*, *has the state of existence changed*, and *does the change implicate disease progression*. Implemented using deep learning method, this automated discourse analysis can learn to output these interpretations in a structured representation.

The interpretations from this automated discourse analysis have various potential applications in the domain of care. At a case level, I would imagine creating a summary snippet and populating key-elements with the outputted knowledge (Note: the proposed summary snippet is an extension of this work, but will not be covered in this thesis.). Based on the contents presented in the snippet, physicians can prioritize existing or previously suspicious finding, and easily reference its prior state to deduce whether the condition has improved, remain normal, or worsen. This can mitigate information overload issue by reducing the time and efforts spent on skimming through irrelevant information, grasping the meaning from texts, providing a quick quality assessment on the treatment outcomes, and planning a better treatment in the future. Moreover, the outputted knowledge (especially state of tumor existence) can be compiled into a patient-level portfolio sorted by documentation time (e.g., a patient's problem list). After compilation, researchers can quickly identify potential candidates with targeted medical conditions. Lastly, patient-level knowledge can also be reused and aggregated into population-level data as well. Thus, it is imperative to conceptualize and auto-populate the key aspects characterizing the tumor status reported in unstructured text in a structured representation for various downstream applications[6].

Table 1.1: Examples of mapping from *'starting existence'* and *'ending existence state'* to *'states of change of existence'*

States in Change of Existence	starting existence state	ending existence state	Examples of Change of Existence
Neg Change	exists	not exists	"mass is resolved/no longer seen"
No change	exists	exists	"the mass is stable/unchanged/again seen"
	not exists	not exists	"No lung nodules suggestive of recurrence or additional disease can be identified"
	uncertain	uncertain	"no new nodules seen"
Pos Change	not exists	exists	"There is a new mass" or "The tumor recurred after a complete resection"

1.1 Representation of Tumor Status and its Progression

Tumor status can indeed be captured under three separate conceptual representations: *existence*, *change of existence* (**C**hange **O**f **S**tates of Existence a.k.a. COS), and *significance of change*. These three conceptual representations can answer the core questions discussed earlier.

Existence is a state of being of an entity or object (e.g., findings) present at a particular time within the space w, where w is defined as a patient that exists in reality. The description on the tumor is the physician's perception of tumor's presence in $\mathbf{w}$ by observing its representation on visuals or examining tissues collected at a particular time. Since descriptions in EHRs include what evidence (i.e., findings) are observed and/or inferred in the medical examination, the discourse analysis can support and confirm the state of existence for an entity (i.e., tumor) at a particular time in w.

Another aspect is *change of existence* where the trends in existence pairs (denoted as $\langle$prior existence state, current existence state$\rangle$) indicate signs of stability between the interval visits (See *Table 1.1*). Here I define two types of change of existence, one viewed on an

instance level (e.g., 'Stable mass in the right crus of the diaphragm') and the other on a collection/patient level (e.g., 'Multiple other ground-glass nodules unchanged'). On an instance-level, the trend is reported as long as one tumor experienced a change in state of existence (COS). Though infrequently said, doctors may also describe collection-level COS when the patient has a set of tumor(s). In patient-level COS, directional COS[1] on instance-level does not affect collection-level COS unless there is an empty set of tumor in the prior visit or current visit. A directional change in the collection-level COS typically connotes with the initiation/re-initiation and termination of the episode of care for this progressive disease.

Lastly, *significance of change* as the highest level of abstraction concludes whether the change(s) signifies tumor progression over the recent visit. Although categorizing information into definite classes within these tasks above seems intuitive, it is an uneasy task.

As a domain specific sublanguage, clinical texts inherit vagueness, expressiveness, and complexity of other natural languages. First, the expression of certainty is uncommon in radiology report[14, 15, 16, 17, 18]. Recalling from before, a reported finding is a transformed representation (e.g., visual evidence, sampled tissue from biopsy or resection) of an entity in the reality. The transformed representation is affected by the uncertainty in the perceptual system, the perceived field and the interpretation of observers. Thus, observers may express various degree of uncertainty ranging from *uncertain/possibly*, *likely*, to *definite* when describing the knowledge for the conceptual frames discussed above. Second, these abstractions need to represent tumor status as a function of time. Tumor as a non-permanent object may exists at one point and not another; while, COS is a determined outcome related to the previous and current state[19]. Third, the interpretation of tumor status is affected by semantic

[1]Directional COS is denoted as either ⟨do not exist, exist⟩ or ⟨exist, do not exist⟩.

compositionality, unstated implications, and high level of phenomenological knowledge.

To address discourse-level interpretation, this work takes into considerations of a wide range of systemic inferences, not only information directly asserted but those derived from cognitive mechanism like logical reasoning and common-sense reasoning. Information can be divided into at least four types of content: assertions[5], entailments, presuppositions[6], and implicatures[7]. As seen in the examples below *Example (1)*, I showed that assertions can entail other information. The examples demonstrate that contents often go beyond the literal meaning of what's asserted. Existence, in particular, is not always asserted in the form of *There*-sentences (e.g., "*There is a tumor*") or existence statements[2] (e.g., "*A tumor exists*"); instead, they are also commonly found in presupposed contents.

(1) Asserted knowledge

a. "Resection of mass"

entail: bulk removal of mass (potentially no clear margin)

b. "mastectomy on mass"

entail: complete removal of mass

c. "nodule becomes a mass during this visit"

entail: interval enlargement of nodule

d. "The mass measures 5.3 x 3.6 cm compared with 4 x 4 cm (I-8 and I-9 respectively)"

entail: interval reduction in size

[5]Asserted content is content explicitly added to the discourse by utterance of a sentence. For instance, doctor asserts "the patient doesn't have edema", then "edema is absence" is true.

[6]Presupposed content is backgrounded information taken for granted by the utterance of a sentence. Suppose the doctor asserts "the patient doesn't have edema near the tumor", this sentence asserts "the patient has no edema near the tumor" and presupposes "there exists a patient with a tumor".

[7]Implicatures are contents not explicitly asserted or presupposed by the utterance of a sentence but the speaker intends the listeners to conclude.

[2]Existence statement refers to subject-predicate statement of exist.

e. "This nodule is not seen on the prior scan"

entail: New lesion developed

f. "Most of these are unchanged, but one lesion measuring 6 mm is not seen on the prior study"

entail: disease progression with a new lesion developed

Equally as important as assertions and entailments, humans readily take presuppositions for granted and accommodated them into common ground knowledge. During communications, readers/hearers naturally assume certain backgrounded information is true and exist without having the speaker/writer be explicit about it. As seen in Examples (2a-d) below, the existence of this tumor is rather unaffected by the diverse linguistics environments (negation, conditionals, and questionable), but explicitly denying it yields an infelicitous proposition (*Example (2e)*). What's more sophisticated yet interesting is exemplified in *Example (3)*. The change of state propositions here **project** the prior state of existence, presupposing a tumor has existed in the past. Most importantly, these presuppositions contain the existence contents that I need to capture as well.

(2) Simple Existential Presuppositions

a. "There is a cavitation within **this mass**" *Presupposing: the mass exists*

b. "There is no cavitation within **this mass**" *Presupposing: the mass exists*

c. "Is there a cavitation within **this mass**?" *Presupposing: the mass exists*

d. "If there is a cavitation within **this mass**, can you take a note of that?"

Presupposing: the mass exists

e. #[3] "There is a cavitation within **this mass**, but there exists no mass."

[3] # is a symbol for infelicity. An utterance can be infelicitous because it is self-contradictory, trivial, irrelevant, or because it is somehow inappropriate for the context of utterance.

Infelicitous statement

(3) Projected Presuppositions

a. "There is still cavitation within this mass"

Presupposing: the cavitation and the mass exist at some point in the past

b. "Edema surrounding this mass no longer seen."

Presupposing: the edema and the mass exist at some point in the past

I argue that these phenomena are worthy of investigation in the clinical field for various reasons. Tumor itself is often communicated as context to supplement the status of another clinical finding. And, the temporality of existence is typically under-specified in clinical texts, thus recognizing, understanding, and modeling contents projected from these fine grained presuppositions can enhance the quantity and quality of retrieved information directly[20, 21, 22].

Linguists have been studying these presuppositional phenomena for decades [23, 24, 25, 26, 27], and they have noticed that not all inferences from these utterances are drawn logically, but rather some degree of common-sense; thus these interpretations may carry some degree of uncertainty depending on the types of presuppositions. Therefore, when we leverage a wide range of systemic inferences, we should carefully categorize information into the appropriate class, certainty, and temporality for the three conceptual frames.

1.2 Automated Summarization of Tumor Status and its Progression

Radiology reports contain information related to tumor status and tumor progression, but these data need to be analyzed before usage. Manually mapping these unstructured context

in radiology reports into a fixed knowledge representation is extremely time consuming. Thus, there are critical needs to automate this process.

Although there has been great progress and development in clinical natural language processing (cNLP) for various applications[10, 28, 29], very few have automated the characterization of conceptual representations discussed in the previous section. Most of the existing works focused on modeling the tumor's presence as an approximation of existence [3, 19, 21, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48] and very few on COS and its significance [49, 50]. Moreover, prior works mainly focus on the literal meaning of a sentence and none have attempted to reliably extract contents from presuppositions within a discourse to enhance the retrieved information on tumor progression. In sum, relevant works are inadequate in capturing the breadth and depth of representation displayed in clinical texts.

To quickly retrieve an accurate knowledge representation on tumor status for various applications, the main objective of this thesis is to develop a cNLP pipeline that automatically output the knowledge representation for the three conceptual frames by leveraging systemic inferences drawn from unstructured radiology reports. This thesis has three major contributions:

1. investigate and formalize relevant semantic and pragmatic presuppositions to conceptualize tumor existence

2. design and implement a hybrid of traditional and deep learning cNLP pipeline that automatically output the defined knowledge representations for existence (prior and current states), change of existence (instance-level and collection-level), and significance of change from unstructured radiology narratives.

3. evaluate and conduct an error analysis on the results to evaluate trends in commonly missed cases where future improvements can be made.

1.3 Thesis Outline

Chapter 2 discusses related works that focus on classifying tumor status automatically. Specifically, I will point out the strengths and weaknesses of current systems that have tackled this problem. **Chapter 3** provides an overview on semantic and pragmatic presuppositions and, more importantly, show that these phenomena can be adapted to discern tumor existence. **Chapter 4** goes over the annotation scheme and the developed cNLP pipeline that automatically output the defined knowledge representations for the five conceptual frames. **Chapter 5** discusses the collected results and performs an error analysis on the predictions. **Chapter 6** concludes with a discussion on the limitations and future opportunities of such approach.

Chapter 2

Related Works

Having introduced the problem statements in **Chapter 1**, the main goal of this chapter is to discuss existing works that have extracted the knowledge on tumor status and tumor progression from the descriptions in unstructured clinical texts. I compare and contrast the types of information extracted instead of the physical frameworks implemented in existing works. Summarized in *Table 2.1*, I have noticed the inadequacies of current systems in capturing the dimensions around the representations of tumor existence and its progression. Lastly, I conclude this chapter by highlighting the differences between our automated discourse-level NLP algorithm from existing works.

2.1 Existence

Prior works primarily focused on modeling tumor existence rather than *COS* and the *significance* aspects. They aim to provide information on how strong the medical professionals believes of the tumor existence at the medical examination. In practice, existence is approximated by mentioned events and dependency relations through keyword matching, regular expression, named entity recognition (NER), and/or relation detections. These dependency

Table 2.1: Compare and contrast existing works with respect to their knowledge representation

Works	Included Other Clinical Findings (yes, no)	Polarity of Event (presence, absence)	Certainty (possibly, likely, definite)	Time Axis (past, present, date-time)	Instance COS (positive, stable, negative)	Significance of Change (deterioration, stable, improvement)	Highest Level of Understanding (surface, semantic, discourse)
[30, 31, 32, 36, 37]	No	Both	Definite	–	–	–	semantic
[3, 19, 33, 34, 35]	Yes	Both	Definite	–	–	–	semantic
Irvin et al. [38]	Yes	Both	Hedged[1], Definite	–	–	–	semantic
[45]	No	Both	Merged Definite and Possible	Either past or current	–	–	semantic
[39]	Yes	Both	Definite, Hedged[1]	date-time	–	–	semantic
[42, 44]	No	Both	all	date-time	–	–	semantic
[43]	Yes	Both	all	Either past or current [2]	–	–	semantic
[51][3]	–	–	–	–	yes	–	semantic
[52]	Yes	–	–	–	yes	yes	semantic
Cheng et al. [49]	–	–	all	–	–	yes	discourse
This work	No	Both	all	Both past and current	yes	deterioration	discourse

[1] Hedge is referring to a merge of 'possible' and 'likely' into an 'uncertain' state.

[2] This paper has not evaluated the performance in this dimension.

3 For general purpose usage, Vanderwende et al. [51] only proposed an annotation schema for labeling the change of existence.

relations modify the mentioned events to provide information on experiencer (e.g., self or family history), assertion types (conditional[1], hypothetical[2], possible[3], negated[4], asserted[5][40, 53, 54, 55]), temporality (e.g., date time, either past or current, or both past and current), certainty[1] (uncertain, likely, definite), et cetera. These two tasks work in conjunction to compose the meaning from surface expressions.

Certainty of Existence

Existing works have characterized specific dimensions of tumor existence through polarity of an event (e.g., exists and not exists), certainty, and time, but they differ from one another in breadth and depth respectively. Works in [3, 19, 30, 31, 32, 33, 34, 35, 36, 37] capture the contexts on factual existence (presence) and non-existence (absence) only. Other works such as Irvin et al. [38] and Roberts et al. [39] also approximate a level of uncertainty between existence and non-existence by merging *conditional*, *hypothetical*, and *possible* states under a speculative/hedged category. Though Irvin et al. [38] and Roberts et al. [39] have already improved upon the works that only characterized factual existence, they have overlooked the various level of uncertainties expressed in the hedge expressions.

Since hedging is very common in clinical texts, it is very important to separate out expressions intended to avoid full commitment from those with low uncertainty and require followups [56, 57, 58]. Works in Ping et al. [42], Mamlin et al. [43] and Coden et al. [44] have included a more comprehensive certainty representation, ranging from *definite*, *probable*, to *possible* states for both existence and non-existence. This is the certainty representation

[1]The expression of certainty is closely related to assertion task.
[1]Expression falling under conditional scope occurs under specific conditions.
[2]Expression falling under hypothetical scope may be present in patient sometimes in the future.
[3]Expression falling under possible scope could have existed in the patient.
[4]Expression included within negated relation are considered as pertinent absence/does not exist.
[5]Without being modified by other asserted relations, asserted entity are assumed to exist by default.

used in the works of [40, 41] for general purpose NLP. However, only a few have considered the complexities introduced by the dimension of time.

Temporality of Existence

Temporality is one of the most important dimension for existence. For instance, "A mass is observed **in the current examination**" is very different from "A mass is **no longer seen**." in the sense that in the first sentence the mass exists, while the second describes that the mass existed before but not now through the meaning conveyed by the bolded aspectual phrase. We can see that the interpretation of temporal expression can affect the existence of a mentioned entity. Roberts et al. [39], Ping et al. [42] and Coden et al. [44] have captured absolute time reference (DateTime) within the text, while Mamlin et al. [43] and Yim et al. [45] have the content labeled with relative past or current with respect to (w.r.t.) the document time (DocTime) instead[2]. There are problems associated with both representations. Because temporal information is often underspecified, significant amount of temporal context is implicit and conveyed through DocTime[21, 46, 47, 48]. The usage of absolute time reference in the works of [39, 42, 44] is nevertheless less effective in handling the complexity and diversity of temporal context.

Though both the LifeCode System from [43] and the work of Yim et al. [45] claim to be able to classify the relative DocTime (either past or current), the first study has not conducted an evaluation on this aspect. The latter work, in this regard, is more comparable to the work presented in this thesis. The information model in the Yim et al. [45] has included other properties such as polarity (exist, not exist) and certainty (uncertain and definite). However, it can only assign one temporal aspect for the referenced tumor. This

[2]Since the primary goal is retrieving the semantic interpretation and not to recover its exact document time to construct a timeline, we can safely assume there exists a post-processing module that can properly link relative DocTime back to its absolute document time.

representation is expected to have problems when two or more temporal aspects are given. A typical example is when the current visit confirms a previously suspicious finding (i.e., *"suspicious tumor seen in the previous visit is now confirmed on the current visit"*). This example illustrates a dissimilar yet acceptable states of existence at two different visits. In this situation, the prior state will be neglected, resulting in an imprecise transformation of unstructured text to structured knowledge. To sum up, existing works vary in both completeness and depth for the existence frame, but most are unable to handle the temporal intricacies associated with existence characteristics.

2.2 Change of Existence and Significance of Change

I have only identified three works focusing on the instance-level COS and significance of change, but they have all omitted the state of existence and collection-level COS frames. Vanderwende et al. [51] and Hassanpour et al. [52] can detect mentioned change of state statements (e.g., 'again', 'recur', 'resolve', 'new'). Vanderwende et al. [51]'s work proposed a general change of state annotation schema but did not further train an automated system to auto-populate these aspects. On the other hand, Hassanpour et al. [52] were able to depict the certainty associated with the state of change and tumor progression by using rule-based expressions. However, this rule based system cannot understand COS knowledge expressed indirectly. To handle COS embedded in situational knowledge (e.g., "There are two nodules previously but now there are three."), this system must hand-code additional symbolic rules to capture these knowledge.

Most importantly, tumor existence, COS, and progression can be conveyed through changes in its attributes. Most of the prior works simplified the complexity of these tasks. Cheng et al. [49] is the only discourse model that use value changes in tumor attributes (e.g., mass effects and size trend) and change in existence (e.g., recurrence, remission, and

development) as indicators of COS, but their work has not utilized multi-step reasoning or other types of intuitive judgements like presupposed contents.

2.3 This Work

Highlighted in *Table 2.1*, my work differs from existing works in three major aspects: coverage of dimensions, depth, and level of understandings. It covers the important dimensions of existence and COS through the descriptive axes of polarity, certainty, and temporality. Given the clinical text, this automated discourse analysis will return the certainty in knowledge perceived and the classes in tumor existence at recent visits (past and current), and the COS during the interval visit. This work is less thorough compared to Cheng et al. [49] in *significance of change*, only capturing whether the tumor has worsen or not. But, the richness of representations in this work is the most comprehensive compared to the those discussed in the earlier section. Most importantly, this is the only work that utilize semantic and discourse level interpretations for these tasks.

Evidence for inferring the intended meaning of existence can be divided into three levels of text understanding: (1) surface level[1], (2) semantic level[2], and (3) discourse level[3]. Prior works discussed in this chapter stayed mostly in a mixture of surface level and semantic level using explicitly mentioned events and dependency relations. The level of details depends on the types of semantic relations modeled. Primarily focusing on factuals, hedges and temporal relations, existing works have overlooked other types of semantic relations affecting the tumor status. Take causal relation as an example, surgery (i.e., surgical removal, biopsy, lesioning

[1]Surface level understanding concerns about what is mentioned, i.e., keyword matching.

[2]Semantic level understanding concerns about the meaning of the sentence from the meanings of its parts.

[3]Discourse level meaning is knowing what the information that speaker is trying to convey, namely situational knowledge. Take this as an example 'micro-nodule becomes a nodule during this visit', the speaker is trying to convey there's an interval enlargement but this is not said explicitly.

therapy) may have a negative (e.g., remove an old finding), neutral (e.g., has no impact), or even positive (e.g., induce a new finding) effect on the existence. Another type of relation is aspectual phrases[4] that modify the temporal reference of an entity, and reveal the states of existence at two different timepoints. In contrast, this work can model a range of semantic relationships including the effects of aspectuals and causal relations on tumor existence.

Beside the explicitly mentioned events and semantic relations, this work utilizes a wide range of inferences to achieve a discourse level interpretation. Similar to Cheng et al. [49], the model can learn to use indicators (e.g., tumor's properties) to determine existence, COS, and progression. Their model utilizes upward entailments (descriptions on the object's attributes entails there exists an object) to enhance the level of understanding. However, their discourse model only applied a limited set of inference rules, still far from drawing systemic inferences. Our work aims to capture other types of inferences involving higher level of semantic and pragmatic meanings based on logical and phenomenological reasoning.

To better approximate human-level understanding, we tap into presuppositional contents to enhance the discourse level interpretation. Briefly discussed in the previous chapter, tumor mentions can be used as background information to supplement another finding in focus. Some of these presupposed contents may come from pragmatic meanings, governed by cognitive mechanism other than logic (i.e., common-sense reasoning) and may inherently carry some degree of uncertainty. This is the only work in the field to adapt, formalize, and apply semantic and pragmatic presuppositions to infer tumor existence (w.r.t. time) and categorize the certainty in that belief.

Our cNLP system is a hybrid model that uses traditional NLP techniques to process the document into sentence(s) and BioBERT-based text classifiers (pre-trained biomedical language model stacked with multi-class classifier layer) to output our desired structured

[4]Aspectual phrases contain a change of state dimensions like '*continue*', '*cease*', '*start*'. For instance in '*Tumor is again*$_{asp:continue}$ *seen*', the aspectual phase 'again' reflects the tumor exists previously and currently as well.

representations. Since the core-component is a supervised text classifier, this data-driven approach does not have to hand-code semantic relations and symbolic rules to mimic human-level systemic inferences. By annotating the clinical texts according to the guidelines (which also uses the formalized presuppositional rules at the end of *Section 3.1*), annotators have infused semantic and pragmatic level of systemic inferences into the system as it learns from the data itself. To the best of my knowledge, no prior works have integrated a discourse level interpretation to characterize the existence of mass finding at relative DocTime, collection-level COS, instance-level COS, and significance of change between the recent interval visits.

Chapter 3

Conceptualizing Tumor Existence from Presuppositions in Medical Texts

3.1 Introduction

Chapter 1 has pointed out that the state of existence is often implied through the context while its temporality is often under-specified, so it is crucial to capture the background knowledge (e.g., presuppositions) that speakers/writers take for granted to support the main communicative intents of the sentence. A common case containing a presupposition is shown here: "There is a cavitation within **the nodule**". Although this example does not explicitly assert whether the nodule exists or not, the readers of this sentence can easily reconstruct that "a nodule exists" from the bolded phrase because the main communicative intents (e.g., there exists a cavitation and it has a location) presuppose there exists a nodule for this statement to be true. The goal of this chapter is to conceptualize the rules for annotating

the state of existence evidenced in presupposed contents, so our NLP system can learn these interpretations from the annotated data.

What is presupposition?

In order to answer *what* is presupposition, let us first start with one characteristic of presuppositions: non-at-issue contents. *Example (4)* shows instances where the existence of a tumor is directly **asserted** by being the subject of a main proposition *(4a)* versus **presupposed** as the main subject (e.g., cavitation) that references the mass during the utterance*(4b)*.

(4) Mass Finding under At-issue and Not At-issue Content

a. **At-issue**: There is **a mass** seen on the current examination.

b. **Not at-issue**: There is a cavitation within **the mass**.

(5) **An Existential Presupposition in *Example (4b)***

a. P_{t_0}**:** There is a cavitation within **the mass** (at time t_0).

b. $\neg\ P_{t_0}$**:** There is no cavitation within **the mass** (at time t_0).

c. P_{t_0} **and** $\neg\ P_{t_0}$ $>>$ **Q:** There exists a mass (at time t_0).

As seen in *Example (5)*, the bolded backgrounded content*(5c)* is unaffected by the negated linguistic environment. That can be attributed to the phenomenon that a change to the linguistic environment of the original proposition*(5b)* targets the meaning of the main communicative content, not the non-at-issue content (e.g., state of tumor existence). Specifically, not at-issue content should hold even when **presupposition "holes"** (predicates that complement the original sentence) transform the linguistics environments of the original proposition using the negated operator ($\neg$P or "It is not the case that P"), probable operator ($\diamond$P or "It is probably the case that P"), question operator (?P or "Is it the case that P?"),

and conditional operator ($IF[P, Z^1]$)[59] or "If P, then Z"). Presupposition is non-at-issue content taken for granted in order for the proposition to hold. For instance, the conventional existential presupposition – there must exists a referent for a referential expression to hold [60, 61] – **survives all the presupposition holes** as seen in *Example (6).*

(6) There is a cavitation within the **nodule** (at time t_0). P_{t_0}

a. There is **no** cavitation within the nodule (at time t_0). $\neg\ \mathrm{P}_{t_0}$

b. There **might be** a cavitation within the nodule (at time t_0). $\diamond\ P_{t_0}$

c. **Is** there a cavitation within the nodule (at time t_0)**?** $?\ P_{t_0}$

d. **If** there is a cavitation within the nodule (at time t_0), **then** doctor will report it. $IF[P_{t_0}, Z]$

Adapted from the definition provided in the works of Stalnaker [25], Potts [27] and Karttunen [59], presupposition Q is formally defined as proposition P presupposes Q ($P >> Q$) if and only if Q must be true in order for P to have a truth value[2], and Q as a non-at-issue content survives the presuppositional holes. If $P_{t'} >> Q_t$, then Q_t should hold true in $O(P_{t'})$, where $O(P_{t'})$ is an operator on the proposition and $O \in \{\neg, \diamond, ?, IF\}$. Putting into context, suppose the proposition P_{t_0} in *Example (5)* is true, then existential presupposition Q_{t_0} has a truth value; or, in another words, the described mass exists.

Uncertainty in contents derived from projected presuppositions

Besides the simple existential presupposition discussed earlier, proposition can occasionally project other types of **non-asserted** semantic presupposition (sem-ps) and pragmatic presupposition (prag-ps) which can rely on cognitive mechanism in additional to logical

[1] where Z is another proposition

[2] or not undefined

reasoning. Sem-ps and prag-ps have been studied intensively; however, their differences are debatable[25, 26, 27, 60]. The general rule of thumb is that sem-ps is not contextually dependent because it is **entailed**[1] by the underlying meaning of asserted proposition. While prag-ps (conversationally-triggered presupposition) is contextually implied and projected by the intents of a rational speaker, assuming this **rational agent**[2] complies to the maxims of rational communications[3] (expanded in *Section 3.3*). Thus, it is reasonable to believe that this conversationally-triggered presupposition is implied given the context of a proposition. Consequently, the certainty of entailed contents is stronger than the certainty of implied contents.

To avoid ambiguity, I strictly define "P entails Q" as "P necessarily results in Q", and "P implies Q" as "P suggests but not necessarily entails Q". Thus, if P **implies** Q and P' is P with additional context, it is not necessarily true that $P' >> Q$. Thus, entailment is not implication and vice versa. Therefore summarized in *Equation (3.1)*, we can safely assume if P presupposes and entails[3] Q ($P >> Q^{\text{sem}}$) and P is true, then Q^{sem} must be true based on logical reasoning. Moreover, if P presupposes and implies Q ($P >> Q^{\text{prag}}$), then Q^{prag} is probably true given the context in P.

$$\text{Certainty}(P_{t'}, Q_{t''}) = \begin{cases} \text{Definite} & P_{t'} >> Q^{\text{sem}}_{t''} \\ \text{Probably} & P_{t'} >> Q^{\text{prag}}_{t''} \end{cases} \tag{3.1}$$

[1]In general, entailment includes at-issue and not at-issue entailment. We will not discuss at-issue entailment because conventional presupposition is a type of not at-issue entailment. If a proposition is an at-issue entailment, then negating that proposition will contradict with the original proposition. Presupposition, on the other hand, will survive that test because background information is assumed true in regardless.

[2]Rational agents are referring to speakers and audiences following the maxims of rational communication.

[3]The four maxims of rational communication proposed by Grice [62] include the maxim of quality, maxim of quantity, maxim of relevance, and maxim of manner.

Conceptualizing State of Existence from Projected Presuppositions

To tap into a discourse level understanding, this chapter is dedicated to conceptualize the state of tumor existence and the certainty of that belief in presupposed contents. Specifically, I targeted prevalent clinical cases where the tumor is a referent of another finding (i.e., edema, cavitation) experiencing a change in the state of existence (COS) as seen in *Example (7)*. There are two types of presuppositions seen; one is the conventional existential presupposition (e.g., (7b, 7d)), and the other is presupposing the prior state of change (e.g., *(7c)*) triggered by the aspectual predicate. The state of existence for the mentioned mass can be evidenced under the original proposition*(7a)* and presupposed content*(7c)*, but categorizing their certainties is rather non-trivial because the certainty of presupposed content is affected by the mode of presuppositions (semantic versus pragmatic).

(7) **Change of Existence Predicates Projected Existential Presupposition in Current Visit and Previous Visit**

a. Directional COS **P**: The cavitation **developed** in the mass (at time t_0).
Stable COS **P'**: No cavitation within the mass **as before**.

b. P and P' >> Q_{t_0} : the mass exists (at time t_0).

c. ? P and P' >> $Q_{t_{-1}}$: Cavitation is not within the **mass previously**.

d. ? $Q_{t_{-1}}$ >> $Q'_{t_{-1}}$: The mass existed **previously**.

To resolve these problems, the goal of this chapter is to model the state of existence and the certainty in that belief from existential presuppositions stated in both the end-state and the preconditions of COS predicates. The rest of the chapter proceeds in the following order. *Section 3.2* goes over the tools to discriminate sem-ps from prag-ps and shows that the prior

and end state of **stable COS predicates** (SCOS) project two existential presuppositions based on pure logical inferences. Next in *Section 3.3*, I reason that clinicians satisfy the assumptions of rational speakers and then show the states of existence are evidenced in the presuppositions projected by **directional COS predicates** (DCOS, e.g., recur, resolve, develop)[3]. The rules from *Section 3.2* and *Section 3.3* are summarized below.

When a statement has a predicate whose subject references the non-at-issue content, tumor, at time t, the statement presupposes there exists a tumor at t, and

(1) there **exists** a tumor in the visit prior to t if the predicate is a stable COS predicate

(2) there is **probably** a tumor in the visit prior to t if the predicate is a directional COS predicate

3.2 Learning from Presuppositions Projected by Stable Change-of-state Predicates

Before verifying the SCOS predicate projects existential sem-ps at two relative time-points and assigning the certainty on those contents, this section further discusses a tool needed to depict the mode of presupposition.

Unlike Q_t^{prag}, Q_t^{sem} is most notable for being entailed and not implied, so it will fail the defeasibility/cancellability test. To clarify, the **defeasibility test** is denoted as "P and **it is not the case that** Q" and it aims to evaluate Q's contextual dependency through cancellability. Provided that P is true and specific word(s) or expression(s) in P **entails** and takes Q_t^{sem} for granted, then Q_t^{sem} must be true and Q_t^{sem} is non-cancellable and not context-dependent. One can show that $Q_t \in Q_t^{\text{sem}}$ if Q_t passes the presupposition holes and fails the defeasibility test using proof by contradiction.

[3]If unclear on the notion of DCOS, please refer to Chapter 1 for more detail definition on positive COS and negative COS.

Certainty in the Existence Information Derived from Conventional Presuppositions Projected from the SCOS Predicates

Using presupposition holes and defeasibility test discussed above, I identified and validated that SCOS projects sem-ps. As seen in (c-d) from *Example (8)* and *Example (10)*, SCOS projected sem-ps, revealing the state of tumor's existence when the main subject of the predicate references the tumor in the prior state.

The general scheme of the proofs in *Example (9)* and *Example (11)* is generalizable for similar cases. I've illustrated two ways to prove the projected existential presupposition in the prior visit and current visit (Examples (9c, 11c)). The shortest method is to identify at-issue entailment, $P_{t_{-1}}$, speaking of the past. Then existential presuppositions within P_{t_0} and $P_{t_{-1}}$– stating the tumor's existence at both the prior (t_{-1}) and current visit (t_0)– can be taken for granted[4]. The alternative method in *Example (9d)* and *Example (11d)* show that $P_{t_{-1}}$ entails the existential presupposition by surviving the presupposition holes and failing the defeasibility test. These properties indicate $Q_{t_{-1}} \in Q^{\text{sem}}$. Finally, using *Equation (3.1)* from *Section 3.1*, I concluded that SCOS predicate projects the tumor existed during the utterance time and the visit prior to that.

(8) **Stable COS in Mass Effect >> *Example (8c)* & *Example (8c)***

a. P_{t_0}*:* The left lateral ventricle is obliterated by the mass, **as before**$_{asp:cont}$

b. P_{t_0} *entails* $P_{t_{-1}}$: The left lateral ventricle is obliterated by the **mass** previously.

c. $P_{t_0} >> Q^{sem}_{t_0}$: The mass exists currently.

d. $P_{t_{-1}} >> Q^{sem}_{t_{-1}}$: The mass existed previously.

[4]without running through the presupposition holes and defeasiblity tests.

(9) **Deriving semantic presuppositions *Example (8c)*& *Example (8d)* for causal effects in *Example (8)***

a. Given: "A left lateral ventricle is obliterated by the mass, **as before**$_{asp:cont}$."

b. **Show $Q_{t_0}^{\mathbf{sem}}$ (The mass exists)**. The existential presupposition $Q_{t_0}^{\text{sem}}$ is evidenced in P_{t_0}, thus the mass exists.

c. **Show $P_{t_0} >> Q_{t_{-1}}^{\mathbf{sem}}$ (The mass existed previously)**.

P entails at-issue proposition $P_{t_{-1}}$ ("A left lateral ventricle is obliterated by the mass previously.")

'As before' entails the ventricle is previously obliterated and also by the same cause, mass. Since we know that P_{t_0} is true, then $P_{t_{-1}}$ is true by entailment.

Show $P_{t_{-1}} >> Q_{t_{-1}}^{\mathbf{sem}}$. The existential presupposition $Q_{t_{-1}}^{\text{sem}}$ is evidenced in $P_{t_{-1}}$, thus the mass existed previously.

d. **Alternatively show $P_{t_0} >> Q_{t_{-1}}^{\mathbf{sem}}$ (U = The mass existed previously)**.

Show U is non-at-issue presupposition by passing the presupposition holes ($O(P_{t_0})$ where $O \in \{\neg, \diamond, ?, IF\}$). $O(P_{t_0})$ does not affect the consistency of U (in parenthesis) appended in the background, suggesting U is non-at-issue content.

(i) "(A mass existed previously.) A left lateral ventricle is not obliterated by **the mass**, as before."

(ii) "(A mass existed previously.) A left lateral ventricle is probably obliterated by **the mass**, as before."

(iii) "(A mass existed previously.) Was the left lateral ventricle obliterated by **the mass**, as before? "

(iv) "(A mass existed previously.) If the left lateral ventricle is obliterated by **the mass** as before, then the doctor should take note of it. "

Use proof by contradiction to show U fails the ***defeasibility test****.* Suppose U passes the defeasibility test (P and ¬ Q), then 'A left lateral ventricle is obliterated by the mass as before, **but the mass did not exist previously**.' is true. Since we know that 'as before' entails the ventricle is obliterated and by the same cause, then the mass must have existed, which contradicts with 'the mass did not exist previously'. Thus, we show that U fails the defeasibility test.

In all, $P_{t_0} >> \mathbf{U}$ and U $\in Q^{\text{sem}}_{t-1}$ because it is an entailed and non-at-issue presupposition.

(10) **Stable COS >> *Example (10c) &Example (10d)***

a. P_{t_0}*:* Mild **atelectasis**$_{another finding}$ is present adjacent to the upper lobe lung mass, and is **slightly worse**$_{asp:continue}$ than on the previous examination

b. P_{t_0} *entails* $P_{t_{-1}}$*:* Mild **atelectasis**$_{another finding}$ is present adjacent to the upper lobe lung mass previously

c. $P_{t_0} >> Q^{sem}_{t_0}$: The upper lobe lung mass exists.

d. $P_{t_{-1}} >> Q^{sem}_{t_{-1}}$: The upper lobe lung mass existed previously.

(11) **Deriving semantic presuppositions *Example (10c)* & *Example (10d)* in *Example (10)***

a. P_{t_0}: Mild atelectasis is present adjacent to the upper lobe lung mass, and is **slightly worse** than on the previous examination "

b. **Show** $P_{t_0} >> Q^{\mathbf{sem}}_{t_0}$ The existential presupposition $Q^{\text{sem}}_{t_0}$ is evidenced in P_{t_0}, thus the mass exists currently.

c. **Show** $P_{t_0} >> Q^{\mathbf{sem}}_{t_{-1}}$ **(The mass existed previously)**.

P entails at-issue proposition $P_{t_{-1}}$ ("Mild atelectasis is present adjacent to the upper lobe lung mass previously, and is slightly better than on the current examination") 'Is slightly worse than on the previous examination' entails 'Mild atelectasis is present adjacent to the upper lobe lung mass previously and it is slightly better than current examination.'. Since we know that P is true, then $P_{t_{-1}}$ is true by at-issue entailment.

Show $P_{t_{-1}} >> Q^{\mathbf{sem}}_{t_{-1}}$. The existential presupposition $Q^{\text{sem}}_{t_{-1}}$ is evidenced in $P_{t_{-1}}$, thus the mass existed previously.

d. **Alternatively, show** $P_{t_0} >> Q^{\mathbf{sem}}_{t_{-1}}$ **(U = The mass existed previously)**.

Show U is non-at-issue presupposition by passing the presupposition holes ($O(P_{t_0})$ where $O \in \{\neg, \diamond, ?, IF\}$). $O(P_{t_0})$ does not affect the consistency of U appended in the background (in parenthesis), suggesting U is non-at-issue content.

(i) "(The mass existed previously.) Mild atelectasis is not present adjacent to the upper lobe lung mass, and is slightly worse than on the previous examination."

(ii) "(The mass existed previously.) Mild atelectasis is probably adjacent to the upper lobe lung mass, and is slightly worse than on the previous examination."

(iii) "(The mass existed previously.) Was the mild atelectasis is present adjacent to the upper lobe lung mass, and is slightly worse than on the previous examination?"

(iv) "(The mass existed previously.) If mild atelectasis is present adjacent to the upper lobe lung mass, and is slightly worse than on the previous examination, then doctor should take note of it."

Use proof by contradiction to show U fails the defeasibility test. Suppose U passes the defeasibility test (P and ¬ Q), then 'Mild atelectasis is present adjacent to the upper lobe lung mass, and is slightly worse than on the previous examination, **but the mass did not exist previously**' is true. We know that 'slightly worse than on the previous examination' entails atelectasis, including its location reference, which has not changed much since the last examination. Since that atelectasis previously referenced the mass's location, then that mass must have existed previously, contradicting 'the mass did not exist previously'. Thus, we show that U fails the defeasibility test.

In all, $P_{t_0} >>$ U and U $\in Q^{\text{sem}}_{t-1}$ because it is an entailed and non-at-issue presupposition.

3.3 Learning from Presuppositions Projected by Directional Change-of-state Predicates

Unlike entailed Q_t^{sem}, the projected Q_t^{prag} from a proposition is rather contextual. With additional contexts to the proposition (e.g., applying the defeasibility test), the projected prag-ps may be cancelled because Q_t^{prag} is not strictly entailed. Stalnaker (1974) therefore *opposes* the idea that semantic presupposition can account for all types of presuppositions and *propose* to use pragmatic notions to account for (general) presuppositions instead[25]. This is heavily debated among the conventionalists and conversationalists, but later works generally agreed that there are various strength of presupposition projection, and some rely on premises **in additional** to logical premises [24, 26, 27, 63].

In the work of [24], Simons (2013) took an additional step to show that some presuppositions share conversational sources. She speculated that some prag-ps exhibits contextual defeasibility and nondetachability properties, suggesting some prag-ps are contextually dependent and prag-ps triggers (e.g., aspectual phrase) sharing the same lexical meaning consistently project these prag-ps. For instances, Examples (12a) >> (12b), but (12b) can be suppressed with additional context *(12c)* or by ignorance *(12d)* on the interval time. Distinct from sem-ps projected from stable COS predicates, directional COS predicates defeasibly and nondetachably[1] presuppose the precondition of COS before the utterance in time t.

(12) A General English Example: change-of-state predicates defeasibly presupposes the precondition of Change-of-state predicates holds (before the utterance in time t) in general language.

[1] I would like to further clarify that COS psup is undetachable because verbs with the lexical meaning of COS (i.e., left, exit, went out) all project the precondition of change of state.

a. P_t: He left/exit/went out the house.

b. $P_t >> Q^{\text{prag}}_{t-}$

"He was inside the house (before the utterance in time t)."

c. $Q^{\textbf{prag}}_{t-}$ **is suppressed in defeasibility test (Passes "P and it is not Q")**

"He left the house. In fact, he was out of the house a long time ago."

d. $Q^{\textbf{prag}}_{t-}$ **is suppressed by ignorance (Passes "P and it is not Q")**

"He left the house, but I'm unsure about the timing."

These contextual defeasibility and nondetachability features are pragmatic properties akin to those seen in conversational implicatures. For these reasons, some prag-ps are arguably licensed by conversational principles which **aim to maximize the information exchange between speaker and hearer/audience** [24]. The conversational principles are first proposed by Grice in the work of [62], which consist of four main/prevalent maxims[5] as seen in *Example (13)*. More importantly, the theoretical foundation in Simons (2013) gives an opportunity for us to constrain the inference spaces by using the conversational principles and show that some prag-ps are contextually entailed as seen in our work in the next section.

(13) **Grice's Maxims of Conversation [62]**

a. Maxim of Quantity: be as informative as possible without being overly informative

b. Maxim of Quality: speaker does not say what he believes is false or lacks evidence for

c. Maxim of Relation: be as relevant as possible

d. Maxim of Manner: be brief and orderly, and avoid obscurity and ambiguity

[5]To clarify, speaker may also follow other maxims (i.e., politeness and aesthetic) not listed in this section as well.

Certainty in the Existence Information Derived from Presuppositions Projected by the Directional Change of Existence Predicates

I built upon the works from Chierchia and McConnell-Ginet [63] and mostly Simons [24] by assuming that some prag-ps, such as those projected by the change-of-state predicates, share conversational basis. Then I adapted it to fit it on medical contexts. Similar to what Simons [24] have observed in the general domain, the prag-ps projected from directional COS in medical domain also share the nondetachable and defeasible characteristics. As seen in *Example (14)*, the **change of existence** verbs (e.g., resolve, develop, recur)– sharing the lexical meaning of the **change of state**– also project the precondition of change (Examples (14a-14d) >> Examples (14e-14f)). Evidenced in *Example (15)*, explicitly denying and hedging on the relative time interval can directly suppress the presuppositions in Examples (14e - 14f) without being infelicitous. *Example (15)* further illustrates the presupposed contents are not logically entailed by the asserted contexts.

In contrast to the general Example *(12)* in the prior section, I revised the reference time in the projected presupposition. Normally, the prior reference time is defaulted as "before the utterance t". But obviously, change of existence does not always occur while the doctor utters the contexts, so the reference time clearly needs to be revised. I assume the doctors compare the observations between the most recent visits, hence it is reasonable to believe that the change in existence occurs in the interval between *the prior visit* and *the current visit*, but such belief is not logically certain. I propose that the existential contents evidenced in the precondition (at the prior visit) is **contextually entailed** and the end-state (at the current visit) is **logically entailed**. In order to show that the directional change of existence predicate projects the proposed prag-ps, I will first explain why clinicians are indeed rational speakers and then use the conversational principles in the derivations.

(14) Nondetachable Presuppositions in Change of Existence Predicates

a. **Positive Change of Existence**

Cavitation$_1$ *developed/recurred*$_{cos:pos}$ *in*$_{pp}$ the *mass*$_2$

b. **Denial of Positive Change of Existence**

Cavitation$_1$ did not *develop/recur*$_{cos:pos}$ *in*$_{pp}$ the *mass*$_2$

c. **Negative Change of Existence**

Cavitation$_3$ *went away/no longer present/vanished/disappeared/resolved*$_{cos:neg}$ *in*$_{pp}$ the *mass*$_4$

d. **Denial of Negative Change of Existence**

Cavitation$_3$ did not *go away/no longer present/vanish/disappear/resolve*$_{cos:neg}$ *in*$_{pp}$ the *mass*$_4$

e. *Example (14a)* & *Example (14b)* >> Q^{prag}_{t-1}(precondition of positive COS):

The *entity*$_1$ **was not** *preposition* **of** *Entity*$_2$ **in the prior visit.**

f. *Example (14c)* & *Example (14d)* >> Q^{prag}_{t-1}(precondition of negative COS):

The *entity*$_3$ **was** *preposition* **of** *Entity*$_4$ **in the prior visit.**

(15) *Example (14a)* passes the defeasibility test

a. P: "*cavitation*$_i$ developed within the *nodule*$_j$"

b. Q: "that *cavitation*$_i$ was not within the *nodule*$_j$ at the prior visit"

c. **Q Passes the Defeasibility Test (P and it is not the case that Q)**

"Cavitation has developed within the nodule. In fact, this cavitation was seen within the nodule at the prior visit." (seems consistent)

d. **Q passes the Defeasibility Test ("P and it is not the case that Q")**

"Cavitation has developed within the nodule. This scan confirms the suspicious

cavitation seen in the prior visit." (seems consistent)

e. **Q passes the Defeasibility Test ("P and it is not the case that Q")**
"Cavitation has developed within the nodule, but we are unsure when did that happen" (seems consistent)

Radiologists as Rational Speakers

I have assumed our speakers, radiologists, are rational and being cooperative within the conversation, thereby abiding to the general conversational principles. This is a rather strong assumption. One may argue that a speaker does not have to be rational and can violate conversational maxims. One situation could be that a criminal fails to cooperate in a police interrogation. The criminal intentionally violates Grice's maxims by deliberately providing irrelevant answer to the question, being non-informative, or even fabricating lies. Another less extreme case could be that a comedian speaking sarcastically so literal meaning should not be taken for granted. So how could I be sure this assumption hold?

Unlike general speakers who may possibly be uncooperative, doctors do comply to conversational maxims due to their work settings, job responsibilities, and potential legal consequences. The fast pace working environment and job burnout drive radiologists to avoid being overly informative by discussing and including pertinent information needed for current diagnosis[64]. Clinical texts often presented with lists of improper and/or ungrammatical fragments for the sake of brevity in our dataset. Though studies suggested this type of communications/writing styles give rise to language ambiguity in clinical practice and potentially lead to poor decision making, there were inadequate evidence to support these claims[65]. Language ambiguity does not affect those working in the domain of care[65], thus medical professionals do comply to maxim of manners. Moreover, US radiologists and doctors tend to be truthful and ethical since they are responsible for providing evidence-based diagnosis.

Effective communications yield a better compliance and greater patients' satisfactions[64, 66]. Misdiagnosis, on the other hand, can delay optimal treatments, cost lives, and even bring on legal consequences. Thus, medical professionals are indeed driven and motivated to be rational agent who follow maxim of relevance, quantity, manner, and quality when dictating and writing a medical report *under normal circumstances*.

Conceptualize existence from conversational and conversational presuppositions in Change of Existence Predicates

Using established assumptions, I can now use logical reasoning and conversational principles to conceptualize the existence contents from sem-ps and prag-ps. I have identified in the projected presuppositions summarized in (c-d) in *Example (16)* and *Example (18)*, and validated them in *Example (17)* and *Example (19)* respectively.

The general scheme of these proofs in *Example (17)* and *Example (19)* for projections from positive and negative COS predicates are generalizable for cases with the same constructs. First, I showed that existential presuppositions are evidenced in the end-state of directional COS predicates. Second, I demonstrated that directional COS contextually entails the precondition of change of existence in U **maximally**[6], because U survives under presuppositional holes, passes the defeasibility test logically but violates the Grice's Maxims stated in *Example (13)* in the prior subsection. Thus, this allow us to say that U $\in Q_{t-1}^{\text{prag}}$. Provided that P is true and speaker is a rational agent and $P >> Q_t^{\text{prag}}$, then Q_t^{prag} is probably true. Third, the existential sem-ps Q_{t-1}^{sem} is sourced from Q_{t-1}^{prag}, so the certainty of that Q_{t-1}^{sem} is capped by certainty of its source. In conclusion, the certainty in the contents denoted in Q_{t-1}^{sem} under Q_t^{prag} is **probably true** and certainty in the contents denoted in $Q_{t_0}^{\text{sem}}$ is **true**.

[6]**Note that the utterance U not only presupposes the prior state of existence for subject of COS predicates, but it maximizes the surrounding context of precondition projected by the directional COS predicate. The prior state of existence for the subject of COS is semantically entailed by the usage of aspectual predicates.**

In lay language, a tumor **probably existed in the prior visit** and **definitely exists in current visit** when it is a referent of another finding (i.e., edema, cavitation) experiencing a directional change in the state of existence (COS).

(16) **Positive COS >> *Example (16c) & Example (16d)***

a. P_{t_0}: A *cavitation*$_i$ has developed within the *nodule*$_j$.

b. $P_{t_0} >> Q^{prag}_{t-1}$: "that *cavitation*$_i$ was not within the *nodule*$_j$ at the prior visit"

c. $P_{t_0} >> Q^{sem}_{t_0}$: The *nodule*$_j$ exists currently.

d. $Q^{prag}_{t-1} >> Q^{sem}_{t-1}$: The *nodule*$_j$ existed at the prior visit

(17) Deriving conventional and conversational presupposition for *Example (16)*

a. Given P_{t_0}: A cavitation has developed within the nodule.

b. **Show $Q^{\mathbf{sem}}_{t_0}$ (The nodule exists).** P_{t_0} contains an existential presupposition, the nodule exists (at t_0).

c. **Established premises**: Doctor is following the Maxim of Quantity, Relevance and Quality

d. **Show $P_{t_0} >> Q^{\mathbf{prag}}_{t-1}$ (U = "*cavitation*$_i$ was not within the nodule at the prior visit")**

Show U is non-at-issue presupposition by passing the presupposition holes ($O(P_{t_0})$ where $O \in \{\neg, \diamond, ?, IF\}$). $O(P_{t_0})$ does not affect the consistency of U (in parenthesis) appended in the background, suggesting U is non-at-issue content.

(i) "(That *cavitation*$_i$ was not within the nodule at the prior visit.) Cavitation did not developed within the nodule. "

(ii) "(That *cavitation*$_i$ was not within the nodule at the prior visit.) Perhaps,

cavitation has developed within the nodule. ”

(iii) “(That $cavitation_i$ was not within the nodule at the prior visit.) If the cavitation has developed within the nodule, please take a note of it. ”

(iv) “(That $cavitation_i$ was not within the nodule at the prior visit.) Did the cavitation develop within the nodule? ”

Show U is contextually dependent by passing defeasibility test (“P and ¬ U” is consistent)

(i) “Cavitation has developed within the nodule. In fact, this cavitation was seen within the nodule at the prior visit.” (seems consistent)

(ii) “Cavitation has developed within the nodule. This scan confirms the suspicious cavitation within the nodule at the prior visit.” (seems consistent)

Using proof by contradiction to show U is contextually entailed. Let's suppose U is false, then there could be 2 causes, (1)“the doctor does not have enough of evidence to assert whether cavitation or nodule was there at the prior visit”, or (2) “that $cavitation_i$ was within the nodule at the prior visit” .

Suppose case (1) is true, “the doctor does not have sufficient evidence to assert whether the cavitation or nodule was there at the prior visit”. Since we know that the doctor is making a quality statement, we assume the claim that “the cavitation has developed” is supported by evidence. Then, the doctor must have had access to the patient's history in order to make a comparison. Therefore, the doctor must have known the existence and location of cavitation prior and during utterance. This contradicts with case (1). Since we know that doctor is following the maxim of manner (making correct statements), then the proposition: “doctor

does not know about the state and location of nodule and cavitation in the prior visit" must be false.

Let's assume case (2) is true, "that $cavitation_i$ was within the nodule at the prior visit". Then, the doctor should not have use 'develop' to convey there is a change because the doctor should have been more relevant and informative when possible (i.e., "cavitation is again seen within the nodule"). Therefore, the doctor believes the change-of-existence is recent, hence that $cavitation_i$ was **not** within the nodule at the prior visit. This contradicts with (2). Hence case (2) must be false.

Since cases (1) and (2) are false, then the doctor believes and contextually implies U, hence U is Q^{prag}_{t-1}.

e. **Show $Q^{\textbf{prag}}_{t-1} >> Q^{\textbf{sem}}_{t-1}$ (The nodule existed at the prior visit)**. From Q^{prag}_{t-1} "That $cavitation_i$ was not within the $nodule_j$ at the prior visit", we have evidenced an existential presupposition, thus the speaker also believes **the $nodule_j$ existed at the prior visit**.

(18) **Negative COS >> *Example (18c)* & *Example (18d)***

a. P_{t_0}*:* The $edema_i$ surrounding the $mass_j$ has disappeared (at t_0).

b. $P_{t_0} >> Q^{prag}_{t_{-1}}$: $Edema_i$ was surrounding the $mass_j$ at the visit prior to t_0.

c. $P_{t_0} >> Q^{sem}_{t_0}$: The $mass_j$ exists (at t_0).

d. $Q^{prag}_{t_{-1}} >> Q^{sem}_{t_{-1}}$: The $mass_j$ existed at the visit prior to t_0.

(19) Deriving conventional and conversational presupposition for *Example (18)*

a. Given P_{t_0}: The edema surrounding the mass has disappeared.

b. **Show $Q^{\mathbf{sem}}_{t_0}$ (The mass exists)**: P_{t_0} contains an existential presupposition, thus the mass exists (at t_0).

c. **Established premises**: Doctor is following the Maxim of Quantity and Maxim of Manner. Therefore, the statement was made 'as informative as possible without breaking the other maxims' and 'making correct statements'.

d. **Show $P_{t_0} >> Q^{\mathbf{prag}}_{t_{-1}}$ (U = 'that $edema_i$ was surrounding the mass at the prior visit.")**

Show U is non-at-issue presupposition by passing the presupposition holes($O(P_{t_0})$ where $O \in \{\neg, \diamond, ?, IF\}$). $O(P_{t_0})$ does not affect the consistency of U (in parenthesis) appended in the background, suggesting U is non-at-issue content.

(i) "($Edema_i$ was surrounding the mass at the prior visit.) Edema surrounding the mass **did not** disappear. "

(ii) "($Edema_i$ was surrounding the mass at the prior visit.) Edema surrounding the mass **seems to have** disappeared."

(iii) "($Edema_i$ was surrounding the mass at the prior visit.) **If** the edema surrounding the mass disappeared, that's a good sign."

(iv) "($Edema_i$ was surrounding the mass at the prior visit.) **Did** the edema surrounding the mass disappear?

Show U is contextually dependent through passing defeasibility test.
(i) "$Edema_i$ surrounding the $mass_j$ disappeared, but we don't know when it disappeared." (seems consistent)
(ii) "$Edema_i$ surrounding the $mass_j$ disappeared. In fact, edema and the mass disappeared a long time ago." (seems consistent)

Using proof by contradiction to show U is contextually entailed. Let's suppose U is not implied, "Edema was surrounding the mass at the prior visit" is false. Then there could be two causes, (1)"doctor does not have enough of evidence to assert whether the edema or mass was there at the prior visit", (2)"the $edema_i$ was not surrounding the $mass_j$ at the prior visit".

Suppose case (1) is true, "doctor does not have sufficient evidence to assert whether the edema or mass was there at the prior visit". Since we know that the doctor is making quality statements, we assume the claim that edema disappeared is supported by evidence. Then, the doctor must have had access to the patient's medical history before concluding there is a change in existence. Therefore, the doctor must have known the prior and current state of edema, as well as their locations. This contradicts with case (1). Since we know that doctor is following the maxim of manner (making correct statements), then "doctor does not know about the mass's or edema's state and location in the prior visit" must be false.
Suppose case (2) "edema was not surrounding the mass at the prior visit" is true. Then the doctor should have been more informative and said "no new edema", or

emphasize it disappeared at another time interval, but the doctor did not. Since we know that doctor is being relevant and informative, the doctor believes that disappearance of edema is recent, contradicting with case (2). Thus, case (2) must be false.

Since cases (1) and (2) are false, the doctor believes that "edema was surrounding the mass at the prior visit" is contextually entailed in the background content, hence U is Q_{t-1}^{prag}.

e. **Show** $Q_{t-1}^{\mathbf{prag}} >> Q_{t-1}^{\mathbf{sem}}$(The mass existed at the prior state).

From Q_{t-1}^{prag} "$Edema_i$ was surrounding the $mass_j$ at the prior visit.", we have evidenced an existential presupposition, thus the speaker also believes the mass existed at the prior visit.

Chapter 4

Materials and Methods

4.1 Materials: Mass Dataset

This work used a corpus of deidentified and unannotated mass sentences (N = 5901) from deidentified UCLA Radiology Reports collected by the UCLA Medical Imaging Informatics group. These sentences mentioned keywords on either 'lesion', 'nodules', or 'mass'.

Two independent biomedical informatics researchers provided sentence-level annotations on the mass sentences following the annotation guideline written by one of the annotator under the supervision of a senior researcher experienced in reading radiology reports. The guideline described the mapping rules transforming the unstructured text into the defined structured knowledge representations for the conceptual frames: *Prior Existence*, *Current Existence*, *Collection-level Change of Existence*, *Instance-level Change of Existence*, and *Significance of Change Tasks*.

After the annotators labeled the same corpus separately, they met together to resolve the inconsistencies in the two sub-corpus. For cases where both parties cannot come to a consensus, one of the annotator received advices from the senior medical researcher to break the tie. The guideline was revised several times to further elaborate on the problematic

Table 4.1: Corpus Statistics: text characteristics and dataset size

	Mass Dataset
Max # of **Words** in Text Sequence	61
Max # of **Tokens** in Text Sequence	115
# Vocabulary Size	3172
# Distinct WordPiece Tokens	3792
# Train	3540
# Dev	1180
# Test	1181
# Total	5901

cases encountered. After the guideline was finalized (Note: **Annotation guideline for this mass corpus** is available upon request), one of the annotator checked the consistency of the finalized corpus with the guideline, yielding gold standard labels on the five conceptual frames.

Characteristics of Mass Dataset

The annotated corpus contains 5901 mass sentences, but there are class imbalances in the five conceptual tasks. To ensure there are sufficient samples for training and evaluation, the mass dataset was split into 60% training set, 20% validation set, and 20% held-out test set using random sampling[1]. *Table 4.1* and *Table 4.2* summarized the corpus statistics on the clinical text characteristics and the number of class instances within the tasks respectively.

4.2 Methodology

This work used a combination of traditional NLP techniques and deep learning models to extract mass sentences from radiology reports and automatically output the discourse analysis

[1]Due to the limitation of computational power and time, future work will validate the model using a k-fold cross validation instead.

Table 4.2: Corpus Statistics: number of class instances in the five conceptual tasks

Tasks	Classes	total (counts)	train (counts)	val (counts)	test (counts)
1. Prior Existence	definitely existed	2288	1354	474	460
	definitely did not exists	64	33	18	13
	definitely bulk removed	215	141	33	41
	definitely inadequate imaging evidence	4	2	2	0
	likely existed	51	22	13	16
	likely did not exists	3	1	2	0
	uncertain	300	186	60	54
	not mentioned	2976	1801	578	597
2. Current Existence	definitely exists	3661	1801	922	938
	definitely did not exists	1592	1354	121	117
	definitely bulk removed	32	22	3	7
	definitely inadequate imaging evidence	8	1	4	3
	likely exists	57	33	16	8
	likely not exists	14	2	6	6
	uncertain	366	186	93	87
	not mentioned	171	141	15	15
3. Collection-level Change of Existence	definitely stable	2236	1320	476	440
	definitely negative COS	38	24	4	10
	definitely positive COS	251	158	43	50
	likely stable	53	24	12	17
	likely negative COS	13	6	3	4
	likely positive COS	23	14	7	2
	uncertain	3287	1994	635	658
4. Instance-level Change of existence	definitely stable	2307	1360	490	457
	definitely negative COS/fewer mass(es)	27	17	2	8
	definitely positive COS/more mass(es)	29	19	8	2
	likely stable	55	25	13	17
	likely negative COS/fewer mass(es)	15	7	4	4
	likely positive COS/more mass(es)	9	5	4	0
	uncertain	3459	2107	659	693
5. Significance of Change	has sign/have signs of disease progression	803	484	158	161
	no signs of disease progression	5098	3056	1022	1020

characterizing the five conceptual frames: *Prior Existence*, *Current Existence*, *collection-level COS*, *instance-level COS*, and *significance of change*. *Figure 4.1* demonstrates the expected outputs given a sample input to this pipeline. In practice, the automated semantic analysis was reformulated as five independent multi-class text classification tasks, where each takes in unstructured text sequence as input and outputs a class for that conceptual frame. Using the annotated mass dataset, the classifiers were trained and evaluated separately using the gold-labels for each of the conceptual frames. *Figure 4.2* shows an overview of our approach and the details are shown in the following subsections.

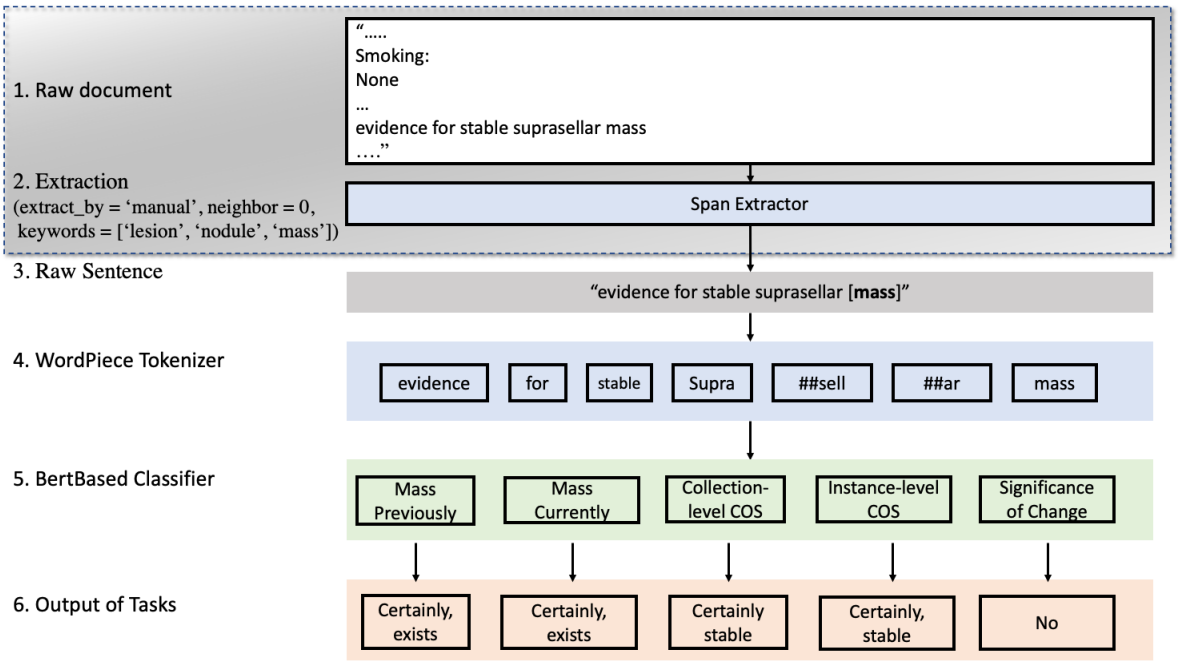

Figure 4.1: Sample output from the pipeline

Span Extractor

Since text classifiers are trained on sentence annotations, our approach provides the option to perform document classification by processing full radiology reports to output relevant sentence(s) before feeding into the information models. This span extractor is comprised of three substrates: identify relevant concepts, isolate hotspot sentence (sentence containing relevant concept) and the neighbor(s) of hotspot sentences.

The user may choose relevant concepts manually or semi-automatically. The manual method requires a provided list of regular expressions. The latter method approximates concept relevance using correlation of mentioned clinical concept to the mentioned tasks[2].

[2]Here I would like to clarify this mentioned task is a binary classification problem, which is a simplification of the original existence problem. If the sentence mentioned existences at either time points, then the sentence has a mentioned mass. This mentioned task will map 'Not-mentioned' label as 0 while the rest of the existence labels as 1 (i.e., certainly exists, certainly not exists)

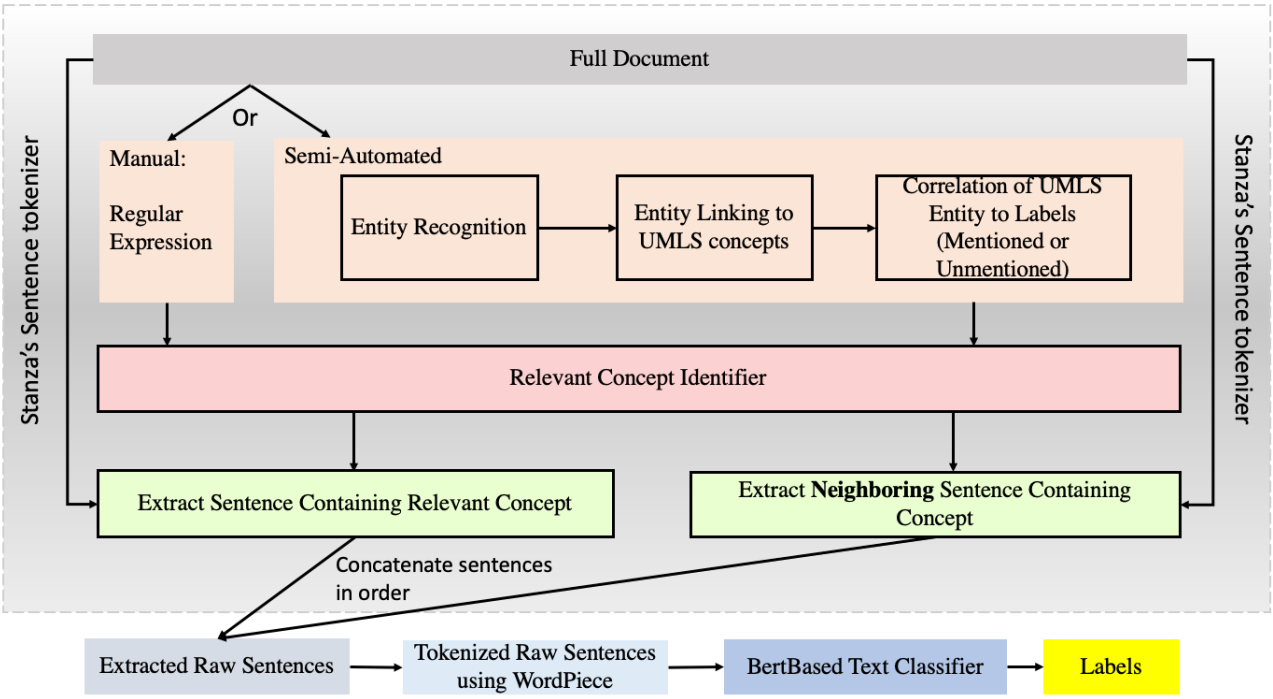

Figure 4.2: Overview of our methodology for characterizing *existence*, *change of existence*, and *significance of change for unstructured radiology reports*

Relevant medical mentions are identified using clinical entity recognition, entity linking to Concept Unique Identifier (CUI) in the Unified Medical Language System (UMLS). UMLS knowledge sources organize biomedical concepts by its meaning and cluster similar terminologies under concept unique identifiers [67]. Concept normalization allows the system to understand a variety of phrases expressing the same clinical concepts and improving the correlation of concepts to the tasks. Using Chi-square tests, these significant top-k UMLS concepts are considered as important concepts. Parameters, such as k, significance level (α = 0.05, 0.025, 0.001), and number of neighboring sentences (typically 0 or 1), are trained, optimized, and tested using train, validation, and test sets respectively.

Since the status of mentioned mass may be potentially modified by its surrounding contexts, sentences containing important concepts are isolated because we assumed topic localiz-

ation under the same sentence and neighboring sentences may contain co-references. Hence, a full-length report is distilled into hotspot sentences and neighboring sentences using the identified relevant concepts and Stanza's Sentence Tokenizer from the work in Zhang et al. [68]. These extracted raw sentence(s) would serve as the inputs to the text classifiers if the user feeds in a full-length unstructured radiology report.

BERT-Based Text Classifiers

Given the recent success of pre-trained BERT model on text classification requiring wide range of systemic inferences, this work used the pre-trained Bio-ClinicalBERT model as our language model (LM) and adapt it for text classification. The clinical text characteristics in our dataset closely resemble those from MIMIC III texts, so Bio-ClinicalBERT from the work of Alsentzer et al. [69] (which was pretrained on MIMIC III notes) was selected to be the baseline language model to output a deep learning representation for the inputted text as our preliminary work. Text sequence was tokenized using WordPiece and padded to max length of 512. To obtain a deep learning (sequence) representation from the text sequence, the text sequence was passed to the Bio-ClinicalBERT LM. Then the sequence representation was obtained from a special token [CLS], which is the first token in the final hidden layer of the BERT architecture.

To adapt it for text classification, the pooling layer (dimension of pooling layer = 768) was concatenated to a fully-connected layer **h** (dimension of h = 768), then a dropout layer for regularization and a ReLu activation layer. Lastly, the fully-connected layer was fed into a soft-max layer to output the probability of label c: $p(c|h) = Softmax(Wh)$ and $c \in$ Task-specific Classes. *Figure 4.3* shows the classifier's architecture.

The semantic analysis on the conceptual frames was treated as separate multi-class problems, thus there were five separate text classifiers corresponding to the frames. For each

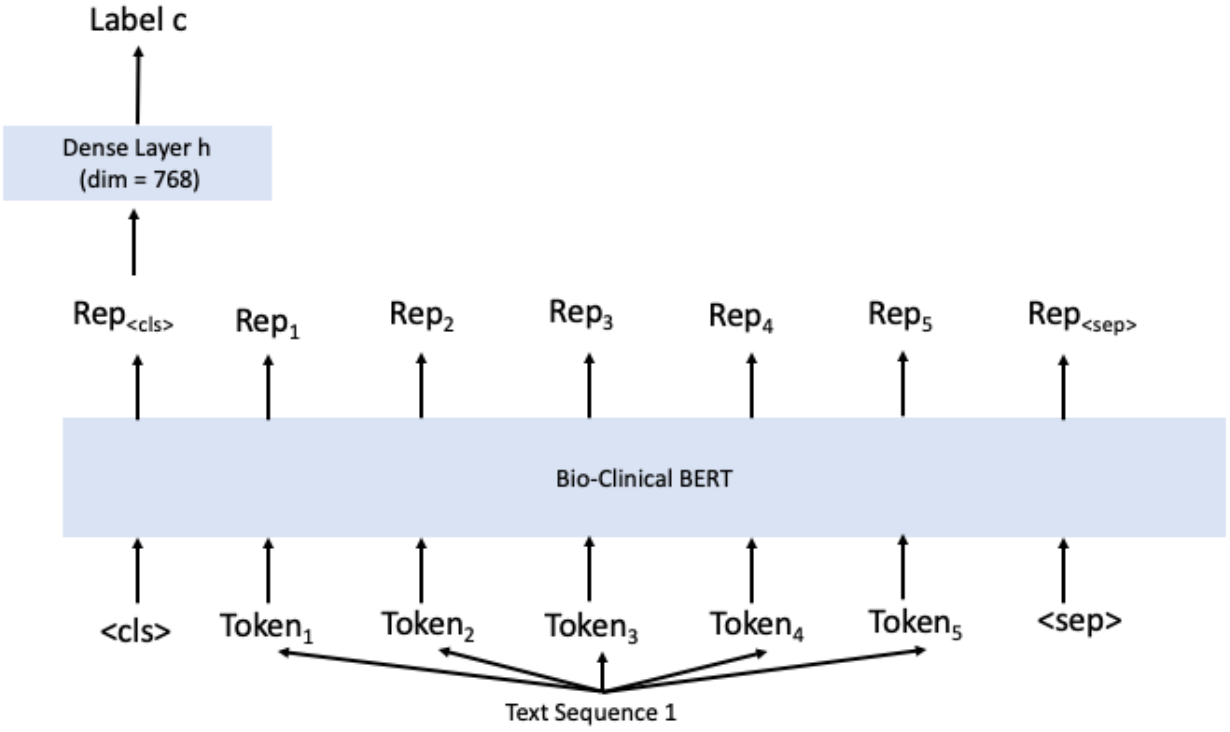

Figure 4.3: BERTbased Text Classification for 1 sub-task within the dataset

classifier, the inputs (mass sentence) and outputs (label for that mass sentence) were feed to the classifier and fine-tuned over all the parameters in BERT, **h**, and **W** jointly by minimizing over the weighted multi-class cross-entropy loss.

Model Training

The entire model was coded in Python and the classifiers were implemented using PyTorch Library. I used the above architecture in *Figure 4.3* for each of the task, and the training process on a single NVIDIA TESLA P100 (12GB) GPU. In order to control the effect of randomness, fixed initial seed of 42 and epoch seed were initialized for reproducibility. The classifiers in the model were fine-tuned over the dataset for 50 epochs using a minibatch

Adam optimizer separately. Hyperparameters such as training batch size (among 10, 16, 32), learning rates (among 1e-4, 1e-5, 5e-5, 5e-6, 1e-6), and B decay (among 0, 0.1, 0.2) were optimized using grid search. For each task, the best classifier was selected based on average of micro-F1 and macro-F1 scores in the Dev set.

Chapter 5

Results and Discussion

5.1 Results

Based on the model's performance on the development set, the best model was selected and evaluated on a held-out test set (N = 1181) using precision, recall, and F_1 scores. The overall performance were evaluated using micro- and macro-averaging F_1 scores. *Table 5.1* in the following page summarizes the main results of our experiment for each of the classes and the overall performances within the conceptual frames.

The results for micro-averaging and macro-averaging F_1 are (0.91, 0.71) for *Prior Existence*, (0.93, 0.55) for *Current Existence*, (0.91, 0.50) for *collection-level COS*, (0.87, 0.52) for *instance-level COS*, and (0.94, 0.88) for *significance of change* tasks respectively. Comparing the overall performances across the tasks, the results indicate that the classifier ranked best in the classification of *significance of change*, then followed the *Prior Existence* task.

Consistent across all the tasks, the micro-F_1 scores are significantly higher than the macro-F_1 scores, especially for *Current Existence*, and *Change of Existence* tasks. Taking a closer look within each task, the model demonstrated a higher recall, precision, and F_1 on majority classes than on minority classes ($N < 20$) in this test set. Specifically, the

Table 5.1: The performance of the BERT-based text classifiers on the test set for five defined tasks in terms of precision, recall, and F_1-score. Macro-average and Micro-average are the average F_1 performance over all the classes within their respective task.

(Certainty, Classes)	N	Precision	Recall	F_1-score
Task 1: Existence Past				
Definitely Exists	460	0.91	0.91	0.91
Definitely Absent	13	0.38	0.23	0.29
Definitely Bulked Removed	41	0.76	0.78	0.77
Definitely Inadequate Technique	0	-	-	-
Likely Exists	16	0.80	0.50	0.62
Likely Absent	0	-	-	-
Uncertain	54	0.75	0.74	0.75
Not Mentioned	597	0.94	0.95	0.95
Macro-Avg	1181	0.76	0.69	0.71
Micro-Avg	1181	0.90	0.91	0.91
Task 2: Existence Now				
Definitely Exists	938	0.96	0.98	0.97
Definitely Absent	117	0.91	0.89	0.90
Definitely Bulked Removed	7	1.00	0.86	0.92
Definitely Inadequate Technique	3	0.00	0.00	0.00
Likely Exists	8	0.17	0.25	0.20
Likely Absent	6	0.25	0.17	0.20
Uncertain	87	0.81	0.77	0.79
NotMentioned	15	0.71	0.33	0.45
Macro-Avg	1181	0.60	0.53	0.55
Micro-Avg	1181	0.93	0.93	0.93
Task 3: Change of Existence (Patient-level)				
Definitely +Change	2	0.00	0.00	0.00
Definitely -Change	8	0.33	0.25	0.29
Definitely Stable	457	0.90	0.91	0.90
Likely +Change	0	–	–	–
Likely -Change	4	1.00	0.25	0.40
Likely Stable	17	0.50	0.41	0.45
Uncertain	693	0.94	0.94	0.94
Macro-Avg	1181	0.61	0.46	0.50
Micro-Avg	1181	0.91	0.91	0.91
Task 4: Change of Existence (Instance-level)				
Definitely +Change	50	0.65	0.74	0.69
Definitely -Change	10	0.43	0.30	0.35
Definitely Stable	440	0.81	0.90	0.85
Likely +Change	2	0.00	0.00	0.00
Likely -Change	4	1.00	0.25	0.40
Likely Stable	17	0.41	0.41	0.41
Uncertain	658	0.96	0.88	0.92
Macro-Avg	1181	0.61	0.50	0.52
Micro-Avg	1181	0.88	0.87	0.87
Task 5: Significance of Change				
Progression	161	0.84	0.74	0.79
No Progression	1020	0.96	0.98	0.97
Macro-Avg	1181	0.90	0.86	0.88
Micro-Avg	1181	0.94	0.94	0.94

classifier was better at determining whether the tumor previously existed ($F_1 = 0.91$) or has not been mentioned ($F_1 = 0.95$) for Task 1, and whether it currently exists ($F_1 = 0.97$), absent ($F_1 = 0.90$), or bulked removed ($F_1 = 0.92$) for Task 2. For the *change of existence* tasks, the model was better at classifying the state of 'uncertain' (0.94 F_1 for patient-level COS, 0.92 F_1 for instance-level COS) and 'definitely stable' (0.90 F_1 for patient-level, 0.85 F_1 for instance-level) states compared to directional COS.

The differences in F_1 between majority classes and minority classes also explain why micro-F_1 scores are significantly higher than the macro-F_1 scores for the conceptual frames. Since micro-averaging F_1 is calculated differently from macro-averaging F_1, micro-F_1 use global precision and recalls (giving higher importance to minority class), while macro-F_1 take the average of F_1 for each class (giving equal importance to each classes). Micro-F_1 biased toward the F_1 contributed by majority class while macro-F_1 got weighted down by the poor performance in minority classes, thereby resulting in high micro-F_1 and low macro-F_1.

Figure 5.1 shows the confusion matrices for the five tasks. The matrices show that the false negatives on minority classes are mostly mistaken as majority classes within these conceptual tasks. The results suggest the errors are caused by issues of lack of sufficient data (few classes have no examples, e.g., ⟨likely, absent⟩ in **Prior Existence** Task) and class imbalances in the conceptual frames, thus the model is biased toward the predictions on majority classes.

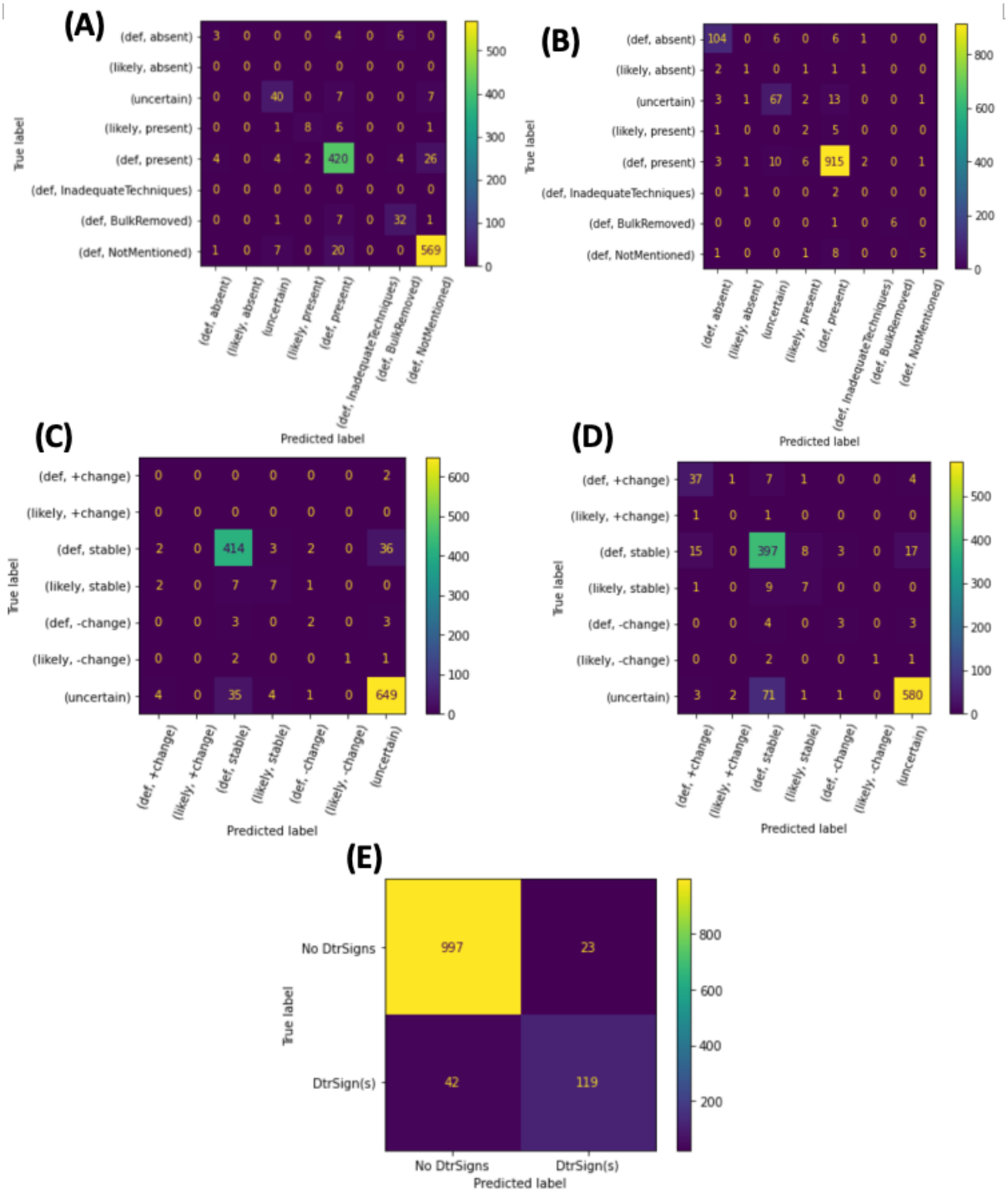

Figure 5.1: **Confusion Matrices for the Task 1 (A), Task 2 (B), Task 3 (C), Task 4 (D), and Task 5 (E).** Predicted and true labels correspond to the x-axis and y-axis. The values on the diagonals of the matrices represent true positive cases.

5.2 Discussion

This section analyzes the errors made by our classifier to reveal directions for future improvements. Recalling from before, aggregated statistics in *Table 5.1* show that our classifier performs poorly in minority classes across the five tasks. This may be explained by the class imbalances and lack of training data, resulting in a biased model that did not learn very well on minority classes. However, these claims have to be further investigated from the error patterns.

With that in mind, I conducted a comprehensive error analysis for these five tasks. An error can either be a false positive or a false negative. For each of the subtasks, misclassified examples were grouped under few common error typology: temporality mismatch, classes mismatch, certainty mismatch. I further investigated and listed out the potential causes of such errors. In summary, error analysis shows that the model has quality reductions on certain types of cases requiring natural language understanding.

Error Analysis for Task 1-2: Existence

To explain *why* existence tasks is challenging for our model, I manually analyzed the errors made by the classifier on the test set for task 1 and task 2. Although it is plausible that the model biased toward the prediction in majority class, exploratory study shows that challenging test instances (i.e., those described in **Chapter 2** and **3**) resulted in quality reduction, especially cases providing tumor existence states in two temporal dimensions. To assess what types of cases the model finds challenging, I simultaneously reviewed the model's predictions on the *Prior Existence* and *Current Existence* for each test instance. Based on the ground truth in existence pairs (denoted as ⟨Prior Existence State, Current Existence State⟩), the test set is stratified into three separate groups: **non-overlap group** (tumor status provided exclusively in past or present), **same-overlap groups** (same tumor

status covered in both temporal descriptions), and **dissimilar-overlap group** (dissimilar existence states within the existence pairs). Examples of existence pairs categorized by temporal groups are shown in *Example (20)*.

(20) Examples of stratifying existence pairs based on their temporal groups

a. Non-overlap group (NO): "There exists a mass in the current visit"

b. Same-Overlap group (SO): "The mass is again seen in the current visit"

c. Dissimilar-overlap group (DO): "The previously suspicious mass is confirmed on the current visit."

As a preliminary study, *Figure 5.2* summarizes the percentages of incorrect predictions within the stratified categories discussed earlier. The inner pie in this figure shows the test set consists of 50.4% non-overlap (**NO**), 38.0% same-overlap (**SO**), and 11.6% dissimilar-overlap (**DO**) groups. 49.6% of test instances have existence states covered in both temporality, suggesting the tumor status is being closely monitored and the descriptions frequently referenced the prior state. SO group in this test set is made up of 95.8% ⟨definitely exists, definitely exists⟩, 2.5% ⟨definitely not exists, definitely not exists⟩, 1.8% ⟨NotMentioned, NotMentioned⟩. NO group is made up of 76.1% ⟨not mentioned, definitely exists⟩, 16.3% of ⟨not mentioned, definitely not exists⟩, 5.4% of ⟨not mentioned, uncertain⟩, and the rest on other distinct pairs. DO group is the most diverse of all and hardest to annotate manually. This group is comprised of 24.8% ⟨uncertain, definitely exists⟩, 24.1% ⟨definitely bulked removed, uncertain⟩, 12.4% ⟨uncertain, uncertain⟩[1], and 10.9% ⟨likely exists, definitely exists⟩.

As seen from the outer circle of *Figure 5.2*, the classifier has a total accuracy of 86% in existence pairs measured by the exact match of existence status (both certainty and class of existence) in both temporal axes. The classifier is predominately tested on cases in NO and

[1] ⟨uncertain, uncertain⟩ existence pairs belong to the DO group because uncertain class in the past and present may have dissimilar status.

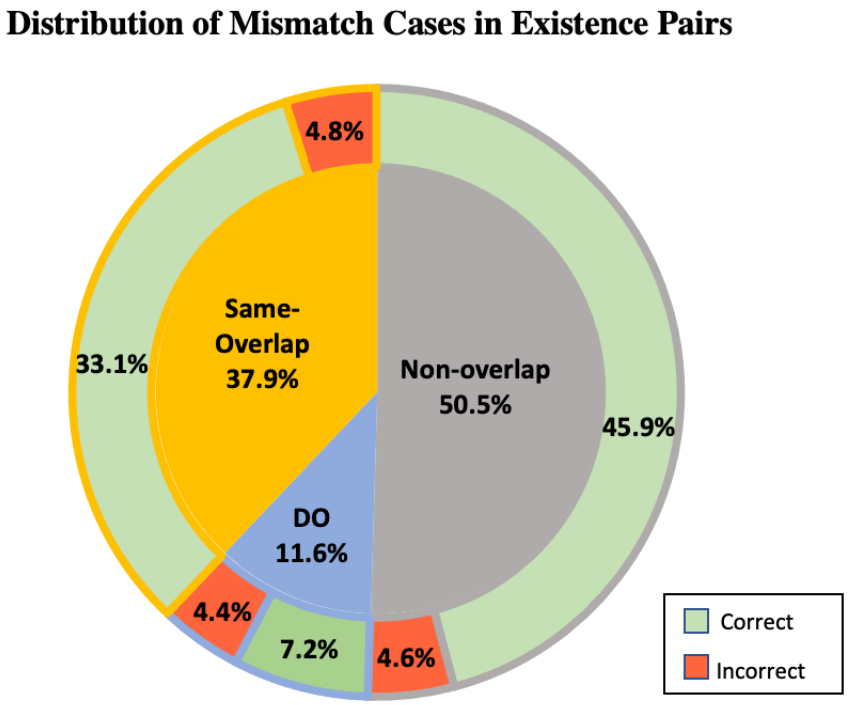

Figure 5.2: **Distribution of Mismatch Instances in Test Set Based on the Types of Existence Pairs.** The inner circle shows the proportion of instances in test set (N = 1181) within these existence pairs stratified by temporal overlaps: non-overlap (grey), same-overlap (yellow), dissimilar-overlap (blue). The outer circle shows the percentage of correct instances (green) and incorrect (red) instances from each category respectively (outlined color based on inner circle's category). The percentage of correctly and incorrectly predicted existence pairs from each category sum up to roughly 86% and 14% respectively.

SO groups; but NO, SO, and DO groups each contributed 4.6%, 4.8%, 4.4% to the total of 13.8% error rate, respectively. Considering the ratio of error cases to the number of instances within each group, the figure shows that the classifier is best at classifying instances from the NO and SO groups, and significantly worst in DO group, substantiating certain types of cases resulted in quality reduction.

To gain additional insights, incorrect existence pairs were then categorized under the most common types of errors: temporality mismatch (predicted the correct certainty and class but failed in temporal distinctions), certainty mismatch (predicted the correct class

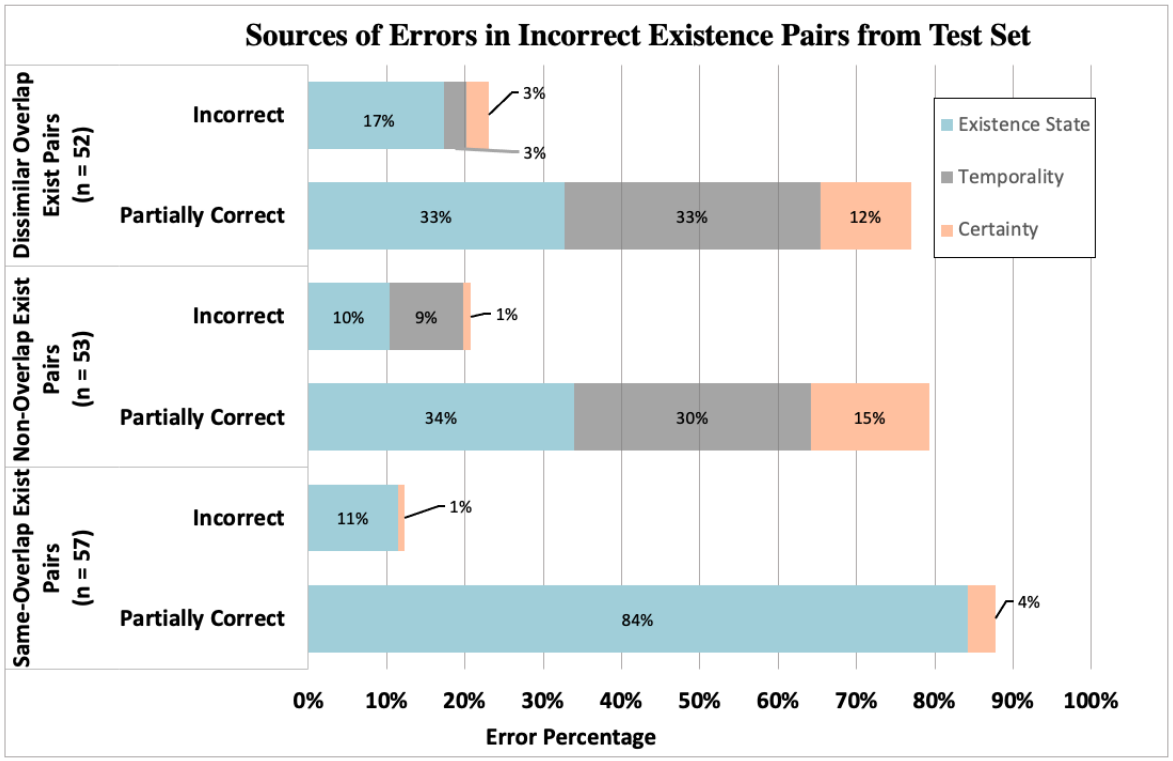

Figure 5.3: **Sources of Errors in Incorrect Existence Pairs (N = 162) from Test Set.** This graph shows the trend of mismatch errors within the three types of existence pairs. The height of the column shows the percentage of existence pairs (w.r.t. its cluster) having the entire existence pair incorrect (incorrect) or partially correct on either temporality (partially correct). The values inside the columns represent the percentage of errors caused by categorical (blue), temporal (grey), and certainty (orange) mismatches.

and temporal distinction but failed in certainty distinction), class mismatch (predicted the wrong category entirely). The types of errors in the incorrect existence pairs are summarized in *Figure 5.3*. The graph shows that roughly 88%, 79%, and 77% of incorrect predictions in SO, NO, DO groups are partially correct on a state within the existence pair respectively. And, the trends in each group consistently show that the mismatches are mainly caused by incorrect class prediction (i.e., bulked removed, inadequate evidence, uncertain), closely followed the temporal distinction, and lastly certainty distinctions.

Starting off with the qualitative error analysis on the least problematic mismatch category, I have observed few trends associated with the incorrect predictions on the expression of certainty for the existence tasks. The model does well in detecting simple hedge or negation triggers but has mistaken double-negative hedge triggers(e.g., 'cannot be excluded') as negated triggers. Misinterpreting the semantics of double-negative hedge cues is a minor problem, the more challenging problem is identifying the scope[2] of hedge cues in sentences with coordinating conjunctions (i.e., "Differential considerations might include primary lung neoplasms with metastastic nodes and nodules, metastatic disease from unknown primary."). As a result, incorrect scope contributed to certainty mismatch by underestimating or overestimating the scope of the cue. Another problem is dealing with challenging instances where the tumor surrogates lie within the hedging scope, but the context suggests clinicians are hedging on the malignancies, diagnosis, or effect of tumor rather than tumor existence. Consider the following example: "This patient's altered mental status is *likely* **due to a combination of the patient's intracranial lesion and recent chemotherapy**". The italicized hedge trigger, *likely*, is modifying the bolded prepositional phrase syntactically, but the context suggests the observer is hedging on the cause of altered mental status and not on the presence of lesion. In such condition, the model had incorrectly predicted "a mass likely exists" because it did not considered contextual information.

Temporal distinction is the second major challenge in addition to the certainty distinction discussed above. Trend suggests the model may have occasionally ignored low frequency temporal modifiers (e.g., "in the comparison study", "at that time"), aspectual relations (e.g., "**again** seen"), and adjectives and adverbs containing additional temporal information (e.g., "**same** mass seen", "**still** seen").

Encouragingly, there were very few cases without explanations for incorrect predictions

[2]The scope is defined as a sequence of one or more words that are affected by the negated or hedged triggers.

in the existence state. The errors in the existence category can be explained by misunderstanding the aspectuals relations (e.g., "no longer identified, disappeared"), the effects of resections on tumor, and/or distinct types of lack of evidence. The classifier predominately failed to realize the effects of types of resections on the state of tumor existence: complete resection (resection without residual margin → absence of tumor), partial resection (resection with residual margins → tumor exists), and resection (resection without additional information on tumor margin → bulked removed). In particular, 'partial resection' and 'complete resection' cases were predicted as bulk removal, suggesting the model was unable to depict status of tumor margin from contexts.

Another challenge is understanding the notion on the absence of evidence of a tumor due to imaging difficulties[3]. Specifically, there are three types of absence of evidence: absence of evidence suggesting no tumor (e.g., no evidence of mass → tumor does not exists), absence of a complete tumor observation but still sufficient to confirm the existence (e.g., the mass is *partially* obscured → tumor exists), and insufficient evidence detected to confirm the existence states at all (e.g., the mass is *completely* obscured → inadequate evidence). Building upon the concept of absence of evidence, there is an unusual case that is worthy of discussion. Context can also contain conflicting tumor status where one lower-quality scan reported the absence of evidence and a higher-quality reported evidence of tumor existence. Lacking the commonsense capability, the classifier was unable to determine which piece of evidence is more diagnostic than the other and incorrectly outputted tumor absence instead. Since clinicians do report and use collected evidence to further confirm and diagnose tumor existence, future improvement should emphasize on cases where the lack of evidence does not necessarily suggest absence. Lastly, improving the understanding of causal relations, temporal understanding, common-sense on the quality of evidence can address these difficult

[3]Note: the absence of evidence is not the same as the evidence of absence. Although the first concept *can* also suggest the evidence concluded the absence of existence of a finding, but it can also suggests non-sufficient evidence to deduce anything at all.

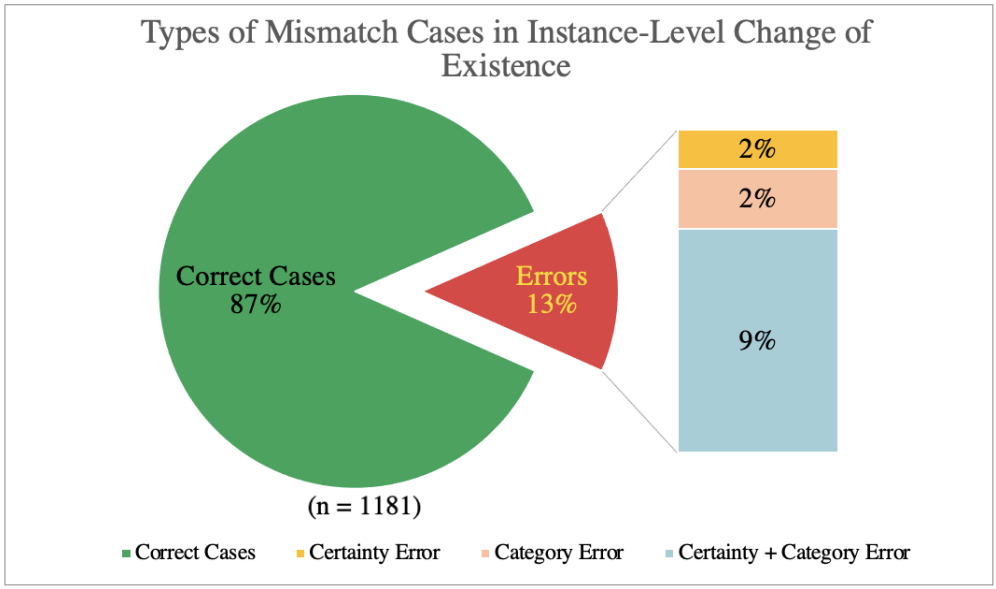

Figure 5.4: **Types of Errors in Test Set for the *Instance-Level COS* Task (N = 1181).** The left circle shows 87% are correctly predicted (green) and 13% are incorrectly predicted (red) out of an entire test set (N = 1181). The column in the right shows percentage of mismatched cases due to certainty error (yellow), category error (pink), or both (blue).

cases in the future.

Error Analysis for Task 3-4: Change of Existence

As a part of the quantitative error analysis for instance-level and collection-level COS, I examined the types of errors in the incorrect predictions tabulated in *Figure 5.4* and *Figure 5.5*. These graphs show 9% were incorrectly predicted for collection-level COS and 13% for the instance-level COS. Consistent across both tasks, errors were due to mismatch in certainty and change category, but majority in a combination of both.

The error patterns in change of existence viewed on patient-level and instance-level share many similarities because test instances mostly described a single mass rather than a collection of masses. However, the error rate was higher on change of existence viewed on an

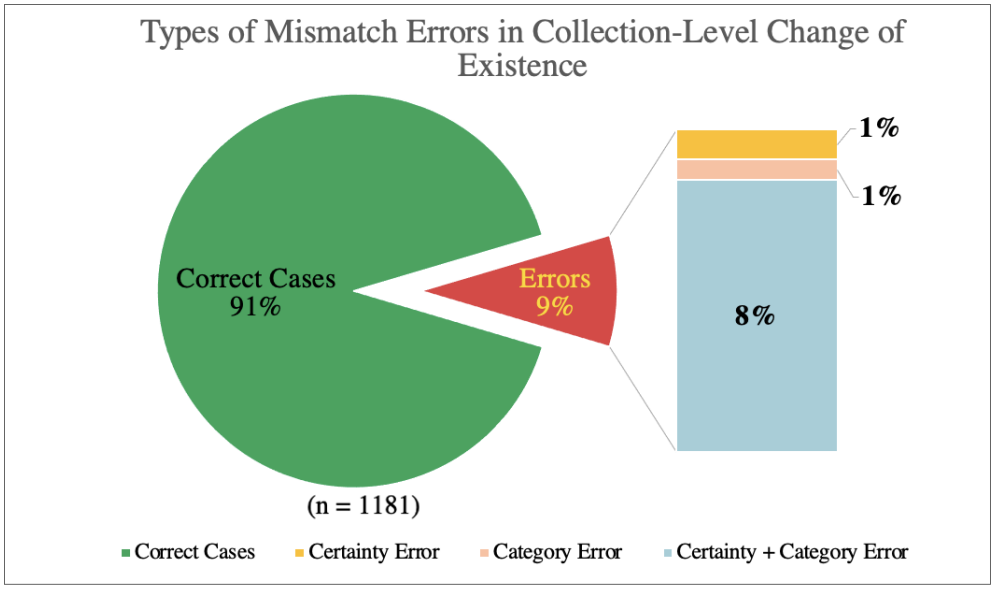

Figure 5.5: **Types of Errors in Test Set for the *Collection-Level COS* Task (N = 1181)**. The left circle shows that 91% are correctly predicted (green) and 9% are incorrectly predicted (red) out of an entire test set (N = 1181). The column in the right shows percentage of mismatched cases due to certainty error (yellow), category error (pink), or both (blue).

instance level because the model occasionally became confused with patient-level prediction when a collection of tumors were described. As illustrations, "Most of these are unchanged, but one lesion measuring 6mm is not seen on the prior study" and "On the current examination, there are three lung lesions in the left lung versus two seen on the previous CT examination", the model incorrectly predicted 'definite stable' when the truth label was 'definitely changed' because there were at least one tumor instance in the set experiencing a change from non-exists to exists.

A qualitative analysis shows that the model is less reliable on cases where COS is underspecified for both of the tasks. In simple 'change of state' sentences, the classifier associated potential hedging cues and its scopes on COS phrases (i.e., 'again seen', 'recurred', 'resolved', 'mostly unchanged', 'likely unchanged'). However, COS can be expressed in a variety of ways.

Table 5.2: Difficult Cases in Change of Existence Tasks

Existence Pairs ⟨Def Exist(t_{-1}), Def Exist(t_0)⟩	**Entailed COS** ⟨Certainty, COS Classes⟩	Examples
⟨Def Absent, Def Exists⟩	⟨Def +Change⟩	"this large left temporal mass was not present on prior 2002 MRI"
⟨Def Exists, Def Absent⟩	⟨Def -Change⟩	"the lesion in the left frontal lobe was about 6mm diameter, but is not visible now."
⟨Likely Exists, Def Exists⟩	⟨Likely Stable⟩	"The mass appears to have been present on the previous study"
⟨Likely Absent, Def Exists⟩	⟨Likely +Change⟩	"this large left temporal mass was not present on prior 2002 MRI"
⟨Def Exists, Likely Absent⟩	⟨Likely -Change⟩	"the brain lesion found is not well seen now."

COS is formally the change with respect to existence over the interval visits. The expressed certainty and direction of change are derived from the expression of certainties and classes in that existence pair (e.g., ⟨likely exists at t_{-1}, definitely exists at t_0⟩ → likely stable). Incorrect predictions were frequently caused by the confusion in the mapping from specific existence pairs to COS states. To elucidate, *Table 5.2* shows examples in support of the mapping from existence pairs to COS states. The model has trouble in realizing the COS from clues on the existence pairs.

Another source of error for incorrect COS predictions is associating the change in the trend of mass properties to specific state of COS. One example is "no tumor has gotten smaller". The description on the size trend presupposes the tumor existed on the prior and current visit. From that, we can further infer that there is no change (equivalent to ⟨definitely stable⟩) in the instance-level and collection-level COS. This example shows that instances where context can map to an overlapped existence pair (both SO and DO groups) as an intermediary, can also map to states in the COS tasks as well. Here, I conclude that there is room for improvement in realizing the states from cases where COS can be derived from deductive reasoning.

Error Analysis for Task 5: Significance of Change

As a part of the analysis, I reviewed the results for the significance of change task manually. The NLP system has a high macro- and micro-F1 scores (0.88, 0.94) in classifying tumor

progression. Reports that were correctly classified by this system were mainly those with explicit mentions of indicators. Recalling from before, reports with classifiable progression associate with indicators such as metastasis, novelty, recurrence, growing size, and worsening mass effects. Error analysis shows the classifier had difficulty in associating the mass effects (i.e., more cavitation, edema, or necrosis) with disease progression. This may be caused by the classifier was unable to learn from these low frequency progression indicators. The classifier mostly predicted progression correctly when features other than mass effects are used. However, it is noteworthy that this is also affected by *how* these markers are expressed.

Most error instances occurred when these main indicators (size trend, metastasis, novel mass) were not mentioned explicitly. A general statement on the main indicators is more likely to be correctly classified compared to those expressed through inferences. But in instances like *Example (21a)*, the system incorrectly predicted no progression when increase size trend is entailed by the specific measurements on prior and current visits. Another case of incorrect prediction is shown in *Example (21b)*. An additional mass was entailed by the incongruity in the number of lesions observed between the two visits. In essence, the system is also having trouble on cases that require multi-step reasoning, specifically numerical reasoning. It may be helpful to first abstract the numerical representation and use symbolic rules to conclude the trends between these numerical representations before associating these indicators with disease progression in the future.

(21) Incorrect Predictions in Deterioration Task

a. *"The left adrenal mass now measures 4.6 x 3.8 cm, which was 1.7 x 1.2 cm on the prior study"*

b. *"On the current examination, there are three lung lesions in the left lung versus two seen on the previous CT examination."*

Chapter 6

Conclusion

Developing an algorithm to automatically perform discourse-level analysis aiming to retrieve the tumor status and tumor progression from unstructured radiology reports has many practical applications in the domain of care, oncology research, and population-level cancer surveillance. This thesis has focused on using NLP techniques to characterize the tumor status and its progression through three conceptual frames: existence, change of existence, and significance of change. Existing NLP works that have conceptualized these knowledge from unstructured EHRs were not as comprehensive and in-depth as the model presented in this thesis.

By learning from the annotations that leverage systemic inferences (including presuppositions and entailments), the developed BERT-based text classification model can reliably output the tumor's status on *existence*, *change of existence*, and *significance of change* tasks. On the test set, the model achieved micro-average F1-measures in the range between 0.88 - 0.94 depending on the tasks. Specifically, this model performs reasonably well in determining whether the tumor has progressed (micro-F1 0.94), further demonstrating the model's capability to utilize a wide range of systemic inferences to come to a conclusion.

A detailed error analysis shows that future improvements can focus on difficult cases

where these status are entailed and not explicitly said. It is not as simple as targeting the interpretation of descriptions on *existence*, *change of existence*, and *significance of change* under negated or uncertainty environments. In fact, these status can be expressed in a variety of ways. To further improve the system's capability in handling challenging cases, future works can prioritize on enhancing the model's cognitive ability – casual reasoning (i.e., effect of treatments on existence), logical reasoning (i.e., mapping clues on states of existence to states in change of existence), numerical reasoning (i.e., comparing size of tumor), and temporal reasoning. In conclusion, the preliminary works presented in this thesis have many practical downstream applications, but can be refined to tackle cases needed the cognitive processing step-by-step.

Bibliography

[1] (CDC.) Centers for Disease Control and Prevention & (NCI)., t. N. C. I. (n.d.). Cancer statistics at a glance.

[2] Pons, E., Braun, L. M. M., Hunink, M. G. M. & Kors, J. A. (2016). Natural Language Processing in Radiology: A Systematic Review. *https://doi.org/10.1148/radiol.16142770*, *279*(2), 329–343. https://doi.org/10.1148/RADIOL.16142770

[3] Demner-Fushman, D., Chapman, W. W. & McDonald, C. J. (2009). What can natural language processing do for clinical decision support? *Journal of Biomedical Informatics*, *42*(5), 760–772. https://doi.org/10.1016/J.JBI.2009.08.007

[4] Kreimeyer, K., Foster, M., Pandey, A., Arya, N., Halford, G., Jones, S. F., Forshee, R., Walderhaug, M. & Botsis, T. (2017). Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. https://doi.org/10.1016/j.jbi.2017.07.012

[5] Savova, G. K., Danciu, I., Alamudun, F., Miller, T., Lin, C., Bitterman, D. S., Tourassi, G. & Warner, J. L. (2019). Use of Natural Language Processing to Extract Clinical Cancer Phenotypes from Electronic Medical Records. *Cancer research*, *79*(21), 5463. https://doi.org/10.1158/0008-5472.CAN-19-0579

[6] Warren, J. L. & Yabroff, K. R. (2015). Challenges and Opportunities in Measuring Cancer Recurrence in the United States. *JNCI: Journal of the National Cancer Institute*, *107*(8), 134. https://doi.org/10.1093/JNCI/DJV134

[7] Weir, C. R. & Nebeker, J. R. (2007). Critical Issues in an Electronic Documentation System. *AMIA Annual Symposium Proceedings*, *2007*, 786.

[8] Hsu, C., Karnwal, S., Mullainathan, S., Obermeyer, Z. & Tan, C. (2020). Characterizing the Value of Information in Medical Notes.

[9] Jabour, A. M., Dixon, B. E., Jones, J. F. & Haggstrom, D. A. (2018). Toward Timely Data for Cancer Research: Assessment and Reengineering of the Cancer Reporting Process. *JMIR Cancer*, *4*(1). https://doi.org/10.2196/CANCER.7515

[10] Merriman, K. W., Broome, R. G., Pozas, G. D. L., Landvogt, L. D., Qi, Y. & Keating, J. (2021). Evolution of the Cancer Registrar in the Era of Informatics. *JCO Clinical Cancer Informatics*, 3–4. https://doi.org/10.1200/CCI.20

[11] Cancer Registry. (2019). *National Cancer Institutes*, *2546*(4), 1–15.

[12] Hiller, J. G., Perry, N. J., Poulogiannis, G., Riedel, B. & Sloan, E. K. (2017). Perioperative events influence cancer recurrence risk after surgery. *Nature Reviews Clinical Oncology 2017 15:4*, *15*(4), 205–218. https://doi.org/10.1038/nrclinonc.2017.194

[13] Lodish, H., Berk, A., Zipursky, S. L., Matsudaira, P., Baltimore, D. & Darnell, J. (2000). Tumor Cells and the Onset of Cancer. *Molecular Cell Biology*, (4).

[14] Langlotz, C. P. (2015). *The Radiology Report: A Guide to Thoughtful Communication for Radiologists and Other Medical Professionals* (1st ed.).

[15] Galloway, M. & Taiyeb, T. (2011). The Interpretation of Phrases Used to Describe Uncertainty in Pathology Reports. *Pathology Research International*, *2011*, 1–6. https://doi.org/10.4061/2011/656079

[16] Saurí, R. & Pustejovsky, J. (2012). *Are You Sure That This Happened? Assessing the Factuality Degree of Events in Text* (tech. rep. No. 2). https://doi.org/10.1162/COLI_a_00096

[17] Poggi, I., D'errico, F. & Vincze, L. (2019). Uncertain Words, Uncertain Texts. Perception and Effects of Uncertainty in Biomedical Communication. *Acta Polytechnica Hungarica*, *16*(2), 2019–2032.

[18] Shinagare, A. B., Lacson, R., Boland, G. W., Wang, A., Silverman, S. G., Mayo-Smith, W. W. & Khorasani, R. (2019). Radiologist Preferences, Agreement, and Variability in Phrases Used to Convey Diagnostic Certainty in Radiology Reports. *Journal of the American College of Radiology*, *16*(4), 458–464. https://doi.org/10.1016/j.jacr.2018.09.052

[19] Yim, W.-w., Yetisgen, M., Harris, W. P. & Kwan, S. W. (2016). Natural Language Processing in Oncology: A Review. *JAMA Oncology*, *2*(6), 797–804. https://doi.org/10.1001/JAMAONCOL.2016.0213

[20] Currie, A.-M., Cohan, J. & Zlatic, L. (2001). Information Retrieval of Electronic Medical Records. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *2004*, 460–471. https://doi.org/10.1007/3-540-44686-9_46

[21] Zhou, L. & Hripcsak, G. (2007). Temporal reasoning with medical data—A review with emphasis on medical natural language processing. *Journal of Biomedical Informatics*, *40*(2), 183–202. https://doi.org/10.1016/J.JBI.2006.12.009

[22] Sun, W., Rumshisky, A. & Uzuner, O. (2013). Temporal reasoning over clinical text: the state of the art. *Journal of the American Medical Informatics Association*, *20*(5), 814–819. https://doi.org/10.1136/AMIAJNL-2013-001760

[23] Simons, M. (2002). Presupposition and Relevance. *Semantics vs. pragmatics* (pp. 329–355). Oxford University Press.

[24] Simons, M. (2013). On the Conversational Basis of Some Presuppositions. *Perspectives in Pragmatics, Philosophy and Psychology*, *2*, 329–348. https://doi.org/10.1007/978-3-319-01014-4_13

[25] Stalnaker, R. (1974). Pragmatic Presuppositions. *Context and content* (pp. 47–62). Oxford University Press.

[26] Kadmon, N. (2001). *Formal Pragmatics: Semantics, Pragmatics, Presupposition, and Focus*. Blackwell. https://doi.org/10.1016/j.pragma.2004.04.009

[27] Potts, C. (2015). Presupposition and Implicature. *The handbook of contemporary semantic theory* (pp. 168–202). https://doi.org/10.1002/9781118882139.ch6

[28] Kehl, K. L., Xu, W., Lepisto, E., Elmarakeby, H., Hassett, M. J., Allen, E. M. V., Johnson, B. E. & Schrag, D. (2020). Natural Language Processing to Ascertain Cancer Outcomes From Medical Oncologist Notes. *https://doi.org/10.1200/CCI.20.00020*, (4), 680–690. https://doi.org/10.1200/CCI.20.00020

[29] Bitterman, D. S., Miller, T. A., Mak, R. H. & Savova, G. K. (2021). Clinical Natural Language Processing for Radiation Oncology: A Review and Practical Primer. *International Journal of Radiation Oncology*Biology*Physics*, *110*(3), 641–655. https://doi.org/10.1016/J.IJROBP.2021.01.044

[30] He, T., Ogunti, R., Puppala, M., Chen, S., Yu, X., Mancuso, J. J. & Wong, S. T. (2016). A smartphone app framework for segmented cancer care coordination. *3rd IEEE EMBS*

International Conference on Biomedical and Health Informatics, BHI 2016, 372–375. https://doi.org/10.1109/BHI.2016.7455912

[31] He, T., Puppala, M., Ogunti, R., Mancuso, J. J., Yu, X., Chen, S., Chang, J. C., Patel, T. A. & Wong, S. T. (2017). Deep learning analytics for diagnostic support of breast cancer disease management. *2017 IEEE EMBS International Conference on Biomedical and Health Informatics, BHI 2017*, 365–368. https://doi.org/10.1109/BHI.2017.7897281

[32] Ritzwoller, D., Carroll, N., Burnett-Hartman, A., Feigelson, H. & Lyons, E. (2016). Performance Of Natural Language Processing In Identifying Lung Nodule Characteristics After Low-Dose CT Lung Cancer Screening. *American Journal of Respiratory and Critical Care Medicine*.

[33] Lacson, R., Harris, K., Brawarsky, P., Tosteson, T. D., Onega, T., Tosteson, A. N., Kaye, A., Gonzalez, I., Birdwell, R. & Haas, J. S. (2015). Evaluation of an Automated Information Extraction Tool for Imaging Data Elements to Populate a Breast Cancer Screening Registry. *Journal of Digital Imaging*, *28*(5), 567–575. https://doi.org/10.1007/s10278-014-9762-4

[34] Acevedo, F., Armengol, V. D., Deng, Z., Tang, R., Coopey, S. B., Braun, D., Yala, A., Barzilay, R., Li, C., Colwell, A., Guidi, A., Cetrulo, C. L., Garber, J., Smith, B. L., King, T. & Hughes, K. S. (2018). Pathologic findings in reduction mammoplasty specimens: a surrogate for the population prevalence of breast cancer and high-risk lesions. *Breast Cancer Research and Treatment 2018 173:1*, *173*(1), 201–207. https://doi.org/10.1007/S10549-018-4962-0

[35] Yala, A., Barzilay, R., Salama, L., Griffin, M., Sollender, G., Bardia, A., Lehman, C., Buckley, J. M., Coopey, S. B., Polubriaginof, F., Garber, J. E., Smith, B. L., Gadd, M. A., Specht, M. C., Gudewicz, T. M., Guidi, A. J., Taghian, A. & Hughes, K. S. (2016). Using machine learning to parse breast pathology reports. *Breast Cancer Research and Treatment*, *161*(2), 203–211. https://doi.org/10.1007/S10549-016-4035-1

[36] Karunakaran, B., Misra, D., Marshall, K., Mathrawala, D. & Kethireddy, S. (2017). Closing the loop - Finding lung cancer patients using NLP. *Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017*, *2018-Janua*, 2452–2461. https://doi.org/10.1109/BIGDATA.2017.8258203

[37] Chapman, W. W., Bridewell, W., Hanbury, P., Cooper, G. F. & Buchanan, B. G. (2001). A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics*, *34*(5), 301–310. https://doi.org/10.1006/jbin.2001.1029

[38] Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., Seekins, J., Mong, D. A., Halabi, S. S., Sandberg, J. K., Jones, R., Larson, D. B., Langlotz, C. P., Patel, B. N., Lungren, M. P. & Ng, A. Y. (2019). CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, 590–597. https://doi.org/10.1609/aaai.v33i01.3301590

[39] Roberts, K., Si, Y., Gandhi, A. & Bernstam, E. V. (2019). A framenet for cancer information in clinical narratives: Schema and annotation. *LREC 2018 - 11th International Conference on Language Resources and Evaluation*, 272–279.

[40] Vincze, V., Szarvas, G., Farkas, R., Móra, G. & Csirik, J. (2008). The BioScope corpus: Biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, *9*(SUPPL. 11), S9. https://doi.org/10.1186/1471-2105-9-S11-S9

[41] Vincze, V. (2010). Speculation and negation annotation in natural language texts: what the case of BioScope might (not) reveal. *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, 28–31.

[42] Ping, X. O., Tseng, Y. J., Chung, Y., Wu, Y. L., Hsu, C. W., Yang, P. M., Huang, G. T., Lai, F. & Liang, J. D. (2013). Information extraction for tracking liver cancer patients' statuses: From mixture of clinical narrative report types. *Telemedicine and e-Health*, *19*(9), 704–710. https://doi.org/10.1089/tmj.2012.0241

[43] Mamlin, B. W., Heinze, D. T. & McDonald, C. J. (2003). Automated Extraction and Normalization of Findings from Cancer-Related Free-Text Radiology Reports. *AMIA Annual Symposium Proceedings*, *2003*, 420.

[44] Coden, A., Savova, G., Sominsky, I., Tanenblatt, M., Masanz, J., Schuler, K., Cooper, J., Guan, W. & de Groen, P. C. (2009). Automatically extracting cancer disease characteristics from pathology reports into a Disease Knowledge Representation Model. *Journal of Biomedical Informatics*, *42*(5), 937–949. https://doi.org/10.1016/J.JBI.2008.12.005

[45] Yim, W.-w., Kwan, S. W. & Yetisgen, M. (2017). Classifying tumor event attributes in radiology reports. *Journal of the Association for Information Science and Technology*, *68*(11), 2662–2674. https://doi.org/10.1002/ASI.23937

[46] Styler, W. F., Bethard, S., Finan, S., Palmer, M., Pradhan, S., de Groen, P. C., Erickson, B., Miller, T., Lin, C., Savova, G. & Pustejovsky, J. (2014). Temporal Annotation in the Clinical Domain. *Transactions of the Association for Computational Linguistics*, *2*, 143–154. https://doi.org/10.1162/tacl_a_00172

[47] Galvan, D., Okazaki, N., Matsuda, K. & Inui, K. (2018). *Investigating the Challenges of Temporal Relation Extraction from Clinical Text* (tech. rep.).

[48] Leeuwenberg, A. & Moens, M.-F. (2019). A Survey on Temporal Reasoning for Temporal Information Extraction from Text. *Journal of Artificial Intelligence Research*, *66*, 341–380. https://doi.org/10.1613/JAIR.1.11727

[49] Cheng, L. T. E., Zheng, J., Savova, G. K. & Erickson, B. J. (2009). Discerning Tumor Status from Unstructured MRI Reports—Completeness of Information in Existing Reports and Utility of Automated Natural Language Processing. *Journal of Digital Imaging 2009 23:2*, *23*(2), 119–132. https://doi.org/10.1007/S10278-009-9215-7

[50] Datta, S., Bernstam, E. V. & Roberts, K. (2019). A frame semantic overview of NLP-based information extraction for cancer-related EHR notes. *Journal of Biomedical Informatics*, *100*, 103301. https://doi.org/10.1016/J.JBI.2019.103301

[51] Vanderwende, L., Xia, F. & Yetisgen-Yildiz, M. (2013). Annotating Change of State for Clinical Events. *Workshop on Events: Definition, Detection, Coreference, and Representation*, (June), 47–51.

[52] Hassanpour, S., Bay, G. & Langlotz, C. P. (2017). Characterization of Change and Significance for Clinical Findings in Radiology Reports Through Natural Language Processing. *Journal of Digital Imaging*, *30*(3), 314–322. https://doi.org/10.1007/s10278-016-9931-8

[53] Wiegand, M., Roth, B., Klakow, D., Balahur, A. & Montoyo, A. (2010). *Negation and Speculation in Natural Language Processing*.

[54] Uzuner, Ö., South, B. R., Shen, S. & DuVall, S. L. (2011). 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, *18*(5), 552–556. https://doi.org/10.1136/AMIAJNL-2011-000203

[55] Mehrabi, S., Krishnan, A., Sohn, S., Roch, A. M., Schmidt, H., Kesterson, J., Beesley, C., Dexter, P., Max Schmidt, C., Liu, H. & Palakal, M. (2015). DEEPEN: A negation

detection system for clinical text incorporating dependency relation into NegEx. *Journal of Biomedical Informatics*, *54*, 213–219. https://doi.org/10.1016/j.jbi.2015.02.010

[56] Lakoff, G. (1973). Hedges: A study in meaning criteria and the logic of fuzzy concepts. *Journal of Philosophical Logic*, *2*(4), 458–508. https://doi.org/10.1007/BF00262952

[57] Wallis, A. & McCoubrie, P. (2011). The radiology report — Are we getting the message across? *Clinical Radiology*, *66*(11), 1015–1022. https://doi.org/10.1016/J.CRAD.2011.05.013

[58] Hanauer, D. A., Liu, Y., Mei, Q., Manion, F. J., Balis, U. J. & Zheng, K. (2012). Hedging their mets: the use of uncertainty terms in clinical documents and its potential implications when sharing the documents with patients. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, *2012*, 321–330.

[59] Karttunen, L. (1973). Presuppositions of Compound Sentences. *Linguistic Inquiry*, *4*(2), 169–193.

[60] Potts, C. (2005). *The Logic of Conventional Implicatures*. Oxford University Press. https://doi.org/https://doi.org/10.1093/acprof:oso/9780199273829.001.0001

[61] Simons, M. (2006). Presupposition without Common Ground.

[62] Grice, P. (1975). Logic and Conversation. *Syntax and Semantics*, *3*(Speech Acts), 43–58. https://doi.org/10.1057/9780230005853_5

[63] Chierchia, G. & McConnell-Ginet, S. (1990). *Meaning and Grammar: An Introduction to Semantics*. MIT Press. https://doi.org/10.2307/415078

[64] Cox, J. & Graham, Y. (2019). Radiology and patient communication: If not now, then when? *European Radiology 2019 30:1*, *30*(1). https://doi.org/10.1007/S00330-019-06349-8

[65] Stallinga, H. A., Ten Napel, H., Jansen, G., Geertzen, J., De Vries Robbe, P. & Roodbol, P. (2014). Does language ambiguity in clinical practice justify the introduction of standard terminology? An integrative review aggregation and reuse of data from electronic patient records for different purposes, including multidisciplinary decision-making and research. *Journal of Clinical Nursing*, *24*, 344–352. https://doi.org/10.1111/jocn.12624

[66] Zolnierek, K. B. H. & DiMatteo, M. R. (2009). Physician communication and patient adherence to treatment: A meta-analysis. *Medical care*, *47*, 826. https://doi.org/10.1097/MLR.0B013E31819A5ACC

[67] Bodenreider, O. (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, *32*(Database issue), D267. https://doi.org/10.1093/NAR/GKH061

[68] Zhang, Y., Zhang, Y., Qi, P., Manning, C. D. & Langlotz, C. P. (2021). Biomedical and clinical english model packages for the stanza python nlp library. *Journal of the American Medical Informatics Association*, *28*(9), 1892–1899.

[69] Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, D., Naumann, T. & McDermott, M. B. A. (2019). Publicly Available Clinical BERT Embeddings. https://arxiv.org/abs/1904.03323v3