A FrameNet for Cancer Information in Clinical Narratives: Schema and Annotation

Kirk Roberts[1], **Yuqi Si**[1], **Anshul Gandhi**[1], **Elmer V. Bernstam**[1,2]
[1] School of Biomedical Informatics
[2] Division of General Internal Medicine, Department of Internal Medicine, McGovern Medical School
The University of Texas Health Science Center at Houston
Houston, TX USA
kirk.roberts@uth.tmc.edu

Abstract

This paper presents a pilot project named Cancer FrameNet. The project's goal is a general-purpose natural language processing (NLP) resource for cancer-related information in clinical notes (i.e., patient records in an electronic health record system). While previous cancer NLP annotation projects have largely been ad hoc resources to address a specific and immediate information need, the frame semantic method employed here emphasizes the information presented in the notes themselves and its linguistic structure. To this end, three semantic frames (targeting the high-level tasks of cancer diagnoses, cancer therapeutic procedures, and tumor descriptions) are created and annotated on a clinical text corpus. Prior to annotation, candidate sentences are extracted from a clinical data warehouse and de-identified to remove any private information. The frames are then annotated with the three frames totaling over thirty frame elements. This paper describes these steps in the pilot project and discusses issues encountered to evaluate the feasibility of general-purpose linguistic resources for extracting cancer-related information.
Keywords: clinical information extraction, cancer, frame semantics

1 Introduction

Important medical information about cancer patients is often only available in free text (natural language) notes in electronic health records (EHRs). This information is frequently needed for research, quality improvement, surveillance, and other important functions. However, the manual abstraction of this information can be incredibly time-consuming and expensive, often making it infeasible for both early-stage research and clinical quality improvement projects. Thus, natural language processing (NLP) approaches can offer a tremendous service in oncology.

Existing NLP systems for cancer-related information generally fall under a type of biomedical NLP task known as *phenotyping*, which is the task of identifying patients that meet a certain set of criteria. Phenotyping specifications are often highly task-specific, which frequently yields one-off NLP datasets and algorithms that do not generalize to similar phenotyping tasks. For example, one phenotype method may identify lung cancer patients with a tumor of at least 2cm in diameter, while another method may identify lung cancer patients with at least two tumors that are 1cm in diameter. Such methods are quite similar, yet often produce incompatible annotations and algorithms.

The key insight is that phenotyping methods often conflate extraction (identifying relevant portions of text) and reasoning (determining if the extracted text fulfills the task's needs). By separating these steps and focusing on general-purpose extraction, it will be possible to easily and rapidly develop phenotyping methods. This is because the bulk of the effort is typically spent in the extraction step (annotating data and developing NLP algorithms), while the reasoning step is often a straightforward set of rules. Continuing the example above, an NLP system capable of extracting all tumor references and their sizes from text would easily meet the needs of both phenotyping methods (along with potentially many others).

The challenge then becomes how to develop general-purpose extraction algorithms for clinical text, especially when it isn't necessarily clear a priori what information needs to be extracted. Luckily, *frame semantics* provides a useful framework for developing such a resource. In frame semantics, a word or phrase evokes a frame of semantic knowledge that describes the characteristic attributes associated with a concept. For example, for *tumor*, the frame would likely contain elements that describe the size, location, and morphology of the tumor. The set of frame elements can either be defined a priori by a subject expert or added iteratively based on the data (this work combines both approaches).

The specification of a set of frames combined with annotated examples is referred to as a FrameNet, the best known being Berkeley FrameNet (Baker et al., 1998). But other FrameNets exist as well, notably domain-specific FrameNets. This paper describes a pilot project to build such a domain-specific resource—referred to hereafter as Cancer FrameNet—that focuses on cancer-related information. The goal of the pilot is to test the feasibility of a much larger resource covering the depth and breadth of cancer information in patient records. Even as a pilot, however, this resource is still sizable, covering three important frames, 22 lexical units, and nearly 8 thousand annotated sentences.

There are several potential pitfalls for frame annotation in clinical notes, thus the need for a feasibility pilot. These issues include the consistency of cancer-related information in clinical notes: is cancer too complex with too many variables to be reliably annotated (both manually and by an automatic NLP system)? Another issue is the overlapping information of frames: are there really semantically distinct

concepts that can be formed into different frames? Finally, frame annotation is typically limited to sentence and clause context (excluding implicit information), but does this apply to cancer-related information in clinical notes?

It should be noted that there is a second major challenge in the generalizability of clinical NLP systems. This involves the portability of algorithms from one institution's clinical notes to another, as oftentimes these can be drastically different. While we acknowledge the critical importance of this problem, we make no attempt to solve it here. All data described in this paper come from a single institution. It is our hope, however, that upon establishing a general-purpose resource, annotated frames from additional institutions can be added in order to improve inter-institutional generalizability.

The rest of this paper is structured as follows: Section 2 outlines previous work in both domain-specific FrameNets and cancer-related information extraction. Section 3 describes the pre-annotation process: where the data came from, how it is extracted and prepared for annotation, including how it is de-identified to protect patient privacy. Section 4 lays out the frames covered in the pilot project: what they are, why they were chosen, and the current frame elements. Section 5 details the annotation process. Finally, Section 6 discusses the potential ramifications of Cancer FrameNet on cancer information extraction, including its strengths and weaknesses, and how it might or might not overcome some of the aforementioned pitfalls.

2 Background

The theory of frame semantics has spawned many practical natural language resources. Most prominent is the Berkeley FrameNet project (Baker et al., 1998; Baker et al., 2015). which is intended to be an open-domain encoding of common knowledge (commercial transactions, transportation, crime, international affairs, etc.). The FrameNet construction methodology has been ported to many other languages (Heppin and Gronostaj, 2014; Lin et al., 2015; Rezhake and Kuerban, 2015; Ohara, 2016). More interesting, many domain-specific FrameNets exist, ranging from soccer (Schmidt, 2006; Torrent et al., 2014) to sentiment (Ruppenhofer, 2013) to disability (Savova et al., 2005) to cellular pathways (Dolbey et al., 2006; Dolbey, 2009). No such resource targets the types of cancer-related information found in EHR notes.

While not explicitly based on a FrameNet-style approach, a significant amount of work has focused on NLP systems for extracting cancer-related information from EHRs. A sampling of the types of information extracted include: a frame-like representation of radiological findings (Taira et al., 2001); procedures, tumor stages, and various biomarker scores (Xu et al., 2004); tumor and node staging (McCowan et al., 2007); Gleason score, tumor stages, and margin status (D'Avolio et al., 2008); histology, site, dimension, and various tumor types (Coden et al., 2009); tumor progression (Cheng et al., 2010); colonoscopy status (Denny et al., 2010); colonoscopy quality measures (Harkema et al., 2010); Gleason score, Clark level, and Breslow depth (Napolitano et al., 2010); cancer history (Wilson et al., 2010); counts of examined and positive tumors and nodes (Martinez and Li, 2011); pancreatic cancer predictors (Zhao and Weng, 2011); tumor staging and biomarkers (Segagni et al., 2012); pain in prostate cancer patients (Heintzelman et al., 2013); highest level of pathology, number of removed adenomas (Imler et al., 2013); tumor, node, metastases, and ACPS stages (Martinez et al., 2013); liver cancer status (Ping et al., 2013); change of event state (Vanderwende et al., 2013); twenty-two staging indicators (Ashish et al., 2014); volume, size, and location (Wang et al., 2014); and diagnosis, hormone receptor status, tumor size, and number of positive nodes (Napolitano et al., 2016). This synopsis understates the number of extracted information types, but it is still clear that there is a significant breadth of information as well as consistent areas of overlap.

A more direct comparison to our work is the recent work in the DeepPhe project (Savova et al., 2017). DeepPhe takes a document-level approach to extracting cancer information, which is more appropriate for certain data types than the sentence-based approach proposed below. Most crucially, it is unknown how well their document-level approach generalizes to other institution's data. On the other hand, while the pilot project discussed in this paper focuses on a single institution as well, we hypothesize that a frame-based method targeting information at the sentence level will result in greater potential for generalization across institutions.

3 Preparing Clinical Narratives

Several steps are necessary to prepare clinical narratives for frame annotation: clinical notes must be retrieved from the clinical data warehouse, lexical units (see Section 4) and their proper context must be extracted, then private patient information must be de-identified. All of this must be done within a secure HIPAA-compliant environment.

The clinical notes are derived from the UT Physicians clinics, a chain of outpatient clinics in the Houston area. The snapshot available from the data warehouse contains more than 260,000 notes with more than 175 million tokens. While by no means a large corpus by EHR standards, it contains sufficient data for a pilot evaluation. This project was deemed exempt by the Committee for the Protection of Human Subjects, the UTHealth Institutional Review Board, under protocol number HSC-SBMI-13-0549.

For each lexical unit (see Section 4), every sentence containing the lexical unit is extracted from the note corpus. Sentence segmentation is by no means a simple task in clinical data (Miller et al., 2015; Zweigenbaum et al., 2016). Stanford CoreNLP (Manning et al., 2014) was used to identify initial sentences, but due to the lack of punctuation in clinical notes, these often constituted multiple (sometimes dozens) of sentences (newlines are often used as end-of-sentence markers, but frequently newlines do not end sentences). As a result, several high-precision rules were used to prune down sentence length. Ultimately, a human annotator was required to remove words not part of the lexical

unit's proper sentence boundary. This was done in conjunction with the de-identification stage below.

Clinical notes contain significant amounts of copy-pasting and templated sentences, so a random sample of sentences might contain substantial numbers of duplicates and near-duplicates. To maximize the diversity of the annotations, the sentences were sorted by TF-IDF cosine distance.

The final preparation step is to de-identify the notes, removing any protected health information (PHI) and replacing it with a placeholder. In the United States, the Health Insurance Portability and Accountability Act (HIPAA) mandates the de-identification of 18 categories of information:

(1) names
(2) geographic localities smaller than a state
(3) dates and ages over 89
(4) telephone numbers
(5) fax numbers
(6) e-mail addresses
(7) Social Security numbers
(8) medical record numbers
(9) health plan numbers
(10) account numbers
(11) certificate/license numbers
(12) vehicle & license numbers
(13) device identifiers
(14) web URLs
(15) IP addresses
(16) biometric identifiers
(17) full face photographic images
(18) any other identifying number/characteristic/code

We expanded on this to cover all ages and geographic localities, as well as other types of identifying information as described in Stubbs and Uzuner (2015). Furthermore, one of our lexical units is frequently a surname in the data: these sentences are completely discarded. Both human annotation and an in-house automatic system (Lee et al., 2017) were used. To reduce bias, the human de-identification occurred first, then the automatic system provided additional suggestions that the human may have missed. The automatic de-identifications were all manually verified to reduce the proliferation of false positives common with de-identification systems. Finally, at future stages of the annotation, as detailed below, annotators always have the option of identifying further PHI missed by this process.

4 Initial Frame Schemas

Based on the existing literature, three common phenotyping tasks were selected: (1) identification of patients with a particular cancer diagnosis, (2) identification of patients with a particular cancer treatment, and (3) identification of patients with particular tumor characteristics. To reduce the complexity of the second task, and to focus on data more likely to be found in outpatient notes, treatments are limited to surgical procedures (i.e., excluding medications, chemotherapy, etc.). These three tasks yield three frames (i) CANCER_DIAGNOSIS, (ii) CANCER_THERAPEUTIC_PROCEDURE, and (iii) TUMOR_DESCRIPTION. We also define an abstract root frame, CANCER_MASTER_FRAME, from which all three inherit elements. This enables frame elements (attributes) that are universal, such as negation and certainty. For each of these frames, an expert in cancer informatics (EVB) helped devise a list of lexical units:

CANCER_DIAGNOSIS: *adenocarcinoma*, *cancer*, *carcinoma*, *leukemia*, *lymphoma*, *malignancy*, *malignant*, *melanoma*, *myeloma*, *sarcoma*

CANCER_THERAPEUTIC_PROCEDURE: *colectomy*, *hysterectomy*, *lymphadenectomy*, *mastectomy*, *palliative*, *pancreatectomy*, *prostatectomy*, *radiation*, *whipple*

TUMOR_DESCRIPTION: *lesion*, *mass*, *tumor*

For each of these lexical units, the process described in Section 3 was followed until up to 500 de-identified sentences were available for each lexical unit. Five of the lexical units (*lymphadenectomy*, *myeloma*, *pancreatectomy*, *sarcoma*, and *whipple*) had fewer than 500 sentences in the corpus.

The elements (attributes) of each frame were determined by an iterative process. First, an initial set of elements was proposed by the cancer expert. During the course of the annotation, new elements were frequently proposed by the annotators. Elements with sufficient frequency and importance–as determined by the cancer expert–were added to the frame schema. For example, FAMILY-HISTORY (for the CANCER_DIAGNOSIS frame), EXTENT (for CANCER_THERAPEUTIC_PROCEDURE), and RECURRENCE (for TUMOR_DESCRIPTION) were added after the start of annotation. It is expected that further annotation will yield additional changes to the schema. The set of frame elements, including brief definitions, is shown in Table 1. Note that some elements (e.g., STATUS, PATIENT) are part of all three frames, but not the CANCER_MASTER_FRAME as these are not expected to necessarily apply to future frames.

5 Annotation

The annotation process largely followed standard linguistic annotation practices (Pustejovsky and Stubbs, 2013). Notably, the sentences containing candidate lexical units were double-annotated then reconciled with the help of a third individual. All frame annotation was performed in Brat (Stenetorp et al., 2012). See Figure 1 for examples.

Two special annotations, whose functionality was briefly mentioned earlier, deserve more attention here. First, as shown in Table 1, there is a special "???" element that annotators can use to indicate potentially useful information that may later result in the creation of a new element. (see Figure 1 for an example). As the disease is so complex, there is simply too much information associated with cancer to include elements for all possible types of information. So the ??? element allows for the prioritization of information based on the actual frequency in the clinical notes. Second, the annotation denoted as ERROR is used by

Frame Element	Description
CANCER_MASTER_FRAME	
CERTAINTY	Certainty/hedging of frame (e.g., *possible*, *likely*)
DATETIME	Temporal information for the frame (often reference to PHI element)
POLARITY	Existence/negation of frame (e.g., *no*, *positive*)
???	Used for other phrases in the text the annotator feels is important, but do not have a corresponding frame element
CANCER_DIAGNOSIS	
DESCRIPTION	Other frame with further information (e.g., TUMOR_DESCRIPTION)
FAMILYHISTORY	Specifies a family member with the diagnosis (as opposed to the PATIENT)
HISTOLOGY	Histological description (e.g., *carcinoma*), can be lexical unit
LOCATION	Part of body associated with the cancer
PATIENT	Reference to the patient (e.g., *patient*, *female*)
QUANTITY	Some quantitative measure of the cancer
STATUS	Diagnostic status (e.g., *history*, *ongoing*)
CANCER_THERAPEUTIC_PROCEDURE	
AGENT	Agent performing the procedure (e.g., *surgeon*)
COMPLICATION	Unexpected, undesirable outcome of procedure (e.g., *nausea*)
EXTENT	Extent of the procedure, often how much of the mass is removed (e.g., *complete*)
LOCATION	Part of body procedure targets
PATIENT	Reference to the patient (e.g., *patient*, *female*)
RESULT	Result of the procedure (e.g., *successful*, *negative*)
STATUS	Procedure status (e.g., *planned*, *postoperative*)
TUMOR_DESCRIPTION	
LOCATION	Part of body tumor is located in, often ambiguous (e.g., *lymph nodes*)
MALIGNANCY	Whether the tumor is benign or malignant
MARGINSTATUS	Description of tumor margin (e.g., *superficial edge*)
METASTATIS	Whether the tumor has metastasized
PATIENT	Reference to the patient (e.g., *patient*, *female*)
QUANTITY	Some quantitative measure of the tumor
RECURRENCE	Whether the tumor has recurred
RESECTABILITY	Indicator of whether tumor is resectable
MORPHOLOGY	Morphology of tumor
SIZE	Diameter/volume of tumor, including unit (e.g., *3-4 mm*)
SIZETREND	Trend in tumor size over time (e.g., *increasing*, *decreasing*, *stable*)
STAGE	Stage number (e.g., *stage IV*)
STATUS	Tumor status (e.g., *present*, *active*)
SUBTUMOR	Link to another TUMOR_DESCRIPTION that further describes this tumor, especially if this is describing a group of tumors

Table 1: Cancer FrameNet pilot frames and their elements.

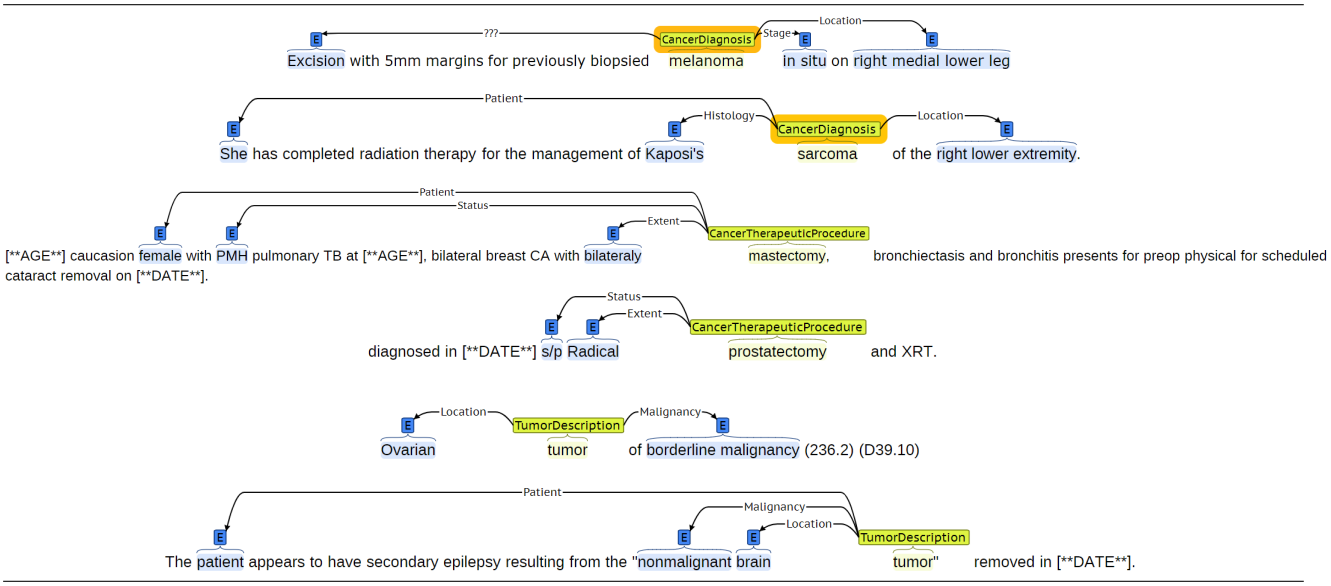

Figure 1: Example annotations.

Information Type	Frequency
Sentences	7,961
Frame Instances	7,163
Average Sentence Length	18
CANCER_DIAGNOSIS	3577
adenocarcinoma	474
cancer	419
carcinoma	495
leukemia	364
lymphoma	473
malignancy	487
melanoma	450
myeloma	191
sarcoma	200
CANCER_THERAPEUTIC_PROCEDURE	2204
colectomy	281
hysterectomy	428
lymphadenectomy	68
mastectomy	459
palliative	133
pancreatectomy	34
prostatectomy	276
radiation	470
whipple	55
TUMOR_DESCRIPTION	1382
lesion	524
mass	352
tumor	506

Table 2: Frequencies of frame instances in the corpus.

annotators to mark both PHI that was missed (thankfully a rare occurrence) and sentence boundaries that should have been removed by the process described in Section 3. Thus ERRORs indicate "sentences" that must be altered before the annotations can be considered final.

Annotation Statistics The annotation was completely performed in 90 hours, taking approximately one minute per sentence, and additional time for reconciliation. Descriptive statistics of the annotated corpus are provided in Table 2. A total of 7,961 sentences are annotated with 7,163 frame-evoking lexical units (out of a total of 8,206 candidate lexical units). Specifically, there are 3,577 CANCER_DIAGNOSIS frames, 2,204 CANCER_THERAPEUTIC_PROCEDURE frames, and 1,382 TUMOR_DESCRIPTION frames. In terms of frame elements (e.g., LOCATION, PATIENT, HISTOLOGY), CANCER_DIAGNOSIS had an average of 3.2 elements per frame instance, CANCER_THERAPEUTIC_PROCEDURE had an average of 1.7, and TUMOR_DESCRIPTION had an average of 1.1 elements. The most common elements which are shared across frames are PATIENT (2,749 instances), STATUS (2,214), LOCATION (2,190), CERTAINTY (999), and POLARITY (638). See Table 3 for more frame element details.

Annotation Agreement Inter-annotator agreement results are shown in Table 4. While observed agreement for frames–shown in Table 4(a)–is high (around 90%) for all three frames, the fairly high levels of expected agreement (73-77%) result in at best moderate κ agreement for CANCER_THERAPEUTIC_PROCEDURE and TUMOR_DESCRIPTION (0.46 and 0.58, respectively). However, κ agreement for CANCER_DIAGNOSIS is excellent (0.84). The reason for the high expected agreements is the lack of ambiguity in many of the lexical units (e.g., *adenocarcinoma* is almost always used to indicate a diagnosis), but each of the frames have at least one lexical unit that has high levels of ambiguity (e.g., *cancer*, *radiation*, and *mass*). Table 4(b) shows the F_1 agreement for frame elements. It is unsurprising to see a wide range of agreements, from the very high (e.g., RESECTABILITY, EXTENT, LOCATION) to the quite low (e.g., RESULT, COMPLICATION, AGENT). Notably, the elements with lowest agreement tend to be relatively rare in the corpus (e.g., 6 AGENTs and 71 COMPLICATIONs compared to 553 EXTENTs and 2,190 LOCATIONs), so low agreement was likely due to a lack of data for calibration.

(a) Frame Agreement

Frame	Observed Agreement	Expected Agreement	κ
CANCER_DIAGNOSIS	0.96	0.73	0.84
CANCER_THERAPEUTIC_PROCEDURE	0.88	0.77	0.46
TUMOR_DESCRIPTION	0.89	0.74	0.58

(b) Frame Element Agreement

Frame Element	Overall F_1
AGENT	0.30
CERTAINTY	0.69
COMPLICATION	0.33
DATETIME	0.69
EXTENT	0.87
FAMILYHISTORY	0.84
HISTOLOGY	0.66
LOCATION	0.82
MALIGNANCY	0.87
MARGINSTATUS	0.61
METASTASIS	0.57
MORPHOLOGY	0.56
PATIENT	0.83
POLARITY	0.77
QUANTITY	0.41
RECURRENCE	0.70
RESECTABILITY	0.88
RESULT	0.34
SIZE	0.83
SIZETREND	0.60
STAGE	0.80
STATUS	0.75

Table 4: Annotator agreement.

6 Discussion & Conclusion

In this paper, we presented Cancer FrameNet, a pilot project focuses on building a cancer-related clinical narrative resource for developing NLP systems. We introduced an annotation schema consisting of three frames (CANCER_DIAGNOSIS, CANCER_THERAPEUTIC_PROCEDURE, and TUMOR_DESCRIPTION), as well as a corpus annotated according to the schema which consists of almost eight thousand sentences. Our primary goal is to inform the developmental process for an extended Cancer FrameNet resource, with the secondary goal of informing the develop-

(a) CANCER_DIAGNOSIS elements

Lexical Unit	CERTAINTY	DATETIME	POLARITY	FAMILY HISTORY	HISTOLOGY	LOCATION	PATIENT	STAGE	QUANTITY	STATUS
adenocarcinoma	112	10	16	46	91	376	119	134	1	84
cancer	61	8	30	85	12	335	197	25	0	101
carcinoma	98	0	24	15	49	419	176	63	5	92
leukemia	37	7	15	167	6	4	96	1	0	91
lymphoma	87	16	28	56	41	77	202	22	0	142
malignancy	192	5	209	9	2	122	195	6	1	46
malignant	6	1	3	0	0	5	3	0	0	1
melanoma	55	17	43	96	19	197	181	49	9	156
myeloma	38	4	18	46	0	10	73	2	0	44
sarcoma	20	3	5	34	2	101	69	3	0	63

(b) CANCER_THERAPEUTIC_PROCEDURE

Lexical Unit	CERTAINTY	DATETIME	POLARITY	AGENT	COMPLICATION	EXTENT	LOCATION	PATIENT	RESULT	STATUS
colectomy	7	7	4	0	4	197	56	77	1	194
hysterectomy	21	14	13	1	9	53	40	183	1	237
lymphadenectomy	3	7	2	0	0	3	48	20	0	38
mastectomy	15	15	5	2	7	191	188	223	16	306
palliative	18	6	0	1	0	0	3	72	6	64
pancreatectomy	1	2	0	0	3	15	25	8	2	24
prostatectomy	9	12	5	0	34	89	65	109	6	208
radiation	25	11	90	2	10	5	117	253	3	149
whipple	3	5	1	0	4	0	2	27	2	41

(c) TUMOR_DESCRIPTION elements

Lexical Unit	CERTAINTY	DATE TIME	POLARITY	MALIG-NANCY	MARGIN STATUS	METAS-TASIS	MORPH-OLOGY	PATIENT	QUANTITY	RECUR-RENCE	RESECT-ABILITY	SIZE	SIZE TREND	STAGE	STATUS
lesion	64	5	42	20	10	9	60	145	14	6	26	51	36	9	44
mass	50	7	48	10	3	10	17	115	0	2	25	54	12	2	16
tumor	77	6	37	42	5	18	24	206	2	11	96	34	27	15	73

Table 3: Frequencies of frame elements in the corpus.

ment of further such corpora in other clinical domains. Additionally, we plan to utilize the corpus as training data for a future NLP system. However, due to the large variety of information types in cancer, and the restriction to a single institution, this corpus is of limited utility in developing complete and robust cancer information extraction methods. For example, important textual information such as post-treatment status, medication, and genomic & molecular testing results are critically important in the "precision medicine" era of cancer treatment.

Apart from the limitations related to the corpus and missing elements in the schema, another complex issue is the granularity of the frames. For example, a reasonable phenotyping task might be to find cases of cancer of any original that have metastasized to the lymph nodes. However, a common phrase used to describe biopsy results is "*with lymph node involvement*", which does not necessarily specify directly whether it is a lymph node cancer (e.g., non-Hodgkin lymphoma) or a metastasis, even if it is likely the latter. Further, encoding hypothetical phrases such as "*if tumor untreated*" presents schematic difficulties, though these could likely be overcome given sufficient data for initial exploration.

On the whole, however, frame semantics provides a flexible means of encoding important cancer information without getting bogged down in the minute details of standardizing clinical representations (a major barrier to interoperability in structured clinical data). Rules can be developed on top of the extracted frames (the "reasoning" step we describe in the Introduction) that make the necessary assumptions to utilize imperfect data. For instance, in the lymph node example, the phenotyping algorithm could exclude patients who only have a diagnosis related to the lymph node, thus likely excluding the non-metastasis cases. It is also important to recognize that, given the presence of structured data in the EHR, NLP is often seen as an (imperfect) means of supplementing (also imperfect) structured data. In this context, given that the vast majority of sentences were represented quite well using the proposed frames, this approach appears quite promising.

Another potential pitfall of using frame semantics for cancer involves the ability of organizing information into compact frames. Put another way, if almost all cancer frames shared the exact same frame elements, then a frame-based schema would be a poor fit. Instead, we found frames to work quite well in this regard. First, three frame elements were used in the CANCER_MASTER_FRAME (CERTAINTY, POLARITY, and DATETIME), but these actually generalize to just about all frames, well beyond cancer. Second, two frame elements were used in all three frames, PATIENT and STATUS. The former might ultimately be a better fit for CANCER_MASTER_FRAME, but the STATUS element has different semantics for each frame (e.g., the status of someone's cancer versus the status of the surgery to remove a tumor). Third, two elements, LOCATION and STAGE, were in two frames (CANCER_DIAGNOSIS and TUMOR_DESCRIPTION), but given the inter-relatedness of having a tumor and being diagnosed with cancer, this is not particularly surprising. Finally, beyond these cases, the remaining 16 elements were unique to a single frame, suggesting that frame-based representations are a good fit for cancer-related EHR text.

A common problem with clinical text is the mixture of structured data and true natural language prose. Commonly, structured descriptions are automatically or semi-automatically integrated with the clinical narrative. While frame semantics can handle such cases (often trivially), they are of limited value given the likely duplication of that same information in the structured part of the EHR data.

However, in this pilot project we limited ourselves to just the sentence-level scope for frame annotation, so determining better ways to handle this kind of data in future projects will be important.

On the other hand, our limitation to sentences provides a useful starting point for cross-institutional data sharing. The privacy concerns surrounding sharing of complete records largely goes away when individual sentences are manually stripped of their PHI and any potential linking information back to the original record. The ability to gather multiple institutions' data, all organized according to the same frame semantic schema, into a single FrameNet would be highly valuable to the clinical NLP community. These frames could then be mapped to existing structured clinical data standards, such as FHIR and OHDSI.

The one notable exception to sentence-based annotation, and something we would strongly consider changing in a future such project, is the use of sentences at the frame identification stage. While individual frame elements can easily be annotated at the sentence level, the word sense disambiguation task of determining whether a phrase invokes a particular frame is sometimes difficult. This is usually limited to cases where the sentence is in fact a fragment (e.g., a single item in a bulleted list), but this is a common phenomenon in EHR text and therefore worth taking careful consideration in how to handle. It may be sufficient to provide a human annotator with the extra-sentential context, but limit the machine to simply the sentence itself for classification. Alternatively, we could relax the notion of sentences in the case of fragments to provide more context. These compromise strategies could overcome one of the primary issues the annotators struggled with, while likely not having that significant an impact on the resource for training NLP systems.

7 Acknowledgements

This work was supported by the U.S. National Library of Medicine, National Institutes of Health (NIH) (awards R00 LM012104 and R01 LM01068); the National Center for Advancing Translational Sciences (NIH awards UL1 TR000371, TL1 TR000369, KL1 TR000370, UL1 TR001105); UTHealth DSRIP; the Cancer Prevention Research Institute of Texas (CPRIT) Data Science and Informatics Core for Cancer Research (RP170668) and Precision Oncology Decision Support Core (RP150535); and the Patient-Centered Outcomes Research Institute (PCORI) grants "Privacy Preserving Interactive Record Linkage (PPIRL) via Information Suppression" and CDRN-1306-04608.

8 Bibliographical References

Ashish, N., Dahm, L., and Boicey, C. (2014). University of California, Irvine-Pathology Extraction Pipeline: the pathology extraction pipeline for information extraction from pathology reports. *Health Informatics Journal*, 20(4):288–305.

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley FrameNet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*.

Baker, C. F., Schneider, N., Petruck, M. R. L., and Ellsworth, M. (2015). Getting the Roles Right: Using FrameNet in NLP. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorial Abstracts*.

Cheng, L., Zheng, J., Savova, G., and Erickson, B. (2010). Discerning tumor status from unstructured MRI reports: completeness of information in existing reports and utility of automated natural language processing. *Journal of Digital Imaging*, 23(2):119–132.

Coden, A., Savova, G., Sominsky, I., Tanenblatt, M., Masanz, J., Schuler, K., Cooper, J., Guan, W., and de Groen, P. (2009). Automatically extracting cancer disease characteristics from pathology reports into a disease knowledge representation model. *Journal of Biomedical Informatics*, 42(5):937–949.

D'Avolio, L., Litwin, M., Rogers, S., and Bui, A. (2008). Facilitating Clinical Outcomes Assessment through the Automated Identification of Quality Measures for Prostate Cancer Surgery. *Journal of the American Medical Informatics Association*, 15(3):341–348.

Denny, J., Peterson, J., Choma, N., Xu, H., Miller, R., Bastarache, L., and Peterson, N. (2010). Extracting timing and status descriptors for colonoscopy testing from electronic medical records. *Journal of the American Medical Informatics Association*, 17(4):383–388.

Dolbey, A., Ellsworth, M., and Scheffzyk, J. (2006). BioFrameNet: A Domain-specific FrameNet Extension with Links to Biomedical Ontologies. In *Proceedings of the Biomedical Ontology in Action Workshop at KR-MED*, pages 86–94.

Dolbey, A. (2009). *BioFrameNet: A FrameNet Extension to the Domain of Molecular Biology*. Ph.D. thesis, UC Berkeley.

Harkema, H., Chapman, W., Saul, M., Dellon, E., Schoen, R., and Mehrotra, A. (2010). Developing a natural language processing application for measuring the quality of colonoscopy procedures. *Journal of the American Medical Informatics Association*, 18(Suppl 1):i150–i156.

Heintzelman, N., Taylor, R., Simonsen, L., Lustig, R., Anderko, D., Haythornthwaite, J., Childs, L., and Bova, G. (2013). Longitudinal analysis of pain in patients with metastatic prostate cancer using natural language processing of medical record text. *Journal of the American Medical Informatics Association*, 20(5):898–905.

Heppin, K. F. and Gronostaj, M. T. (2014). Exploiting FrameNet for Swedish: Mismatch? *Construction and Frames*, 6(1):52–72.

Imler, T., Morea, J., Kahi, C., and Imperiale, T. (2013). Natural language processing accurately categorizes findings from colonoscopy and pathology reports. *Clinical Gastroenterology and Hepatology*, 11(6):689–694.

Lee, H.-J., Wu, Y., Zhang, Y., Xu, J., Xu, H., and Roberts, K. (2017). A hybrid approach to automatic deidentification of psychiatric notes. *Journal of Biomedi-*

cal Informatics.

Lin, L., Chen, H., and Bi, Y. (2015). The Designing and Construction of Domain-oriented Vietnamese-English-Chinese FrameNet. In *Proceedings of Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, pages 53–65.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

Martinez, D. and Li, Y. (2011). Information extraction from pathology reports in a hospital setting. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM)*, pages 1877–1882.

Martinez, D., Cavedon, L., and Pitson, G. (2013). Stability of Text Mining Techniques for Identifying Cancer Staging. In *Proceedings of the 4th Workshop on Health Document Text Mining and Information Analysis*.

McCowan, I., Moore, D., Nguyen, A., Bowman, R., Clarke, B., Duhig, E., and Fry, M. (2007). Collection of cancer stage data by classifying free-text medical reports. *Journal of the American Medical Informatics Association*, 14(6):736–745.

Miller, T. A., Finan, S., Dligach, D., and Savova, G. (2015). Robust Sentence Segmentation for Clinical Text. In *Proceedings of the AMIA Annual Symposium*, pages 112–113.

Napolitano, G., Fox, C., Middleton, R., and Connolly, D. (2010). Pattern-based information extraction from pathology reports for cancer registration. *Cancer Causes & Control*, 21(11):1887–1894.

Napolitano, G., Marshall, A., Hamilton, P., and Gavin, A. (2016). Machine learning classification of surgical pathology reports and chunk recognition for information extraction noise reduction. *Artificial Intelligence in Medicine*, 70:77–83.

Ohara, K. H. (2016). Universality of Frames: A View from Japanese FrameNet. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (Tutorial)*.

Ping, X.-O., Tseng, Y.-J., Chung, Y., Wu, Y., Hsu, C., Yang, R., Huang, G., Lai, F., and Liang, J. (2013). Information extraction for tracking liver cancer patients' statuses: from mixture of clinical narrative report types. *Telemedicine Journal and e-Health*, 19(9):704–710.

Pustejovsky, J. and Stubbs, A. (2013). *Natural Language Annotation for Machine Learning*. O'Reilly Media.

Rezhake, M. and Kuerban, A. (2015). Design of the Uyghur FrameNet Desktop. *Lecture Notes on Software Engineering*, 3(1):53–56.

Ruppenhofer, J. (2013). Anchoring Sentiment Analysis in Frame Semantics. In *Veredas*, pages 66–81.

Savova, G., Harris, M., Pakhomov, S., and Chute, C. (2005). Frame Semantics and the Domain of Functioning, Disability and Health. In *Proceedings of the AMIA Annual Symposium*, page 1006.

Savova, G. K., Tseytlin, E., Finan, S., Castine, M., Miller, T., Medvedeva, O., Harris, D., Hochheiser, H., Lin, C., Chavan, G., and Jacobson, R. S. (2017). DeepPhe: A Natural Language Processing System for Extracting Cancer Phenotypes from Clinical Records. *Cancer Research*, 77(21).

Schmidt, T. (2006). Interfacing Lexical and Ontological Information in a Multilingual Soccer FrameNet. In *Proceedings on OntoLex*, pages 75–81.

Segagni, D., Tibollo, V., Dagliati, A., Zambelli, A., Priori, S., and Bellazzi, R. (2012). An ICT infrastructure to integrate clinical and molecular data in oncology research. *BMC Bioinformatics*, 13(Suppl 4):S5.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstration Session at EACL 2012*, pages 102–107.

Stubbs, A. and Uzuner, Ö. (2015). Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus. *Journal of Biomedical Informatics*, 58:S20–S29.

Taira, R. K., Soderland, S. G., and Jakobovits, R. M. (2001). Automatic Structuring of Radiology Free-Text Reports. *Radiographics*, 21:237–245.

Torrent, T., Salomão, M. M., Campos, F., Braga, R., Matos, E., Gamonal, M., Gonçalves, J., Souza, B., Gomes, D., and Peron, S. (2014). Copa 2014 Brazil: a frame-based trilingual electronic dictionary for the Football World Cup. In *Proceedings of the International Conference on Computational Linguistics (COLING): System Demonstrations*, pages 10–14.

Vanderwende, L., Xia, F., and Yetisgen-Yildiz, M. (2013). Annotating Change of State for Clinical Events. In *Proceedings of the 1st Workshop on EVENTS: Definition, Detection, Coreference, and Representation*.

Wang, H., Zhang, W., Zeng, Q., Li, Z., Feng, K., and Liu, L. (2014). Extracting important information from Chinese Operation Notes with natural language processing methods. *Journal of Biomedical Informatics*, 48:130–136.

Wilson, R., Chapman, W., Defries, S., Becich, M., and Chapman, B. (2010). Automated ancillary cancer history classification for mesothelioma patients from free-text clinical reports. *Journal of Pathology Informatics*, 1:24.

Xu, H., Anderson, K., Grann, V. R., and Friedman, C. (2004). Facilitating cancer research using natural language processing of pathology reports. In *Studies in Health Technology and Informatics*, volume 107(Pt 1), pages 565–572.

Zhao, D. and Weng, C. (2011). Combining PubMed Knowledge and EHR Data to Develop a Weighted Bayesian Network for Pancreatic Cancer Prediction. *Journal of Biomedical Informatics*, 44(5):859–868.

Zweigenbaum, P., Grouin, C., and Lavergne, T. (2016). Supervised classification of end-of-lines in clinical text with no manual annotation. In *Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM)*, pages 80–88.