CREATING DOMAIN SPECIFIC RESOURCES FOR BUILDING SEMANTIC
ROLE LABELING SYSTEM FOR OPERATIVE NOTES

A DISSERTATION
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

YAN WANG

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

ADVISER: SERGUEI V. PAKHOMOV, PH.D.

APRIL 2015

Acknowledgements

I would like to express the deepest appreciation to both Dr. Serguei Pakhomov, my adviser, for his insightful and invaluable contributions to my academic progress, and Dr. Genevieve Melton-Meaux, my research mentor, for providing exceptional guidance, inspiration and motivation to successfully complete my dissertation.

I would like to thank my committee members, Dr. Terrence Adam and Dr. Rui Zhang for their valuable advice and comments. Also, I would like to thank Dr. Stuart Speedie (Formal committee member), Dr. Connie Delaney (IHI Acting Director), Dr. David Pieckiewicz (IHI Director of Graduate Studies), Jessica Whitcomb-Trance (IHI Academic Programs Manager) and Elizabeth Madson (IHI Administrator) for all their support.

Finally, I would like to thank the University of Minnesota affiliated Fairview Health Services for providing access to relevant patient records, and the invaluable time and efforts of medical interns, residents at the University of Minnesota who participated in my studies.

Dedication

To my husband, Tingda Lu and my son, Michael Lu, whose love and encouragement

made this thesis possible.

Abstract

Operative notes contain rich information about techniques, instruments, and materials used in surgeries. With widespread electronic health record (EHR) system adoption throughout healthcare, operative reports are increasingly accessible in electronic format and are potential information sources which may be valuable for a wide variety of secondary functions including new medical knowledge development, decision support, and clinical research. But manual review of large number of reports is time consuming and limits our ability to provide timely evidence-based guide in clinical environment. Automatic extraction of techniques, instruments, materials, and other factors surrounding operative procedures from operative notes can provide an efficient way for physicians to acquire valuable information distilled from diverse experiences reported by clinicians and decide optimal technique approach for patients.

To automate the representation and extraction of the rich information from operative notes, the goal of this research is to create domain specific resources needed for creating a semantic role labeling (SRL) system to extract information from operative notes. The coverage of existing domain-specific resources and general English resources for building a SRL system for operative notes were evaluated on a corpus obtained from the Fairview Health Services and the sublanguage used to describe surgical actions in operative notes was investigated. The results from the study show that general English resources are not adequate for building a SRL system for clinical text. Also the study shows some sublanguage characters of operative notes that can be used for parser

adaption. Next, an existing unlexicalized probabilistic context-free grammar (PCFG) parser, the Stanford PCFG parser, was adapted to clinical text for better syntactic parsing performance. Finally, domain specific predicate argument structure (PAS) frames were created for operative notes, as existing semantic frames for general English are not enough for operative notes. The domain specific resource created in this research can be used to build a SRL system for automatically extracting detailed information from operative notes.

Table of Contents

List of Tables

List of Figures

CHAPTER 1 INTRODUCTION

1.1 Surgery and Operative Notes

Surgery is the branch of medicine concerned with treatment of injuries or disorders of the body by incision or manipulation(1). It can be minor operations such as an appendectomy, hernia repair or major procedures such as a coronary artery bypass grafting or solid organ transplant. There is a wide range of surgical specialties providing treatment in all areas of the human body including the heart, brain, bones, and visceral organs. For example, for more than a decade, laparoscopic cholecystectomy has been utilized as a surgical treatment of gallbladder diseases and has been accepted as the gold standard for uncomplicated cholecystectomies(2).

As shown in a large body of research(2-8), various elements such as incision length, supplies used (e.g., mesh type and prosthetic), or parameters like pneumoperitoneum pressure can affect surgical patient outcomes. For example, P.J. O'Dwyer et al. showed(5) that of all patients in their study following open cholecystectomy, the postoperative hospital stay was significantly shorter in a 6 cm incision group than in a 15 cm incision group. The study results suggest that the surgery performed through shorter and less traumatic incisions may offer a cost-effective alternative to laparoscopic cholecystectomy. To determine the best way to perform a surgery, surgeons rely on their own clinical experience, formal training, interaction with mentors and peers, and case series reports published in the literature. These sources are either limited to small groups that with similar treatment approaches, or not feasible for

large amount of manually reviewing. New evidence-based support mechanisms and automated techniques are needed to help assist clinicians in their decision- making process to improve patient outcome.

Operative notes are documents created by surgeons in or after a procedure to document an intervention based the surgeon's recollection of the detail of the procedure. As defined in the standards set by the Joint Commission on Accreditation of Healthcare Organizations (JCAHO)(9) and the Accreditation Association for Ambulatory Health Care (AAAHC)(10), each operative note contains sections that describe the pre- and post-procedure diagnoses, name of the procedure, the detailed description of the procedure and other information. Table 1-1 enumerates the sections of operative notes.

In particular, the procedure description section within each procedure note contains detailed information on the interventional techniques, instruments, materials, and other details used to perform a procedure as in example (1) . As shown in Table 1-1 and example (1), operative notes contain detailed description of surgical procedures. Information contained in the operative notes is critical to better understanding and improving of clinical practice. But manual review of large number of reports is time consuming and limits our ability to provide timely evidence-based guide in clinical environment.

Table 1-1. Operative note sections.

Section	Description
Pre-op diagnosis	Reason for surgery
Post-op diagnosis	Actual finding at surgery
Name of procedure	
Procedure description	Description of procedure
Indication	What brought the patient to see the doctor, the patient's name, age and sex, how long the problem has been a concern and other relevant information.
Complication	Complications like injury to structure, myocardial infarction.
Anesthesia	Type i.e., general, spinal, epidural.
Surgeon	Attending physician
Assistant(s)	Resident/medical student/other surgeon/physician assistant
Estimated Blood Loss	Volume of blood loss
IV Fluids	Volume of IV fluids
Urine output	Volume of urine output
Findings	In detail what was found at surgery. Size of intraoperative pathology, adhesions and other relevant anatomy.
Pathology	What was sent to the pathologist for evaluation.
Disposition	Where patient is going from operating room (e.g. preoperative anesthesia care unit)

Automatic extraction of techniques, instruments, materials, and other factors surrounding operative procedures from operative notes can provide an efficient way for physicians to acquire valuable information distilled from diverse experiences reported by clinicians and decide optimal technique approach for patients. Applications for this include a wide variety of secondary functions including automated summarization and other clinical research. With the cumulating of large volume machine-readable operative notes, there is an increasingly demand for computational nature language processing

(NLP) applications to extract and provide necessary information from clinical narrative documents such as operative reports to help clinicians in their decision making process.

> (1) "After adequate anesthesia, the patient in the dorsal lithotomy position was prepped and draped in the usual manner. A 28 French continuous flow resectoscope sheath was inserted. Inspection showed that the patient had significant regrowth of his prostatic tissue. This patient in the past had undergone transurethral resection of the prostate elsewhere. The verumontanum and both ureteral orifices were noted to be intact. All the prostatic chips were irrigated from the bladder. A total of 46 grams of prostate was resected. Good hemostasis was obtained. A 22 French three way Foley catheter was inserted and continuous bladder irrigation was started. Sponge and needle correct X 2. The patient tolerated the procedure well."

1.2 Information Extraction (IE) and Semantic Role Labeling (SRL)

Information extraction (IE) is a prominent sub-domain of NLP used in text mining. The overall goal of IE is to extract predefined types from interested text(11). Academic research groups have largely investigated extraction of findings, problems, medications, and other items from medical reports using a range of techniques including basic pattern matching techniques or systems based on full or partial parsing(12-16). In Long's study(15), a program was developed to extract diagnoses and procedures from discharge summaries with regular expressions. Turchin et al. used regular expressions to identify

and extract instances of documented blood pressure values and anti-hypertensive treatment intensification from the text of physician notes(17). In another application, Meystre et al. developed an automated problem List system to extract problem list information from multiple free-text electronic documents(16). This system is also designed to propose the extracted problems to physicians with the official problem list.

Joined with clinical information systems, IE systems can be used to assist and improve the process of healthcare. The past decades has seen an increase of interest in using IE for surveillance of a broad range of adverse events, for enriching the content and utility of electronic health record (EHR) systems (e.g. support computerized decision-making) and supporting clinical research. Most of the research studies on IE concentrate on developing methods for processing clinical visit notes (inpatient or outpatient), radiology reports, discharge summaries, and pathology reports. Little work has been done on extracting from operative notes, though information about procedural interventions is critical to better understanding and improving many aspects of clinical practice, including interventional radiology, surgical subspecialties, cardiology, gastroenterology, oncology, and pulmonology.

IE systems often employ pattern matching that exploits basic patterns over structures such as text strings, part-of-speech tags, semantic pairs, and dictionary entries(18). Some successful IE systems are built around domain dependent relevant linguistic patterns based on select verbs (e.g. inhibit, activate for relations between bio-entities; gain, lose for the "market change" topic). These patterns are matched against

domain text for identifying and extracting interested relevant information. Although these pattern-based approaches are simple and work well for each interested question, it is difficult to extend from one domain of interest to the next since do not operate at these scales, since they focus attention on a well-defined small set of relations.

Semantic role labeling (SRL) is the task of detecting semantic roles associated with predicates, which mostly are the verbs in a sentence such as "incise", "place", or "dissect". Semantically labeled arguments in a sentence always correspond to the arguments in IE problems. In addition, semantic roles are less domain-specific than slots such as "TO AIRPORT" or "JOINT VENTURE COMPANY" used in IE system. The slot values for a given predicate is defined at the level of semantic frames of the type introduced by Fillmore(19), which describes abstract actions or relationships, along with their participants. For example, in following sentence (2) , the predicate is the verb "place". The semantic frame for "place" contains roles "placer" – who place, "thing placed" – what is placed, and "location" – where it's placed. Labeling semantic roles like above for predicates in text of interest answers the questions such as "Who", "When", "What", "Where", and "Why" and can be used for IE(20-23), question answering (QA), summarization(24-26), and other NLP tasks that required some kind of semantic interpretation. Example semantic roles include agent, patient, instrument and adjunctive arguments indicating other meaning such as locative and temporal. PropBank(27) is a project which has defined semantic roles for thousands of verbs using a corpus annotated

with semantic roles for each verb in the corpus. PropBank has defined 18 modifier roles independent of verbs and a set of core semantic roles for each of its included verb.

(2) "[**A0** We] [V placed] [**A1** a double stranded Mersilene tape] around [**A2** the coracoid]."

A0: placer

A1: thing placed

A2: location

In general, SRL can be addressed using classification (supervised machine learning). Given a predicate and each constituent in a syntactic parsed output, the task is to assign a semantic role from a pre-defined set of roles for the predicate. A typical automatic SRL design is to extract machine learning features for each constituent, train a machine learning classifier on the annotated training set and then predict the label for unlabeled constituents with the given features.

Several key components are required for building a SRL system including a deep parser and semantic frames, which play key roles in automatic SRL systems for both general English and the scientific domain. The following sections will provide some information on each of these components.

1.3 Deep Parser

Statistical deep parsers, although computationally expensive, provide important syntactic information on sentence structure for semantic interpretation(28). Full syntactic parsing of interested text provides deep linguistic features such as predicate lemma, POS tag, voice, phrase type, position and path, which perform considerably better than surface-oriented features for IE(29).

Figure 1-1 shows the parsing tree of a sentence generated by a deep parser.
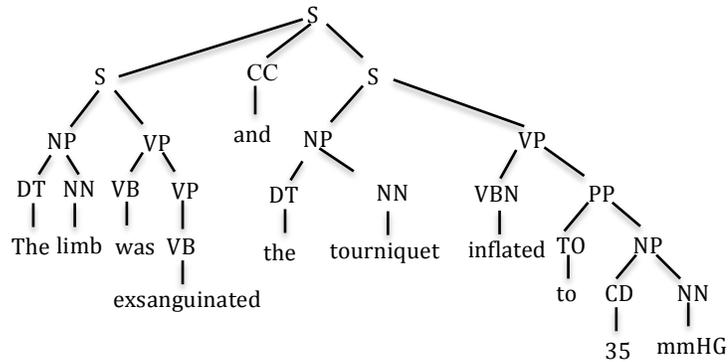
Figure 1-1. Parsing tree for "The limb was exsanguinated and tourniquet inflated to 350 mmHg".

Domain adaption

Parser performance is an unsolved issue for NLP tools in the medical domain since medical/clinical language has different characteristics than general English(30, 31). Unidentified domain lexicons, special grammar and lexical statistics all contribute to the limited performance of existing established parser like Stanford parser(32) and Link Grammar Parser(33). As a critical component to SRL, existing deep parsers like Stanford parser and OpenNLP parser have limited performance on medical text and a number of parser adapting methods were proposed in the past(31, 34-37) to address this problem:

(A) Lexicon augment - adding new entries from domain lexicons like SPECIALIST lexicon(38) in the medical and biomedical domain to existing parser through direct

expansion, heuristic mapping and use of morphological clues, etc.(31, 35, 37).

(B) POS tag disambiguation - providing POS tag information for domain lexical elements to the parser to avoid inconsistencies between domain POS tag and parser lexicon POS tag(34).

(C) Statistics adjustment – adjusting the statistics used by parsers. Existing parsers were trained on general English copra like the *Wall Street Journal*. The frequencies associated with each of the various syntactical categories for each token, used by a statistical parser to generate the most likely parsing, are different from a specific domain text(37). For this reason, it is expected that better parsing will be obtained through adjusting of syntactical category statistics for important domain lexical elements like verbs in procedure description and other lexical elements that have unusual usage.

In this research work, we modified the Stanford parser lexicon with more accurate statistics of the SPECIALST lexicon from our clinical corpus to improve the parsing performance for operative notes. In addition to extending the lexicon of Stanford unlexicalized parser with new entries in SPECIALIST lexicon that occurred in our operative notes corpus, we also modified the parser grammar.

1.4 Semantic Frames

Traditionally, most work for IE focused on surface level patterns(39, 40), which can be learned from annotated text or hand built. Such approaches seem to be unable to achieve high completeness and accuracy for IE due to the complexities of language. As the

following examples shows, the same task in a procedure can be narrated with a number of distinctive combinations of verbs, nominal and voice.

(3) "Subsequent *curettement* of the bone edge was performed."

(4) "Gentle curetting was done."

(5) "The canal of the humerus was carefully curetted."

(6) "We curetted the sockets."

(7) "Gentle, sharp curettage was performed."

Due to these linguistic complexities, syntactic roles and semantic roles are often necessary to extract information from narrative text. Several computational linguistics projects, such as PropBank(27), FrameNet(41), and VerbNet(42), have been developed to provide semantic frames for predicates. Tasks that requiring semantic processing such as SRL, QA and text categorization have largely benefited from these resources(29, 43, 44).

The Berkley FrameNet project is an online resource for general English semantics that has been frequently utilized in SRL(45-47). It provides frame-semantic descriptions of several thousand English lexical items and with these descriptions includes semantically annotated attestations from contemporary English corpora(41). Currently a total of 12423 distinct lexical predicates, or target words are annotated: 5075 verbs, 4768 nouns, and 2205 adjectives. Table 1-2 shows an example of FrameNet Frame for the concept "placing".

The PropBank is a corpus that is annotated with verbal propositions and their arguments. Compared with FrameNet, it commits to annotating only verbs as predicates

and the syntactic constituents that form the semantic arguments of the verbs. In PropBank, semantic roles are divided into two classes: core roles (A0, A1, A2, A3, A4, A5), whose semantic value is defined by the predicate syntactic frame, and adjunct roles (ArgM e.g., AM-TMP, AM-LOC) which are a closed set of semantic labels accounting for predicate aspects such as temporal, locative, manner and purpose. Table 1-3 shows a PropBank frame for the verb "irrigate".

Table 1-2. FrameNet frame for "place".

Frame:	Placing
Definition:	Generally without overall (translational) motion, an Agent places a Theme at a location, the Goal, which is profiled.
Lexical Units:	*place.v, placement.n, plant.v, plunge.v, pocket.v, position.v, pot.v, put.v, rest.v, etc.*
Core Arguments:	Agent, Cause, Goal, Theme
Examples 1:	"The waiter PLACED the food on the table."
Examples 2:	"David PLACED his briefcase on the floor."

Table 1-3. PropBank frame for "irrigate".

Verb:	Irrigate
Arguments:	
	Arg0: Provider of water
	Arg1: Recipient of water
Examples:	
	"Tolek Alterman had returned from the colonies in Palestine and, before the national leadership, exalted the miracles of drying up swamps and irrigating the desert."

The VerbNet project maps PropBank verb types to their corresponding Levin classes. It is the largest on-line verb lexicon currently available for English that incorporates both semantic and syntactic information about its contents. In VerbNet, verbs are organized into extended verb classes through refinement and addition of subclasses. Each verb class is completely described by thematic roles, selection restrictions on the arguments, frames consisting of a syntactic description and semantic predicates with a temporal function as exemplified in Table 1-4.

Table 1-4. VerbNet frame for "put".

Verb Group:	Put-9.1-2
Verb Members:	Place, Put, Set, Stick
Arguments:	
	Agent
	Theme
	Destination
Syntax:	Agent V (on upon) Destination Theme
Examples:	"They put upon me a brilliant, red helm."
Semantics:	Motion (during(E), Theme) not (Prep(start(E), Theme, Destination))
	Prep(end(E), Theme, Destination) cause(Agent, E)

As presented above, FrameNet, PropBank and VerbNet are the three large semantic resources for general English. Among the three, FrameNet and PropBank are the most frequently used for SRL in general English(46, 48-50) (51-53)and the scientific domain(54-57). As presented in the work of Zapirain et al.(58), the PropBank role set is more robust due to the lack of verb–specific semantic information and generalizes better to infrequent and unseen predicates compared with VerbNet. In Wattarujeekrit et al.'s

work(59), the authors showed that in contrast to VerbNet and FrameNet, PropBank defines verb-specific PAS frames for each distinct verb. Those Frames are more likely to contain detailed specifications of arguments than are possible for verb groupings as in VerbNet. In addition, analyzing semantic frames in a verb-specific manner would help to discover rules for mapping from surface syntactic structures to underlying semantic propositions.

Domain adaption

Past research on SRL system for scientific domain demonstrated that significant difference exists on both predicate sense and syntactic construction between domain text and general English(57, 59-61). In a scientific domain the predicates may have one of three properties:

1) Have the same senses with the general English.
2) Have same sense but different semantic roles from general English.
3) Have a sense that not exists in general English.

Senses for each predicates can be collected by a survey on the usage of predicates on several sample sentences(59) from the domain text or with the help of an automated approach(62). For the predicates that have same senses and semantic roles as in the general English, existing frames can be directly borrowed. For the predicates that have the same sense but has different semantic roles, new frames can be built based on modification of the existing frame. For example, the frame for verb "irrigate" in Propbank has two roles "Provider of water" and "Recipient of water" as shown in Table

13

3. When used to describe an action in operative notes, the verb is often modified by following roles exemplified by examples. A new frame as shown in Table 1-5 can be created by modifying the existing frame for the predicate.

(a) Material used to irrigate like "antibiotic saline", "amphotericin solution".

(8) "The wound was irrigated with **25% Betadine solution**."

(b) The path to irrigate.

(9) "Fluorescein was irrigated through the nasolacrimal duct system."

(10) "Fluorescein was then irrigated through the superior punctum."

For predicates that have a sense that does not exist in general English, such as "exsanguinate", new frames must be build from scratch.

Table 1-5. New frame for "irrigate".

Verb:	Irrigate
Arguments:	
Arg0:	Provider of liquid material
Arg1:	Recipient of liquid material
Arg2:	Path
Arg3:	Material used to irrigate
Examples:	
	Fluorescein was irrigated with 25% Betadine solution through the nasolacrimal duct system.

In this work, we will choose the PropBank frame sets as a starting point. The PropBank frame sets are freely available from the PropBank website. For each frequently occurring verb in procedure descriptions, we will gather a set of sample sentences to analyze the semantic senses and roles of each sense of the verb and compare the verb

arguments with the Propbank frame of the same verb in operative notes. In PropBank a verb may have more than one frame if the verb has more than one sense. For example, the verb 'put' has 3 senses as shown in following examples:

(11) "Wilder has *put* the onus on Cole."

(12) "Let's *put* it bluntly."

(13) "The pilots *put* the amount as high as $100 million."

Different senses for a verb reflect the fundamental assumption that syntactic frames are directly related to the underlying semantics. It is expected that in our work some verbs will have the same arguments as in PropBank. For these verbs we will borrow the PropBank frames.

1.5 Significance

Evidence-based practice is one of the most important underlying principles in modern health care as quality requires the use of best evidence. Health professionals are becoming more accountable within clinical governance structures for the care they provide. With large amount of health data have been collected in electronic format over the past several decades as the nation's hospitals and health systems increasingly implement EHRs, automated computational approaches to exploit patient data from EHR documents represent an important opportunity to synthesize the massive clinical reports and provide comprehensible representation of clinical practices. Automated IE system based on SRL is a promising computational technology for answering medical questions such as "what instruments", "what condition", "how to act".

15

A large body of research work has been conducted for automatic SRL system in general English including semantic resources, annotations, algorithms and deep parsing techniques, etc. While state-of-the-art SRL systems perform well in general English, they are not capable of dealing with the special languages used in scientific domains. Thus, many researchers have committed to explore the sublanguage variation, domain adaption, new semantic resources including semantic frames and annotations. In medical domain, little effort has been put on automatic SRL system for actions in operative notes.

This work would be the first work on automatic SRL system for IE of surgical information by adapting existing PCFG parser for operative notes and creating domain specific PropBank style semantic frames.

1.6 Specific Aims

While the long-term goal is to understand how much and what types of information is expressed in surgical reports through predications and how to formalize and automate the representation and extraction of this information with NLP. Towards this long-term goal, short-term goal is to fill the current gap in existing resources required to build an SRL system are not good enough for clinical text in operative notes. Domain specific resources are needed to create a SRL system for extracting information from operative notes. Towards these goals, I propose to address the following 3 specific aims:

1) Evaluate the coverage of existing domain-specific and general English resources for building a semantic role labeling system for operative notes and investigate the sublanguage used to describe surgical actions in operative reports.

2) Adapt an existing unlexicalized probabilistic context-free grammar (PCFG) parser, the Stanford PCFG parser, to clinical text for better syntactic parsing performance.

3) Create domain specific predicate structure (PAS) frames for operative notes as existing semantic frames for general English are not enough for operative notes. The created frames will be used to create training corpus and golden standards, which is an important component for building an automatic SRL system.

CHAPTER 2 A STUDY OF ACTIONS IN OPERATIVE NOTES

2.1 Background

In linguistics, besides verbs as productive predicates, nominalization, gerunds, and relative nouns are also used to express predicative relations and can take the same arguments as the corresponding verbs(63). Several syntactic structures have been described for action descriptions, as depicted in Table 2-1. Part of the motivation of this study was to understand the language (i.e., syntactic structures) of action sentences utilized for depicting actions, including the use of 'activity verbs' to show when a nominal, indefinite verb, or gerund introduces an action.

Table 2-1. Action description forms.

Form	Examples	Activity verbs
Action verb	1. The medial edge of the cleft was *incised* sharply with a knife. 2. It was *incised* just above the level of the bladder flap.	
Activity verb + gerund	1. The scope was removed and the *curetting* was *performed.* 2. We then *did* a *lengthening* of the lateral aspect of the quad approximately 5 mm.	perform, carry out, apply, carry, do, fashion, begin, undertake, continue
Activity verb + verbal nominals	1. The *incision* was *carried* through the subcutaneous tissue. 2. We *carried* the *dissection* down through dartos muscle. 3. A sagittal split *incision* and subperiosteal dissection was *accomplished*.	perform, carry out, apply, carry, do, fashion, undertake, begin, continue, achieve, gain, get, obtain, provide, etc.
Activity verb + Indefinite verb	1. We *began* to *lift* the gland up and away from essential anatomy. 2. A rongeur was *used* to *remove* the hyaline cartilage.	begin, continue, use, etc
Activity verb + deverbal nominals	1. A box *cut* was *made* to substitute for the PCL. 2. I *made* an *incision* paralleling the acromioclavicular joint	make, create, develop, etc
Activity verb + deverbal nominals	1. General *anesthesia* was *administered*. 2. *Dressings* were *applied*, drapes removed.	administer, apply, etc

The same action can be narrated with different verb combinations, nominals and voice, as exemplified below. Understanding this better has practical importance in the construction of NLP systems to process those notes.

(14) "Subsequent *curettement* of the bone edge was performed."

(15) "Gentle *curetting* was done."

(16) "The canal of the humerus was carefully *curetted*."

(17) "We *curetted* the sockets."

(18) "Gentle, sharp *curettage* was performed."

Traditionally, most work with medical IE has focused on surface level patterns(39, 40), which can be learned from annotated text or hand built. Such approaches seem to be unable to achieve high completeness and accuracy for IE due to the complexities of language. As shown in examples in Table 2-1 and examples (14) - (18), verbs are subject to syntactic variation and nominalization, which can be used to describe the same event. As a consequence, a wide range of syntactic patterns could potentially express the same operation action.

Despite ongoing progress of clinical IE methodologies, there has been realization that resources and NLP tools, which may perform well on text from one source, may fail to perform well on text from a new domain or source. A number of researchers have explored the linguistic differences between different sublanguages associated with clinical and biomedical domains(64-67). In 2002, Friedman et al.(64) surveyed features of sublanguages, documented two biomedical-domain sub-languages (clinical reports and

molecular biology) and discussed the similarities and differences between them. Lippincott et al.(65) showed that rich variation exists across a variety of linguistic dimensions (lexical, syntactic, sentential and discourse features). The authors also support an awareness of such variation when deploying NLP systems for use in single or multiple subdomains. Kilicoglu et al.(66) explored the task of interpretation of nominalizations and developed a set of linguistic generalizations for effective interpretation of a wide range of patterns used to express arguments of nominalization in clinically-oriented biomedical text.

Despite research that has looked at the general topic of sublanguages, limited work has been done examining the sublanguage of surgical procedures. In this study, we aimed to investigate the surface patterns of action descriptions, the action predicates, and distribution of different predicates usage. We also aimed to evaluate the adequacy of existing domain-specific and general English resources to extract action information from procedure descriptions. Finally, we also offer the top action predicates along with the mapping information as a knowledge resource.

2.2 Methods

A total of 362,310 operation narratives obtained from University of Minnesota-affiliated Fairview Health Services, with data from 4 metropolitan hospitals in the Twin Cities including both community and tertiary-referral settings were used for this study. The corpus includes operative reports created by 2,300 surgeons with 4,333 different procedure types defined by Current Procedural Terminology (CPT) codes.

2.2.1 Automated Section Extraction with Locally Defined Header Hierarchy

From the data repository, the 'procedure description' section was first extracted from each note. Most operative notes organize their document content into sections and sub-sections such as 'Procedure description', 'Pre-operative diagnosis', and 'Anesthesia' and are specific to the type of the note (e.g., an Admission note has sections corresponding to a standard history and physical examination). Typically, each section will have a section header string that includes words that provide context for the encapsulated text. For example, a section with a header string 'Procedure description' provides detailed and step-by-step description of a surgery. The text with these sections provides important information about surgeries.

While clinical notes typically organized into sections, clinicians often label the sections with frequently used but non-standardized terms based on use of acronyms, abbreviations or synonyms. Review of clinical notes shows that procedure description section could be labeled as 'Procedure details', 'technique procedure', 'OP report', 'incisions', 'case details' etc. Sections can have subsections, such as 'HEENT', 'vital', etc. in 'physical examination' section. Occasionally, a section, like the 'procedure description' section, may not be associated with a section header. In this case, a human reader needs to infer the existence of a specific section by semantic content of the text. Automated extraction of sections from clinic notes is challenging. In this study, we examined the section headers and sections in a subset of operative notes and developed a NLP tool to extract procedure sections.

21

Potential sections headers were extracted from the data repository using a random set of 3,000 operative notes. One of two surgeons (GM and NB) reviewed the 300 most used section headers along with headers from relevant note templates and grouped them into a hierarchy of headers. We developed a tool based on this hierarchy to extract the description section by combining features such as header string matching, header format pattern, section length, and section-specific terms. An evaluation of 200 operative notes with 1,594 sections demonstrated an accuracy of 95% for correct extraction of 'procedure description' section text.

2.2.2 Sentence Categorization

All sentences within 10 random operative notes 'procedure description' sections were reviewed to categorize sentences, revealing that sentences could be classified into three categories based on the semantic content of the event described (Table 2-2). In our dataset, most sentences fall into the action category.

Table 2-2. Sentence Categories.

Category	Examples
Perception/Report	1. I could feel no full thickness tear. Visualized no full thickness tear. 2. Sponge and needle counts were reported as correct. 3. There appeared to be a simple cyst within.
Action	1. We placed a double stranded Mersilene tape around the coracoid. 2. A box cut was made to substitute for the PCL. 3. We continued mobilization up to the hepatic flexure.
Other	1. This array of components allowed for full extension with minimal recurvatum and easy flexion. 2. She wanted to proceed with the right knee.

2.2.3 Categorization of Actions, Expansion of Nominals, and Distributions in Operative Notes

Parsing results of the adapted Stanford parser for 'procedure description' text were used to collect the most frequently used verbs. For each parsed sentence, the top-level main verb of each sentence was collected based on the syntactic tree. A random set of 50 notes (964 sentences) was used to evaluate the accuracy of the approach. A trained linguist and an informatitist annotated the main verbs of each sentence for the entire evaluation set (JR, YW). Kappa statistic indicates reasonable inter-rater agreement (0.78) and proportion agreement (0.94). The approach demonstrated a recall of 90.2% for detecting main verbs from all 13,095 tokens in the evaluation set.

For the entire set of 'procedure description' sections, verbs, including phrasal verbs, and their frequency were collected. We focused on verbs providing coverage for over 92% of the corpus. From this, verbs were categorized into action verbs, activity verbs, and verbs with a perception/report or other non-action (Table 2-2). Since gerunds and other nominals derived from an action verb are also used to describe actions, potential nominals of each verb were collected through automatic and manual approaches from existing resources including the SPECIALIST lexicon, the WordNet lexicon,(68) New Oxford American Dictionary(1), and Stedman's Medical Dictionary(69). From this, the incidence of verbs and their nominals used to describe actions were collected from the overall corpus.

2.2.4 Semantic and Domain Knowledge Resource Coverage Evaluation

Since semantic resources derived from general English, lexical resources, and domain-knowledge play important roles in IE, the adequacy of existing resources to facilitate the usage of verb predicates and their nominals was evaluated with the UMLS, SPECIALIST Lexicon, WordNet, and FrameNet.

In the biomedical and clinical domains, the UMLS Metathesaurus is a large, multi-purpose database built from over 100 disparate terminology sources in patient care, health services billing, public health statistics, and biomedicine. It is designed to support a broad range of biomedical research and includes rich information. For example, the UMLS concept '[C0677554] Anastomosis – action' has a semantic type 'Therapeutic or Preventive Procedure' and the entry provides detailed definition of the action from several sources like 'CHV/PT | surgical connection between two hollow organs'. The SPECIALIST lexicon includes the syntactic, morphological, and orthographic information for each lexicon term and is, as previously described, a resource for improving the performance of NLP tasks.

WordNet and FrameNet are two notable general English semantic resources repeatedly used in biomedical and clinical research. WordNet is a repository of hierarchically organized English words that are organized into sets of synonymous terms (verbs, nouns, adjectives, and adverbs), called synsets, each of which represents one lexical concept. The database contains about 150,000 lexical items organized in over 115,000 synsets. The Berkley FrameNet project is an online resource for general English

semantics. As introduced in the background section, it is an essential lexical semantic resource providing predicate frames that can aid in natural language understanding.

2.3 Results

Application of the Stanford parser results on the 'procedure description' section demonstrated that the 200 most frequent top-level verbs in the entire corpus covered 92% of all verbs in the 'procedure description' section. To test the coverage of the verbs selected in several related surgical domains, Prostatectomy, Colectomy, and Total Abdominal Hysterectomy datasets were created, each with 1,000 randomly selected operative notes with corresponding CPT codes. These datasets demonstrated 89%, 90%, and 92% coverage of verbs with the top 200 verbs for the entire corpus, respectively. Each verb was individually examined and 147 verbs were classified as action verbs, while 15 were activity verbs. Table 3 shows a partial list of these action verbs:

Table 2-3. Action verb examples.

place	drape	bring	dissect
take	close	Identify	open
remove	give	divide	tolerate
Irrigate	insert	close	undergo
prep	tie	cauterize	transect

Using the SPECIALIST lexicon, WordNet, New Oxford American Dictionary, and Stedman's Medical Dictionary, a total of 97 unique nominals (median 0, range (0-2)) were extracted, several of which are listed in Table 2-4.

Table 2-4. Action verbs and their nominals.

Verb	Nominals
anaesthetiz	anaesthetization,
anastomose	anastomosis
approximat	approximation
cannulate	cannulation
curette	curettage, curettement
drain	drainage
debride	debridement
expose	exposure
withdraw	withdrawal

Table 2-5 shows the distribution of verbs, gerunds, and nominals used to describe actions. As shown in Table 2-5, physicians tend to use verbs to describe actions and prefer using a passive voice, as in: *"A computer plan was developed for placement of 75 palladium-103 seeds"*. The predominant use of the passive voice was also true for nominal action predicates.

Table 2-5. Predicate distributions.

Predicate form	Total	Passive voice	Active voice
Verb	3,808,845 (94,4%)	3,306,300 (86.8%)	502,545 (13.2%)
Gerund	13,425 (0.3%)	12,820 (95.3%)	605 (4.7%)
Nominal	211,102 (5.2%)	184,509 (87.4%)	26,593 (12.6%)

Table 2-6 shows the distribution of several top-, middle-, and low-incidence actions and nominals of each action verb. Most actions are expressed using verb predicates with the exception of '*incision*' and '*dissection*', which were commonly described with the pattern of 'verb + nominal'.

Table 2-6. Usage of verbs, gerunds, and nominals to describe surgical actions.

Action	Total action mentions	Categorized action mentions			Nominals
		Verb	Gerund	Nominals	
place	431,576	430,871 (99.84%)	0	703 (0.16%)	placement
close	260,450	253,243 (97.23%)	14 (0.01%)	7,193 (2.76%)	closure
drape	176,439	174,883 (99.12%)	13 (0.01%)	1,543 (0.87%)	drape
take	167,126	167,125 (100.00%)	0	0	-
prep	165,522	164,459 (99.36%)	21 (0.01%)	1,427 (0.86%)	prep
incise	163,007	37,032 (22.72%)	0	125,973 (77.28%)	incision
remove	156,487	156,078 (99.74%)	0	408 (0.26%)	removal
bring	129,445	129,444 (100.00%)	0	0	-
irrigate	92,689	90,171 (97.28%)	0	2517 (2.72%)	irrigation
dissect	82.450	52,185 (63.2%)	5	32,260 (36.7%)	dissection

As summarized in Figure 2-1, which shows coverage of the top 147 actions, the SPECIALIST Lexicon had very good coverage of both verb predicates (89.9%) and nominal predicates (100%), although it missed some phrasal verbs (e.g., '*bring back*', '*dissect out*', '*carry down*'). WordNet also had good coverage for predicates, specifically 89.9% for verbs and 93.8% for nominals. Since it is a resource addressing general English, WordNet missed some domain-specific terms like '*curette*', '*exsanguinate*', '*extubate*', and '*free up*'. The UMLS Metathesaurus, which contains important domain knowledge, covered only 11.5% of action verb predicates and 58.8% of nominal predicates. As a semantic resource, the FrameNet also had poor coverage of nominals (36.1%) and fair coverage for verbs (64.2%).
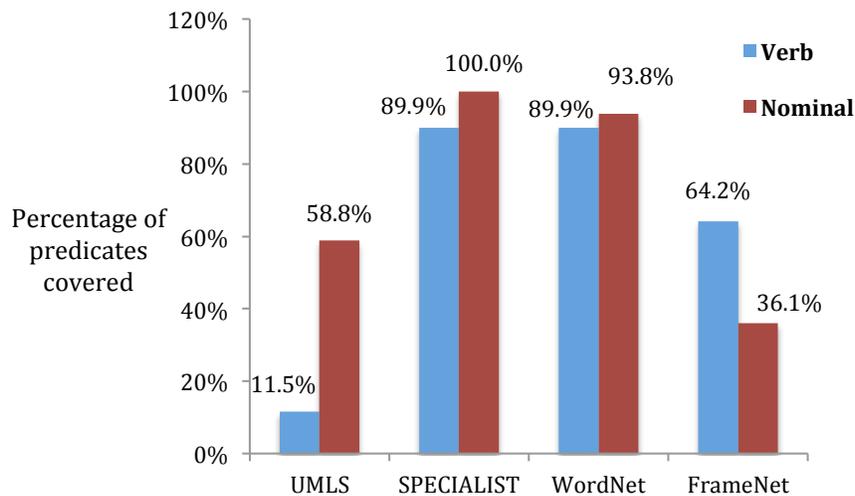
Figure 2-1. Coverage of operative note verbs and nominals by semantic resources.

2.4 Discussion

Information contained in operative notes is critical to better understanding and improving surgical clinical practice and potentially has many secondary uses for surgical research and quality improvement. Traditionally, many of the decisions made by surgeons about how to optimally perform a particular surgical procedure are made on the basis of the clinician's clinical experience, opinions from colleagues, or available case series reports. These information sources are often limited to small groups, and, unfortunately, randomized controlled trials in surgery are rare and difficult to conduct due to ethical and financial barriers. Likewise, manual review of large numbers of operative reports is not a scalable solution. With the accumulation of large volumes of machine-readable operative notes, there is an opportunity for developing tailored clinical NLP methods to extract and provide necessary information from these narratives. The features specific to the surgical domain sublanguage in operative notes have important implications for the development

28

of effective computerized NLP systems for operative note IE. In this work, we studied and characterized the 'procedure description' sublanguage for surgical actions to lay the groundwork for accurate and effective IE from operative notes. Specifically, we studied prevalent predicates, including 147 action verbs and 97 action nominals. We believe that the set is broad enough (92% coverage of verbs in our operative note repository) to support IE from operative notes and deep enough to deal with the syntactic variability that exists in the sublanguage.

Overall, actions within operative notes in the 'procedure description' section were mostly verb predicates along with associated semantic arguments. Nominal predicates were uncommonly used with some notable exceptions, like 'incision' and 'dissection'. Also, a great majority of action-verb predicates (86.8%) were found to be in the passive voice. Similarly, verb nominals, which occurred less often (5.2%), were also predominantly (87.4%) in the passive voice. Only a very small portion of actions was described by gerund predicates. Our coverage evaluation demonstrated that the SPECIALIST lexicon had entries for all the nominal predicates and 89.9% of the verbs, with the exception of some phrasal verbs (e.g., '*free up*'). As a general English resource, WordNet misses some medical terms (e.g., '*extubate*'). Despite this, WordNet still covered 93.8% of the nominal predicates and 89.9% of the verbs.

Interestingly, the domain resource (UMLS) and the semantic resource (FrameNet) showed unsatisfactory coverage to both verb predicates and nominal predicates. Incorporated with different types of biomedical vocabularies, the UMLS encompasses

terms and codes in a wide range of categories including diagnosis, procedures, disease, anatomy, drugs, genetics, nursing and others. In the 2011AB version, the UMLS includes 215,327 Therapeutic or Preventive Procedure concepts, 31,826 Diagnosis Procedure concepts, 8,175 other Health Care Activity concepts and 451 Daily or Recreational Activity concepts. It is somewhat unexpected that such a large vocabulary covered only 11.5% percent of the action verbs. The evaluation of the mapping results shows that all the phrasal verbs like '*take down*' or '*free up*' were not covered by the UMLS. Also, a large number of prevalent and domain specific verbs such as '*incise*', '*expose*' and '*close*' were also not defined. Nominal predicates, on the other hand, had fair coverage (58.8%) by the UMLS. Since the UMLS provides linkage to biomedical terminologies and FrameNet had potential for semantic processing with frames, improvement and expansion of both resources for the surgical domain is a needed step in future system development.

Besides actions within the main clauses of sentences (e.g., '*The 20-French rigid cystoscope with blade was removed and an attempt was made to place the 24-French rigid resectoscope*'), phrases also contain actions as with following four examples:

(19) "The patient was taken to the operating room ***where general anesthetic was administered***",

(20) "***After the successful induction of spinal anesthesia***, she was placed supine on the operating table",

(21) "***Prior to removing the trocar***, cystoscopy was again performed".

30

(22) "An attempt was made *to place the 24-French rigid resectoscope"*.

Although these phrases were not systematically analyzed in this study, we did observe that actions expressed in phrases tended to have fewer semantic arguments compared with the actions described in a main clause. Additionally, in some cases phrases can be used to describe an event that may or may not be an actual action performed in a procedure. For example in (22) , it is difficult to determine if the action was performed or not. Due to the large syntactic variability of the sentence structures of these phrases, in this work we focused on the verbs in the main clauses. However, we realize that for many NLP tasks or applications, such as procedure summarization, it will be critical to effectively extract these actions as well.

One important discussion point surrounds our use of the Stanford parser and its augmentation with the SPECIALIST lexicon. As we presented before, the analysis of actions in this work was based on the deep parsing output of 'procedure description' text with the Stanford parser expanded with the addition of the SPECIALIST lexicon. Since the Stanford parser was trained on a general English corpus, the parser's grammar statistics are collected from a much different text than the operative report text that we are interested in. Consequently, the adapted parser may not be capable of resolving many of the complex or unusual sentences found in the 'procedure description' section. It is also possible that better parsing accuracy can be achieved by retraining the parser on an annotated corpus from the medical domain. However, we found that the parsing output

from the current adapted Stanford parser showed good parsing accuracy on 'procedure description' text in this study.

Examination of the most prevalent action predicates and their usage in operative reports also gave certain insights into knowledge sources for frame semantics. Analysis of operative note predicates revealed that existing resources are not fully adequate for effective IE from operative notes. This is an important consideration for future work that could build upon semantic frames in operative note summarization. Our results also indicate that further work may be needed for creating new frames and adapting existing frames, as the frame resource (FrameNet) had significant coverage gaps to action predicates. Moreover, in operative notes some predicates are used for a different meaning than in general English. For example, the phrasal verb '*come across*' means '*meet*' in general English, but in the following example, the phrasal verb means '*go through*'.

(23) "We ***came across*** the liver parenchyma using the Helix device."

This example is one that demonstrates the need for large annotated corpora for both semantic frame generation and also for the related and subsequent semantic role labeling process. Besides expansion and adaption of current frames and operative note annotations, we anticipate needing to build robust algorithms to define how to transform the relevant constituents of a surface sentence to the semantic arguments in frames. To facilitate sophisticated text mining applications, a lexicon that describes real, observed usage of predicates and other domain terms in operative notes and a domain knowledge resource that provide domain information are also required.

The overall work of this study gives insight into the language used by surgeons to communicate action events in the operating room. This study provides an understanding of the relative variability of action expressions. The action verbs, their nominals, and mappings are available to other researchers on request. Our next step is to work towards development of new frames and extension of existing frames in a pilot study to assess the feasibility of this as a methodology for operative note IE related to surgical techniques.

CHAPTER 3 DOMAIN ADAPTION OF PARSING FOR OPERATIVE NOTES

3.1 Background

3.1.1 Unlexicalized Parsing and Lexicalized Parsing

Full syntactic parsing results in a hierarchical tree-like representation of the syntactic structure of a piece of text according to some formal grammar such as, for example, a constituency grammar(70). Figure 3-1 shows the constituency parse tree of the sentence: "The eye was patched with hyoscine ophthalmic drops."
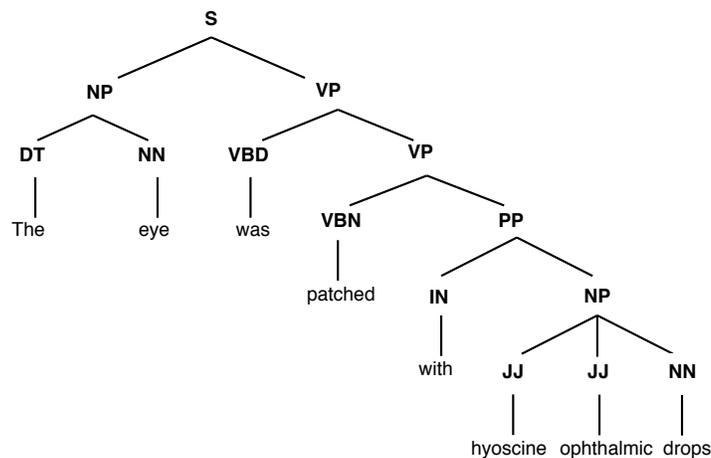
Figure 3-1. Constituent (phrase structure) tree for the sentence: "The eye was patched with hyoscine ophthalmic drops." *S: Sentence; NP: Noun phrase; VP: Verb phrase; DT: Determiner; NN: Noun, singular or mass; VBD: Verb, past tense; IN: Preposition or subordinating conjunction; JJ: Adjective; VP: Verb phrase.

As shown in Figure 3-1, the tree representation of the input sentence from a parser conveys useful information such as the constituent boundaries, the grammatical relationship between constituents, which is expressed by the path from one constituent to another, the head word of each candidate constituent and a number of other features.

34

In formal linguistics, Context Free Grammars (71) (CFG) are formal systems used to model natural language. CFGs contain a set of production rules (or recursive rewrite rules) that are used to generate linguistic expressions from underlying constituent building blocks. Formally, a CFG is represented as a 4-tuple consisting of 4 sets: G = (N, Σ, R, S) where:

N is a finite set of non-terminal symbols.

Σ is a finite set of terminal symbols.

R is a finite set of rules of the form $X \rightarrow Y_1 Y_2 ... Y_n$, where $X \in N$, $n \geq 0$, and $Y_i \in (N \cup \Sigma)$ for $i = 1...n$.

$S \in N$ is a distinguished start symbol.

For an input sequence of words, a parse tree can be derived according to the CFG production rules. Figure 3-2 exemplifies a set of simple production rules. For an input sentence 'The patient left the OR', a parse tree can be derived from the production rules as shown below in Figure 3-2.
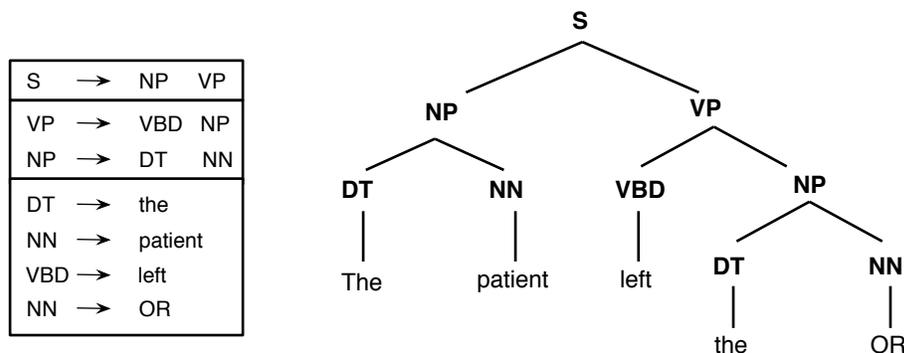
Figure 3-2 Production rules example. *S: Sentence; NP: Noun phrase; VP: Verb phrase; DT: Determiner; NN: Noun, singular or mass; VBD: Verb, past tense; OR=Operating room.

35

When dealing with complex natural language text, more than one production rule may apply to a sequence of words, which results in syntactic ambiguity. Figure 3-3 shows two syntactic trees derived for the same sentence "The I&A removed the viscoelastic with a tip….".
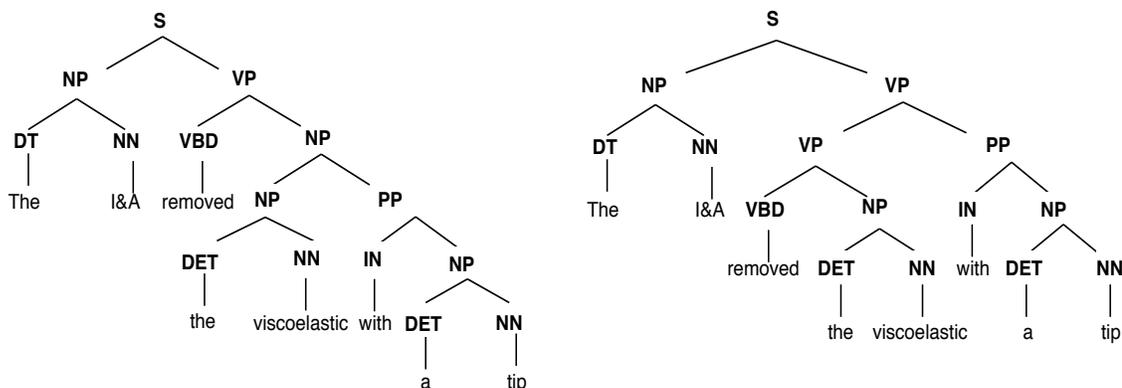
Figure 3-3. Two syntactic trees for the sentence: 'The I&A removed the viscoelastic with a tip.' *I&A=Irrigation and aspiration.

The sentences in Figure 3-3 illustrate the classic phenomenon of prepositional attachment ambiguity where the interpretation of the sentence depends on whether the prepositional phrase "with a tip" attaches to the verb phrase node "removed …" or the lower noun phrase node "the viscoelastic."

Probabilistic context-free grammars (PCFGs) are an attempt to deal with this ambiguity encountered when applying CFG production rules on complex natural language text. Thus, PCFG is a probabilistic version of CFG where each production has a probability, as shown in Figure 3-4. In PCFG, the probability of a parse tree is the product of the probabilities of its re-write rules productions. The parse tree with the greatest probability will be picked from a number of alternatives with varying

likelihoods. Probabilities of a PCFG model are typically estimated from a set of training texts (e.g., Penn Treebank (72)). Formally, a PCFG is defined as follows:

A context-free grammar G = (N, Σ, R, S)

Parameters $q(\alpha \to \beta)$, which is the conditional probability of choosing rule $\alpha \to \beta$

Given a PCFG with all parameters estimated from a corpus such as the Penn Treebank, a parse tree for a sentence s is chosen from all possible alternative parse tress by finding the parse tree with maximum likelihood:

$$\arg\max_{t \in T(s)} p(t)$$

Here t is a parse tree for s; $T(s)$ is a set of all possible parse trees for sentence s; $p(t)$ is the probability of parse tree t calculated based on parameters collected from corpus. Out-of-the-box and unenhanced PCFGs usually do not perform optimally on text from new domains (73). Unlexicalized PCFGs with special linguistic annotations (74) and lexicalized PCFGs are two approaches that have been used to address the weaknesses of basic PCFGs.
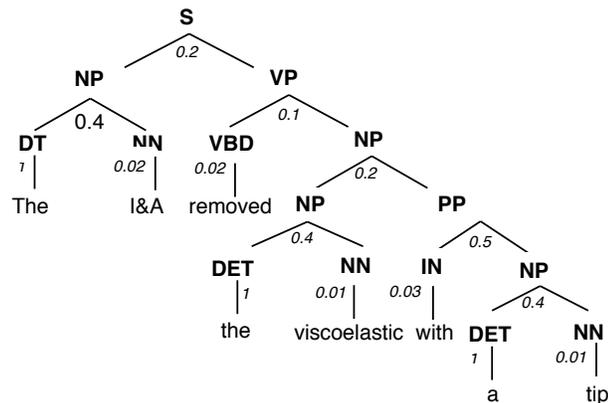
Figure 3-4. A syntactic tree with production probabilities for sentence 'The I&A removed the viscoelastic with a tip.' *I&A=Irrigation and aspiration.

Klein and Manning utilized a set of linguistic annotations to construct an unlexicalized PCFG parser using the probabilities associated with different syntactic categories to include vertical and horizontal history of tree nodes (74). For example, the UNARY-INTERNAL annotation was used to mark any nonterminal node in Penn Treebank with only one child. Similarly, the TAG-PA annotation is used to mark all preterminals with their parent category as shown in Figure 3-5. As shown in Klein's work, the TAG-PA annotation significantly improves parsing accuracy (74). Here, the unlexicalized Stanford PCFG parser was trained on the Penn Treebank corpus and enriched with additional annotations and achieved similar performance to the start-of-the-art lexicalized PCFG parser without relying heavily upon lexical dependencies.
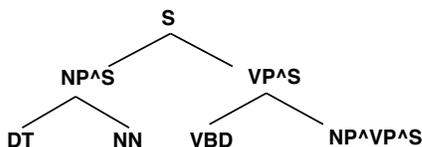
Figure 3-5. Adding parent annotation to trees

The lexicon of an unlexicalized PCFGs parser trained on treebanks with the additional annotations, as a result, stores not only lexical entries, but also the statistics that a lexical is associated with an POS tag as well as the parent tag such as "NN^NP" - a noun with a noun phrase as parent and "VBN^ADJP" - a past participle verb with an adjective phrase as parent. The grammars of an unlexicalized PCFG parser also incorporate these additional annotations. For example, a unary rule "NP^S-U -> PRN^NP" that specifies that the node has only one child. One advantage of using the

unlexicalized Stanford parser is that the text format of the lexicon and grammar can be easily extended and reloaded into original parser.

A lexicalized PCFG specializes its production rules for specific words by including their head-word in the trees as shown in Figure 3-6. In this way, a lexicalized PCFG largely resolves ambiguities such as the prepositional phrase (PP) attachment problem. Additionally, Collins(75) and Charniak(76) used a discriminative re-ranking technique to obtain better parse from a list of parses generated from original parsers for each sentence. However, the performance of lexicalized PCFGs is limited by the sparseness of lexical dependency information available in Penn Treebank. Also, modeling word-to-word dependencies is difficult, especially if these dependencies are domain-specific.
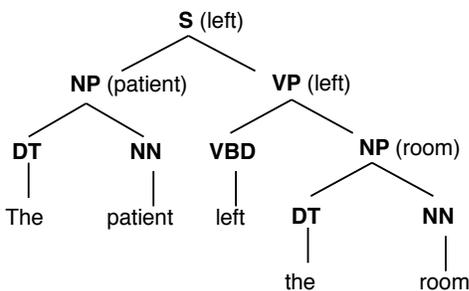
Figure 3-6. Adding headtags to trees

3.1.2 Domain Adaption for Unlexicalized Parsing and Lexicalized Parsing

A number of groups have reported and evaluated methods to improve parsing performance of existing unlexicalized parsers. Xu and colleagues (77) reported that the use of POS tags from manual annotation could be used to produce a POS tagger for the medical domain with improved Stanford parser performance of between 2 to 4% with a

small set of sentences from clinical reports. The evaluation of these enhancements revealed an improvement on the high level NLP task of noun phrase identification. Similarly, Huang et al.(37) enriched the Stanford lexicon with unambiguous entries in the SPECIALIST lexicon and customized the Stanford parser grammar based on the review of clinical reports although no formal evaluation of these modifications was performed.

We observed from preliminary experiments on clinical text particularly with operative reports that sometimes even with correct POS tags, general English parsers were not able to give correct parse tree. Figure 3-7 shows parse trees of a POS tagged sentence (24) produced by the Stanford parser with and without enriched lexicons. Parse tree (b) is produced by the original Stanford parser with correct POS tagging provided (a) is a parse tree produced by the enriched Stanford parser.

(24) "The/DT wound/NN was/VBD extended/VBN proximally/RB and/CC distally/RB."

```
(S
  (NP (DT The) (NN wound))
  (VP (VBD was)
    (VP (VBN extended)
      (ADVP (RB proximally)
      (CC and)
      (ADVP (RB distally)))))
```

```
(S
  (NP (DT The)
    (NP
      (NP
        (NP (NN wound))
        (SBAR
          (S
            (VP (VBD was)
              (VP (VBN extended)
                (ADVP (RB proximally))))
            (NP (CC and))))
      (ADVP (RB distally)))))
```

(a). Parse tree of sentence 1 from enriched Stanford parser

(b). Parse tree of sentence 1 from original Stanford parser with pre-tagged input

Figure 3-7. Parse trees of a POS tagged sentence (1) produced by Stanford parser (a) with and (b) without enriched lexicon.

Self-training is a technique used to adapt a lexicalized parser to a new target domain. It creates a new parser by training an existing parser with data parsed by the

existing parser as extra training data(78, 79). As shown in McClosky's work(79), the parser is re-trained with the new training data set, which includes large in-domain corpus that parsed with original parser. While some early reports on self-training for parsing reported negative results, McClosky (78, 79) and Bacchiani (80) have shown that this technique can improve parsing performance of the new parser on a target domain. In McClosky's work, the standard Charniak/Johnson parser was trained on a corpus of biomedical abstracts that were labeled with the existing Charniak/Johnson parser, along with Penn-Treebank. The resulted new parser showed performance improvement on a standard test set, the GENIA Treebank (81).

3.1.3 GENIA Corpus

GENIA corpus is a collection of articles on biological reactions of transcription factors in human blood cells. The articles are extracted from MEDLINE database with the MeSH terms, human, blood cell and transcription factor. Each article was annotated with parse trees following the Penn Treebank II (PTB) bracketing guidelines. The following text in Figure 3-8 shows an example of GENIA syntactic annotation.

```
<sentence id="S2"><cons cat="S"><cons cat="NP" id="i2" role="SBJ"><cons cat="NP"><tok cat="NN">Resistance</tok>
    </cons> <cons cat="PP"><tok cat="TO">to</tok> <cons cat="NP"><tok cat="NN">glucocorticoid</tok> <tok
    cat="NN">therapy</tok></cons></cons></cons> <cons cat="VP" syn="COOD"><cons cat="VP"><tok cat="VBZ">has</
    tok> <cons cat="VP"><tok cat="VBN">been</tok> <cons cat="VP"><tok cat="VBN">observed</tok> <cons cat="NP"
    ref="i2" null="NONE"></cons><cons cat="PP"><tok cat="IN">in</tok> <cons cat="NP"><cons cat="NP"><tok
    cat="NNS">patients</tok></cons> <cons cat="PP"><tok cat="IN">with</tok> <cons cat="NP"><tok cat="JJ">
    autoimmune/inflammatory</tok> <tok cat="NNS">diseases</tok></cons></cons></cons></cons></cons></
    cons> <tok cat="CC">and</tok> <cons cat="VP"><tok cat="MD">may</tok> <cons cat="VP"><tok cat="VB">be</tok>
    <cons cat="VP"><tok cat="JJ">related</tok> <cons cat="NP" ref="i2" null="NONE"></cons><cons cat="PP"><tok
    cat="TO">to</tok> <cons cat="NP"><cons cat="NP"><tok cat="DT">the</tok> <tok cat="JJ">inflammatory</tok>
    <tok cat="NN">process</tok></cons> <cons cat="NP"><tok cat="PRP">itself</tok></cons></cons></cons></cons></
    cons></cons></cons><tok cat="PERIOD">.</tok></cons>
</sentence>
```

Figure 3-8. GENIA syntactic annotation example.

3.1.4 SPECIALIST lexicon

The SPECIALIST Lexicon consists of a set of lexical entries including multi-word terms with spelling variants, part(s) of speech, and other information for biomedical domain terms. SPECIALIST consists of over 200,000 biomedical terms, as well as common English words. It has been successfully used to adapt parsers for general English to the biomedical domain as it contains important syntactic, morphological, and orthographic information for each entry (31, 35, 37). For instance, a lexical record for a term in SPECIALIST contains base forms of the term, the part-of-speech, a unified identifier, spelling variants, and inflection for nouns, verbs and adjectives. As presented in our previous work (82), the SPECIALIST lexicon has very good coverage of both verb predicates (89.9%) and nominal predicates (100%) occurring in operative notes. Table 3-1 shows the number of entries of four important POS categories in SPECIALIST lexicon and Stanford lexicon, demonstrating that the SPECIALIST lexicon contains many more word entries than the Stanford lexicon.

Table 3-1. Entries of 4 POS categories in SPECIALIST lexicon and Stanford lexicon.

POS category	SPECIALIST	Stanford
Verb	56859	8477
Noun	280482	27832
Adjective	90884	9032
Adverb	12467	1422

In the clinical domain, only a small amount of research has focused on parser adaption for clinical text, with previous work not focusing on operative notes. Therefore in this paper we will describe our experiments on adapting the Stanford parser for the

clinical text of operative reports. We hypothesized that the addition of more accurate statistics from our clinical corpus of operative reports and use of the SPECIALST lexicon could improve the parsing performance of the Stanford parser for operative notes. We extended the lexicon of Stanford unlexicalized parser with new entries in SPECIALIST lexicon that occurred in our operative notes corpus and modified the parser grammar. We also tested the performance of parsers augmented with statistics collected from corpus POS tagged with two start-of-art POS taggers, GENIA tagger and Medpost tagger.

3.2 Methods

Figure 3-9 provides an overview of this study. Overall, we enriched the Stanford lexicon with SPECIALIST lexicon and with statistics collected from POS-tagged operative reports from our clinical note repository and customized the Stanford grammar to the special syntactic structure of operative report text. The resulting enhanced Stanford parser output was then evaluated and compared with POS-tagged corpus with different POS taggers using a set of manually annotated operative report sentences.

3.2.1 Dataset and Overview

A total of 362,310 operative reports from University of Minnesota-affiliated Fairview Health Services in the Twin Cities including both community and tertiary-referral settings were used for this study. The corpus includes operative reports created by 2,300 surgeons with 4,333 different procedure types defined by Current Procedural Terminology (CPT) codes. The procedure description was extracted from each note and split into sentences with a locally developed heuristically-based text-processing tool (See
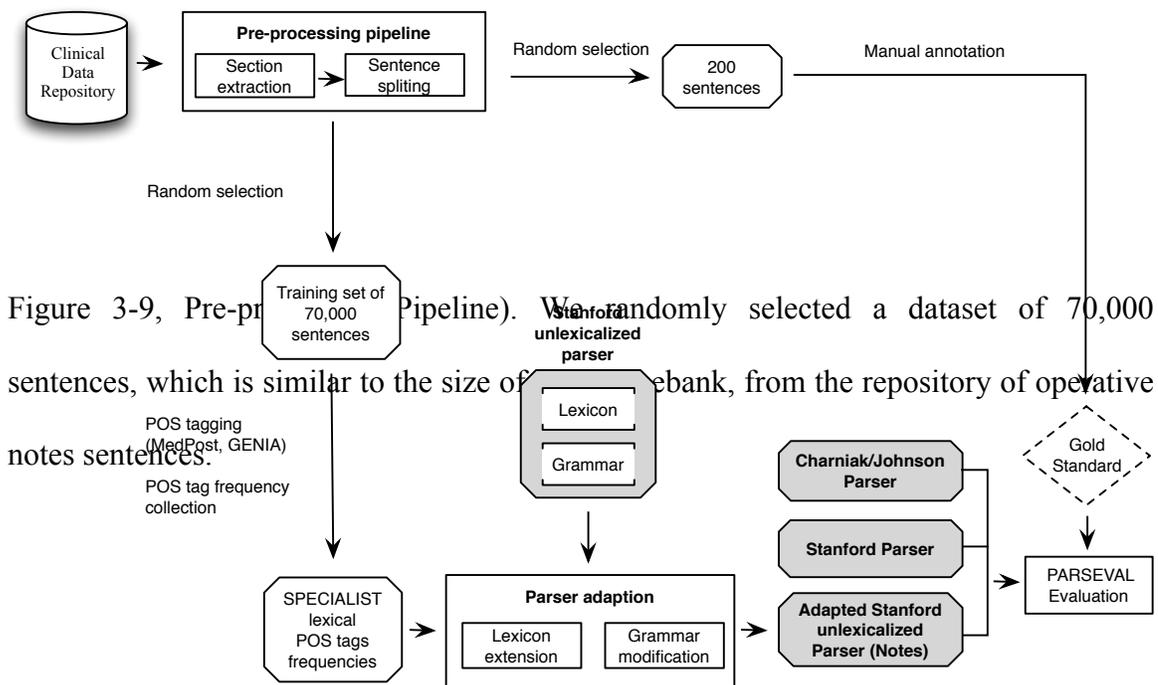
Figure 3-9, Pre-processing Pipeline). We randomly selected a dataset of 70,000 sentences, which is similar to the size of the treebank, from the repository of operative notes sentences.

44

Figure 3-9. Overview of operative notes parser adaption.

3.2.2 Stanford Unlexicalized PCFG Parser Adaption for Operative Notes

The SPECIALIST Lexicon contains far more entries than the Stanford lexicon as shown previously in Table 3-1. To selectively expand the Stanford lexicon for operative notes, we added only SPECIALIST Lexicon entries (single word entries in this study) contained within the overall operative note corpus. This approach was taken since words that were not within the operative note corpus do not have associated frequency statistics and also to decrease the associated computational overhead encountered with loading the parser and parsing the text associated with adding a large lexicon.

In adding entries to the Stanford Lexicon, we had to take into account that the SPECIALIST Lexicon uses a set of syntactic categories that are different from the Penn Treebank tags for its entries. For unambiguous entries in the SPECIALIST lexicon, the same set of mapping rules used in Huang's work (37) were used to convert the SPECIALIST Lexicon syntactic categories into Penn Treebank tags. For ambiguous entries in the SPECIALIST lexicon, we converted those entries with multiple syntactic categories (about 20,000 words) into Stanford entries using statistics collected from the

tagged corpus combined with several heuristic rules. As introduced above, an unlexicalized PCFG model requires statistics for usage of each POS tag under different parent for parsing. For instance, the word "callus" can be both a noun and a verb. To collect frequencies for the tags of each word, we first created a corpus with a similar size to the Penn Treebank from 70,000 randomly selected sentences in the operative note "procedure description" section text. Heuristic rules based on the Stanford lexicon were also used, where we observed that some parents for a particular POS tag were more frequent than others. Using adjectives as an example, in the Stanford lexicon the incidence of adjectives (68,090 in total) used within an adjective phrase (11,498) or a noun phrase (54,211) was significantly greater than other phrase types. The sentence set was then tagged using the five Stanford POS taggers. For example, in the Stanford lexicon, frequencies for each POS tag with a different parent for the word "inject" are given in Table 3-2. To decide the frequency distribution of each possible parent, we collected the frequency from POS tagged sentences.

Table 3-2. Frequency of each POS tag of word "inject" with different parents in the Stanford lexicon.

POS tag	Parent tag	Frequency
VBD	VP	2
VBN	VP	2
JJ	ADJP	7
JJ	NP	2
JJ	WHADJP	1
JJ	WHNP	1
JJ	UCP	1
JJ	QP	1

We also observed that for some POS tag and parent combinations, only one or a few specific words were associated. For example, the word "only" in sentence (25)   is the only adjective word that could be used in a conjunction phrase:

(25) "The biceps tendon, long head intra-articular portion, was not only

split, but remarkably frayed."

Thus, for each POS tag such as "JJ", "NN" and "VBD", we defined a heuristic parent distribution for it and split the collected frequency based on these distributions. For example, for POS tag "JJS" (superlative adjective), we define a distribution as shown in Figure 3-10. From each POS tagged corpus, the frequency of POS tags associated with each SPECIALIST lexicon entry within the set of 70,000 sentences was collected and used to adapt the Stanford lexicon and create a new adapted lexicon. For example, the new Stanford lexicon extended with the MedPost lexicon contained 172,636 entries while the original Stanford lexicon had 101,703 entries.

Using our previous observation that physicians tend to use passive voice to narrate the procedure description section (82), we manually adjusted the frequencies of VBD (verb, past tense) and VBN (verb, past participle) tags for verb entries that could be both a past tense verb and a past participle. Also, the POS tag of some verbs, such as "appeared", "tolerated" and "revealed", can be either VBD and VBN in the SPECIALIST lexicon, but after review of a random set of sentences with these words, we found that the POS tags of these verbs were mostly VBD as opposed to other verbs such as "incised" and "dissected" which tended to mostly be used in text as VBN. To assign frequencies

47

that could better reflect actual usage of verbs, we used the 200 verbs previously reported

that covers 92% of all verbs from operative notes to help provide reasonable frequencies

of potential ambiguous POS tags.
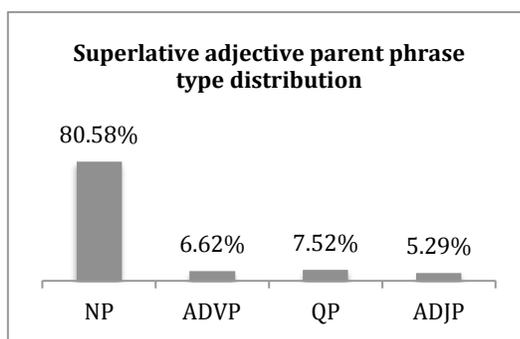
Figure 3-10. Parent phrase type distribution of the POS tag superlative adjective.

Finally, we were able to omit auxiliary verbs as this was another feature

previously observed in the sublanguage of operative notes. For example, in following

sentences (26) and (27) , the auxiliary verb "was" is omitted in the operative note text.

(26) "A transverse incision was made in the popliteal fossa and the lesser

saphenous vein identified, ligated proximally."

(27) "Good hemostasis obtained."

Syntactical information such as the voice of verbs is also critical for many NLP

tasks such as semantic role labeling. To address this problem in operative notes, we

modified the grammar of the Stanford parser by including more productions rules. For

example, given sentence "Good hemostasis obtained" original sentence will give a parse

tree as (28) . After adding a new rule "VP^S-VBF-v -> VBN^VP", the parser gives

correct parse as (29) . The new parse assigns correct phrase tags and POS tags for the

verbs, which are very important to NLP tasks such as semantic role labeling (28). As shown in Gildea's work the phrase tags and POS tags are used to extract voice and parse tree path for semantic role calculating.

(28) (ROOT (S (NP (NNP Gelfoam)) (VP (VP (VBD applied)) (CC and)
(ADVP (RB hemostasis)) (VP (VBD confirmed)))))

(29) (ROOT (S (NP (NP (NNP Gelfoam)) (VP (VP (VBN applied)) (CC
and) (VP (NN hemostasis)))) (VP (VBN confirmed))))

3.3 Evaluation

To evaluate the performance of parsers adapted from the corpus POS-tagged using different POS taggers, we created a reference standard with 200 manually annotated parse trees of randomly selected operative notes sentences. The reference standard parse trees were annotated by two separate annotators with both a linguistics and informatics background and experience in clinical NLP. Annotations followed the Penn Treebank II Bracketing guidelines (83). To compare parse results of adapted parsers with the parse trees produced by the Charniak/Johnson parser, parse trees generated by the original Stanford parser and parse trees generated by the original Stanford parser with POS tags from MedPost were examined. In addition, we tested the performance of the parser on a random set of GENIA parse trees. Since the GENIA corpus is from a slightly different domain, we wanted to evaluate the same technique for parser adaption on this domain.

Parsing performance was evaluated following the PARSEVAL standards (84) for parsing accuracy evaluation. Each constituent in the parse was represented as a labeled

49

span. A constituent is counted as correct only if the label and text span is correct. Given two parses, the precision and recall of constituents were calculated. Precision and recall can be formally defined in terms of the number of true positive (TP), false positive (FP) and false negatives (FN) as below. F-score is the weighted harmonic mean of precision and recall. Syntactic annotations from two annotators for the same evaluation set of a 10% sample of the full evaluation set were compared and the proportion agreement of annotations was computed at the sentence level.

To evaluate the significance of parsing performance differences between the parsers, a pair-wise Wilcoxon Signed-Rank test with Bonferroni adjustment was conducted on the F-scores of the parsers evaluated on the test set sentences. As the F-score differences between parsers severely deviated from a Gaussian distribution, the Wilcoxon Signed-Rank test was used for a statistical evaluation since it does not require a normal distribution of differences between data pairs as required by the pair-wise t-test.

To evaluate the proposed parser adaption technique, a similar approach to the parser adaption for operative notes was used to adapt Stanford unlexicalized PCFG parser for the GENIA corpus. We used 14,325 training trees from the GENIA Treebank as a training corpus and collected statistics from it. Since we did not have enough biology domain knowledge, the words that occurred in GENIA were simply ported into the Stanford unlexicalized PCFG parser lexicon. Since GENIA trees have parent labels for each word, we tested our approach with two sets of lexicons, one with the accurate parent statistics and the other one with parent statistics generated from heuristics rules. We

removed old entries in the original Stanford lexicon when the entry exists in the GENIA corpus. A simple grammar was added into the Stanford lexicon for testing resulting in about 129,600 entries for the new parser.

3.4 Results

The inter-rater agreement between the two annotators for the syntactic tree annotation task was 85%. For most sentences, the two annotators agreed on all the phrase tags and POS tags in the syntactic tree. In the six sentences where the annotators did not agree, there were minor differences in annotations in three sentences and major differences in three sentences. The three sentences with major differences in annotations tended to be complex sentences such as the following sentence: "Following induction of general anesthesia, intubation with a bronchial blocker, positioning in the right lateral decubitus position, the left chest was prepped and draped and a total, ultimately, of 3 port incisions were made.".

The precision, recall, and f-score means for each of the parsers evaluated are summarized in Table 3-3 for parsers adapted for operative notes and in Table 3-4 for those adapted for the GENIA corpus. As shown in Table 3-3, at baseline, the Charniak/Johnson parser had slightly better parsing performance for operative notes compared to the Stanford parser. The expansion of the lexicon yielded moderate improvement in parsing performance. Grammar modification combined with statistics adjustment also resulted additional performance gain. The f-score of the final adapted Stanford parser on the operative notes test set improved from 87.64% to 89.90%.

The pair-wise t Wilcoxon Signed-Rank Test with Bonferroni adjustment for F-scores of the parsers shows parsing performance improvement of the best-adapted parser to the baseline parser (Stanford unlexicalized parser) (p-value < 0.001).

Table 3-3. Evaluation results of parser adaption for operative notes.

Evaluation of parser adaption for operative notes

Parser	Precision	Recall	F-score
Baseline (Stanford unlexicalized parser)	87.54%	87.74%	87.64%
Charniak/Johnson	88.43%	88.46%	88.45%
Adapted Stanford unlexicalized parser (New grammars)	87.73%	87.94%	87.83%
Adapted Stanford unlexicalized parser (Lexicon expansion)	88.82%	89.28%	89.04%
Adapted Stanford unlexicalized parser (New grammars + lexicon expansion)	89.27%	89.84%	89.55%
Adapted Stanford unlexicalized parser (New grammars + lexicon expansion + statistics adjustment)	89.65%	90.13%	89.90%

Table 3-4 shows the performance of the parser adapted on the GENIA corpus, when apply same technique on GENIA corpus, the parsing result of adapted parser on the GENIA test set improved from 75.78% to 79.59% with parent distribution from rules and to 81.25% with parent distribution collected from GENIA Treebank annotations.

Table 3-4. Evaluation results of parser adaption for GENIA.

Evaluation of parser adaption for GENIA corpus

Parser	Precision	Recall	F-score

Baseline (Stanford unlexicalized parser)	78.18%	73.52%	75.78%
Adapted Stanford unlexicalized parser (New lexicon with parent statistics by rules and new grammar)	82.92%	76.52%	79.59%
Adapted Stanford unlexicalized parser (New lexicon with actual parent statistics and new grammar)	84.08%	78.60%	81.25%

3.5 Discussion

Full syntactic parsing of text provides deep linguistic information (e.g. voice, phrase type) useful for many NLP tasks. Parsers developed for general English text have benefited from a large tree bank and training corpus (e.g, Penn Treebank) and have achieved high parsing performance. Clinical documents are known to have special sub-language features (e.g. domain vocabulary, telegraphic text, special grammar), which often require adaptation of general English NLP tools. Parsers often have a decrement in performance when applied to scientific texts (85). Domain NLP experts have investigated methods to adapt parsers trained on general English to new target domains (31, 34, 35, 78, 85-88). However, these approaches have been attempted to only a limited extent in some types of clinical texts. In this work, we investigated the adaptation of a general unlexicalized PCFG parser to a specific type of clinical text - operative reports using tag statistics collected from operative reports and other sublanguage features of operative notes. We applied the approach on two different domains, clinical operative notes and the GENIA corpus. The results show that this approach can improve parsing performance on both domains. Though an increase of 2.26% of the parsing performance on operative notes is not large in absolute performance, this improvement is still noteworthy as the

baseline performance of the unlexicalized PCFG parser very was good at operative notes. As shown in our results, domain adaptation was helpful in improving parser performance further. We plan to incorporate the adapted parser into our NLP system, the biomedical information collection and understanding system (BioMedICUS(89)).

To compare our results with previously work on parser domain adaption, we applied our approach on the GENIA corpus, which is a public available corpus. Our evaluations show that the performance of the new parser adapted to GENIA corpus is close to the state of the art parser performance 80.7% without parser training using domain parse trees (90), which requires a large annotated corpus and is not feasible for parser adaption in most cases.

To extend the Stanford parser lexicon, we incorporated only the SPECIALIST entries that existed in our corpus. Another option to consider with future enhancements would be to add all tokens in the operative notes corpus, which would not limit us to the ones contained in the SPECIALIST lexicon. We observed that out of all the tokens in our corpus, about 75% of them were contained in the SPECIALIST lexicon. Some tokens in our corpus are not counted as in SPECIALIST lexicon because that the first letters of these words are capitalized since that the Stanford unlexicalized PCFG parser treat upper cased word and lower cased word differently. Of all of the tokens not in the SPECIALIST lexicon, a large portion of them (about 85%) were nouns. Since the Stanford parser treats unknown words as nouns by default, we chose to ignore these tokens. However, we did include adjective and adverb tokens, which are in our corpus

but not in the SPECIALIST lexicon because of capitalization of the first letter when these words appear at the beginning of a sentence. In this study, only single words entries in SPECIALIST lexicon were incorporated into the Stanford lexicon. More research and experiments will needed to incorporate multiword entries in future study.

In this work, we used a set of heuristic rules to specify the parent distribution of each entry depend on the POS tag of the token as shown in section 3.2. As shown in Table 3-4, when use real parent phrase tag distribution collected from GENIA tree bank, the adapted parser performance improved another 1.68%. However, real parent phrase tag distribution is not always available for other domain such as the clinical text. To acquire a better estimation of the statistics on parent distribution, some features such as the POS tag of the word before and after the interested word may help to decide the parent phrase tag. More work will be needed to analysis the algorithm for parent distribution in the future. When tested the new unlexicalized PCFG parser adapted with clinical text on GENIA tree bank, as we expected, we found no performance improvement. As the GENIA corpus is a domain with very different sublanguage features, the statistics of GENIA text have differences from clinical text.

Since the Stanford PCFG parser is unlexicalized, no head word information is incorporated in the associated production rules. Thus, we observed that the adapted Stanford parser was unable to solve the prepositional phrase (PP) attachment ambiguity, which an issue often observed in general English. In the text for procedure description, we observed that the average sentence length (86 characters) is less than that of the Wall

Street Journal sentences (126 characters). As shown in the example procedure description in the introduction, surgeons tend to describe actions, which occurred during a procedure using short and simple sentences. Thus, the ambiguity is potentially less of a problem in operative notes than in general English and other clinical texts.

In addition, since procedures in operative notes are usually described with short and simple sentences, the parsing performance of regular parsers is better than that of some other types of clinical text such as the corpus presented in Xu's work (77). Other areas where we might consider further study include increasing the parse tree training set, which we purposefully did not do here with the goal of enhancing the parser with corpus statistics and other sublanguage characteristics. Subjectively, the overall parsing performance improvement observed with these enhancements was good despite the small magnitude of increase observed since the baseline performance of the unenhanced Stanford parser was fairly high. Furthermore, the magnitude of increase in performance accuracy found in this study is consistent with that found in other similar studies of parser adaptation (34, 35, 77, 86).

While the operative notes dataset is relatively small and is a limitation of the study, the dataset is unique in nature and labor intensive to create. Other publicly available labeled clinical corpora for research contain few operative notes, such as the MiPACQ (91) corpus which contains only one operative note. We also evaluated our parser adaption technique using the GENIA tree bank for biology text and observed similar results. As additional publically available tree banks are become available, it

would be value to perform other parallel, independent evaluations to test if this approach is more generalizable.

In placing this study in the overall context of clinical NLP, we only concentrated on the clinical text for the procedure description of operative notes. Additional work will be needed to determine if the approach used here with operative reports will be generalizable to other types of clinical texts such as discharge summaries and radiology reports. These approaches may require a good understanding and consideration of other unique syntactic structures and language features seen in clinical documents, such as the irregular sentence structures observed in Xu's work (77). We suspect that by including additional grammars for irregular structures into the Stanford parser and extending the parser lexicon to the lexicon specific to those texts that the performance of the Stanford parser can similarly be improved on other clinical text in an analogous manner.

CHAPTER 4 PREDICATE ARGUMENT STRUCTURE FRAMES FOR MODELING INFORMATION IN OPERATIVE NOTES

4.1 Background

The language used in medical reports can be quite distinct from general English(64, 92, 93). In addition to sublanguage features (e.g., the distinct set of domain terms, omission of information), the set of action verbs in operative notes is quite different from general English. Existing semantic resources have limited coverage of the action verbs that frequently occur in operative notes. FrameNet covers only two-thirds of the most frequently used action verbs in operative notes(82), while PropBank covers approximately 85% of these verbs. Moreover, action verbs in operative notes often take different semantic arguments from general English and may have additional special meanings. Thus, existing semantic PAS resources like PropBank may not reflect the true usage of action verbs in operative notes.

The objective of this study was to create PropBank style PAS frames for a set of frequently occurring action verbs used in a subset of operative notes and then to confirm or expand upon the PAS frames with a separate set of operative notes as a foundational task in building an IE system to automatically extract clinically-relevant surgical elements from operative notes.

4.2 Methods

4.2.1 Dataset

The documents used in this study were gathered from 362,310 operative notes of University of Minnesota-affiliated Fairview Health Services. From this data repository, we randomly selected 3,000 Laparoscopic Cholecystectomy (ICD-9-CM Procedure code 51.23) notes based on the procedure code associated with each note and a keyword search of the procedure name. This dataset was used to study and create PropBank style PAS frames.

To examine the completeness of the PAS frames generated from operative notes of laparoscopic cholecystectomy, we created six evaluation datasets from our data repository from gastrointestinal, specifically colorectal, surgical procedures (Figure 4-1). The first evaluation dataset consists of 3,000 randomly selected operative notes of a wide range of colorectal surgeries(including all procedures with ICD-9-CM Procedure Code 45.7, 45.8 and 45.9). From this dataset, 20 sample sentences from notes of 17 colorectal procedures were randomly selected for each verb.. A second group of five datasets from all the operative notes for Lower Anterior Resection (ICD-9-CM Procedure Code 48.63), Right Colectomy (ICD-9-CM Procedure Code 45.73), Left colectomy (ICD-9-CM Procedure Code 45.75), Transverse Colectomy (ICD-9-CM Procedure Code 45.74) and Sigmoid Colectomy (ICD-9-CM Procedure Code 45.76) was created along with 10 random sample sentences for each verb in the verb list.

4.2.2 Pre-processing of Datasets

Datasets were initially pre-processed to generate deep parse trees for sentences in the procedure descriptions in the following manner. First, the procedure description section within each operative note was extracted with a locally developed rule-based NLP tool. Each section was then split into sentences with a sentence splitter. The Stanford parser, which was enriched with a biomedical lexicon – the SPECIALIST lexicon(38), was used to generate deep parses for sentences within the procedure description sections. All verbs were collected based on the verb specific Penn Treebank(72) tags such as VBD, VBN, and VB, the former two indicating the past tense and past particle tense of verbs respectively. The collected verbs were normalized to the base form with lexical variant generation (LVG) - a SPECIALIST lexical tool. Deep parse trees were also later used for sample selection.

4.2.3 Selection of Predicates and Samples

Our main interest in this study was to examine verbs either commonly used in operative notes or verbs denoting certain important surgical actions. Twenty action verbs were selected from a verb list encompassing the most frequently occurring verbs(82). In addition, a surgeon (GM) handpicked 10 additional surgery-specific verbs such as "suture", "aspirate", and "clip". For each selected verb, a set of sample sentences was collected from the deep parse trees. To determine a good sample size for verbs, we analyzed the PropBank annotation statistics on the *Wall Street Journal*. In PropBank annotation, each verb has about 26 instances on average. Because of the high data quality

requirement of medical NLP applications, we chose a sample size of 40 for each verb thinking that the semantic argument analysis on these samples would provide enough information on semantic arguments of those verbs in operative notes and could guide PAS creation for other actions verbs in operative notes (Figure 4-1).

4.2.4 Creation of PropBank Style PAS

We followed PropBank's guidelines to define and create PAS frames for each verb. A survey was made on the usage of each verb in our sample sentences to determine verb senses and semantic arguments of each sense. We firstly divided the sample sentences into a set of coarse-grained senses or meanings. For example, the sense or meaning of verb "leave" is different in following two examples. In the first sentence, "left" means "moved away from" while in the second sentence, it means "left behind".

(30) "The patient left the operating room."

(31) "2 pieces of Surgicel were left in this area."

Different senses of a verb usually require different semantic arguments to complete the meaning of the sense. In PropBank, the roleset for verb "left" in the first sentence includes arguments: Arg0 - "entity leaving", Arg1 - "place, person, or thing left" and Arg2 - "attribute of Arg1". The ProbBank roleset for the same verb in the second sentence has the same number but different semantic arguments: Arg0 – "giver/leaver", Arg1 – "thing given" and Arg2 – "benefactor".

61
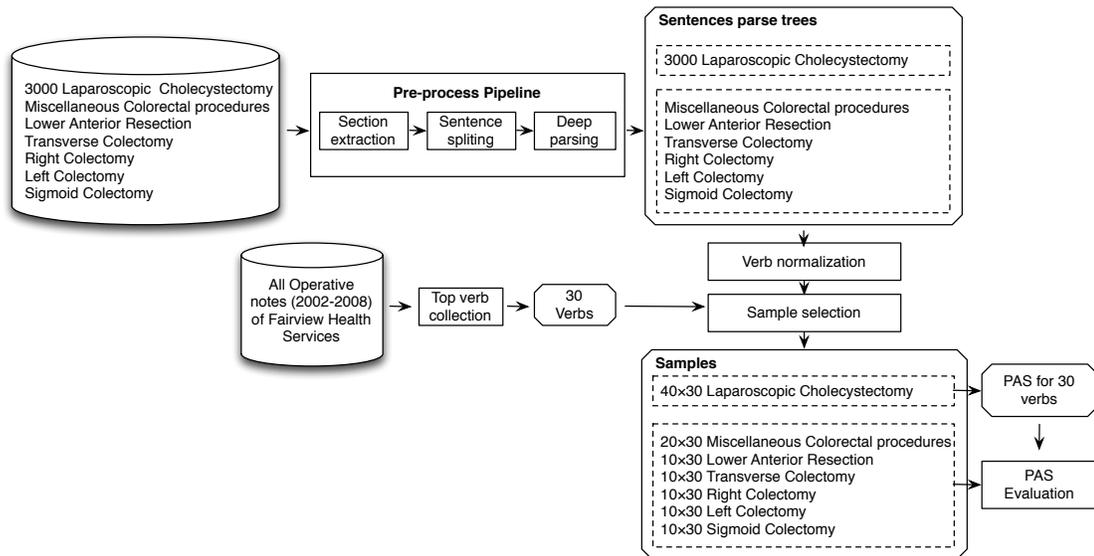
Figure 4-1. Pre-processing, predicates selection, samples selection and PAS creation.

In PropBank, the semantic arguments of a PAS frame are labeled as numbered roles: Arg0, Arg1,... ArgM. These specific roles correspond to the various valences, such as subject and object, controlled by the verb. Table 4-1 shows the PropBank PAS for verb "incise", which has only one sense. In general, each numbered argument in a PropBank PAS frame corresponds to a specific semantic role. For example, Arg0 often represents the agent - the cause or initiator of an event of a predicate. Arg1 is mostly the patient - undergoer of an action.

Table 4-1. PropBank PAS for verb "incise"

Roleset id: incise.01, *cut, carve*
Arg0: carver
Arg1: surface carved
Arg2: thing created on the surface

In examining of verb senses in operative notes, we observed that a large portion of the verb senses had a PAS frame defined in PropBank but with slightly different core arguments. To define the core arguments of these verb senses, we combined the usage of a verb sense in the sample sentences with the PropBank examples. The core arguments were defined based on whether a semantic role is required to describe the event denoted by the predicate. If a semantic role occurred with high frequency in the sample sentences or with the PropBank PAS frame core arguments of the verb sense, then we treated it as a core argument. For example, the verb "make" has 3 senses in the 40 sample operative note sentences for "make". The first sense "make.01" means "creation of objects" (e.g., serial pedicle, hole and pocket). The second sense "make.02" means "cause (to be)" and the last sense "make.03" is a light verb sense of "make" – "make sure". In PropBank, the first sense takes 4 semantic arguments as shown in Table 4-2. After examining the sample sentences from Laparoscopic Cholecystectomy notes, we found that 35 out of 40 sentences contain the first sense "make.01". Among the 35 sample sentences representing this sense, 19 contained a prepositional phrase (e.g., "*at the inferior margin of the umbilicus*", "*in the cystic duct*") that indicates the location of the action. Therefore, we inherited the original 4 arguments from PropBank PAS and included an additional core argument, Arg4 – "where the object is created", as shown in Table 4-3. For other verbs,

We carefully compared PropBank style PAS with our operative note specific PAS arguments, along with differences in senses, as described in our results.

63

Table 4-2. PropBank PAS frame for one sense of verb "make"

Roleset id: make.01, *create*
    Arg0: creator
    Arg1: creation
    Arg2: created-from, thing changed
    Arg3: benefactive

Table 4-3. Modified PAS frame for one sense of verb "make"

Roleset id: make.01, *create*
    Arg0: creator
    Arg1: creation
    Arg2: created-from, thing changed
    Arg3: benefactive
    Arg4: *where the creation (object) is created*

Verbs with exactly the same usage in operative notes and PropBank and where the semantic arguments of the PropBank PAS frame for the verb sense adequately described their usage in operative notes, the PropBank PAS was directly re-used. For verbs in operative notes with no PAS frame defined in PropBank, we analyzed the sample sentences, collected necessary semantic arguments and created a new PAS following PropBank framing guidelines. For example, the verb "prep" is a medical word with no entry in PropBank. In all the sample sentences, it means "To prepare for a medical examination or surgical procedure". We collected the semantic arguments and created a PAS for it.

For verbs with new senses in medical domain, we examined the sample sentences of the new sense, collected semantic arguments and created new PAS for them. For example, in PropBank the verb "clip" has only one sense "clip.01", which means, "to cut, cut off". In our sample sentences, the verb is frequently used as "to fasten, hold tightly" (e.g., "*The cystic duct was clipped twice toward the common duct and divided.*"). In PropBank, there is no PAS for this sense of the verb. Therefore, we created a new PAS for it as shown in Table 4-4.

Table 4-4. PAS frame for a sense of verb "clip"

Roleset id: clip.m01, *fasten, hold tightly*
Arg0: clipper, agent
Arg1: thing clipped
Arg2: location clipped

For the purposes of this study, several phrasal verbs such as "*dissect off*", "*dissect out*" and "*irrigate out*" were not defined in PropBank. Instead of creating new PAS frames we treated these phrasal verbs the same as the main verb (e.g., "dissect", "irrigate"). In examples where modifier phrases such as temporal modifier phrases occurred (e.g., "*at this point*", "*then*", "*followed by sterile dressings*" and "*next*") or adverbial phrases (e.g., "*in satisfactory condition*", "*without difficulty*" and "*in satisfactory condition*"), we did not consider them as core arguments since they were not indispensable to deliver the meaning of the sense.

4.2.5 PAS Evaluation

In order to validate the PASs derived from the Laparoscopic Cholecystectomy notes for completeness, we randomly selected sets of sentences for each verb from operative notes of different colorectal surgeries, including 10 sample sentences for each verb from Left Colectomy, Right Colectomy, Lower Anterior Resection, Transverse Colectomy, and Sigmoid Colectomy notes. Twenty sample sentences were also selected from a miscellaneous set of major abdominal colorectal surgery procedures. In addition to adding core arguments and senses from these notes to our overall PAS, we formally evaluated the coverage of the Laparoscopic Cholecystectomy verb senses and core arguments of each sense of the clinical PAS on those new sample sentences.

4.3 Results

The overall PAS of the 30 action verbs consisted of PAS frames for 40 verb senses. As shown in Table 4-5, 26 out of the 40 senses in operative notes, such as "apply", "aspirate" and "bring", had the same exact core arguments as defined in the PropBank PAS frame. There were 11 out of the 40 verb senses, such as "make", "enter" and "irrigate", which required additional arguments to completely describe the action and its arguments. Finally, there were several verb senses, such as the verbs "prep" which were completely absent from the PropBank framesets, along with one sense of "clip" and "dissect". In sample sentences, we observed 12 phrasal verbs such as "free up", "take down" and "irrigate out".

Table 4-5. Action verb senses

Verbs	Total senses	Same as in PropBank	Modified from PropBank	New senses
30	40	26 (65%)	11 (27.5%)	3 (7.5%)

For the 11 senses that with PAS adapted from PropBank, we added 11 core arguments in total based on additional semantic roles observed in sample sentences along with adding their semantic meaning. For example, Table 4-6 shows the adapted PAS for the verb "close" along with a sample sentence. The verb "close" was observed in our sample set to have only one sense (Roleset id: close.01). Overall, 37 out of 40 sample sentences for this sense were described with a phrase to present the "manner" such as "*with double layer of running absorbable suture*" and "*with 4-0 subcuticular Vicryl stitches*". In surgical procedures, the manner used to close incisions, defects and other body structure was a very important piece of information to the action "close". Therefore, we included a new core argument "manner" into the PAS of this sense of "close".

The coverage of action verb senses collected from Laparoscopic Cholecystectomy notes when compared to the other types of colorectal surgery notes was very good when compared to the instances collected from colorectal surgery operative notes. Of the overall 2,100 sample sentences from the 6 datasets, only 2 sentences were not covered by the verbs senses from the Laparoscopic Cholecystectomy notes. In the evaluation datasets, two verb senses (one for "clip" and one for "pass") had missing verb senses in the original Laparoscopic Cholecystectomy dataset. However, both of these senses were

already defined in PropBank. Moreover, in examining sample sentences from the 6 colorectal surgery datasets, the core arguments of each PAS frame derived from the Laparoscopic Cholecystectomy notes completely reflected the usage of these arguments in these sentences. We therefore did not find new semantic arguments that need to be included as core arguments into the PAS frames.

Table 4-6. PAS and an example sentence for a sense of "close"

Action verb: Close

Roleset id: close.01 , shut
   Arg0:   person doing the closing
   Arg1:   thing closing
   Arg2:   anti-beneficiary
   Arg3:   **manner**

Example sentence in an operative note:

1. The fascia at the umbilical incision was ***closed*** with interrupted 2-0 Vicryl sutures.
 	
  Rel : closed
  Arg1: The fascia at the umbilical incision
  Arg3: with interrupted 2-0 Vicryl sutures

4.4 Discussion

Operative notes contain critical information for surgeons to make decisions on the optimal surgical treatment for patients. Development of better automated systems customized for these notes are needed to extract this information in a high-throughput manner to facilitate surgical clinical research. This study focuses upon understanding the utility of PropBank PAS frames for gastrointestinal surgery operative notes and constructs a resource of operative note PAS frames. We envision that the major

68

application of this resource will be a tool for facilitating extraction of structured information from operative notes enabling surgical clinical research and surgical decision support for applications that require deep semantic knowledge of operative note details.

In this study, PAS frames were generally created following PropBank guidelines for PAS creation. For this study, we treated phrasal verbs the same as their main verb, however phrasal verbs are often treated as separate verb senses. There are about 500 phrasal verbs in PropBank. In operative notes, we also observed a large number of phrasal verbs such as "dissect out", "dissect off" and "clip off". Different from PropBank phrasal verbs, most of these phrasal verbs in operative notes are of the same or similar meaning as the main verb contained in them. For example "dissect out", "dissect down" and "dissect off" all mean the same as the sense "dissect.m01" (i.e., "separate"). In this work, we treated these phrasal verbs as having the same sense as the main verb, instead of considering them as new senses. Similarly, our previous work looking specifically at operative note actions also demonstrated a significant number of phrasal verbs in operative notes(82). We anticipate continuing to look at the issue of phrasal verbs in operative notes going forward, understanding that there may be semantic differences requiring that some of these may require special treatment.

When creating PAS frames, we noticed that some core arguments in the PropBank PAS frames did not occur in our corpus of operative notes. As the most prominent example, Arg0 of most verbs denotes the agent. In most cases, the agent did not occur in our sample sentences as most actions in operative notes are described in a

passive voice and the agent in operative notes (typically the surgeon) is omitted from the text. Another example was the omission of an argument is Arg2 of verb sense "make.01", which means "created-from, thing changed". This argument was not used at all in our sample sentences. While for the purposes of this work we kept all core arguments for PAS frames because some studies showed that semantic role labeling (SRL) tools for general English could be used in the creation of SRL systems for a scientific domain(54), we recognize that this could be a potential limitation of this study, and the frequency of use of each argument had been maintained. Work will be needed to determine whether not removing arguments not used in this set of gastrointestinal surgery notes is more generalizable or degrades the performance of our future operative note automated SRL system.

We found in our study a total of 40 senses for 30 action verbs and derived a gastrointestinal surgery-specific PAS and related frame arguments. To obtain fuller coverage of all verb senses for the larger body of operative notes, one option is to include all the verb senses that exist in PropBank into our operative note PAS frames. However, in PropBank, there are a total of 158 verb senses for these 30 verbs. At this point, since we only discovered an additional two senses in our validation phase with 2,100 samples in colorectal notes, it is our belief that keeping the PAS frames simpler with these more prominent operative note senses has a higher likelihood of maximizing the performance of our automated operative note SRL system. While this paper focused upon gastrointestinal surgery notes to develop PAS frames and used a set of verbs previously

found to have good coverage for overall with operative notes, further work in other surgical subspecialties would be helpful for validating the generalizability of our findings and understanding potential differences in PAS frame structures and content. As we proceed with expanding our PAS resource and validating it with other operative note corpora, we will be able to determine the best approach to construct PAS frames for operative notes.

In this pilot work, we were able to expand upon PropBank PAS frames for the top 30 action verbs in our larger operative note corpus. Future work includes creating additional PAS frames for other verbs in operative notes through automatic or semi-automatic methods to inductively create PAS frames. Furthermore, we will utilize these PAS frames to create training and test corpora for building a SRL system component which will be part of an IE system customized for operative notes. Also, we intend to extend our PAS frames to include nominal action predicates.

CHAPTER 5 SUMMARY AND FUTURE DIRECTIONS

As we discussed, to build an automatic SRL system for the medical domain we need to first address several issues caused by the language variation between the general English and the medical domain. We investigated previous works on building automatic SRL systems for general English and scientific domain. In this work, we created two domain specific resources, a PCFG deep parser and PropBank style semantic frames.

The final goal of the study is to develop an automatic SRL system to facilitate operative notes information extraction, which can be used for a wide range of clinical applications. One immediate application of the IE system would be representation and summarization of procedures. For this purpose, several issues need to be addressed in advanced. First, actions in procedures need to be categorized based on the semantic meaning. For example, verbs like "divide", "excise", "remove" carry similar meanings in procedures. The categorized actions could assist for standardized procedure representation. In addition, the relation between predicates is another area requiring more investigation. In procedure descriptions, it is very common that some physicians prefer detailed description of each single step whereas others favor a summary style. For example, sentences (32)   (32)   and (33)   both convey the message that a "pneumoperitoneum" is developed, but with different details. We need a method to properly count the relations between those predicates (like "insufflate" and "pneumoperitoneum") for procedure representation and summarization.

(32) "The abdomen was then insufflated with carbon dioxide."

(33) "After obtaining a pneumoperitoneum, three additional trocars were

placed."

Our future work will focus on procedure representation and summarization through more efforts on addressing the issues presented above. The successful construction of the procedure representation and summarization system could in turn serve for clinical applications in a higher level such as evidence searching and clinical outcome predication.

BIBLIOGRAPHY

1. The New Oxford American Dictionary. 2nd ed: Oxford University Press; 2005.

2. Sefr R, Puszkailer K, Jagos F. Randomized trial of different intraabdominal pressures and acid-base balance alterations during laparoscopic cholecystectomy. *Surgical endoscopy*. 2003 Jun;17(6):947-50.

3. Staehr-Rye AK, Rasmussen LS, Rosenberg J, Juul P, Lindekaer AL, Riber C, et al. Surgical space conditions during low-pressure laparoscopic cholecystectomy with deep versus moderate neuromuscular blockade: a randomized clinical study. *Anesth Analg*. 2014 Nov;119(5):1084-92.

4. Joshipura VP, Haribhakti SP, Patel NR, Naik RP, Soni HN, Patel B, et al. A prospective randomized, controlled study comparing low pressure versus high pressure pneumoperitoneum during laparoscopic cholecystectomy. *Surg Laparosc Endosc Percutan Tech*. 2009 Jun;19(3):234-40.

5. O'Dwyer PJ, McGregor JR, McDermott EW, Murphy JJ, O'Higgins NJ. Patient recovery following cholecystectomy through a 6 cm or 15 cm transverse subcostal incision: a prospective randomized clinical trial. *Postgrad Med J*. 1992 Oct;68(804):817-9.

6. Bordelon BM, Hobday KA, Hunter JG. Laser vs electrosurgery in laparoscopic cholecystectomy. A prospective randomized trial. *Arch Surg*. 1993 Feb;128(2):233-6.

7. Jakob SM, Knuesel R, Tenhunen JJ, Pradl R, Takala J. Increasing abdominal pressure with and without PEEP: effects on intra-peritoneal, intra-organ and intra-vascular pressures. *BMC Gastroenterol*. 2010;10:70.

8. Lee KC, Kim JY, Kwak HJ, Lee HD, Kwon IW. The effect of heating insufflation gas on acid-base alterations and core temperature during laparoscopic major abdominal surgery. *Korean J Anesthesiol*. 2011 Oct;61(4):275-80.

9. Joint Commission on Accreditation of Healthcare Organizations 2012; Available from: http://www.jointcommission.org.

10. Accreditation Association for Ambulatory Health Care. 2011; Available from: http://www.aaahc.org/eweb/dynamicpage.aspx?webcode=about_accred.

11. DeJong GF. An Overview of the FRUMP System. In: Lehnert WG, Ringle MH, editors. Strategies for Natural Language Processing. Hillsdale, NJ: Lawrence Erlbaum; 1982. p. 149--76.

12. Clark C, Good K, Jezierny L, Macpherson M, Wilson B, Chajewska U. Identifying smokers with a medical extraction system. *Journal of the American Medical Informatics Association*. 2008 Jan-Feb;15(1):36-9.

13. Denecke K, Bernauer J. Extracting Specific Medical Data Using Semantic Structures. In: Bellazzi R, Abu-Hanna A, Hunter J, editors. Artif Intell Med: Springer Berlin Heidelberg; 2007. p. 257-64.

14. Hripcsak G, Kuperman GJ, Friedman C. Extracting findings from narrative reports: software transferability and sources of physician disagreement. *Methods Inf Med*. 1998;37(1):1-7.

15. Long W. Extracting diagnoses from discharge summaries. *AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium*. 2005:470-4.

16. Meystre S, Haug PJ. Automation of a problem list using natural language processing. *BMC Med Inform Decis Mak*. 2005;5:30.

17. Turchin A, Kolatkar NS, Grant RW, Makhni EC, Pendergrass ML, Einbinder JS. Using regular expressions to abstract blood pressure and treatment intensification information from the text of physician notes. *J Am Med Inform Assoc*. 2006 Nov-Dec;13(6):691-5.

18. Pakhomov S, Buntrock J, Duffy P. High throughput modularized NLP system for clinical text. Proceedings of the ACL 2005 on Interactive poster and demonstration sessions; Ann Arbor, Michigan. 1225760: Association for Computational Linguistics; 2005. p. 25-8.

19. Fillmore CJ. FRAME SEMANTICS AND THE NATURE OF LANGUAGE*. *Annals of the New York Academy of Sciences*. 1976;280(1):20-32.

20. Abulaish M, Dey L. Biological relation extraction and query answering from MEDLINE abstracts using ontology-based text mining. *Data Knowl Eng*. 2007;61:228 - 62.

21. Schlaefer N. Deploying Semantic Resources for Open Domain Question Answering: Carnegie Mellon University, Pittsburgh, USA; 2007.

22. Pizzato LA, Mollá D. Indexing on semantic roles for question answering. Coling 2008: Proceedings of the 2nd workshop on Information Retrieval for Question Answering; Manchester, UK. 1641461: Association for Computational Linguistics; 2008. p. 74-81.

23. Christensen J, Mausam, Soderland S, Etzioni O. Semantic role labeling for open information extraction. Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading. Los Angeles, California2010. p. 52-60.

24. Suanmali L, Salim N, Binwahlan MS. A Hybrid Approach based on Semantic Role Labeling and General Statistic Method for Text Summarization. *Journal of Applied Sciences*. 2010;10(3):166-73.

25. Vickrey D, Koller D, editors. Sentence Simplification for Semantic Role Labeling2008: Association for Computational Linguistics.

26. Trandabăț D. Using semantic roles to improve summaries. Proceedings of the 13th European Workshop on Natural Language Generation; Nancy, France. 2187708: Association for Computational Linguistics; 2011. p. 164-9.

27. Kingsbury P, Palmer M, editors. From TreeBank to PropBank2002: European Language Resources Association (ELRA).

28. Gildea D, Palmer M. The necessity of parsing for predicate argument recognition. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. 2002:239-46.

29. Surdeanu M, Harabagiu S, Williams J, Aarseth P. Using Predicate-Argument Structures for Information Extraction. *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*.

30. Cheng LT, Zheng J, Savova GK, Erickson BJ. Discerning tumor status from unstructured MRI reports--completeness of information in existing reports and utility of automated natural language processing. *J Digit Imaging*. 2010 Apr;23(2):119-32.

31. Szolovits P. Adding a Medical Lexicon to an English Parser. AMIA Annual Symposium proceedings 2003.

32. Stanford Parser. Avaiable from: http://nlpstanfordedu/software/lex-parsershtml.

33. Sleator DD, Temperly D. Parsing English with a Link Grammar. *Third International Workshop on Parsing Technologies*. 1993.

34. Rimell L, Clark S. Porting a lexicalized-grammar parser to the biomedical domain. *J Biomed Inform*. 2009 Oct;42(5):852-65.

35. Pyysalo S, Salakoski T, Aubin S, Nazarenko A. Lexical adaptation of link grammar to the biomedical sublanguage: a comparative evaluation of three approaches. *BMC Bioinformatics*. 2006;7 Suppl 3:S2.

36. Jiang J, Zhai C. Instance Weighting for Domain Adaptation in NLP. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. 2007:264-71.

37. Huang Y, Lowe HJ, Klein D, Cucina RJ. Improved identification of noun phrases in clinical radiology reports using a high-performance statistical natural language parser augmented with the UMLS specialist lexicon. *J Am Med Inform Assoc*. 2005;12(3):275-85.

38. SPECIALIST Lexicon. Available from:

    http://wwwnlmnihgov/pubs/factsheets/umlslexhtml.

39. Alphonse E, Aubin S, Bisson G, Hamon T, Lagarrigue R, Nazarenko A, et al. Event-based information extraction for the biomedical domain: the Caderige project. *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*. 2004:43-9.

40. Riloff E. Automatically generating extraction patterns from untagged text. *Proceedings of the thirteenth national conference on Artificial intelligence - Volume 2*. 1996:1044-9.

41. Baker CF, Fillmore CJ, Lowe JB. The Berkeley FrameNet Project. *Proceedings of the 17th international conference on Computational linguistics - Volume 1*. 1998:86-90.

42. Kingsbury P, Palmer M, Marcus M. Adding Semantic Annotation to the Penn TreeBank. *In Proceedings of the Human Language Technology Conference*. 2002.

43. Giuglea A-M, Moschitti A. Semantic role labeling via FrameNet, VerbNet and PropBank. Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics; Sydney, Australia: Association for Computational Linguistics; 2006. p. 929-36.

44. Abend O, Reichart R, Rappoport A. Unsupervised argument identification for Semantic Role Labeling. Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1; Suntec, Singapore. 1687884: Association for Computational Linguistics; 2009. p. 28-36.

45. Thompson C, Levy R, Manning C. A Generative Model for Semantic Role Labeling Machine Learning: ECML 2003. 2003;2837:397-408.

46. Gildea D, Jurafsky D. Automatic labeling of semantic roles. *Comput Linguist*. 2000;28(3):245-88.

47. Hacioglu K, Ward W. Target word detection and semantic role chunking using support vector machines. Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003--short papers - Volume 2; Edmonton, Canada. 1073492: Association for Computational Linguistics; 2003. p. 25-7.

48. Xue N, Palmer M. Calibrating Features for Semantic Role Labeling. *In Proceedings of EMNLP 2004*. 2004:88-94.

49. Johansson R, Nugues P. Dependency-based semantic role labeling of PropBank. Proceedings of the Conference on Empirical Methods in Natural Language Processing; Honolulu, Hawaii. 1613726: Association for Computational Linguistics; 2008. p. 69-78.

50. Toutanova K, Haghighi A, Manning CD. Joint learning improves semantic role labeling. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics; Ann Arbor, Michigan. 1219913: Association for Computational Linguistics; 2005. p. 589-96.

51. Fleischman M, Kwon N, Hovy E. Maximum entropy models for FrameNet classification. Proceedings of the 2003 conference on Empirical methods in natural language processing. 1119362: Association for Computational Linguistics; 2003. p. 49-56.

52. Gildea D, Palmer M. The necessity of parsing for predicate argument recognition. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics; Philadelphia, Pennsylvania: Association for Computational Linguistics; 2002. p. 239-46.

53. Pradhan S, Hacioglu K, Krugler V, Ward W, Martin JH, Jurafsky D. Support Vector Learning for Semantic Argument Classification. *Mach Learn*. 2005;60(1-3):11-39.

54. Tsai RT-H, Chou W-C, Lin Y-C, Sung C-L, Ku W, Su Y-S, et al. BIOSMILE: adapting semantic role labeling for biomedical verbs: an exponential model coupled with automatically generated template features. Proceedings of the Workshop on Linking Natural Language Processing and Biology: Towards Deeper Biological Literature Analysis; New York City, New York. 1567629: Association for Computational Linguistics; 2006. p. 57-64.

55. Dahlmeier D, Ng HT. Domain adaptation for semantic role labeling in the biomedical domain. *Bioinformatics*. 2010;26(8):1098-104.

56. Barnickel T, Weston J, Collobert R, Mewes H-W, Stümpflen V. Large Scale Application of Neural Network Based Semantic Role Labeling for Automated Relation Extraction from Biomedical Texts. *PLoS One 2009*. 2009;4(7).

57. Kogan Y, Collier N, Pakhomov S, Krauthammer M. Towards Semantic Role Labeling & IE in the Medical Literature. 2005.

58. Zapirain B, Agirre E, Màrquez L. A preliminary study on the robustness and generalization of role sets for semantic role labeling. Proceedings of the 9th international conference on Computational linguistics and intelligent text processing; Haifa, Israel. 1787601: Springer-Verlag; 2008. p. 219-30.

59. Wattarujeekrit T, Shah PK, Collier N. PASBio: predicate-argument structures for event extraction in molecular biology. *BMC Bioinformatics*. 2004 Oct 19;5:155.

60. Thompson P, McNaught J, Montemagni S, Calzolari N, Gratta RD, Lee V, et al. The BioLexicon: a large-scale terminological resource for biomedical text mining. *BMC Bioinformatics*. 2011;12(1).

61. Dolbey A, Ellsworth M, Scheffczyk J. BioFrameNet: A Domain-Specific FrameNet Extension with Links to Biomedical Ontologies. Proceedings of KR-MED2006.

62. Kawahara D, Kurohashi S. Acquiring Reliable Predicate-argument Structures from Raw Corpora for Case Frame Compilation. *LREC 2010*. 2010.

63. Cohen KB, Palmer M, Hunter L. Nominalization and Alternations in Biomedical Language. *PLoS One*. 2008 09;3(9):e3158.

64. Friedman C, Kra P, Rzhetsky A. Two biomedical sublanguages: a description based on the theories of Zellig Harris. *J Biomed Inform*. 2002;35(4):222-35.

65. Lippincott T, aghdha DO, Korhonen A. Exploring variations across biomedical subdomains. Proceedings of the 23rd International Conference on Computational Linguistics; Beijing, China: Association for Computational Linguistics; 2010. p. 689-97.

66. Kilicoglu H, Fiszman M, Rosemblat G, Marimpietri S, Rindflesch TC. Arguments of nominals in semantic interpretation of biomedical text. Proceedings of the 2010 Workshop on Biomedical Natural Language Processing; Uppsala, Sweden: Association for Computational Linguistics; 2010. p. 46-54.

67. Wang Y, Melton GB, Pakhomov S. It's about This and That: A Description of Anaphoric Expressions in Clinical Text. *AMIA Annu Symp Proc*. 2011;2011:1471-80.

68. Miller GA. WordNet: A Lexical Database for English. *Communications of the ACM*. 1995;38(11):4.

69. Stedman's Medical Dictionary. 28th ed: Lippincott Williams & Wilkins; 2005. p. 2100.

70. Sipser M. Introduction to the Theory of Computation: International Thomson Publishing; 1996.

71. Chomsky N. Syntactic structures: Mouton de Gruyter; 2002.

72. Marcus MP, Marcinkiewicz MA, Santorini B. Building a large annotated corpus of English: the penn treebank. *Comput Linguist*. 1993;19(2):313-30.

73. Collins M. Head-Driven Statistical Models for Natural Language Parsing. *Comput Linguist*. 2003;29(4):589-637.

74. Klein D, Manning C. Accurate unlexicalized parsing. *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics: 7-12 July 2003; Sapporo*. 2003:423 - 30.

75. Collins M, Koo T. Discriminative Reranking for Natural Language Parsing. *Comput Linguist*. 2005;31(1):25-70.

76. Charniak E, Johnson M. Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. *ACL'05 Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. 2005:173-80.

77. Hua X, AbdelRahman S, Min J, Jung-wei F, Yang H. An initial study of full parsing of clinical text using the Stanford Parser. *Proceedings of the 2011 IEEE International Conference on Bioinformatics and Biomedicine Workshops*. 2011 12-15 Nov. 2011:607-14.

78. McClosky D, Charniak E. Self-Training for Biomedical Parsing. Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers; Columbus, Ohio2008. p. 101-4.

79. McClosky D, Charniak E, Johnson M. Reranking and self-training for parser adaptation. Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics; Sydney, Australia2006. p. 337-44.

80. Bacchiani M, Riley M, Roark B, Sproat R. MAP adaptation of stochastic grammars. *Comput Speech Lang*. 2006;20(1):41-68.

81. Genia treebank. Available from:
http://www.nactem.ac.uk/genia/genia-corpus/treebank.

82. Wang Y, Pakhomov S, Burkart NE, Ryan JO, Melton GB. A study of actions in operative notes. *AMIA Annu Symp Proc*. 2012;2012:1431-40.

83. the Penn Treebank II Bracketing Guide. Available from:
    ftp://ftp.cis.upenn.edu/pub/treebank/doc/manual/.

84. Abney S, Flickenger S, Gdaniec C, Grishman C, Harrison P, Hindle D, et al. Procedure for quantitatively comparing the syntactic coverage of English grammars. *Proceedings of the workshop on Speech and Natural Language*. 1991:306-11.

85. Clegg AB, Shepherd AJ. Evaluating and Integrating Treebank Parsers on a Biomedical Corpus. *In Proceedings of the ACL Workshop on Software*. 2005:14-33.

86. Rush AM, Reichart R, Collins M, Globerson A. Improved parsing and POS tagging using inter-sentence consistency constraints. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. 2012:1434-44.

87. Lease M, Charniak E. Parsing biomedical literature. *the Second International Joint Conference on Natural Language Processing (IJCNLP-05)*. 2005:58-69.

88. Miller JE, Torii M, Vijay-Shanker K, editors. Subdomain adaptation of a POS tagger with a small corpus. Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing; 2006.

89. BioMedICUS (March 1, 2015).

90. McClosky D. Any Domain Parsing: Automatic Domain Adaptation for Natural Language Parsing: Brown University; 2010.

91. Cairns BL, Nielsen RD, Masanz JJ, Martin JH, Palmer MS, Ward WH, et al. The MiPACQ Clinical Question Answering System. *AMIA Annu Symp Proc*. 2011;2011:171-80.

92. Redd D, Divita G, SamahJarad, Brandt C, Nebeker JR. Characterizing Clinical Text and Sublanguage: A Case Study of the VA Clinical Notes. *Journal of Health & Medical Informatics*. 2011;10(12).

93. Zweigenbaum P. Natural Language Processing in the Medical and Biological Domains: a Parallel Perspective. 3 rd International Symposium on Semantic Mining in Biomedicine, Turku, Finland2008. p. 3-4.