Structuring Operative Notes using Active Learning

Kirk Roberts*
National Library of Medicine
National Institutes of Health
Bethesda, MD 20894
kirk.roberts@nih.gov

Sanda M. Harabagiu
Human Language Technology Research Institute
University of Texas at Dallas
Richardson, TX 75080
sanda@hlt.utdallas.edu

Michael A. Skinner
University of Texas Southwestern Medical Center
Children's Medical Center of Dallas
Dallas, TX 75235
michael.skinner@childrens.com

Abstract

We present an active learning method for placing the event mentions in an operative note into a pre-specified event structure. Event mentions are first classified into action, peripheral action, observation, and report events. The actions are further classified into their appropriate location within the event structure. We examine how utilizing active learning significantly reduces the time needed to completely annotate a corpus of 2,820 appendectomy notes.

1 Introduction

Operative reports are written or dictated after every surgical procedure. They describe the course of the operation as well as any abnormal findings in the surgical process. Template-based and structured methods exist for recording the operative note (DeOrio, 2002), and in many cases have been shown to increase the completeness of surgical information (Park et al., 2010; Gur et al., 2011; Donahoe et al., 2012). The use of natural language, however, is still preferred for its expressive power. This unstructured information is typically the only vehicle for conveying important details of the procedure, including the surgical instruments, incision techniques, and laparoscopic methods employed.

The ability to represent and extract the information found within operative notes would enable powerful post-hoc reasoning methods about surgical procedures. First, the completeness problem may be alleviated by indicating gaps in the surgical narrative. Second, deep semantic similarity methods could be used to discover comparable operations across surgeons and institutions. Third, given information on the typical course and findings of a procedure, abnormal aspects of an operation could be identified and investigated. Finally, other secondary use applications would be enabled to study the most effective instruments and techniques across large amounts of surgical data.

In this paper, we present an initial method for aligning the event mentions within an operative note to the overall event structure for a procedure. A surgeon with experience in a particular procedure first describes the overall event structure. A supervised method enhanced by active learning is then employed to rapidly build an information extraction model to classify event mentions into the event structure. This active learning paradigm allows for rapid prototyping while also taking advantage of the sub-language characteristics of operative notes and the common structure of operative notes reporting the same type of procedure. A further goal of this method is to aid in the evaluation of unsupervised techniques that can automatically discover the event structure solely from the narratives. This would enable all the objectives outlined above for leveraging the unstructured information within operative notes.

This paper presents a first attempt at this active learning paradigm for structuring appendectomy reports. We intentionally chose a well-understood and relatively simple procedure to en-

*Most of this work was performed while KR was at the University of Texas at Dallas.

Proceedings of the 2014 Workshop on Biomedical Natural Language Processing (BioNLP 2014), pages 68–76,
Baltimore, Maryland USA, June 26-27 2014.

sure a straight-forward, largely linear event structure where a large amount of data would be easily available. Section 3 describes a generic framework for surgical event structures and the particular structure chosen for appendectomies. Section 4 details the data used in this study. Section 5 describes the active learning experiment for filling in this event structure for operative notes. Section 6 reports the results of this experiment. Section 7 analyzes the method and proposes avenues for further research. First, however, we outline the small amount of previous work in natural language processing on operative notes.

2 Previous Work

An early tool for processing operative notes was proposed by Lamiell et al. (1993). They develop an auditing tool to help enforce completeness in operative notes. A syntactic parser converts sentences in an operative note into a graph structure that can be queried to ensure the necessary surgical elements are present in the narrative. For appendectomies, they could determine whether answers were specified for questions such as "*What was the appendix abnormality?*" and "*Was cautery or drains used?*". Unlike what we propose, they did not attempt to understand the narrative structure of the operative note, only ensure that a small number of important elements were present. Unfortunately, they only tested their rule-based system on four notes, so it is difficult to evaluate the robustness and generalizability of their method.

More recently, Wang et al. (2014) proposed a machine learning (ML) method to extract patient-specific values from operative notes written in Chinese. They specifically extract tumor-related information from patients with hepatic carcinoma, such as the size/location of the tumor, and whether the tumor boundary is clear. In many ways this is similar in purpose to Lamiell et al. (1993) in the sense that there are operation-specific attributes to extract. However, while the auditing function primarily requires knowing whether particular items were stated, their method extracts the particular values for these items. Furthermore, they employ an ML-based conditional random field (CRF) trained and tested on 114 operative notes. The primary difference between the purpose of these two methods and the purpose of our method lies in the attempt to model all the events that characterize a surgery. Both the work of Lamiell et al. (1993) and Wang et al. (2014) can be used for completeness testing, and Wang et al. (2014) can be used to find similar patients. The lack of understanding of the event structure, however, prevents these methods from identifying similar surgical methods or unexpected surgical techniques, or from accomplishing many other secondary use objectives.

In a more similar vein to our own approach, Wang et al. (2012) studies actions (a subset of event mentions) within an operative note. They note that various lexico-syntactic constructions can be used to specify an action (e.g., *incised*, *the incision was carried*, *made an incision*). Like our approach, they observed sentences can be categorized into actions, perceptions/reports, and other (though we make this distinction at the event mention level). They adapted the Stanford Parser (Klein and Manning, 2003) with the Specialist Lexicon (Browne et al., 1993) similar to Huang et al. (2005). They do not, however, propose any automatic system for recognizing and categorizing actions. Instead, they concentrate on evaluating existing resources. They find that many resources, such as UMLS (Lindberg et al., 1993) and FrameNet (Baker et al., 1998) have poor coverage of surgical actions, while Specialist and WordNet (Fellbaum, 1998) have good coverage.

A notable limitation of their work is that they only studied actions at the sentence level, looking at the main verb of the independent clause. We have found in our study that multiple actions can occur within a sentence, and we thus study actions at the event mention level. Wang et al. (2012) noted this shortcoming and provide the following illustrative examples:

- *The patient was* **taken** *to the operating room where general anesthesia was* **administered**.
- *After the successful* **induction** *of spinal anesthesia, she was* **placed** *supine on the operating table.*

The second event mention in the first sentence (*administered*) and the first event mention in the second sentence (*induction*) are ignored in Wang et al. (2012)'s study. Despite the fact that they are stated in dependent clauses, these mentions may be more semantically important to the narrative than the mentions in the independent clauses. This is because a grammatical relation does not necessarily imply event prominence. In a further study, Wang et al. (2013) work toward the creation of an automatic extraction system by annotating

PropBank (Palmer et al., 2005) style predicate-argument structures on thirty common surgical actions.

3 Event Structures in Operative Notes

Since operations are considered to be one of the riskier forms of clinical treatment, surgeons follow strict procedures that are highly structured and require significant training and oversight. Thus, a surgeon's description of a particular operation should be highly similar with a different description of the same type of operation, even if written by a different surgeon at a different hospital. For instance, the two examples below were written by two different surgeons to describe the event of controlling the blood supply to the appendix:

- *The 35 mm vascular Endo stapler device was* **fired** *across the mesoappendix...*
- *The meso appendix was* **divided** *with electrocautery...*

In these two examples, the surgeons use different lexical forms (*fired* vs. *divided*), syntactic forms (mesoappendix to the right or left of the EVENT), different semantic predicate-argument structures (INSTRUMENT-EVENT-ANATOMICALOBJECT vs. ANATOMICALOBJECT-EVENT-METHOD), and even different surgical techniques (stapling or cautery). Still, these examples describe the same step in the operation and thus can be mapped to the same location in the event structure.

In order to recognize the event structure in operative notes, we start by specifying an event structure to a particular operation (e.g., mastectomy, appendectomy, heart transplant) and create a ground-truth structure based on expert knowledge. Our goal is then to normalize the event mentions within a operative note to the specific surgical actions in the event structure. While the lexical, syntactic, and predicate-argument structures vary greatly across the surgeons in our data, many event descriptions are highly consistent within notes written by the same surgeon. This is especially true of events with little linguistic variability, typically largely procedural but necessary events that are not the focus of the surgeon's description of the operation. An example of low-variability is the event of placing the patient on the operating table, as opposed to the event of manipulating the appendix to prepare it for removal. Additionally, while there is considerable lexical variation in how an event is mentioned, the terminology for event mentions is fairly limited, resulting in reasonable similarity between surgeons (e.g., the verbal description used for the dividing of the mesoappendix is typically one of the following mentions: *fire*, *staple*, *divide*, *separate*, *remove*).

3.1 Event Structure Representation

Operative notes contain event mentions of many different event classes. Some classes correspond to actions performed by the surgeon, while others describe findings, provide reasonings, or discuss interactions with patients or assistants. These distinctions are necessary to recognizing the event structure of an operation, in which we are primarily concerned with surgical actions. We consider the following event types:

- ACTION: the primary types of events in an operation. These typically involve physical actions taken by the surgeon (e.g., creating/closing an incision, dividing tissue), or procedural events (e.g., anesthesia, transfer to recovery). With limited exceptions, ACTIONs occur in a strict order and the i^{th} ACTION can be interpreted as enabling the $(i+1)^{\text{th}}$ ACTION.
- P_ACTION: the peripheral actions that are optional, do not occur within a specific place in the chain of ACTIONs, and are not considered integral to the event structure. Examples include stopping unexpected bleeding and removing benign cysts un-connected with the operation.
- OBSERVATION: an event that denotes the act of observing a given state. OBSERVATIONs may lead to ACTION (e.g., the appendix is perforated and therefore needs to be removed) or P_ACTIONs (e.g., a cyst is found). They may also be elaborations to provide more details about the surgical method being used.
- REPORT: an event that denotes a verbal interaction between the surgeon and a patient, guardian, or assistant (such as obtaining consent for an operation).

The primary class of events that we are interested in here are ACTIONs. Abstractly, one can view a type of operation as a directed graph with specified start and end states. The nodes denote the events, while the edges denote enablements. An instance of an operation then can be represented as some

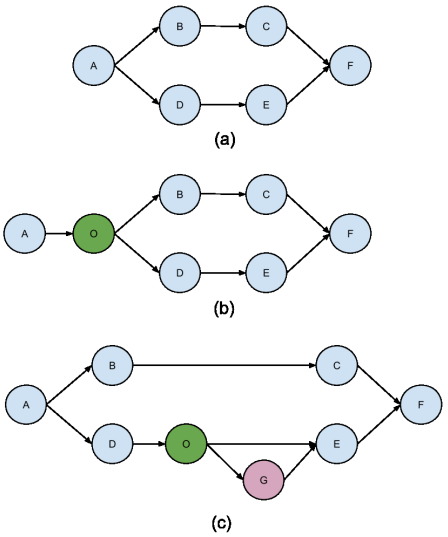

Figure 1: Graphical representation of a surgical procedure with ACTIONS A, B, C, D, E, and F, OBSERVATION O, and P_ACTION G. (a) strict surgical graph (only actions), (b) surgical graph with an observation invoking an action, (c) surgical graph with an observation invoking a peripheral action.

path between the start and end nodes.

In its simplest form, a surgical graph is composed entirely of ACTION nodes (see Figure 1(a)). It is possible to add expected OBSERVATIONS that might trigger a different ACTION path (Figure 1(b)). Finally, P_ACTIONS can be represented as optional nodes in the surgical graph, which may or may not be triggered by OBSERVATIONS (Figure 1(c)). This graphical model is simply a conceptual aid to help design the action types. The model currently plays no role in the automatic classification. For the remainder of this section we focus on a relatively limited surgical procedure that can be interpreted as a linear chain of ACTIONS.

3.2 Appendectomy Representation

Acute appendicitis is a common condition requiring surgical management, and is typically treated by removing the appendix, either laparoscopically or by using an open technique. Appendectomies are the most commonly performed urgent surgical procedure in the United States. The procedure is relatively straight-forward, and the steps of the procedure exhibit little variation between different surgeons. The third author (MS), a surgeon with more than 20 years of experience in pediatric surgery, provided the following primary ACTIONS:

- APP01: transfer patient to operating room
- APP02: place patient on table
- APP03: anesthesia
- APP04: prep
- APP05: drape
- APP06: umbilical incision
- APP07: insert camera/telescope
- APP08: insert other working ports
- APP09: identify appendix
- APP10: dissect appendix away from other structures
- APP11: divide blood supply
- APP12: divide appendix from cecum
- APP13: place appendix in a bag
- APP14: remove bag from body
- APP15: close incisions
- APP16: wake up patient
- APP17: transfer patient to post-anesthesia care unit

In the laparoscopic setting, each of these actions is a necessary part of the operation, and most should be recorded in the operative note. Additionally, any number of P_ACTION, OBSERVATION, and REPORT events may be interspersed.

4 Data

In accordance with generally accepted medical practice and to comply with requirements of The Joint Commission, a detailed report of any surgical procedure is placed in the medical record within 24 hours of the procedure. These notes include the preoperative diagnosis, the post-operative diagnosis, the procedure name, names of surgeon(s) and assistants, anesthetic method, operative findings, complications (if any), estimated blood loss, and a detailed report of the conduct of the procedure. To ensure accuracy and completeness, such notes are typically dictated and transcribed shortly after the procedure by the operating surgeon or one of the assistants.

To obtain the procedure notes for this study, The Children's Medical Center (CMC) of Dallas electronic medical record (EMR) was queried for operative notes whose procedure contained the word "appendectomy" (CPT codes 44970, 44950, 44960) for a preoperative diagnosis of "acute appendicitis" (ICD9 codes 541, 540.0, 540.1). At the time of record acquisition, the CMC EMR had been in operation for about 3 years, and 2,820 notes were obtained, having been completed by 12 pediatric surgeons. In this set, there were 2,757

Surgeon	Notes	Events	Words
$surgeon_1$	8	291	2,305
$surgeon_2$	311	16,379	134,748
$surgeon_3$	143	6,897	57,797
$surgeon_4$	400	8,940	62,644
$surgeon_5$	391	15,246	114,684
$surgeon_6$	307	9,880	77,982
$surgeon_7$	397	10,908	74,458
$surgeon_8$	34	2,401	20,391
$surgeon_9$	2	100	973
$surgeon_{10}$	355	9,987	89,085
$surgeon_{11}$	380	14,211	135,215
$surgeon_{12}$	92	2,417	19,364
Total	2,820	97,657	789,646

Table 1: Overview of corpus by surgeon.

laparoscopic appendectomies and 63 open procedures. The records were then processed automatically to remove any identifying information such as names, hospital record numbers, and dates. For the purposes of this investigation, only the surgeon's name and the detailed procedure note were collected for further study. Owing to the complete anonymity of the records, the study received an exemption from the University of Texas Southwestern Medical Center and CMC Institutional Review Boards. Table 1 contains statistics about the distribution of notes by surgeon in our dataset.

5 Active Learning Framework

Active learning is becoming a more and more popular framework for natural language annotation in the biomedical domain (Hahn et al., 2012; Figueroa et al., 2012; Chen et al., 2013a; Chen et al., 2013b). In an active learning setting, instead of performing manual annotation separate from automatic system development, an existing ML classifier is employed to help choose which examples to annotate. Thus, human annotators can focus on examples that would prove difficult for a classifier, which can dramatically reduce overall annotation time. However, active learning is not without pitfalls, notably sampling bias (Dasgupta and Hsu, 2008), re-usability (Tomanek et al., 2007), and class imbalance (Tomanek and Hahn, 2009). In our work, the purpose of utilizing an active learning framework is to produce a fully-annotated corpus of labeled event mentions in as small a period of time as possible. To some extent, the goal of full-annotation alleviates some of the active learning issues discussed above (re-usability and class imbalance), but sampling bias could still lead to significantly longer annotation time.

Our goal is to (1) distinguish event mentions in one of the four classes introduced in Section 3.1 (event type annotation), and (2) further classify actions into their appropriate location in the event structure (on this data, appendectomy type annotation). While most active learning methods are used with the intention of only manually labeling a sub-set of the data, our goal is to annotate every event mention so that we may ultimately evaluate unsupervised techniques on this data. Our active learning experiment thus proceeds in two parallel tracks: (i) a traditional active learning process where the highest-utility unlabeled event mentions are classified by a human annotator, and (ii) a batch annotation process where extremely similar, "easy" examples are annotated in large groups. Due to small intra-surgeon language variation, and relatively small inter-surgeon variation due to the limited terminology, this second process allows us to annotate large numbers of unlabeled examples at a time. The batch labeling largely annotates unlabeled examples that would not be selected by the primary active learning module because they are too similar to the already-labeled examples. After a sufficient amount of time being spent in traditional active learning, the batch labeling is used to annotate until the batches produced are insufficiently similar and/or wrong classifications are made. After a sufficent number of annotations are made with the active learning method, the choice of when to use the active learning or batch annotation method is left to the discretion of the annotator. This back-and-forth is then repeated iteratively until all the examples are annotated.

For both the active learning and batch labeling processes, we use a multi-class support vector machine (SVM) using a simple set of features:

F1. Event mention's lexical form (e.g., *identified*)
F2. Event mention's lemma (*identify*)
F3. Previous words (*3-the*, *2-appendix*, *1-was*)
F4. Next words (*1-and*, *2-found*, *3-to*, *4-be*, *5-ruptured*)
F5. Whether the event is a gerund (*false*)

Features F3 and F4 were constrained to only return words within the sentence.

To sample event mentions for the active learner, we combine several sampling techniques to ensure a diversity of samples to label. This meta-sampler chooses from 4 different samplers with differing probability p:

1. UNIFORM: Choose (uniformly) an unlabeled instance ($p = 0.1$). Formally, let $\mathcal{L}$ be the

set of manually labeled instances. Then, the probability of selecting an event e_i is:

$$P_{\text{U}}(e_i) \propto \delta(e_i \notin \mathcal{L})$$

Where $\delta(x)$ is the delta function that returns 1 if the condition x is true, and 0 otherwise. Thus, an unlabeled event has an equal probability of being selected as every other unlabeled event.

2. JACCARD: Choose an unlabeled instance biased toward those whose word context is least similar to the labeled instances using Jaccard similarity ($p = 0.2$). This sampler promotes diversity to help prevent sampling bias. Let W_i be the words in e_i's sentence. Then the probability of selecting an event with the JACCARD sampler is:

$$P_{\text{J}}(e_i) \propto \delta(e_i \notin \mathcal{L}) \min_{e_j \in \mathcal{L}} \left[\left(1 - \frac{W_i \cap W_j}{W_i \cup W_j} \right)^{\alpha} \right]$$

Here, α is a parameter to give more weight to dissimilar sentences (we set α = 2).

3. CLASSIFIER: Choose an unlabeled instance biased toward those the SVM assigned low confidence values ($p = 0.65$). Formally, let $f_c(e_i)$ be the confidence assigned by the classifier to event e_i. Then, the probability of selecting an event with the CLASSIFIER sampler is:

$$P_{\text{C}}(e_i) \propto \delta(e_i \notin \mathcal{L})(1 - f_c(e_i))$$

The SVM we use provides confidence values largely in the range (-1, 1), but for some very confident examples this value can be larger. We therefore constrain the raw confidence value $f_r(e_i)$ and place it within the range [0, 1] to achieve the modified confidence $f_c(e_i)$ above:

$$f_c(e_i) = \frac{\max(\min(f_r(e_i), 1), -1) + 1}{2}$$

In this way, $f_c(e_i)$ can be guaranteed to be within [0, 1] and can thus be interpreted as a probability.

4. MISCLASSIFIED: Choose (uniformly) a *labeled* instance that the SVM mis-classifies during cross-validation ($p = 0.05$). Let $f(e_i)$ be the classifier's guess and $\mathcal{L}(e_i)$ be the manual label for event e_i. Then the probability of selecting an event is:

$$P_{\text{M}}(e_i) \propto \delta(e_i \in \mathcal{L})\delta(f(e_i) \neq \mathcal{L}(e_i))$$

Event Type	Precision	Recall	F_1
ACTION	0.79	0.90	0.84
NOT_EVENT	0.75	0.82	0.79
OBSERVATION	0.71	0.57	0.63
P_ACTION	0.66	0.40	0.50
REPORT	1.00	0.58	0.73
Active Learning Accuracy: 76.4%			
Batch Annotation Accuracy: 99.5%			

Table 2: Classification results for event types. Except when specified, results are for data annotated using the active learning method, while the batch annotation results include all data.

The first annotation was made using the UNIFORM sampler. For every new annotation, the meta-sampler chooses one of the above sampling methods according to the above p values, and that sampler selects an example to annotate. For each selected sample, it is first assigned an event type. If it is assigned as an ACTION, the annotator further assigns its appropriate action type. The CLASSIFIER and MISCLASSIFIED samplers alternate between the event type and action type classifiers. These four samplers were chosen to balance the traditional active learning approach (CLASSIFIER), while trying to prevent classifier bias (UNIFORM and JACCARD), while also allowing mis-labeled data to be corrected (MISCLASSIFIED). An evaluation of the utility of the individual samplers is beyond the scope of this work.

6 Results

For event type annotation, two annotators single-annotated 1,014 events with one of five event types (ACTION, P_ACTION, OBSERVATION, REPORT, and NOT_EVENT). The classifier's accuracy on this data was 75.9% (see Table 2 for a breakdown by event type). However, the examples were chosen because they were very different from the current labeled set, and thus we would expect them to be more difficult than a random sampling. When one includes the examples annotated using batch labeling, the overall accuracy is 99.5%.

For action type annotation, the same two annotators labeled 626 ACTIONs with one of the 17 action types (APP01–APP17). The classifier's accuracy on this data was again a relatively low 72.2% (see Table 3 for a breakdown by action type). However, again, these examples were expected to be difficult for the classifier. When one includes the examples annotated using batch labeling, the overall accuracy is 99.4%.

Action Type	Precision	Recall	F_1
APP01	0.91	0.77	0.83
APP02	1.00	0.67	0.80
APP03	1.00	0.67	0.80
APP04	0.95	0.95	0.95
APP05	1.00	1.00	1.00
APP06	0.79	0.72	0.76
APP07	0.58	0.58	0.58
APP08	0.65	0.75	0.70
APP09	0.82	0.93	0.87
APP10	0.63	0.73	0.68
APP11	0.50	0.50	0.50
APP12	0.61	0.56	0.58
APP13	0.94	0.94	0.94
APP14	0.71	0.73	0.72
APP15	0.84	0.79	0.82
APP16	0.93	0.81	0.87
APP17	0.84	0.89	0.86
Active Learning Accuracy: 71.4%			
Batch Annotation Accuracy: 99.4%			

Table 3: Classification results for action types.

7 Discussion

The total time allotted for annotation was approximately 12 hours, split between two annotators (the first author and a computer science graduate student). Prior to annotation, both annotators were given a detailed description of an appendectomy, including a video of a procedure to help associate the actual surgical actions with the narrative description. After annotation, 1,042 event types were annotated using the active learning method, 90,335 event types were annotated using the batch method, and 6,279 remained un-annotated. Similarly, 658 action types were annotated using the active learning method, 35,799 action types were annotated using the batch method, and 21,151 remained un-annotated. A greater proportion of actions remained un-annotated due to the lower classifier confidence associated with the task. Event and action types were annotated in unison, but we estimate during the active learning process it took about 25 seconds to annotate each event (both the event type and the action type if classified as an ACTION). The batch process enabled the annotation of an average of 3 event mentions per second.

This rapid annotation was made possible by the repetitive nature of operative notes, especially within an individual surgeon's notes. For example, the following statements were repeated over 100 times in our corpus:

- *General anesthesia was* **induced**.
- *A Foley catheter was* **placed** *under sterile conditions.*
- *The appendix was* **identified** *and seemed to be acutely* **inflamed**.

The first example was used by an individual surgeon in 95% of his/her notes, and only used three times by a different surgeon. In the second example, the sentence is used in 77% of the surgeon's notes while only used once by another surgeon. The phrase "*Foley catheter was placed*", however, was used 133 times by other surgeons. In the context of an appendectomy, this action is unambiguous, and so only a few annotations are needed to recognize the hundreds of actual occurrences in the data. Similarly, with the third example, the phrase "*the appendix was identified*" was used in over 600 operative notes by 10 of the 12 surgeons. After a few manual annotations to achieve sufficient classification confidence, the batch process can identify duplicate or near-duplicate events that can be annotated at once, greatly reducing the time needed to achieve full annotation.

Unfortunately, the most predictable parts of a surgeon's language are typically the least interesting from the perspective of understanding the critical points in the narrative. As shown in the examples above, the highest levels of redundancy are found in the most routine aspects of the operation. The batch annotation, therefore, is quite biased and the 99% accuracies it achieves cannot be expected to hold up once the data is fully annotated. Conversely, the active learning process specifically chooses examples that are different from the current labeled set and thus are more difficult to classify. Active learning is more likely to sample from the "long tail" than the most frequent events and actions, so the performance on the chosen sample is certainly a lower bound on the performance of a completely annotated data set. If one assumes the remaining un-annotated data will be of similar difficulty to the data sampled by the active learner, one could project an overall event type accuracy of 97% and an overall action type accuracy of 89%. This furthermore assumes no improvements are made to the machine learning method based on this completed data.

One way to estimate the potential bias in batch annotation is by observing the differences in the distributions of the two data sets. Figure 2 shows the total numbers of action types for both the active learning and batch annotation portions of the data. For the most part, the distributions are similar. APP08 (insert other working ports), APP10 (dissect appendix away from other structures), APP11 (divide blood supply), APP12 (di-

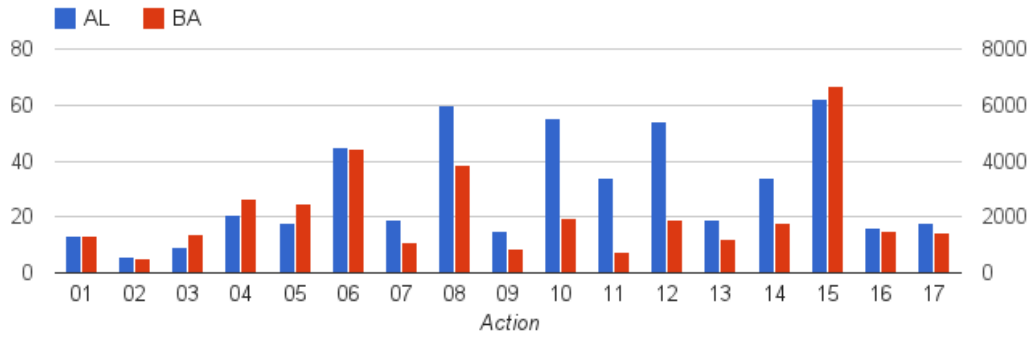

Figure 2: Frequencies of action types in the active learning (AL) portion of the data set (left vertical axis) and the batch annotation (BA) portion of the data set (right vertical axis).

vide appendix from cecum), and APP14 (remove bag from body) are the most under-represented in the batch annotation data. This confirms our hypothesis that some of the most interesting events have the greatest diversity in expression.

In Section 2 we noted that a limitation of the annotation method of Wang et al. (2012) was that a sentence could only have one action. We largely overcame this problem by associating a single surgical action with an event mention. This has one notable limitation, however, as occasionally a single event mention corresponds to more than one action. In our data, APP11 and APP12 are commonly expressed together:

- *Next, the mesoappendix and appendix is* **stapled**$_{\text{APP11/APP12}}$ *and then the appendix is* **placed**$_{\text{APP13}}$ *in an endobag.*

Here, a coordination ("*mesoappendix and appendix*") is used to associate two events (the stapling of the mesoappendix and the stapling of the appendix) with the same event mention. In the event extraction literature, this is a well-understood occurrence, as for instance TimeML (Pustejovsky et al., 2003) can represent more than one event with a single event mention. In practice, however, few automatic TimeML systems handle such phenomena. Despite this, for our purpose the annotation structure should likely be amended so that we can account for all the important actions in the operative note. This way, gaps in our event structure will correspond to actual gaps in the narrative (e.g., dividing the blood supply is a critical step in an appendectomy and therefore needs to fit within the event structure).

Finally, the data in our experiment comes from a relatively simple procedure (an appendectomy). It is unclear how well this method would generalize to more complex operations. Most likely, the difficulty will lie in actions that are highly ambiguous, such as if more than one incision is made. In this case, richer semantic information will be necessary, such as the spatial argument that indicates where a particular event occurs (Roberts et al., 2012).

8 Conclusion

With the increasing availability of electronic operative notes, there is a corresponding need for deep analysis methods to understand the note's narrative structure to enable applications for improving patient care. In this paper, we have presented a method for recognizing how event mentions in an operative note fit into the event structure of the actual operation. We have proposed a generic framework for event structures in surgical notes with a specific event structure for appendectomy operations. We have described a corpus of 2,820 operative notes of appendectomies performed by 12 surgeons at a single institution. With the ultimate goal of fully annotating this data set, which contains almost 100,000 event mentions, we have shown how an active learning method combined with a batch annotation process can quickly annotate the majority of the corpus. The method is not without its weaknesses, however, and further annotation is likely necessary.

Beyond finishing the annotation process, our ultimate goal is to develop unsupervised methods for structuring operative notes. This would enable expanding to new surgical procedures without human intervention while also leveraging the increasing availability of this information. We have shown in this work how operative notes have linguistic characteristics that result in parallel structures. It is our goal to leverage these characteristics in developing unsupervised methods.

Acknowledgments

The authors would like to thank Sanya Peshwani for her help in annotating the data.

References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of ACL/COLING*.

Allen C. Browne, Alexa T. McCray, and Suresh Srinivasan. 1993. The SPECIALIST Lexicon. Technical Report NLM-LHC-93-01, National Library of Medicine.

Yukun Chen, Hongxin Cao, Qiaozhu Mei, Kai Zheng, and Hua Xu. 2013a. Applying active learning to supervised word sense disambiguation in MEDLINE. *J Am Med Inform Assoc*, 20:1001–1006.

Yukun Chen, Robert Carroll, Eugenia R. McPeek Hinz, Anushi Shah, Anne E. Eyler, Joshua C. Denny, , and Hua Xu. 2013b. Applying active learning to high-throughput phenotyping algorithms for electronic health records data. *J Am Med Inform Assoc*, 20:e253–e259.

Sanjoy Dasgupta and Daniel Hsu. 2008. Hierarchical Sampling for Active Learning. In *Proceedings of the International Conference on Maching Learning*.

J.K. DeOrio. 2002. Surgical templates for orthopedic operative reports. *Orthopedics*, 25(6):639–642.

Laura Donahoe, Sean Bennett, Walley Temple, Andrea Hilchie-Pye, Kelly Dabbs, Ethel MacIntosh, and Geoff Porter. 2012. Completeness of dictated operative reports in breast cancer–the case for synoptic reporting. *J Surg Oncol*, 106(1):79–83.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Rosa L. Figueroa, Qing Zeng-Treitler, Long H. Ngo, Sergey Goryachev, and Eduardo P. Wiechmann. 2012. Active learning for clinical text classification: is it better than random sampling? *J Am Med Inform Assoc*, 19:809–816.

I. Gur, D. Gur, and J.A. Recabaren. 2011. The computerized synoptic operative report: A novel tool in surgical residency education. *Arch Surg*, pages 71–74.

Udo Hahn, Elena Beisswanger, Ekaterina Buyko, and Erik Faessler. 2012. Active Learning-Based Corpus Annotation – The PATHOJEN Experience. In *Proceedings of the AMIA Symposium*, pages 301–310.

Yang Huang, Henry J Lowe, Dan Klein, and Russell J Cucina. 2005. 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*, 12:275–285.

Dan Klein and Christopher D. Manning. 2003. Accurate Unlexicalized Parsing. In *Proceedings of ACL*, pages 423–430.

James M Lamiell, Zbigniew M Wojcik, and John Isaacks. 1993. Computer Auditing of Surgical Operative Reports Written in English. In *Proc Annu Symp Comput Appl Med Care*, pages 269–273.

Donald A.B. Lindberg, Betsy L. Humphreys, and Alexa T. McCray. 1993. The Unified Medical Language System. *Methods of Information in Medicine*, 32(4):281–291.

Martha Palmer, Paul Kingsbury, and Daniel Gildea. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–106.

Jason Park, Venu G. Pillarisetty, Murray F. Brennan, and *et al.* 2010. Electronic Synoptic Operative Reporting: Assessing the Reliability and Completeness of Synoptic Reports for Pancreatic Resection. *J Am Coll Surgeons*, 211(3):308–315.

James Pustejovsky, José Castano, Robert Ingria, Roser Saurí, Robert Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir Radev. 2003. TimeML: Robust Specification of Event and Temporal Expressions in Text. In *Proceedings of the Fifth International Workshop on Computational Semantics*.

Kirk Roberts, Bryan Rink, Sanda M. Harabagiu, Richard H. Scheuermann, Seth Toomay, Travis Browning, Teresa Bosler, and Ronald Peshock. 2012. A Machine Learning Approach for Identifying Anatomical Locations of Actionable Findings in Radiology Reports. In *Proceedings of the AMIA Symposium*.

Katrin Tomanek and Udo Hahn. 2009. Reducing Class Imbalance during Active Learning for Named Entity Annotation. In *Proceedings of KCAP*.

Katrin Tomanek, Joachim Wermter, and Udo Hahn. 2007. An Approach to Text Corpus Construction which Cuts Annotation Costs and Maintains Reusability of Annotated Data. In *Proceedings of EMNLP/CoNLL*, pages 486–495.

Yan Wang, Serguei Pakhomov, Nora E. Burkart, James O. Ryan, and Genevieve B. Melton. 2012. A Study of Actions in Operative Notes. In *Proceedings of the AMIA Symposium*, pages 1431–1440.

Yan Wang, Serguei Pakhomov, and Genevieve B Melton. 2013. Predicate Argument Structure Frames for Modeling Information in Operative Notes. In *Studies in Health Technology and Informatics (MEDINFO)*, pages 783–787.

Hui Wang, Weide Zhang, Qiang Zeng, Zuofeng Li, Kaiyan Feng, and Lei Liu. 2014. Extracting important information from Chinese Operation Notes with natural language processing methods. *J Biomed Inform*.