Extraction of Clinical Lab Results: Design and Evaluation of a cTAKES Component

Andrew Wen

School of Medicine

Oregon Health & Science University

Certificate of Approval

This is to certify that the Master's Thesis of

Andrew Wen

"Extraction of Clinical Lab Results: Design and Evaluation of a cTAKES Component"

Has been approved

_____

Dr. Steven Bedrick, Thesis Advisor

_____

Dr. Annette Totten

_____

Dr. Stephen Wu

Table of Contents

ABSTRACT

Objective: To algorithmically extract lab test and value pairs and to evaluate the performance of a machine learning-based solution to this task as compared to a rule-based solution.

Materials and Methods: Feature construction and annotation input/output were implemented as part of the cTAKES pipeline. The final feature space used in our implementation comprised of a combination of token and contextual word contents for both the lab test mention and lab value, as well as features identifying structure, such as distance between the test mention and value. A support vector machine (SVM) classifier was trained on 200 documents from the 2010 Informatics for Integrating Biology & the Bedside (i2b2) data set using the aforementioned feature space, a linear kernel, 10-fold cross validation, and a grid search for hyperparameters. The trained classifier was then run on a 50 document holdout data set, also from the 2010 i2b2 data set.

Results: We obtained extraction results of an F_1-score of 0.874 with precision 0.861 and recall 0.888. A rule-based approach to this extraction task designed for extraction from biomedical text obtained an F_1-score of 0.0897, with a precision of 0.0473 and recall of 0.8597.

Discussion: The performance of the support vector machine was significantly better than its rule-based counterpart when applied to extraction from clinical narratives, for which the rules in the rule-based solution were not specifically designed. We note several possible expansions on this work through which the performance of the support vector machine could be improved, as well as several important limitations within our study, namely the relatively small data set as well as limitations within the gold standard training/evaluation document set.

Conclusion: The additional generalizability offered by the machine learning approach along with its overall relative performance may be sufficient to warrant practical usage in applications where portability amongst different systems and input data sets may be of significant concern.

I. BACKGROUND AND SIGNIFICANCE

A. Introduction

Information contained in clinical records has been shown to be extremely useful both for supporting clinical tasks (1) and for research purposes (2). However, a vast portion of said information is stored electronically as unstructured free-text, which requires translation into semantic structures prior to computational use.

The problem of translating unstructured clinical free-text, particularly in the form of clinical narratives, into computable semantic structures has been one of the most significant barriers to making use of the wealth of information contained in the medical record. Historically, this extraction task was accomplished through manual reading and annotation of values of interest, but with the increasing volume of input data being available and necessary for practical applications, manual extraction is no longer a feasible approach (3). An active field of research has thus been the development of algorithmic extraction methods to support natural language processing (NLP), which is essential to making meaningful computational use of the information contained within unstructured medical text (4,5,6).

In this paper, we evaluate the performance of a machine learning-based approach to algorithmically extract one such example of useful information present in unstructured clinical text, patient laboratory test results.

B. Natural Language Processing

a. Introduction to NLP and Methods

NLP is a rapidly growing field dealing with computational modelling and understanding of human languages. This is done through the algorithmic extraction of useful syntactic and

semantic features so as to replicate features as used during the process of human understanding of language.

Approaches to NLP information extraction can be divided into two overarching paradigms: a symbolic rule-based method using hard-coded rules regarding the presence and structure of various features of the text (e.g. the textual content itself and/or other syntactic or semantic features as extracted by some other NLP extractor such as parts of speech tags), or a machine learning-based method through which statistical or logical models are used to dynamically generate the capability to extract features of interest.

Machine learning approaches can likewise be divided into two overarching implementation methods depending on what they seek to learn: rules or statistical models. Rule learning implementations seek to algorithmically identify extraction rules using supplied correct examples which are then used to perform future extraction tasks (7,8). Several common examples are learning classifier systems (9), association rule learning (10), and inductive logic programming (11).

In the case of model learning, either the parameters of various statistical models are identified and trained to tune the model so as to correctly identify items of interest, or the model itself is learned (12). Examples of this approach include Bayesian classifiers, conditional random fields, the support vector machine (SVM) (13), decision trees (14), and the *k*-nearest neighbors algorithm (15).

Both symbolic rule-based extraction and machine learning extraction have their own strengths and weaknesses. In their paper on the importance of rule-based approaches in industry applications, Chiticariu et al. acknowledge that that while offering excellent performance within

the specific domains and document corpora for which they are encoded, rule-based approaches to information extraction tend to suffer from either a rapid decrease in performance or an exponential increase in cost due to encoding cost of the sheer volume of new rules needed to maintain existing performance levels (16), a problem that is likewise reflected in biomedical rule-based extraction implementations (17). A solution for the problem of generalizability in rule-based approaches is the use of machine-learning techniques (18,19); rather than relying on encoded domain-specific rules, extraction tasks can be done by either algorithmically learning the rules or through statistical models (13,20).

On the other hand, machine learning approaches tend to suffer from worse performance compared to a properly specialized rule-based approach, and there exists some opaqueness in terms of human attempts at debugging should performance issues arise. Similarly, an adjustment to a rule-based approach simply involves modification or encoding of a rule, whereas modification for a machine learning approach involves retraining on a new data set, which takes time. More importantly, as rule-based methods are encoded by domain experts, they do not require input data to function properly, whereas machine learning does, which can be an issue in regards to performance in data-scarce environments (See: **§I.E - Data Sources**) (21).

There have also been proposals to implement a hybrid approach for some extraction tasks, where the strengths of a rule-based approach are used to extract some features that are a prerequisite to support machine learning implementations where representations may differ widely amongst different document corpora (22). In this study, we seek to evaluate the performance of the machine learning portion of one such hybrid approach.

As clinical NLP tends to be done in a pipeline, with additional extractors building upon the extracted features from previous extractors in the pipeline on a given text, several NLP pipelines

exist with plug-and-play extractor functionality to simplify interoperability between different components in the pipeline, notably Apache's Unstructured Information Management Architecture (UIMA) (23).

b. Challenges in Clinical NLP

Several challenges exist for clinical natural language processing applications that are not present in natural language processing for general biomedical language. Notably, biomedical text is typically written with clarity to a general audience in mind, whereas a clinical narrative is typically written by clinicians for clinicians. As a partial consequence to this distinction, a study by Leaman et al. (24) outlined several common complications in clinical narratives as compared to typical biomedical text, notably commonly missing punctuation, parenthetical expressions, significant format differences, unusual part-of-speech combinations (e.g. "Head, eyes, ears, nose, and throat examination revealed normocephalic and atraumatic."), missing words for the sake of brevity, and a much richer vocabulary, including significant use of jargon and acronyms.

These issues are significant as many clinical feature extraction implementations rely on other NLP features first being accurately extracted (25,26). For this reason, it would be beneficial for NLP extractors of clinical information to be built off of other NLP components and pipelines that have specifically been tailored for extraction from the clinical domain.

To that end, several NLP pipelines with components specifically tailored for clinical extraction have been developed for research and practical purposes, including the Apache Clinical Text Analysis and Knowledge Extraction System (cTAKES) (27), which utilizes the Apache UIMA framework. Other existing medical NLP pipelines that were not used as part of this study include

MedLEE (28), MedKAT/p (29), BioMedICUS (30), and ONYX (31). With some adaptation, similar implementation should be possible on many of these systems.

C. Laboratory Test Result Extraction – Significance and Past Work

One of the medical semantic features that have been found to be potentially useful in both clinical and research applications are patient laboratory test mentions and their result values. For instance, within the sentence fragment "…observed a hct of 43.3 mg/dL and platelet count of 450,000…", we would be interested in extracting lab test mention/value pairs, in this case "hct" of "43.3 mg/dL" and "platelet count" of "450,000". This information can then be used in a variety of both research and clinical applications, such as improving past medical history parsing in clinical decision support tools, as well as supporting cohort discovery and EHR analytics tasks.

Past work has been done on the problem of algorithmic extraction and relation of lab values: Kang et al. demonstrated an approach to solving this task using a rule-based approach on FDA decision summaries evaluating the performance or proposed laboratory/diagnostic devices (3). Hao et al. similarly explored using a rule-based method of value extraction and association for lab values as part of identifying HbA1c comparison statements (32).

While these rule-based approaches all reported excellent performance, with F_1-scores (which are a harmonic mean of recall and precision, intended to provide a general representation of overall performance) (33) in the 0.9+ range (3,32), it is important to note that these rulesets were not created for the purpose of extracting from clinical text. As both implementations are rule-based approaches, they are likely to have the aforementioned limitations in terms of generalizability, or the lack of code reusability in different applications. For this reason, we would expect a decline

in performance should these same rulesets be used in a document set for which they are not designed.

In this study, we seek to evaluate a hybrid approach to the problem of lab test/value extraction. We do not believe there is sufficient variance in lab test and numeric value representations between document sets to warrant extensive rule re-encoding. As such, we continue to extract individual mentions and numeric values using existing rule-based methods that have been shown to have high performance. On the other hand, syntax and structure that would be used to associate a test with its respective value tends to differ greatly from text to text (or be completely absent in some cases), and as such the generalizability issue present in rule-based approaches poses a problem to successful extraction implementations. We seek to solve this generalizability problem by instead using a machine learning approach in the association step.

D. Lab Value Extraction as Relation Extraction

The lab value extraction problem presented in this study consists of two sub-problems: first candidate lab tests and candidate lab values must be extracted, and then the two must be associated back to one another. As an example, consider the example sentence fragment used in **§I.C**, "…observed a hct of 43.3 mg/dL and platelet count of 450,000…". The first step in an algorithmic solution would thus be to extract the lab test ("hct" and "platelet count") and value ("43.3 mg/dL" and "450,000" mentions, and the second step would then be to perform association, e.g. determining that "43.3 mg/dL" corresponds to "hct" and not "platelet count".

For the purposes of this paper we assume that the first of these two sub-problems is a solved problem (Kang et al. demonstrated that a rule-based approach for entity/numeric value extraction tended to fare better than machine-learning based counterparts (3)), and instead focus on the

relation extraction task of associating the corresponding extracted lab test and lab value annotations with one another.

Machine learning approaches to relation extraction typically cast the problem into a classification task, where candidate relation instances are determined to either be related or not-related. There are several approaches to machine learning common in literature; broadly speaking they can be divided into supervised, semi-supervised, and unsupervised learning approaches, with the distinguishing factor being the level of annotated data available to the machine learning algorithm for training. As we wish to perform a classification task, the appropriate approach to take would be the supervised approach, where annotations denoting correct relations are provided to the algorithm for training purposes (34).

A common supervised learning approach to relation extraction, and the one used in this study, is the support vector machine (SVM). SVMs function by attempting to find a discriminatory "line", often referred to as a maximum-margin hyperplane, which can be drawn through the input data points to divide them into their corresponding classes (13). This division can then be used to predict the classification of future data points by finding its position relative to the classifying hyperplane.

For relation extraction purposes, this classification task is first supported by extraction of other syntactic or semantic features, e.g. part-of-speech tags and raw text of the two relation arguments and the contents in-between, which are extracted for a given argument pair and sent along with their associated classification, i.e. related or not related, as training data points to the classifier (17). In cases where the input data is not suitable for linear division, as is often the case in information extraction with complex feature spaces, the dimensionality of the input is increased via transformation of the input data points with a kernel function in the hopes of eventually

rendering the data separable (35). The relation extraction task then becomes a matter of evaluating the feature space of every possible combination of the two relation arguments within a given document and classifying the relation pair as either being related or not related.

E. Data Sources

While machine learning does not have the generalizability and cost of maintenance issues that are present with rule-based approaches, supervised and semi-supervised approaches do come at the cost of requiring annotated training data to function, which can be sparse and expensive to obtain. Efforts have thus been made by various organizations to support the production of annotated public data sets of biomedical/clinical text.

Of particular note is the organization Informatics for Integrating Biology & the Bedside (i2b2), which has hosted yearly NLP challenges related to some aspect of producing clinical document sets with annotated information, ranging from automatic de-identification of clinical notes to render them suitable for release (36), to challenges involving extracting syntactically or semantically useful features from clinical text such as medical concepts, assertions, relations (37,38), co-references (39), and various clinically relevant tasks (40,41). The document sets along with gold standards for all produced annotations resulting from i2b2 challenges were then released to the wider research community for public use.

F. Study Roadmap

This paper evaluates the performance of a support vector machine applied to part of the lab value extraction problem as a solution to the generalizability issues inherent to rule-based extraction approaches. Specifically, we evaluate the performance of the SVM classifier in performing the

relation extraction task of relating numeric values back to laboratory test mentions, and compare its performance to a rule-based method both in and outside of its intended domain.

In addition, we will validate that the decline in performance for an exclusively rule-based approach does actually occur when the input document set is changed to one for which the encoded ruleset is not designed.

II. METHODOLOGY

The pipeline used in this study is best described as four separate phases: generation of a training data set and gold standard for evaluation, generation of numeric value candidates, SVM relation extraction, and performance evaluation.

The Clinical Text Analysis Knowledge Extraction System (cTAKES) was selected as the engine used to generate NLP artifacts used in feature generation because its NLP components are specifically tailored for clinical text semantic artifact extraction (27,42). This may improve performance extracting syntactic and semantic features that we use during feature generation compared to other NLP solutions not tailored for the medical domain (See: **§I.B.b - Challenges in Clinical NLP**). As cTAKES is based on the Apache UIMA framework (23), we opted to use the ClearTK library (43) to support feature generation tasks.

For reference and reproduction purposes, code and limited portions of the annotated gold standard/training dataset (as permitted by data use agreements) are publicly supplied at https://www.github.com/andrew2060/ctakes-lab-value-extraction .

A. Phase I: Training Data and Gold Standard Construction

The i2b2 2010 dataset (37) was selected as a baseline data set from which to construct training annotations for this study. This particular dataset was selected due to the presence of laboratory test annotations generated as part of the original NLP challenge for that year (although respective values were not labeled) as well as the broad coverage in different fields of medicine within the provided clinical discharge summaries, ranging from ICU to neonatal care to routine visits. Construction of lab test and lab value relation annotations was done manually and stored in the same format as that of i2b2's original dataset, that is to say the textual value followed by a string

of the following format $annStartLine{:}\ annStartWord\ annEndLine{:}\ annEndWord$, with word indexes starting at 0 (refer to **Supplement 1 - Example Annotation Strings**). The resulting annotated i2b2 document set consisting of 250 doucments was then split into two data sets of 200 and 50 documents for training and testing purposes respectively.

The validity of the manually constructed gold standard and training data was obtained through a determination of inter-annotator agreement: a set of 20 documents was sent to a domain expert for annotation and the resulting annotations were compared with those of our manually created gold standard/training data so as to determine accuracy. Disagreements were noted but not corrected so as to preserve consistency as part of our evaluation. The relative number of agreements and disagreements were used to calculate an inter-annotator agreement percentage, which gave an estimate of how closely the gold standard and training data annotations reflected results that would actually be desired in a real-world application for that particular document set.

These manual annotations were then imported into the UIMA Java Common Analysis System for further processing (i.e. feature generation), and later for evaluation purposes, via a UIMA CollectionReader.

B. Phase II: Candidate Numeric Value Generation

The default cTAKES clinical pipeline was run twice: once prior to the training phase and again prior to the testing phase. The purpose of running this pipeline was to generate appropriate NLP artifacts for later usage.

Candidate lab values were extracted on a per-sentence basis after some textual preprocessing for normalization purposes. Numeric values were then detected using the regular expression filter

"$-?(?:\d*\.)?\d+$" while potential unit denotations were detected using the unit regular expression filter as described by Hao et al. (32)

Candidate lab values were then scored on their likelihood of being a lab value through a combination of the existence of an associated unit and the distance to the closest lab test annotation. Specifically, we selected values that had a maximum distance of five words to the closest lab test annotation or contained a measurement unit (see **§IV.C - Study Limitations and Error Analysis** for effects of this extraction method on overall performance metrics).

C. Phase III: Relation Extraction via Support Vector Machine

The SVM implementation used was that of LibSVM (44), as bundled within the ClearTK library for compatibility with cTAKES' UIMA based framework. Selected model parameters were obtained via grid-based search and 10-fold cross-validation (See: **Table 1 - SVM Training Parameters**) (45). Of the 256 documents in the i2b2 dataset, 200 were used for training purposes with an additional 50 set aside as an evaluation set. The remaining 6 documents were omitted due to incompatibilities with the manual annotation software used during training set creation (See: **Table 2 - Document Sets**).

Classification was a simple true/false problem where true (+1) was used to indicate the existence of a relation between a given lab test and numeric result, while a classification of false (-1) indicated no relation. To compensate for SVM performance being poor in cases where training and test data have unbalanced output classes (46), i.e. there are significantly more false classification examples being supplied than true classifications, SVM parameter weights of 1.0 for true and 0.02188 for false were selected. This selection was done based on the relative frequencies of true and false training feature sets within the i2b2-based training data set by

placing a significantly lesser penalty on incorrect classifications of false (which significantly outnumber true classifications in our dataset) so as to correspond to placing equal penalties on both true and false relation classifications were we to use a balanced data set. (47)

The training data fed to the SVM consisted of a set of instances of true and false relation instances. An instance was constructed for every possible lab test and lab value pair within a given document and consisted of a constructed feature space from the two relation arguments as well as a true or false classification on the existence of a relation between them. The features constructed for each (lab test, lab value) instance are as follows (refer to **Supplement 2 - Example Features** for examples as applied to actual text):

- Token Text Features:
    - The first and last word present in each lab test and value annotation
    - A bag of words of the covered text of each lab test and value annotation
    - The preceding and following 3 words for each annotation, stored by their relative positions and not as a bag
- Window Features:
    - The distance in words between a candidate lab test and lab value
    - The number of candidate lab values that are closer to the candidate lab test than the value being tested
    - A bag of the words within the window between the candidate lab test and lab value
- Syntax Features:
    - A bag of part of speech tags (as generated by cTAKES' preexisting components (27,48)) for every word in both candidate lab test and lab value

- Directionality: whether a lab test or lab value appeared first in the sentence of the lab test in the relation being evaluated (-1 if lab value, 1 if lab test)

Candidate lab test/value pairs that were classified as being relevant to one another from the holdout testing 50 document set were then stored using the existing cTAKES type system as a LabMention and MeasurementAnnotation with the relation stored as a ResultOfTextRelation (42).

D. Phase IV: Evaluation

Evaluation was done through position matching where a positive hit for an identified lab test/lab value pair within the test data set had the same or overlapping word index in the document as the respective gold standard test/lab annotations.

Generally speaking, this matching procedure is done by checking for overlaps in character spans, with an overlap counting as a match. This method was chosen as opposed to full word matching due in part to architectural limitations within cTAKES itself, as well as due to certain irregularities within the i2b2 document generally resulting from spacing issues. For example, lab test and their results may be mashed together into a single word and extraneous formatting elements like bullet points becoming part of the result value word. In such cases, it is possible that a positive match can correspond to only part of a word in the gold standard set, or even that an extracted lab test and lab value both correspond to the same word within the gold standard text (see **§IV.C - Study Limitations and Error Analysis PP. 1**)

To generate a positive hit, for each extracted relation instance, the argument lab test must collide with a gold standard lab test, the argument lab value must collide with a gold standard lab value, and a relation must exist within the gold standard between the lab test and lab values that were

identified from the collision check. From this we generate the true positive (TP), false negative (FN), and false positive (FP) statistics.

Performance evaluation was done through the use of recall (A), a representation of the ability of the extractor to successfully extract an existing relation, precision (B), the positive predictive value or a measurement of how many predicted relations actually existed, and f_1-score (C), a harmonic mean of both precision and recall intended to model overall retrieval ability.

$$A.\ Recall = \frac{TP}{TP + FN}$$

$$B.\ Precision = \frac{TP}{TP + FP}$$

$$C.\ F_1 - Score = 2 * \frac{(Precision * Recall)}{Precision + Recall}$$

Result statistics were then compared to existing rule-based systems by running the withheld 50 document test data set through a reimplementation of the rule-based solution proposed by Hao et al. (32) with the same ruleset (see **§II.F - Comparative Evaluations**).

E. Feature Discovery Iterations

While not strictly necessary to reproduce results, we wish to note that the final feature set as was proposed above was not the initial feature set used when we first began conducting this study. Rather, we started with simply the token text features and gradually added new features through an iteration process between phase III and IV, identifying potentially useful features by examining false negax1tive and false positive extracted instances. Such a process would likely be similarly useful should one wish to further expand on this work by identifying additional useful features to improve performance.

F. Comparative Evaluations

An additional evaluation was run on the withheld 50 document data set to establish a comparative baseline with a rule-based approach. While Hao et al. did report excellent performance with an f_1-score of 0.98, it must be noted that said performance metrics were obtained from running the encoded rules upon clinical trial eligibility criteria: a document corpus for which the encoded rules were specifically designed. To establish a baseline for comparison, we re-ran those same rules on the withheld 50 document evaluation data set from the 2010 i2b2 corpus using lab tests as annotated in the data set and candidate lab values as extracted during phase II. The performance of the reimplementation of this rule-based extractor was then evaluated using the same procedure as that for the SVM approach (see **§II.D - Phase IV: Evaluation**).

III. RESULTS

Inter-annotator agreement as determined from **§II.A - Phase I: Training Data and Gold Standard Construction** to validate the manually created gold standard/training annotations was 98.13%, as found by an automated position matching scheme between the results from each different annotator and using the same criteria for positive hits as in the evaluation section.

Grid search during SVM training yielded a linear kernel as having the best performance as compared to the other library options of polynomial, RBF and sigmoid kernels.

Precision, recall, and F_1-score statistics as compared with gold standard annotations are summarized in **Table 3 - Results and Evaluation Comparison** for both the SVM approach using a linear kernel and for the rule-based approach presented by Hao et al (32), both as reported on trial eligibility criteria as well as on the i2b2 document set.

We found that the support vector machine approach applied to the i2b2 document set attained performance metrics of a f_1-score of 0.874, recall of 0.888, and precision of 0.861. We also found that the same ruleset used by Hao et al. (32) for our simulation of a rule-based approach suffered a significant decrease in performance when applied to the i2b2 data set with a f_1-score of only 0.0897, recall of 0.8597 and precision 0.0473, as opposed to the reported performance metrics of reported f_1-score of 0.980 with precision 0.989 and recall 0.971 when applied to the intended document corpus (clinical eligibility criteria text).

IV. DISCUSSION

A. Result Analysis

As the evaluation results were calculated based on agreement with a constructed gold standard, it was important that said gold standard contained all possible lab test and value pairs so as to model the performance of the extractor in a practical application. In our evaluation of the accuracy of our gold standard data set, inter-annotator agreement was sufficiently high such that a high level of confidence in the accuracy of the gold standard data set exists; the percentage of non-matching annotations, with the assumption a similar rate throughout the rest the data set, is sufficiently low so as to have had a minimal impact on the model training process. Nevertheless, we believe that any errors made in annotating the gold standard and training data/discrepancies in inter-annotator agreement should be viewed as insignificant so long as the errors are consistent throughout the dataset: it is expected that any errors made here will have a minimal impact as extraction of erroneously marked positives and negatives in the training and evaluation data, would still demonstrate the feasibility of a SVM in retrieving variable/value associations in a practical context with non-erroneous training data.

We found that Hao et al's rule-based methodology suffered a significant decline in performance when applied to the i2b2 data set. In particular we found that Hao et al's ruleset suffered in precision due to a significantly higher density of lab test label/value pairs in close proximity to one another as well as a lack of separating words and punctuation in between that would be present in more formal text. For example, the segment "BLOOD WBC 17.1 RBC 3.37 Hgb 10.1 Hct 30.1 MCV 89 MCH 30.0 MCHC 33.6 RDW 18.8 Plt Ct 526" presented significant difficulty to Hao et al's ruleset due to the lack of structure and formatting clues that would be present in most other texts, e.g. "WBC of 17.1, RBC of 3.37…" or "WBC: 17.1 RBC: 3.37…".

On the other hand, we also found that the support vector machine was able to perform the relation extraction task reasonably well, with a f_1-score of 0.874, recall of 0.888, and precision of 0.861. While this is slightly inferior performance compared to the results obtained by Hao et al when running the rule-based extractor on the clinical trial eligibility text for which it was originally designed, we believe that the performance of the machine learning approach is still high enough to warrant practical use in many applications. More importantly, the drastic increase in performance compared to a rule-based approach applied out of domain highlights the benefits of leveraging the re-usability of the machine learning approach; its flexibility allows it to be applied to a much wider variety of input corpora without the need for additional encoding, and as such would be more suited for general purpose applications.

In terms of classification errors, we found no discernable common pattern amongst false positive identifications, but we found that many of the false negative identifications occurred in situations where there were scarce examples of the syntactic structure involved amongst the overall dataset, for instance cases where the laboratory test mention and resulting value occurred on separate lines of text. In such cases, an expanded training document set with more examples of these scarce structures would likely improve performance.

It is worth noting that in cases where sparse amounts of source material are available, as is often the case for biomedical and clinical texts, it may be preferable to sacrifice precision in favor of recall as unextracted instances are undesirable. It may therefore be worthwhile for practical applications of this work to modify the regularization parameters (**§II.C - Phase III: Relation Extraction via Support Vector Machine, Table 1 - SVM Training Parameters**) to reflect this preference by inducing less of a penalty for false positives on the training algorithm so as to

reduce the false negative rate, as opposed to the current settings used in this study that places equal importance on both true and false classifications.

B. Feature Space Analysis

Overall, we feel that our selected feature space is fairly discriminatory for the task at hand. In this section, we outline several ways in which our selected feature set can help discriminate between relations and the lack thereof.

For the token text feature space, we found that both the contents of the lab test mention and values were useful: the contents of the annotations were useful as certain lab test annotations, e.g. “hct”, “bp”, or “t”, would generally have an associated numeric lab value in our data set while others, e.g. “an electrocardiogram”, would not. Contextual information surrounding the lab test and lab value annotations also served to identify structural information. Words such as “respectively” and “each” help indicate the likely structure and correct associations within a sentence.

Window features were especially important for successful extraction: typically, lab tests and their respective values would have a small window size, with rare exceptions that could be gleaned from other features within our used feature space. Usage of the window size feature removed a majority of the initial false positive associations. Furthermore, a portion of the remaining false positives were eliminated using the number of closer values feature. With the exception of certain lab tests and/or the presence of the word “respectively” in the immediate context, the correct lab value associated with a given lab test tended to be the closest or second closest.

The sign of the window was also useful for distinguishing tasks by providing information on whether the value preceded or followed the lab test mention (see comment on directionality further in this section). Furthermore, we used the textual contents of the window as indicators for the existence of a relation, e.g. "{test} of {value}".

Finally, we had several useful features taking advantage of syntactic features. For instance, as lab tests tend to be nouns, values tend to be cardinal numbers and nouns/symbols (for units), the presence of certain parts of speech as part of the lab test mention (such as verbs) decreased the probability that this is a lab test mention associated with a numeric value. The directionality of test/value pairs tended to be preserved throughout a sentence: for instance, "{value} {lab}, {value} {lab}, …. " as opposed to "{lab}: {value}, {lab}: {value} … ". This feature was potentially useful in situations where multiple lab test/value relations existed in close proximity to one another with similar window sizes in either direction, for instance denotations of a patient's vital signs, where readings were listed one after another with no punctuation in between.

C. Study Limitations and Error Analysis

One of the key limitations of this study were occasional flaws in the gold standard data set as provided by i2b2 due to inconsistent formatting. The annotation representation used could only denote whole words but occasionally supplied lab test would also contain the respective value. For instance, instead of "* WBC – 140,000", the representation supplied from I2B2 would actually be "*WBC-140,000". In these cases, the gold standard and training document sets would contain annotations covering the same (whole) word for both the laboratory test mention and respective value. Additionally, some entities, like blood pressure, may correspond to multiple values within the same "word" as defined by i2b2 annotators. We believe that our overlap-based

system of evaluation should have mitigated most of these incompatibilities but it is likely that a small portion of these errors due to limitations in formatting exist. That being said, these occurrences comprise a small (<1%) portion of the total correct annotations so we anticipate this having a minimal effect on the overall results.

It should be noted that our evaluation method did not account for any errors in numeric lab value extraction: this is actually done purposefully as these same errors will likewise be present in the rule-based baseline from Hao et al. that we use for evaluating relative performance compared to the baseline. The results in an ideal situation with gold-standard lab values being used for the relation extraction task are included separately in **Table 3 - Results and Evaluation Comparison** and was shown to have marginally inferior precision but better recall and overall performance than with the extracted lab values. We would expect a perfect extraction system to offer superior recall for the overall extraction task in combination with any other relation extraction method, albeit having a minor effect (if any) on precision.

One possible concern with this study was the size of the document set used: the training and evaluation 250 document set could be considered small enough such that external factors may have played a role in overall performance of the SVM relation extractor, notably overfitting issues on both for both the training and evaluation data sets. We compensated for this possibility by using 10-fold cross-validation during parameter search as well as running evaluation on a holdout data set.

Furthermore, while the 2010 i2b2 document set covered a breadth of topics ranging from neonatal care to chronic condition management, they were all similar in format because they were all clinical encounter notes (discharge summaries and progress reports). This again calls into question the issue of possible overfitting as we would expect that any algorithmic solution

be applied to a wide variety of possibilities in input document types beyond just clinical notes, e.g. the clinical trial eligibility criteria text used by Hao et al (32). However, one of the main strengths to the machine learning approach is its relative flexibility: simply adding documents of other types to the training set and retraining should, in theory, resolve this issue.

The observed performance of the support vector machine approach was potentially inferior to its actual potential performance due to limitations within the training dataset. The i2b2 document set was sparse in its inclusion of units for lab values (e.g. mg/dL, kg, etc). Because a relationship does exist between lab tests and the units associated with the value (a birth weight will typically be in kilograms or pounds, blood urea nitrogen and creatinine will typically be mg/dL or mmol/L, etc.), it is likely that rerunning SVM training on a document set that contains more examples of correct unit associations will yield better performance on the sparse instances where unit values are included, and would especially increase performance when prediction is run on documents sets that make ample use of unit values.

Finally, comparative statements were not extracted as part of this study, despite them containing clinically useful information. Information on whether a measurement is less than, greater than, or equal to a certain value can be considered clinically significant. There was an insufficient amount of comparative statements within the i2b2 data set to perform an evaluation on whether a SVM classifier is able to perform such a task. Nevertheless, it should be possible to perform this task by retraining using the same or very similar feature space with additional emphasis on the window contents feature, as it is essentially the same relation extraction task with additional classification classes involved beyond a simple true/false as determined by the content between the variable and the value.

D. Expansions on this Work

That reasonable performance was attained with a machine learning approach indicates that the same codebase is likely reusable with similar performance on other document sets for the same purpose, given appropriate training examples. Nevertheless, one of the greatest weaknesses of a supervised learning approach such as a SVM is its need for annotated training data, which can often be rather limited in volume. Creation of a greater variety of annotated document sets from a greater variety of sources/scopes would thus be helpful for additional validation purposes with an expanded dataset. Alternatively, another possibility would be the expansion of the supervised learning approach taken in this study to a semi-supervised learning approach by including unannotated documents as part of the training data. Such an approach has been shown to potentially improve the performance of supervised learning methods (49,50) and also mitigates the issue of data sparseness.

Beyond additional experimentation with an expanded document set to further validate results, this task of lab test and lab value association can be generalized to a relation extraction problem between any variable and its respective value. It may therefore be useful to investigate a SVM classifier being used for variable and value association in general as opposed to merely for lab tests/values, one possible application being drug named entity and prescription dosage association.

Finally, feature discovery is an open problem. While the feature space selected in this study were sufficient to attain reasonable results, further investigation into possible features may yield worthwhile performance gains and thus narrow the gap between the performance of in-domain rule-based extraction and machine learning approaches. Several promising features that were not

investigated in this study were dependency graphs and word2vec skip-grams (51), both of which have shown promising results for other, more complex, relation extraction tasks.

V. SUMMARY AND CONCLUSIONS

The sheer volume of health data and the ever-increasing need for information alongside exorbitant time and cost requirements for manual extraction of clinical semantic structures has rendered manual annotation unsustainable as a production-level solution to modern information needs. Algorithmic extraction of semantic features in support of NLP has been the proposed solution to this problem. Broadly speaking, NLP follows two general strategies, rule-based and machine learning-based extraction, or some combination thereof. While rule-based methods tend to offer excellent in-domain performance, they have a high associated cost for rule encoding and lack the ability to be re-used for similar problems outside of their intended domain.

In this paper, we evaluated using a machine learning approach termed a support vector machine to associate numeric lab values back to their respective lab tests, and found that despite a relative decrease in performance compared to a rule-based approach for in-domain extraction tasks, the performance was still sufficiently high for many practical applications. The drastic improvement in performance compared to rule-based methods out of domain, however, may render a SVM based approach for lab test and lab value association more desirable for applications designed to accept a wide variety of potential inputs. Furthermore, we note that this relation extraction task can very likely be generalized to a wide variety of similar variable association tasks, such as association of drug labels with dosages thus reducing costs associated with encoding domain-specialized rulesets.

VI. APPENDICES

Table 1 - SVM Training Parameters

SVM Parameter Name (-argName)	*Parameter Value*
Kernel Type (-t)	Linear (0)
k-Fold Cross-Validation Setting (-v)	10
SVM Type (-s)	C-SVC (0 – default)
Imbalanced dataset penalty/regularization parameters (C in C-SVC algorithm (44))	True(-w+1): 1.0 False(-w-1): 0.02188
Shrinking Heuristics (-h)	Disabled (0 – Library Recommended for Small Iteration Count)

Table 2 - Document Sets

Document Set	*I2b2 Document IDs*
Training	0001 through 0410 (except omitted)
Evaluation	0412 through 0477
*Omitted (**§II.C**)*	0033, 0125, 0149, 0257, 0313, 0393

Table 3 - Results and Evaluation Comparison

	Precision	**Recall**	**F_1-Score**
Hao et al. – Reported Rule-based Results (32) (Clinical Trial Eligibility Criteria)	0.989	0.971	0.980
Hao et al. – Rule-based Results (i2b2 data)	**0.0473**	**0.8597**	**0.0897**
Support Vector Machine – Linear Kernel (i2b2 data)	**0. 8615**	**0. 8882**	**0. 8746**
Support Vector Machine – Linear Kernel w/ Gold Standard Lab Values	0.8576	0.9870	0.9177

Supplement 1 - Example Annotation Strings

Line 111: wbc was 9.1

c="wbc" 111:0 111:0||v="9.1" 111:2 111:2

Line 62: o2sat of 100%

c="o2sat." 62:0 62:0||v="100%" 62:2 62:2

Line 65: w0 w1 w2 w3 w4 w5 w6 w7 phosphate of 1.6

c="phosphate" 65:8 65:8||v="1.6" 65:10 65:10

Supplement 2 - Example Features

Document Text:

- On admission included BUN and creatinine of 33 and 2.1 mg/dL respectively

Correctly extracted relations:

- BUN <-> 33
- Creatinine <-> 2.1 mg/dL

Extracted feature space for Creatinine <-> 2.1 mg/dL

Token/Covered Text Features:

- ARG1_FIRST: Creatinine, ARG1_LAST: Creatinine, ARG2_FIRST: 2.1, ARG2_LAST: dL
- BAG_ARG1: {Creatinine}, BAG_ARG2: {2.1, mg, /, dL}
- ARG1_PRECEDING_3: included, ARG1_PRECEDING_2: BUN, ARG1_PRECEDING_1: and ARG1_FOLLOWING_3: and, ARG1_FOLLOWING_2: 33, ARG1_FOLLOWING_1: of ARG2_PRECEDING_3: of, ARG2_PRECEDING_2: 33, ARG2_PRECEDING_1: and ARG2_FOLLOWING_3: null, ARG2_FOLLOWING_2: null, ARG2_FOLLOWING_1: respectively

Window Features:

- WINDOW_SIZE: 3
- CLOSER_VALUE_COUNT: 1
- WINDOW_CONTENT_BAG: {of, 33, and}

Syntax Features:

- ARG1_POS_BAG: {NN}, ARG2_POS_BAG {CD, NN, SYM}
- DIRECTIONALITY: 1

VII. REFERENCES

1. Berner ES, Detmer DE, Simborg D. Will the Wave Finally Break? A Brief View of the Adoption of Electronic Medical Records in the United States. Journal of the American Medical Informatics Association. 2005 January; 12(1).

2. Miller RA. Medical Diagnostic Decision Support Systems—Past, Present, and Future: a Threaded Bibliography and Brief Commentary. Journal of the American Medical Informatics Association. 1994 January; 1(1).

3. Kang YS, Kayaalp M. Extracting laboratory test information from biomedical text. Journal of Pathology Informatics. 2013 August; 4(23).

4. Cao H, Stetson P, Hripcsak G. Assessing Explicit Error Reporting in the Narrative Electronic Medical Record Using Keyword Searching. Journal of Biomedical Informatics. 2003 February-April; 36(1-2).

5. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. Journal of Biomedical Informatics. 2001 October; 34(5).

6. Chowdhury G. Natural language processing. Annual review of information science and technology. 2003; 37(1): p. 51-89.

7. Riloff E. Automatically generating extraction patterns from untagged text. In Proceedings of the 13th National Conference on Artificial Intelligence; 1996. p. 1044-1049.

8. Bassel GW, Glaab E, Marquez J, Holdsworth MJ, Bacardit J. Functional Network Construction in Arabidopsis Using Rule-Based Machine Learning on Large-Scale Data Sets. The Plant Cell. 2011 September; 23(9).

9. Urbanowicz RJ, Moore JH. Learning Classifier Systems: A Complete Introduction, Review, and Roadmap. Journal of Artificial Evolution and Applications. 2009 June; 2009.

10. Agrawal R, Imieliński T, Swami A. Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD international conference on Management of data; 1993; Washington, D.C.: Springer-Verlag. p. 207-216.

11. Muggleton S, Raedt Ld. Inductive Logic Programming: Theory and methods. The Journal of Logic Programming. 1994 May-July; 19-20(Supplement 1).

12. He Y, Kayaalp M. Biological Entity Recognition with Conditional Random Fields. In AMIA Annual Symposium Proceedings; 2008. p. 293-297.

13. Cortes C, Vapnik V. Support-vector networks. Machine Learning. 1995 September ; 20(3).

14. Schmid H. Probabilistic part-of-speech tagging using decision trees. In New methods in language processing; 2013. p. 154.

15. Altman NS. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. The American Statistician. 1992 February; 46(3).

16. Chiticariu L, Li Y, Reiss FR. Rule-based information extraction is dead! long live rule-based information extraction systems. In Empirical Methods in Natural Language Processing; 2013. p. 827-832.

17. Zhou D, Zhong D, He Y. Biomedical Relation Extraction: From Binary to Complex. Computational and Mathematical Methods in Medicine. 2014; 2014.

18. Bunescu R, Ge R, Kate R, Marcotte E, Mooney R, Ramani A, et al. Comparative experiments on learning information extractors for proteins and their interactions. Artificial Intelligence in Medicine. 2005 February; 33(2).

19. Mooney R, Bunescu R. Mining knowledge from text using information extraction. ACM SIGKDD Explorations Newsletter - Natural language processing and text mining. 2005 June; 7(1).

20. Lafferty J, McCallum A, Pereira F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proceedings of the 18th International Conference on Machine Learning; 2001; San Francisco. p. 282-289.

21. Langley P, Simon HA. Applications of machine learning and rule induction. Communications of the ACM. 1995 November; 38(11).

22. Jiang M, Chen Y, Liu M. Hybrid approaches to concept extraction and assertion classification - vanderbilt's systems for 2010 I2B2 NLP Challenge. In Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data; 2010; Boston.

23. Ferrucci D, Lally A. UIMA: an architectural approach to unstructured information processing in the corporate research environment. Journal of Natural Language Engineering. 2004 September; 10(3-4).

24. Leaman R, Khare R, Lu Z. Challenges in Clinical Natural Language Processing for Automated Disorder Normalization. Journal of Biomedical Informatics. 2015 October; 57(2015).

25. Savova G, Ogren P, Duffy P, Buntrock J, Chute C. Mayo Clinic NLP System for patient smoking status identification. Journal of the American Medical Informatics Association. 2008; 15(1).

26. Sun W, Rumshisky A, Uzuner O. Annotating temporal information in clinical narratives. Journal of Biomedical Informatics. 2013 December; 46(0).

27. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. Journal of the American Medical Informatics Association. 2010 September; 17(5).

28. Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated Encoding of Clinical Documents Based on Natural Language Processing. Journal of the American Medical Informatics Association. 2004; 11(5).

29. Coden A, Tanenblatt M. The MedKAT Pipeline. [Online]. [cited 2017 9 7. Available from: http://ohnlp.sourceforge.net/MedKATp/.

30. Knoll B, Finley G, Wang Y, Pakhomov S. nlpie/biomedicus: 1.7.0. [Online].; 2017 [cited 2017 July 24. Available from: http://doi.org/10.5281/zenodo.834326.

31. Christensen L, Harkema H, Haug P, Irwin J, Chapman W. ONYX: a system for the semantic analysis of clinical text. In Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing; 2009. p. 19-27.

32. Hao T, Liu H, Weng C. Valx: A system for extracting and structuring numeric lab test comparison statements from text. Methods of Information in Medicine. 2016; 55(3).

33. Manning CD, Schutze H, Raghavan P. Introduction to information retrieval: Cambridge University Press; 2008.

34. Alpaydin E. Introduction to Machine Learning. 2nd ed.: MIT Press; 2010.

35. Vert J, Tsuda K, Schölkopf B. A Primer on Kernel Methods. In Kernel Methods in Computational Biology.; 2004. p. 35-70.

36. Uzuner Ö, Juo Y, Szolovits P. Evaluating the state-of-the-art in automatic de-identification. Journal of the American Medical Informatics Association. 2007 June; 14(5).

37. Uzuner Ö, South BR, Shen S, Duvall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. Journal of the American Medical Informatics Association. 2011 June; 18(5).

38. Sun W, Rumshisky A, Uzuner Ö. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. Journal of the American Medical Informatics Association. 2013 December; 20(5).

39. Uzuner Ö, Bodnari A, Shen S, Forbush T, Pestian J, South BR. Evaluating the state of the art in coreference resolution for electronic medical records. Journal of the American Medical Informatics Association. 2012 February; 19(5).

40. Stubbs AKC, Xu H, Uzuner Ö. Identifying risk factors for heart disease over time: Overview of 2014 i2b2/UTHealth shared task Track 2. Journal of Biomedical Informatics. 2015 July; Supplement(S67-77).

41. Uzuner Ö, Goldstein I, Luo Y, Kohane I. Identifying patient smoking status from medical discharge records. Journal of the American Medical Informatics Association. 2008; 15(1).

42. Wu ST, VCK, Dligach D, Masanz JJ, Chen P, Becker L, et al. A common type system for clinical natural language. Journal of Biomedical Semantics. 2013 January; 4(1).

43. Bethard S, Ogren P, Becker L. ClearTK 2.0: Design Patterns for Machine Learning in UIMA. In Proceedings of the Ninth International Conference on Language Resources and Evaluation; 2014; Reykjavik. p. 3289-3293.

44. Chang CC, Lin CJ. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology. 2011; 2(3).

45. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In International Joint Conference on Artificial Intelligence; 1995. p. 1137-1145.

46. Wu G, Chang E. Class-Boundary Alignment for Imbalanced Dataset Learning. In ICML 2003 workshop on learning from imbalanced datasets II; 2003; Washington, DC. p. 49-56.

47. Morik K, Brockhausen P, Joachims T. Combining statistical learning with a knowledge-based approach - a case study in intensive care monitoring. In International Conference on Machine Learning; 1999; Bled. p. 268-277.

48. Marcus MP, Marcinkiewicz MA, Santorini B. Building a large annotated corpus of English: The Penn Treebank. Computational linguistics. 1993 June; 19(2).

49. Li Y, Hu X, Lin H, Yang Z. Learning an enriched representation from unlabeled data for protein-protein interaction extraction. BMC Bioinformatics. 2010; 11(Supplement 2).

50. Mintz M, Bills S, Snow R, Jurafsky D. Distant supervision for relation extraction without labeled data. In Proceedings of the 2009 Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP; 2009; Suntec. p. 1003-1011.

51. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed Representations of Words and Phrases and Their Compositionality. In Proceedings of the 26th International Conference on Neural Information Processing Systems; 2013; Lake Tahoe. p. 3111-3119.