J Pathol Inform

Editor-in-Chief:
Anil V. Parwani , Pittsburgh, PA, USA
Liron Pantanowitz, Pittsburgh, PA, USA

OPEN ACCESS
HTML format

For entire Editorial Board visit : www.jpathinformatics.org/editorialboard.asp

Research Article

Support patient search on pathology reports with interactive online learning based data extraction

Shuai Zheng[1], James J. Lu[1], Christina Appin[2], Daniel Brat[2], Fusheng Wang[3]

[1]Department of Mathematics and Computer Science, Emory University, [2]Department of Pathology and Laboratory Medicine, School of Medicine, Emory University, Atlanta, GA 30322, [3]Departments of Biomedical Informatics and Computer Science, Stony Brook University, Stony Brook, NY 11794, USA

E-mail: * Dr. Fusheng Wang - fusheng.wang@stonybrook.edu and Dr. James J. Lu - jlu@mathcs.emory.edu
*Corresponding author

Received: 03 April 2015, Accepted: 06 June 2015, Published: 28 September 2015

Abstract

Background: Structural reporting enables semantic understanding and prompt retrieval of clinical findings about patients. While synoptic pathology reporting provides templates for data entries, information in pathology reports remains primarily in narrative free text form. Extracting data of interest from narrative pathology reports could significantly improve the representation of the information and enable complex structured queries. However, manual extraction is tedious and error-prone, and automated tools are often constructed with a fixed training dataset and not easily adaptable. Our goal is to extract data from pathology reports to support advanced patient search with a highly adaptable semi-automated data extraction system, which can adjust and self-improve by learning from a user's interaction with minimal human effort. **Methods**: We have developed an online machine learning based information extraction system called IDEAL-X. With its graphical user interface, the system's data extraction engine automatically annotates values for users to review upon loading each report text. The system analyzes users' corrections regarding these annotations with online machine learning, and incrementally enhances and refines the learning model as reports are processed. The system also takes advantage of customized controlled vocabularies, which can be adaptively refined during the online learning process to further assist the data extraction. As the accuracy of automatic annotation improves overtime, the effort of human annotation is gradually reduced. After all reports are processed, a built-in query engine can be applied to conveniently define queries based on extracted structured data. **Results:** We have evaluated the system with a dataset of anatomic pathology reports from 50 patients. Extracted data elements include demographical data, diagnosis, genetic marker, and procedure. The system achieves F-1 scores of around 95% for the majority of tests. **Conclusions:** Extracting data from pathology reports could enable more accurate knowledge to support biomedical research and clinical diagnosis. IDEAL-X provides a bridge that takes advantage of online machine learning based data extraction and the knowledge from human's feedback. By combining iterative online learning and adaptive controlled vocabularies, IDEAL-X can deliver highly adaptive and accurate data extraction to support patient search.

Key words: Controlled vocabularies, data extraction, online machine learning, pathology reports, patient search

Access this article online

Website: www.jpathinformatics.org

DOI: 10.4103/2153-3539.166012

Quick Response Code:

This article may be cited as:
Zheng S, Lu JJ, Appin C, Brat D, Wang F. Support patient search on pathology reports with interactive online learning based data extraction. J Pathol Inform 2015;6:51.

Available FREE in open access from: http://www.jpathinformatics.org/text.asp?2015/6/1/51/166012

INTRODUCTION

Pathology reports contain valuable research information embedded in narrative free text. The same information in structured format can be used to support clinical findings, decision making and biomedical research. Synoptic reporting[1-3] has become a powerful tool for providing summarized findings through predefined data element templates such as CAP Cancer Protocols.[4] Meanwhile, standard groups such as IHE are proposing structured reporting standards such as Anatomic Pathology Structured Reports[5] in Health Level Seven. While there is a major trend for structured reporting, the vast amount of pathology reports remain unstructured in legacy systems. And standardization efforts only capture major data elements, leaving a substantial amount of valuable information in free text that is difficult to process and search.

Information extraction is a technique that can generate structured representation of important information from pathology reports. The transformed data may be used to search easily for patient groups with certain traits as in, for example, find all patients with an age above 40 years old and that have a diagnosis glioma. Figure 1 shows a typical workflow of data extraction from pathology reports.

Previous work on data extraction from pathology reports addresses various tasks and different research problems. caTIES supports coding for surgical pathology reports.[6] A regular expression is used to mine specimens and related information in,[7] MedTAS/P extracts and represents cancer diseases from pathology reports with the hierarchical model.[8] Lupus represents extracted information with Semantic Web techniques.[9] NegEx is adopted to detect negation for annotating surgical pathology report.[10] These systems either employ rules engineered to specific topics and domains or they use statistical models learned in batch from manually annotated training data. The first approach lacks generalizability; new rules need to be designed and developed for each domain. The second approach based on machine learning is more flexible. But obtaining accurate training data can be costly and time-consuming.

We present a system, IDEAL-X, which combines online machine learning and customizable vocabularies to provide a generic, easy-to-use solution for clinical information extraction. Online machine learning[11-13] takes an iterative learning approach through interactive human intervention, the data extraction engine of IDEAL-X automatically predicts answers to annotate reports, gradually learns from human's feedback, and incrementally improves its accuracy. Compared to traditional batch training based algorithm, which requires pretraining with a reasonably large dataset, online learning based algorithms can significantly reduce human effort on labeling training data and provide the possibility of updating the learning models dynamically to fit a continually changing data environment. To enhance its performance, IDEAL-X supports adaptive vocabulary to support data extraction. A user can customize a controlled vocabulary, which could be continuously adjusted during online learning process. Once structured data elements are extracted, a query interface is provided to support patient search with filtering conditions on data elements.

METHODS

IDEAL-X consists of five major parts: Extraction user interface, data extraction engine, online learning model, query engine, and interface. In general, the user interface resembles an ordinary data extraction and data entry

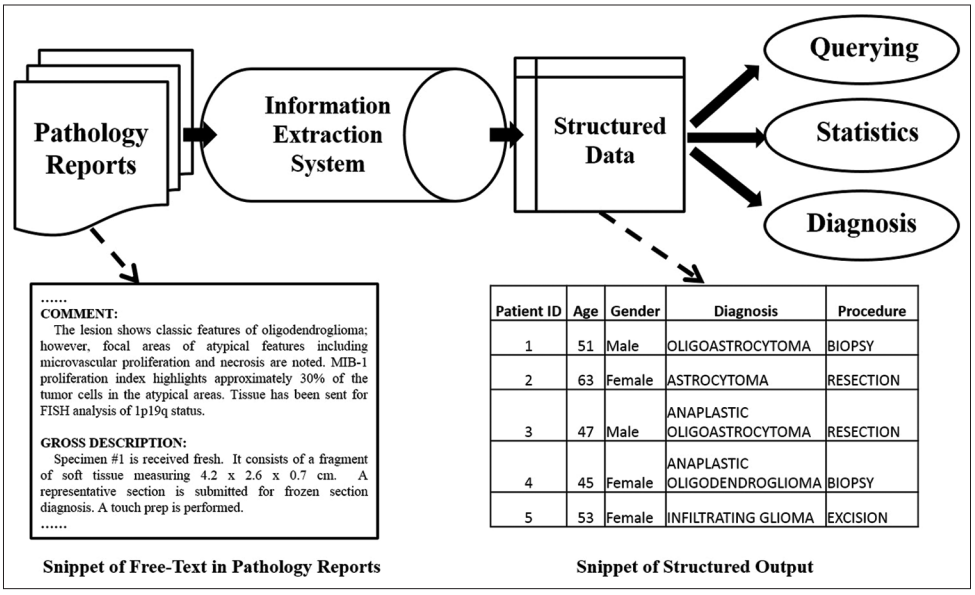

Figure 1: Common pipeline of processing free text medical report

system. It is unique, however, in its ability to transparently analyze and quickly learn from users' interactions the desired values for the data fields with online learning model. Additional user feedbacks incrementally refine the data extraction engine, as well as the vocabulary, in real-time, thus further reduce users' interaction effort thereafter. Processed reports are indexed by the query engine and made searchable by the query interface.

A demo video can be found from the following link.[14]

Interface of Extraction and Workflow

The workflow of IDEAL-X and the user interface are shown in Figures 2 and 3, respectively. A user begins by specifying the input folder that contains the collection of report documents to be extracted. This is followed by an iterative process through the collection, in which for each document the value of interest is extracted, inspected, and verified. The resulting set of all processed documents are coalesced into a final output file. To work with each document, the left panel of the interface displays the report being processed and the right panel is the output of the extraction organized as a list of index-attribute-value triples. The "index" column uses colors to highlight locations of values in the report. The "attribute" and "value" columns show the data element names and the extracted values, respectively. The "previous" and "next" buttons at the bottom of the right side allow users to navigate through the document collection.

When a report is loaded, the system attempts to predict and prefill the values for as many data elements as possible. The user fills in any remaining data element (through click and drop) that the system leaves blank. The user may fill multiple terms if given data element is a multiple value field such as diagnosis. For a prefilled data element, the user may review and update its value if it is incorrect or incomplete. This simple and intuitive interface makes the workflow of the system easily accessible to any user.

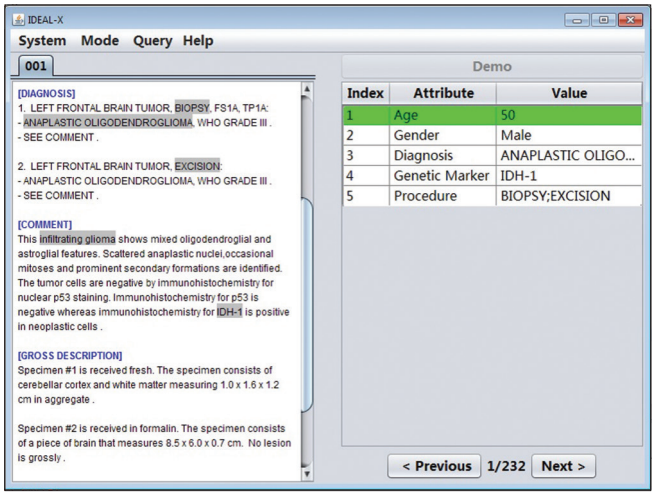

Figure 2: A screen shot of IDEAL-X's interface

At the beginning, the system is unable to predict values to most of the data elements. Through the combined manual extraction and revision process, the system learns the important contexts for the values and updates its decision model accordingly. As we will show in results later, the amount of information that the system is able to prefill correctly grows quickly.

Data Extraction Engine

The data extraction engine consists of the following major components: Preprocessing, vocabulary, answer predicting. Figure 4 shows relationships of the components with respect to the data flow. A parallelogram indicates results or inputs to components and procedures. The preprocessing component converts input texts and output forms into internal data structures used by the answer predicting component. The vocabulary component imports domain specific vocabulary to support information extraction. The answer predicting component extracts values from input texts to fill the output data elements. The online learning model utilizes judgments from users, in the form of edits on generated values, to update the decision model of the answer predicting component, which consists of vector space model, hidden Markov model (HMM) model, and rule induction model (see Adaptive Online Learning Model for details).

Automatic population of the output form is performed in three steps [Figure 5]. First, candidate sentences – those that are likely to contain values of interest, are detected by a combination of vector space model,[15] keywords, and location matching. Candidate values, consisting of phrase chunks are then extracted with the HMM algorithm[16] or user-defined vocabulary. Lastly, constraints such as string patterns and numerical ranges are learned through rule induction,[17] and applied to narrow the set of candidate values. For example, the system may only select candidate chunks with first letter capitalized. In this step, negation and uncertainty detection, which are performed based on predefined rules, are also applied to filter candidates.

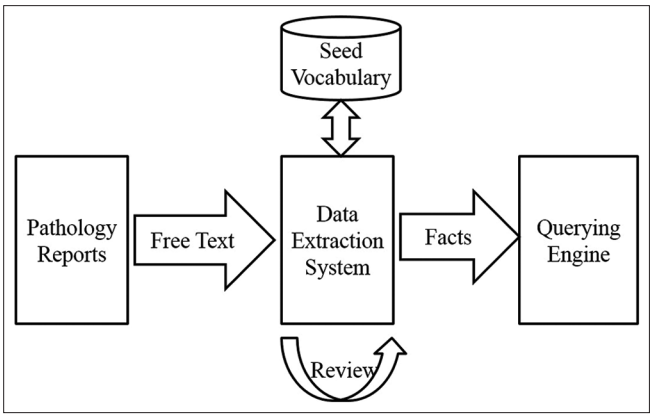

Figure 3: The workflow of the IDEAL-X

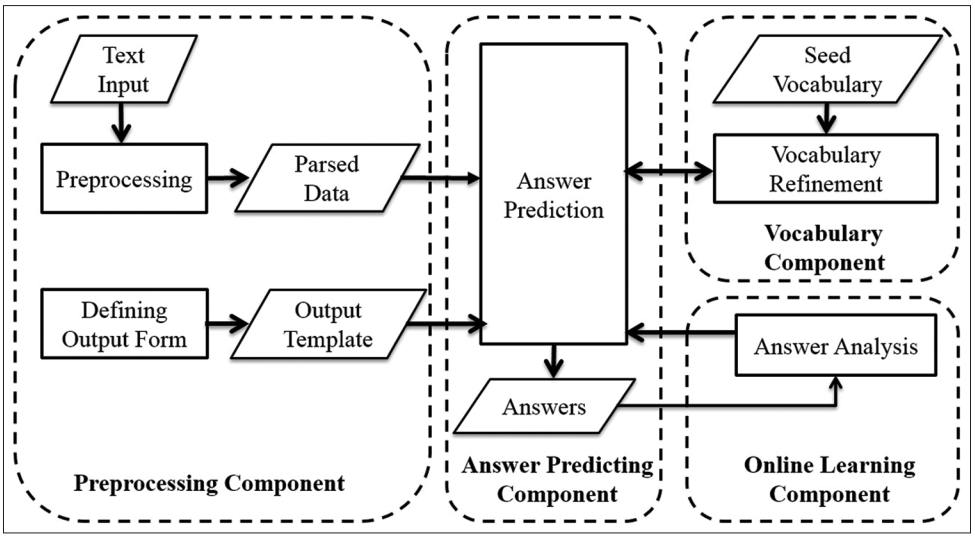

Figure 4: System components and dataflow

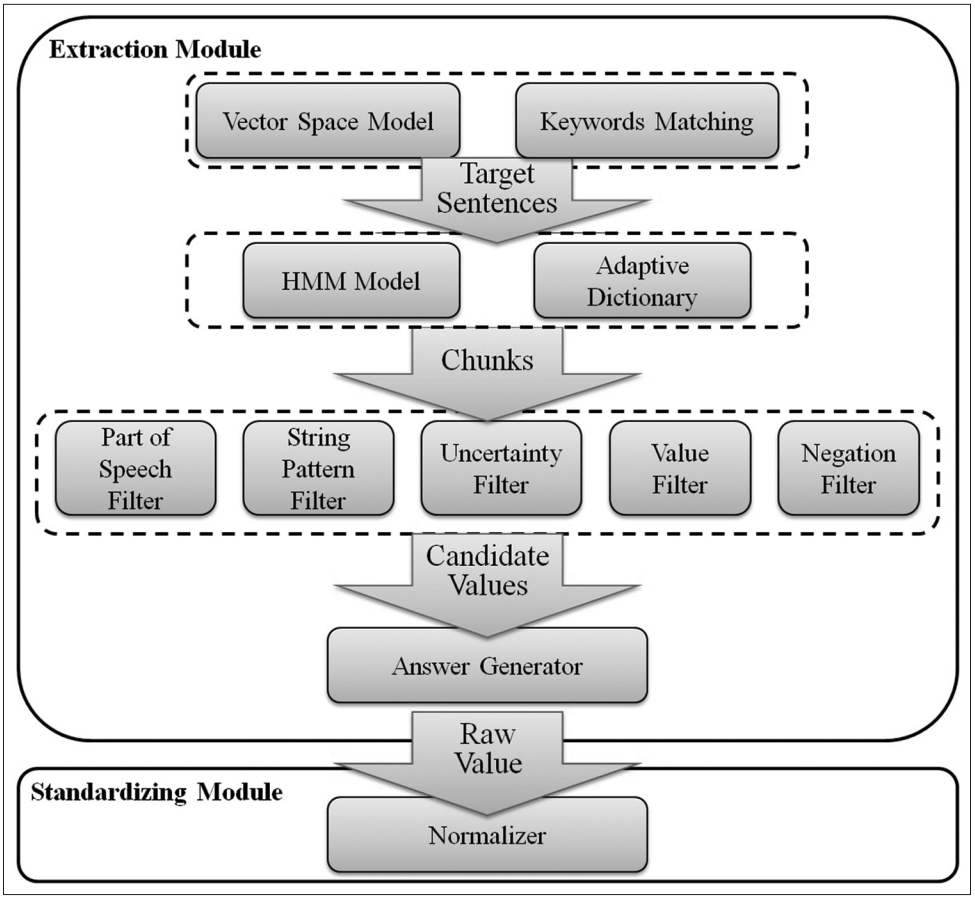

Figure 5: Modules of answer predicting component

Candidates that receive confidence scores above the threshold are used to fill the output form, which could be later transformed into a single structured view compatible with a database table or excel spreadsheet. If required, extracted values may be normalized based on user-defined mappings.

Adaptive Online Learning Model

IDEAL-X employs online supervised learning,[12,13] in which updating system is conducted after processing each report in order to minimize the cumulative gap between prediction and correct answer. As input documents are processed, the algorithm incrementally improves its learning parameters

based on user feedbacks. The feedbacks come in the form of user selection and correction to system predicted values. Text fragments, which are either highlighted by the system or selected by user are treated as answer values. The absence of any user action on a system-generated value is a positive feedback, and reinforces the learning model. If a revision occurred, user revised value is learned as positive feedback, which also indicates that the system's prediction is incorrect. Linguistic features associated with the value, such as part-of-speech tag, located section and co-occurring words in a sentence, are analyzed to improve the three steps performed by data extraction engine through vector space model,[15] HMM model,[16] and rule induction model[17] respectively. Through this interaction, IDEAL-X transparently learns the linguistic features of values to be extracted, and user doesn't have to predefine any constrains or thresholds.

Adaptive Vocabularies

The system allows the user to customize a domain-specific vocabulary such as drug names related to certain disease. In general, to create a seed vocabulary, a standard ontology such as the SNOMED clinical terms[18] or the National Cancer Institute (NCI) Thesaurus[19] is a useful starting point. But the vocabulary may not be complete and miss certain terminologies specific to local reporting domains. When a mismatch occurs between the vocabulary and an extracted value, IDEAL-X refines the vocabulary by adding the extracted terms and removing unneeded terms. This way, the vocabulary converges to a lexicon that is consistent with the extraction task. The vocabulary is also reusable: It can be exported for reproducibility, other extraction projects, and ontology construction. If desired, a postprocessing step to standardize the extracted values can be performed.

Query Interface and Engine

The extracted data are organized and indexed to facilitate querying. The system provides a built-in query interface [Figure 6] that allows the user to search for patients or reports based on user-specified conditions. The interface is split into three main panels. The right panel shows the search condition. For each attributes, the user may specify a value from the list of available values that the system has collected during extraction. The "search" button finds all reports that match all of the search criteria, and displays the results as a directory tree in the left panel. Selecting a node in this tree loads the content of the corresponding report into the text area of the second panel. When the user specifies multiple search conditions, the system searches for results that satisfy all criteria, in other words, the intersection set. When the user selects multiple values from the same condition, the union set will be generated.

Evaluation Metrics

We compared the system's output with the manually annotated ground truth with respect to precision, recall and F-1 measure. Precision estimates the correctness of extraction, recall estimates the completeness, and F-1 measure is the weighted average of these two.

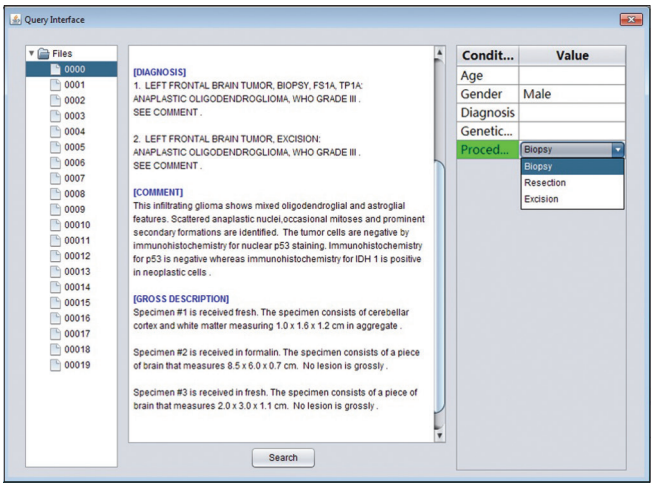

Figure 6: Query interface

RESULTS

To test the performance of information extraction, we perform two experiments: Experiment 1 to examine the effectiveness of online learning, and experiment 2 to examine the importance of adaptive vocabulary. The development of this system is driven by the demands of brain tumor research, in which pathology reports need to be queried based on demographic data, disease, procedure, among others, in order to locate patients with certain traits.

Datasets

We randomly selected and annotated 50 anatomic pathology reports manually as a testing dataset for this study. The guideline for annotation is provided by a pathologist, who also verifies annotation results to resolve the disagreement. These pathology reports were from patients that had been diagnosed with a Grade II or Grade III infiltrating glioma and had their tumors resected at Emory University Hospitals. Another 50 reports, which are disjoint with the testing dataset, were used as development dataset.

Experiment Setup

In order to study the system's learning effectiveness, all experiments began with an empty model, without prior training or predefined constraints.

We perform tests on extracting demographic information such as age and gender, and commonly queried medical information such as diagnosis, genetic marker (both gene and protein) and therapy/procedure [Table 1]. Note that, these attributes may be available in the structured format in some reporting systems or databases. For the purpose of the experiment, here we assume structured data is not available

and use these typical attributes to examine the effectiveness of IDEAL-X. To support extraction, we employ a seed vocabulary consisting of diagnosis, gene and procedure lexicons, obtained from the Human Disease Ontology,[20] the Cell Cycle Ontology[21] and the NCI Thesaurus[19] Ontology respectively. We use these prevalence ontologies for experimental evaluation, other seed vocabulary may be more appropriate for other extraction tasks.

Results of experiment 1 are shown in Table 2. Age and gender typically appear in report headers with limited contextual variation. For these, the system achieved very high precision and recall. Values related to diagnosis, genetic marker, and therapy appear in the text with larger structural and narrative variation. With the support of the seed vocabulary, the system achieved F1 scores of 88%, 93%, and 97%, respectively. To study the effectiveness of learning, for each test case, we divided the 50 reports into two groups in sequential order: The first 20 reports (as they appear in the directory), and the second 30 reports. The improvement of accuracy from the first group to the second group was significant, reflecting a high-rate of learning. For the four classes of attributes, F1 scores between the first and second groups increased from 94.7%, 82.1%, 90.0% and 95.3% to 100%, 91.2%, 95.3% and 99.5%, respectively.

The results of experiment 2 show the ability to refine the vocabulary can have a major effect on the accuracy of data extraction. For diagnosis, genetic marker, therapy and procedure, Table 3 shows the difference of results with and without refining the seed vocabulary. When the system used the seed vocabulary directly without further refinement, the performance of the extraction relies on how closely the vocabulary content aligns with the extraction task. For a genetic marker, a very small difference in the F1 score was observed. For diagnosis and procedure, on the other hand, the downloaded ontology subsets contain considerable irrelevant information for pathology reports. This impacted the precision by 3.1% for diagnosis, and 10.3% for the procedure. Moreover, many terms were missing, which negatively affected the recall by 46% for diagnosis and 30% for procedure. Comparing these results with Table 2, we notice that when adaptive vocabulary is enabled, the system captures important terms quickly. If these terms could not be added in time, the system will keep on missing these values in following report therefore largely impairs recall rate in some cases. These results show that there could be a large discrepancy between standard ontology and the controlled vocabulary of a specific domain, and the benefits adaptable vocabularies could be substantial.

Table 1: Test cases of data extraction

Attributes	Seed vocabulary sources	Value amount
Age and gender	None	100
Diagnosis	Human Disease Ontology	147
Gene and protein	Cell Cycle Ontology	146
Therapy and procedure	NCI Thesaurus	324

NCI: National Cancer Institute

Table 2: Test result of experiment 1: Study of online learning

Attributes	Subsets	Precision (%)	Recall (%)	F-1 (%)
Age and gender	First 20	100	90.0	94.7
	Last 30	100	100	100
	Overall 50	100	96.0	97.9
Diagnosis	First 20	90.6	75.0	82.1
	Last 30	94.3	88.2	91.2
	Overall 50	93.1	83.5	88.0
Genetic marker	First 20	90.0	90.0	90.0
	Last 30	94.8	95.8	95.3
	Overall 50	93.1	93.8	93.5
Therapy and procedure	First 20	97.4	93.3	95.3
	Last 30	100.0	99.0	99.5
	Overall 50	99.0	96.9	97.9

Table 3: Test result of experiment 2: Study of adaptive vocabulary

Attributes	Adaptive vocabulary	Precision (%)	Recall (%)	F-1 (%)
Diagnosis	Off	90.0	36.9	52.4
	On	93.1	83.5	88.0
Genetic marker	Off	84.1	86.9	85.5
	On	93.1	93.8	93.5
Therapy and procedure	Off	89.3	66.9	76.5
	On	99.0	96.9	97.9

DISCUSSIONS

In the experiments, we have selected attributes that appear in different types of text that are found in pathology reports. Our goal is for testing the effectiveness of online machine learning and the importance of refining the seed vocabulary during the extraction process. Though extracting value in generic domain is challenging, given particular task or research topic, which has value of limited domain, the system could be customized easily to meet specific purpose of given task, for example, identify patient for brain tumor research. In follow-up studies, we will consider a broader set of attributes and enrich the supported data types. Values that the system can manage currently are limited to numerical and nominal values. Extracting temporal information, for example, will improve the utility of the system. In pathology research, medical events such as procedures are time sensitive. Augmenting the output with timelines would contextualize and help to connect the extracted values in important ways.

A goal of IDEAL-X is to provide a generic solution for information extraction across medical domains. The system employs techniques that we believe are domain agnostic, which could be validated with use cases in other medical domains. We are collaborating with Emory Clinical Cardiovascular Research Institute and Rutgers University Radiation Oncology Department to validate the broader utility of IDEAL-X. One characteristic of the cardiology research project is the existence of multiple types of reports for each patient. This adds complexity to the interface and the learning algorithm. We also plan to study the advantage of using IDEAL-X, as a software assistant for annotation over manual annotation both on efficiency and accuracy.

Finally, we will study the use of IDEAL-X to ease structured reporting such as providing machine generated automated hints to create synoptic pathology reports. Though answers may be inputted using a combo box to guarantee structured reporting, most existing medical report systems still allow for free-format text and uncontrolled vocabulary. In some cases, direct access to the structured database may not be available, therefore, requires extracting information from text pathology report directly.

CONCLUSIONS

IDEAL-X employs iterative online machine learning and adaptive controlled vocabulary for information extraction from clinical reports. It automatically predicts annotations for values to be extracted and utilizes human feedback as knowledge to improve continuously the performance of extraction. Experimental results demonstrate that both online learning and adaptive vocabulary are highly effective. Extracted data are indexed by the built-in query engine and can be conveniently queried with a graphic interface. The adaptability and usability of the system make IDEAL-X a powerful data extraction tool to support patient search from pathology reports.

Acknowledgments
This study is supported in part by grants from the Centers for Disease Control and Prevention 200-2014-M-59415.

Financial Support and Sponsorship
Nil.

Conflicts of Interest
There are no conflicts of interest.

REFERENCES

1. Srigley JR, McGowan T, Maclean A, Raby M, Ross J, Kramer S, *et al.* Standardized synoptic cancer pathology reporting: A population-based approach. J Surg Oncol 2009;99:517-24.
2. Gill AJ, Johns AL, Eckstein R, Samra JS, Kaufman A, Chang DK, *et al.* Synoptic reporting improves histopathological assessment of pancreatic resection specimens. Pathology 2009;41:161-7.
3. Leslie KO, Rosai J. Standardization of the surgical pathology report: Formats, templates, and synoptic reports. Semin Diagn Pathol 1994;11:253-7.
4. CAP Cancer Protocols. Available from: http://www.cap.org/web/home/resources/cancer-reporting-tools/cancer-protocols. [Last accessed on 2015 Jul 12].
5. Anatomic Pathology Structured Reports. Available from: http://www.wiki.ihe.net/index.php?title=Anatomic_Pathology_Structured_Reports. [Last accessed on 2015 Jul 12].
6. Crowley RS, Castine M, Mitchell K, Chavan G, McSherry T, Feldman M. caTIES: A grid based system for coding and retrieval of surgical pathology reports and tissue specimens in support of translational research. J Am Med Inform Assoc 2010;17:253-64.
7. Schadow G, McDonald CJ. Extracting Structured Information from Free Text Pathology Reports. In AMIA Annual Symposium Proceedings. American Medical Informatics Association; 2003.
8. Coden A, Savova G, Sominsky I, Tanenblatt M, Masanz J, Schuler K, *et al.* Automatically extracting cancer disease characteristics from pathology reports into a disease knowledge representation model. J Biomed Inform 2009;42:937-49.
9. Schlangen D, Stede M, Bontas EP. Feeding Owl: Extracting and Representing the Content of Pathology Reports. In Proceeedings of the Workshop on NLP and XML (NLPXML-2004): RDF/RDFS and OWL in Language Technology. Association for Computational Linguistics; 2004.
10. Mitchell KJ, Becich MJ, Berman JJ, Chapman WW, Gilbertson J, Gupta D, *et al.* Implementation and evaluation of a negation tagger in a pipeline-based system for information extraction from pathology reports. Medinfo 2004;2004:663-7.
11. Smale S, Yao Y. Online learning algorithms. Found Comut Math 2006;6:145-70.
12. Shalev-Shwartz S. Online learning and online convex optimization. Found Trends in Mach Learn 2011;4:107-94.
13. Shalev-Shwartz S. Online Learning: Theory, Algorithms, and Applications. 2007. Available from: http://ttic.uchicago.edu/~shai/papers/ShalevThesis07.pdf. [Last accessed on 2015 Jul 12].
14. IDEAL-X Demo Video. Available from: https://youtu.be/Q-DrWi31nv0. [Last accessed on 2015 Jul 12].
15. Manning CD, Raghavan P, Schütze H. Introduction to Information Retrieval. Vol. 1. Cambridge: Cambridge University Press; 2008.
16. Elliott RJ, Aggoun L, Moore JB. Hidden Markov Models: Estimation and Control. Science & Business Media: Springer; 2008.
17. Fürnkranz J. Separate-and-conquer rule learning. Artif Intell Rev 1999;13:3-54.
18. SNOMED Clinical Terms. Available from: http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html. [Last accessed on 2015 Jul 12].
19. NCI Thesaurus. Available from: http://www.ncit.nci.nih.gov/. [Last accessed on 2015 Jul 12].
20. Human Disease Ontology. Available from: http://www.disease-ontology.org/. [Last accessed on 2015 Jul 12].
21. Cell Cycle Ontology. Available from: http://www.cellcycleontology.org/. [Last accessed on 2015 Jul 12].