Puya: A Method of Attracting Attention to Relevant Physical Findings

Walter Duque de Estrada, MD, MS, Shawn Murphy, MD, PhD, G. Octo Barnett, MD
Laboratory of Computer Science
Massachusetts General Hospital
Boston, Massachusetts

Puya is a method that compares the physical exam in an electronic clinical note with a set of stereotypical physical exam sentences that have been previously classified as "normal". The note is then displayed in a web browser with normal findings clearly delineated. The list of stereotypical sentences comes from a set of physical findings found within extensive electronic medical record. This list is then screened to select only those that represent "normal" findings, a process that yields 96% total agreement among 4 clinicians surveyed. This final list of stereotypical "normal" sentences accounts for 64% of the clinical narrative text. Sentences in the clinical note that do not match sentences in the "normal" list are assumed to be "abnormal". Puya screened 98 clinical notes consisting of 610 individual sentences. Puya achieved a sensitivity of 100%, a specificity of 63%, a positive predictive value of 44% and a negative predictive value of 100%. This leads to an application that reduces informational noise.

INTRODUCTION

The clinical record is steadily migrating from paper to electronic media. The advantage of easy retrieval and display beyond the original document text has been realized only in a limited fashion. The goal of this project is to bring added-value text retrieval through component-based client-server applications in an accurate, intuitive and cost-effective fashion.

There has been great interest in using Natural Language Processing (NLP) for recognizing meaning in clinical narrative text. Sager[1] described their efforts at organizing the clinical narrative text into substrings within a semantic network defined by a subgrammar. They applied this technique to create an Information-Format graph that represents the inherent structure of a clinical narrative sentence. Each substring is semantically typed and is considered a node. These nodes can be connected by edges only if the grammar explicitly allows it. An exact match between the clinical narrative text and the graph would produce an unambiguous interpretation. An obstacle to this approach is that a large body of knowledge has to be explicitly coded; this might require expert resources , such as clinical subspecialists, that may not be readily available.

Text indexing is another approach to identifying content in clinical narrative text; only specific substrings which have been previously semantically typed are extracted from the clinical narrative. Text-indexing assumes that meaning can be extracted by only looking at specific phrases inside of the clinical narrative text. Spackman and Hersh[2] used two industrial NLP systems to identify noun phrases in medical discharge summaries. The programs have a reported sensitivity between 69% and 77%; the authors noted that the NLP systems had not been modified to detect medical vocabulary. Hersh[3] noted that automatic discovery of phrases was efficient, but semantic typing of the phrases was time-consuming. Another limitation of this method comes from modifying phrases that may lie outside of the expected scope of the parser; the modifying phrase might change connotation of the original phrase, leading to semantic error.

A better approach might be to use text-retrieval and matching. In previous work[4], it has been shown that a generic (clinician independent) phrase library, made up of the top 131 phrases in the physical exam, could account for 30% of the text found in physical exam notes. This phrase library could be matched to a clinical narrative text with matched sentences being marked as normal and unmatched sentences assumed to be abnormal. The results would be sent to the user for review. A similar markup approach has been successfully implemented[5].

METHOD

The COSTAR database at Massachusetts General Hospital serves two outpatient internal medicine groups: Bulfinch Medical Group and Internal Medicine Associates. The Bulfinch Medical Group is a semi-private practice while the Internal Medicine Associates forms part of the teaching environment at the Massachusetts General Hospital. A subset of this database was selected, consisting of 162,863 encounters with 41,585 patients; these are all the notes in the database which mention a physical exam being done. Each clinical encounter has a physical exam section with at most 14 sections corresponding to organ systems. Each section has one or more sentences; in this context a sentence is defined as a string of one or more words separated by periods.

The sentences within the physical exam section of the clinical narrative text were rank-ordered according to decreasing frequency. Using cost-benefit threshold function, a list of most frequent sentences for each section of the physical exam was determined. As seen in equation 1 the benefit-cost function is the number of sentences to be added divided by the number of sentences already added; a threshold of 0.1% was arbitrarily used. The benefit is increased accuracy of detection; the cost is increased coding effort.

$$\text{cost - benefit ratio} = \frac{\text{freq. of sentence to be added}}{\text{sum of freq. of sentences already added}} \quad (1)$$

These sentences were then reviewed by one clinician and the sentences that represented abnormal findings were removed from the lists. This review required about 60 minutes to complete. What remained was a database of common, stereotypical sentences that represented normal findings in various sections of the physical exam. To test inter-physician consistency in this scenario, four additional clinicians from the initial one were asked to review the normal sentence databases and determine their level of agreement with it. For each level of agreement on findings a grade was assigned: **total agreement** for 4 physicians agreeing; **partial agreement** for 3 physicians agreeing ; and **total disagreement** for 2 physicians agreeing.

A Java component ("Puya") was chosen to implement the project because it could be integrated with other components in a complete electronic medical record. Puya, a server-side applet, compares the database of normal physical exam sentences to the clinical narrative text of a physical exam. If a sentence from a physical exam matched a sentence within the database of sentences representing normal findings, it was considered to be a normal finding. Sentences not matching within the database were assumed to be abnormal findings. Tags were added to the text to identify the finding as normal or abnormal. The text could then be displayed in something as simple as an HTML page.

Puya runs as a server-based applet (servlet) to which a web-based client can send the clinical narrative text. Puya compares the clinical narrative text text with the physical exam database which has the normal sentences. Puya appropriately marks up the physical exam section of the clinical note and then returns the text for display on the browser client. A prototype client page can be seen in Figure 1.

A random set of 100 clinical notes were extracted from the COSTAR database; these notes were from clinical encounters that happened after the control subset was extracted. Two text notes were garbled and were not used. After running the experimental group through Puya, the researchers then corroborated the program's accuracy by looking at each sentence and agreeing or disagreeing with the program's assessment (Table 1).

Table 1
Definition of Performance Parameters

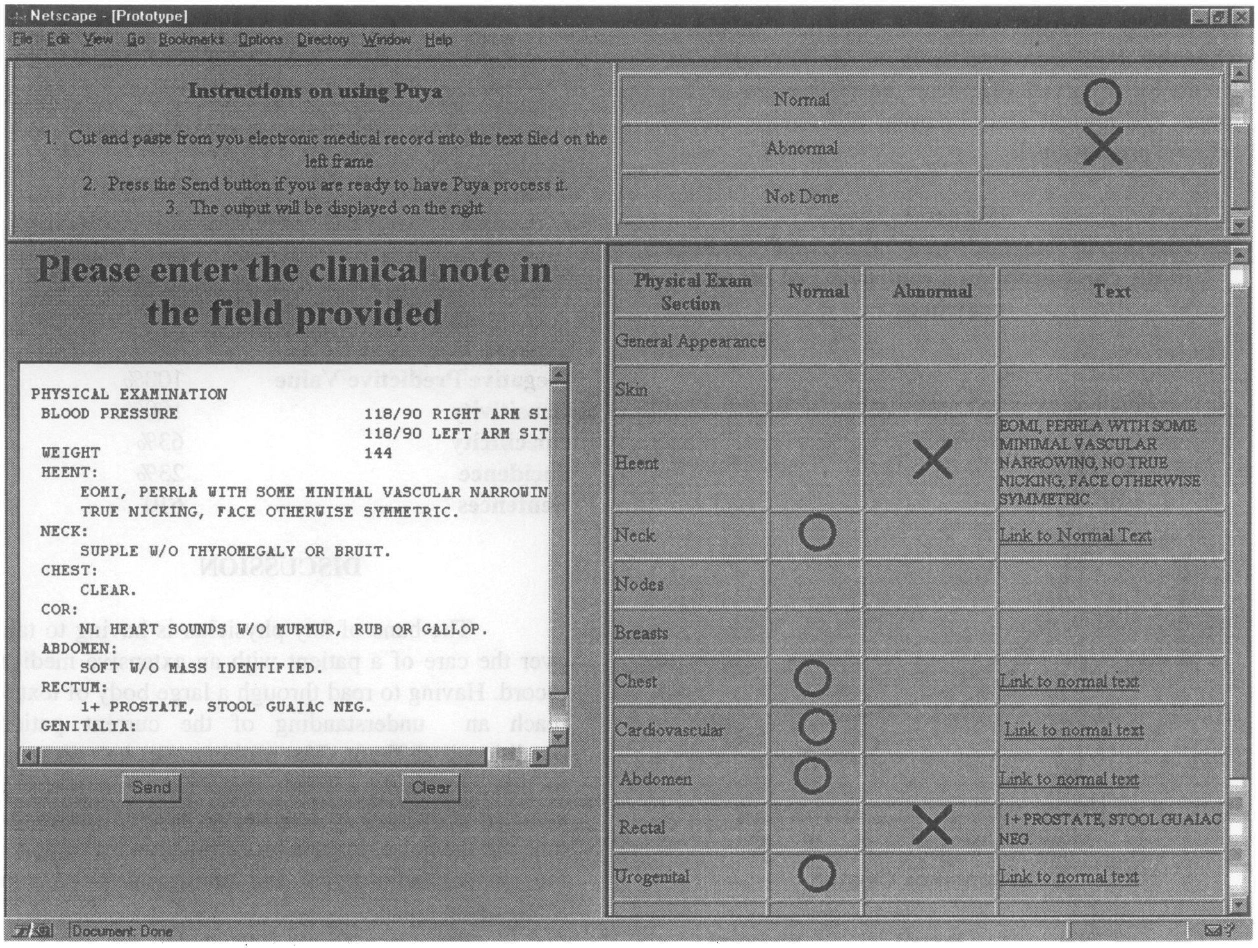
	Sentence Tag Normal	Sentence Tag Abnormal
Normal Finding	True Negative	False Positive
Abnormal Finding	False Negative	True Positive

RESULTS

The cost-benefit function was applied to each physical exam section sentence population. For each section two quantities were determined: the number of individual sentences in the subset; and the percentage of the total text covered by the subset. Table 2 has a summary of these values for all sections of the physical exam. Overall , 64% of the clinical narrative text is covered by a small number of stereotypical sentences such as, within the abdominal exam, "Soft, non-tender, normal bowel sounds."

The survey of the four physicians is summarized in figure 2. There was an overall 96% agreement in classification of physical findings, as "normal" or "abnormal", among all four physicians. There was partial disagreement in 3% of the cases and total disagreement in the remaining 1%. Disagreement among the physicians centered along issues such as the level of a normal JVP or whether obesity is an abnormal finding. Average time of review for each physician was 60 minutes.

Figure 1
Typical Puya Web Page



Instructions on using Puya

1. Cut and paste from you electronic medical record into the text filed on the left frame
2. Press the Send button if you are ready to have Puya process it.
3. The output will be displayed on the right

Normal	○
Abnormal	╳
Not Done	

Please enter the clinical note in the field provided

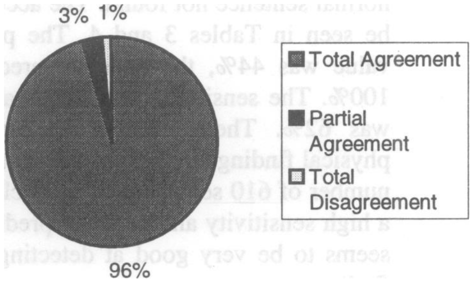
```
PHYSICAL EXAMINATION
 BLOOD PRESSURE                       118/90 RIGHT ARM SI
                                      118/90 LEFT ARM SIT
 WEIGHT                               144
 HEENT:
     EOMI, PERRLA WITH SOME MINIMAL VASCULAR NARROWIN
     TRUE NICKING, FACE OTHERWISE SYMMETRIC.
 NECK:
     SUPPLE W/O THYROMEGALY OR BRUIT.
 CHEST:
     CLEAR.
 COR:
     NL HEART SOUNDS W/O MURMUR, RUB OR GALLOP.
 ABDOMEN:
     SOFT W/O MASS IDENTIFIED.
 RECTUM:
     1+ PROSTATE, STOOL GUAIAC NEG.
 GENITALIA:
```

Physical Exam Section	Normal	Abnormal	Text
General Appearance			
Skin			
Heent		╳	EOMI, PERRLA WITH SOME MINIMAL VASCULAR NARROWING, NO TRUE NICKING, FACE OTHERWISE SYMMETRIC.
Neck	○		Link to Normal Text
Nodes			
Breasts			
Chest	○		Link to normal text
Cardiovascular	○		Link to normal text
Abdomen	○		Link to normal text
Rectal		╳	1+PROSTATE, STOOL GUAIAC NEG.
Urogenital	○		Link to normal text

Table 2
Frequency distribution of most typical sentences

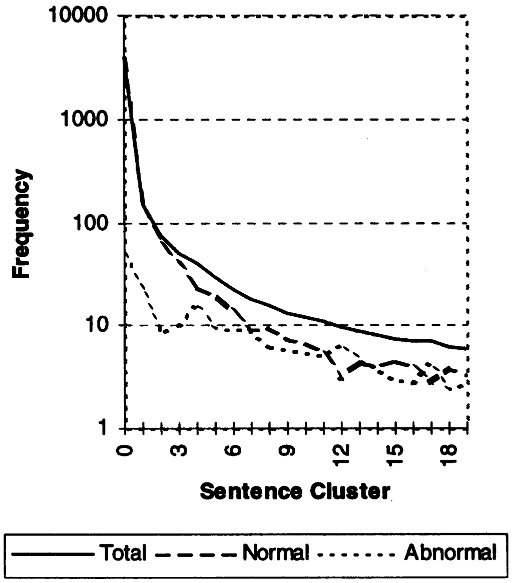
Physical Exam Section	Stereotypical sentences	Percent of total text database	Total text in database
abdomen	126	68	114353
breast	58	62	69983
cardiovascular	123	58	169220
chest	44	78	126653
extremities	72	69	133504
heent	127	67	168823
general	89	61	45923
genitourinary	172	45	81156
musculoskeletal	141	35	46373
neck	98	79	134799
neurological	149	49	72020
nodes	21	89	45198
rectal	88	61	95677
skin	74	53	59046

Figure 2
Physical Findings Survey Among Four Clinicians



In order to investigate the distribution of normal and abnormal findings in the rank-ordered list, the top 500 most frequent sentences of the "chest" physical exam section were selected. This categorized sentence list was clustered in groups of twenty-five and the average frequency of normal and abnormal sentences inside the cluster was then calculated (figure 3). At this point, 86% of the total text database of the chest exam has been included. The other 14% is covered by about 14,000 sentences. It can be observed that near the right hand of the graph, the normal and abnormal distribution curves become proportional.

Figure 3
Distribution of "normal" and "abnormal" findings in the chest exam on a rank-ordered list of sentences



In the case of Puya, the goal is to detect abnormal conditions by **not matching** with the database of normal sentences, i.e. true detection = normal sentence not found. The accuracy of Puya can be seen in Tables 3 and 4. The positive predictive value was 44%, the negative predictive value was 100%. The sensitivity was 100% and the specificity was 62%. There was an incidence of abnormal physical findings in 22% of the sentences with a total number of 610 sentences over 98 clinical notes. With a high sensitivity and negative predictive value, Puya seems to be very good at detecting normal physical findings.

Table 3
Raw Accuracy Results

	Normal Tag	**Abnormal Tag**	**Total**
Normal Finding	295	177	***472***
Abnormal Finding	0	138	***138***
Total	*295*	*315*	**610**

Table 4
Analyzed Performance Results

Positive Predictive Value	44%
Negative Predictive Value	100%
Sensitivity	100%
Specificity	63%
Incidence	23%
Sentences	610

DISCUSSION

The bane of any physician is having to take over the care of a patient with an extensive medical record. Having to read through a large body of text to reach an understanding of the current patient condition is difficult. The problem can be compared to that of finding a weak radio signal in a sea of static; it is difficult to tune in the desired music and tune out the noise. Puya is an application for tuning in the clinical radio signal and tuning out extraneous information noise.

Puya is a server-side applet that helps the user by searching through an extensive electronic clinical text and marking information that might be of interest to the user. In the current implementation Puya highlights and contrasts those physical findings that it has determined to be "normal" or "abnormal" in a browser client.

With a negative predictive value of 100%, the user could assume that Puya is probably right when it displays that a physical exam section is normal. This helps the user concentrate on determining if the sentences declared "abnormal" are truly so.

The accuracy of the method can be improved by reviewing more sentences from the rank-ordered list. One important hypothesis can be made by looking at the graph in Figure 3: after a sufficiently large number of sentences are added to the population, the proportion of "normal" and "abnormal" sentences is constant. As more sentences are reviewed, the specificity of Puya increases. If a certain specificity is desired, then the number of sentences that need to be reviewed can be estimated. For the chest exam frequency distribution a 95% specificity could be reached after reviewing about 9000 sentences out of 14500. With a specificity of 95%, the PPV would be at 85%, much better than 63%. Since the average physician in our survey reviewed 1400 sentences in an hour, it would take less than eight person-hours to reach that specificity.

Another method of improving on Puya is by enhancing the user interface. The interface can be enhanced such that, instead of a table, a human figure would graphically highlight the areas of interest from the report.

Natural language processing and Puya method are complementary techniques for viewing electronic medical data. NLP is able to extract information from clinical text databases by matching the clinical text to a schema that has been previously constructed. With this approach, the goal is in detection of abnormal findings.

The goal of the Puya method is noise reduction. The Puya method matches the information the user is definitely does not consider interesting and removes it from view. This allows Puya to display the information of greater interest to the user.

The value of Puya can be increased by adapting it to search multiple records. Searching through multiple medical records increases the functionality of the program by proportionally reducing the amount of noise the user has to filter.

FUTURE WORK:

At present Puya works with one text file sent by the user at a time. It is intended to have Puya retrieve multiple clinical notes at a time and present them to the user in an unified manner. A more functional web-based client based on Java is also in development. Increasing the specificity of Puya is extremely important. The ultimate goal is to integrate Puya as a component in a complete electronic medical record.

The goal of this project is to bring added-value text retrieval through component-based client-server applications in an accurate and intuitive fashion. Java allows use to create the application while the Puya method has the inherent potential in both accuracy and intuitive display.

This work was supported in part by the National Library of Medicine Training Grant NLM LM 07092-03 and in part by NLM research grants LM05854 and HS06575, and AHCPR grant HS06575.

BIBLIOGRAPHY

1. Sager N, Lyman M, Bucknall C, Nhan N, Tick LJ, Natural Language Processing and the Representation of Clinical Data. J Am Med Informatics Assoc 1994; 1(2): pp 142-160.

2. Spackman KA, Hersh WR, Recognizing Noun Phrases in Medical Discharge Summaries: An Evaluation of Two Natural Language Parsers, Cimino JJ(editor) Proc. 1996 AMIA Annual Fall Symposium (formerly SCAMC), 1996, Washington, pp. 155-158.

3. Hersh WR, Campbell EH, Evans DA, Brownlow ND, Empirical Automated Vocabulary Discovery Using Large Text Corpora and Advanced Natural Language Processing Tools, Cimino JJ(editor) Proc. 1996 AMIA Annual Fall Symposium (formerly SCAMC), 1996, Washington, pp 159-163.

4. Murphy SN, Barnett GO, Achieving Automated Narrative Text Interpretation Using Phrases in the Electronic Medical Record, Cimino JJ(editor) Proc. 1996 AMIA Annual Fall Symposium (formerly SCAMC), 1996, Washington, pp. 532-536.

5. Sager N, Ngo TN, Lyman M, Tick LJ, Medical Language Processing with SGML, Cimino JJ(editor) Proc. 1996 AMIA Annual Fall Symposium (formerly SCAMC), 1996, Washington, pp. 547-551.