Published in final edited form as:
J Biomed Inform. 2010 April ; 43(2): 218–223. doi:10.1016/j.jbi.2009.08.016.

An interactive and user-centered computer system to predict physician's disease judgments in discharge summaries

Jonathan P. DeShazo, MPH, PhD[1] and **Anne M. Turner, MD, MLIS, MPH**[2,3,4]

[1]Department of Health Administration, Virginia Commonwealth University, Richmond, VA

[2]Center for Public Health Informatics, University of Washington, Seattle, WA

[3]Department of Health Services, School of Public Health, University of Washington, Seattle, WA

[4]Division of Biomedical and Health Informatics, University of Washington, Seattle, WA

Abstract

PURPOSE—This article describes a formative natural language processing (NLP) system that is grounded in user-centered design, simplification, and transparency of function.

METHODS—The NLP system was tasked to classify diseases within patient discharge summaries and is evaluated against clinician judgment during the 2008 i2b2 Shared Task competition. Text classification is performed by interactive, fully supervised learning using rule-based processes and support vector machines (SVMs).

RESULTS—The macro averaged F-score for textual (t) and intuitive(i) classification were .614(t) and .629(i), while Micro averaged F-scores were recorded at .966(t) and .954(i) for the competition. These results were comparable to the top 10 performing systems.

DISCUSSION—The results of this study indicate that an interactive training method, de novo knowledge base with no external data sources, and simplified text-mining processes can achieve a comparably high performance in classifying health-related texts. Further research is needed to determine if the user-centered advantages of a NLP system translate into real world benefits.

Keywords

Natural Language Processing; Patient Discharge; Medical Records; Medical Records Systems, Computerized; Obesity

INTRODUCTION

Narrative text documents contain a wealth of information and are frequently found in the biomedical and health services domain. However, as electronic documents expand in length and collections of electronic documents grow in size, accessing the information contained in free text can be difficult and prohibitively time consuming. Text mining is a subset of natural

Reprint requests and corresponding author: Jonathan P. DeShazo, PhD, MPH, Assistant Professor of Biomedical Informatics, Department of Health Administration, Virginia Commonwealth University, Grant House Room 205, 1008 East Clay Street, P.O. Box 980203, Richmond, Virginia 23298-0203, Phone: 804-828-5509, Fax:804-828-1894, jpdeshazo@vcu.edu.

language processing (NLP), which attempts to quickly collect pre-identified concepts from texts, as opposed to analyzing the syntactic and semantic structure and meaning of the entire text [1]. Text mining is ideally suited for finding answers to simple specific questions within large corpora of text [2] and is increasingly applied as a text classification tool for biomedical and public health research [3–5]. It may also be used in clinical environments wherever rapid access to quantities of free text is needed, such as generating problem lists, notes, and quality assurance. The automated identification of diseases in free texts such as medical records is a challenging component of potential improvements in areas of public health, clinical care delivery, and administrative functions [6–8].

Much of the current NLP research within biomedicine focuses on evaluating system output, instead of improving 'real world' tasks shared between the system and individuals [1]. NLP research activities are only recently designing and evaluating NLP systems at the user level; in particular establishing user needs and conducting user centered evaluations of the system [9]. Furthermore Zweigenbaum et al. considers user driven systems (including user needs assessments, attention to user interfaces, and user centered designs) to be one of the six 'new frontiers' in biomedical text mining[10]. Recent attention to user needs within biomedical text mining literature is sparse but growing, and includes improving results visualization and user interface design, and functionality[11,12,13]. In addition, nationally sponsored industry events such as TREC are focusing to some extent on the information needs and applications of actual users. In addition, conceptual models such as "interactive NLP" have been coined to describe the two-way relationship some text mining systems have with their users and the impact the system has on the overall (human) task environment. In contrast to the frequent engineering and theory laden experimental activities of NLP research, the concept of "interactive NLP" is based in practical, value added applications and real time interaction. In their influential paper, Manaris and Slator considered interactive NLP the most useful and mature approach to automatically extracting useful knowledge from text[14]. The authors paired this perspective with user-centered design practices to create a new tool for health related document classification.

The design philosophy adopted by the authors incorporates interactive NLP concepts, user-centered design and user-centered system evaluation. Our aim is to build a flexible, user-centered computer tool that satisfies many of the basic document classification problems found in biomedicine and health services. Three core design objectives were identified in prior focus groups and interviews with users: The need for 1) domain expert supervision of the knowledge base, 2) transparency of the classification process, and 3) simplicity of use.

This work details the design of a user-driven text mining system, reports the performance results of the system from health text-mining competition, and discusses possible implications for interactive, user-centered design in other contexts and tasks.

The data used for this study were collected and de-identified to support the 2008 i2b2 Shared Obesity Challenge. Two physician obesity experts from the Massachusetts General Hospital Weight Center were asked to classify narrative patient records using two different methods: 1) Textual- For each of the sixteen diseases, the clinicians provided judgments based explicitly on the text in the patient narrative.

These judgments were "Yes"(Y), "No"(N), "Questionable"(Q), and "Unmentioned"(U). 2) Intuitive- For each of the sixteen diseases they provided judgments based intuitive information found in the patient narrative. These judgments were "Yes"(Y), "No"(N), and "Questionable"(U). Participants in the challenge were asked to create a system to automatically classify concepts within the narrative of the patient records and compare the results of their system with the manual classification (i.e. "ground truth").

The objectives of this system, which was entered into the 2008 i2b2 Shared Obesity Challenge [15], is to classify the patient based on discharge summaries as having obesity or any of fifteen other related co-morbidities. The diseases of interests are listed below.

Diseases and co-morbidities included in the study:

- Obesity
- Diabetes mellitus (DM)
- Hypercholesterolemia
- Hypertriglyceridemia
- Hypertension (HTN)
- Peripheral vascular disease (PVD)
- Heart failure (CHF)
- Osteoarthritis (OA)
- Venous insufficiency
- Atherosclerotic CV disease (CAD)
- Obstructive sleep apnea (OSA)
- Asthma
- GERD
- Gallstones / Cholecystectomy
- Depression
- Gout

The task is further divided into two parts as reflected by the instructions given to the clinicians when creating the “ground truth”, or designated correct classifications used to evaluate the system. The two tasks were to: 1) find the diseases that are textually stated in the discharge summaries, and 2) predict the diseases that are intuitively inferred from the text.

METHODS

We chose metaphors to describe the user classification judgment processes to help achieve our user-friendly design objectives. ‘Textual’ judgments are used when users can point to specific text that explicitly indicates a particular judgment classification. This classification method was supported by a simple rule based classifier that felt transparent to users. A rule-based classification process is a series of “if then” operations based on words and phrases found in the text that are stored in a knowledge base. The knowledge base for the textual classification is interactively developed by expert users who annotate the training documents and subsequently view the results of their changes on document classifications. In contrast, when a user determines a particular judgment is appropriate yet cannot point to anything specific, they would use the ‘intuitive’ judgment classification. The intuitive classification process is a Support Vector Machine (SVM) approach where the SVM operates on the classification results of the rule-based system. In contrast to a rule-based approach, SVMs are a type of linear classifier that represents the texts as vectors in an n-dimensional space. The SVM classification process is more difficult for users to conceptualize and maps loosely to the ‘intuitive’ metaphor.

Another user-identified design consideration is the highly subjective nature of professional judgments. The users expressed a desire for the system to imitate their own judgments.

Accommodating a rapidly changing vocabulary and context was another concern in the design of the system. To address these concerns we explored using an independent, de novo knowledge base approach. This method uses no external libraries or resources and relies exclusively on user interaction to establish the knowledge base.

Textual Classification

For textual classification, this system uses a simple keyword and rule-based process to find and identify disease names in the text. This method is easy for users to understand and similar processes have achieved reasonable success in comparable tasks [16]. The rule-based model is composed of simple logic that operates on four basic concepts. The four concepts:

- **Features**. Used as class identifiers, features are pre-identified labels representing a relevant textual concept. Documents will be assigned to classifications based on these features. (e.g. “Asthma”, ”Congestive Health Failure”)
- **Textual Evidence**. Textual evidence are key words or phrases that when present indicate or contraindicate a particular feature (e.g. “is asthmatic”, “is not asthmatic” is evidence for and against Asthma respectively, and “Atrovent” is evidence for the feature Asthma Tx (treatment))
- **Negation**. If negation elements are found before textual evidence(e.g. “no evidence of”, “does not have”), the feature is negated and the entire new phrase (including the negation) are added to the knowledge base as a new textual evidence phrase. (e.g. “no evidence of HTN”).
- **Referents**. Referents indicate who the evidence is referring to (e.g. “Family history: ”, “Patient’s Mother has ”). For this challenge, the presence of any referent other than the patient was a trigger to ignore the textual evidence immediately following

For example, a discharge summary may contain the text “The patient does not have diabetes”. If not already automatically annotated by the system, the user would train the system to recognize these in the future. To do this, the user highlights the token “patient” and designates it as a referent. Similarly, “does not have” and “diabetes” are negation and evidence tokens, respectively. These new tokens are added to the knowledge base and are used by the rule based classifier in the future.

Figure 1 illustrates the interactive nature of the supervision process. Knowledge domain experts train the system by annotating evidence for textual classification judgments, in addition to reviewing the automated textual and intuitive judgments.

The discharge summaries provided for system training were initially classified by clinicians recruited by the organizers of the 2008 i2b2 challenge. However, the textual evidence supporting the classifications still needed to be identified and annotated, as well as the identification of the negation elements and referent elements found in the corpus. To accomplish this supervised training, roughly half (n=300) of the provided training set was used to build the evidence base for rule-based classification and to train the SVM classifier. The discharge summaries were reviewed in tandem with the previously assigned disease classifications by one of the authors and two physicians recruited to supervise the knowledge base creation. Textual evidence that appeared to support the provided classifications was manually identified and annotated through a simple graphic interface. Treatment is also frequently used by experts to determine disease from reading charts [17,18]. Therefore, the authors hypothesized the presence of specific treatments (Tx) may increase the accuracy the intuitive disease classifier (described in detail below). In addition to the provided classification of disease from the text of the patient record, because While the training set was reviewed for evidence of the provided disease occurrences, evidence indicating disease treatments was also

annotated. In addition, negation and referent indicators were found in the training set and identified. This process and the textual evidence supporting each feature were reviewed by the author and two physicians using a consensus method of agreement. The identified textual evidence was considered sufficient to support the classification of a disease or a treatment, yet was not a comprehensive list of all evidence that may be found in the individual discharge summary. For example, the first occurrence(s) of textual evidence of a disease found may have been deemed sufficient and additional textual evidence found elsewhere in the summary overlooked.

To train the evidence base, the user identified and annotated textual evidence for each feature by browsing the text documents, and highlighting textual evidence indicating (or negating) a particular predefined classification, or feature of the text (see Figure 1). All of the textual indicators of concepts were identified in this manner from within the text.

When textual evidence is identified, it is added to the evidence base for a particular feature and propagated throughout the corpus of discharge summaries. This cascading process updates classifications for all messages that have not been reviewed and flagged as 'confirmed' according to the new evidence. One consequence of this is that system users can view the results of their training annotations as they work through the corpus. During this brief review, textual evidence, negation, and referent indicators that cause unintended classifications in other texts can be easily removed through the same interface, thus updating the evidence base and message classifications again. The textual classifier makes judgments on texts primarily on the occurrence of patterns of textual evidence. If no textual evidence is found in a text or the evidence refers to someone other than the patient, the textual system output for that discharge summary is "U" (Unknown). If textual evidence is found and it refers to the patient, a judgment of either "Y" or "N" is made based on negation rules.

Intuitive Classification

The intuitive classifier uses a Support Vector Machine (SVM) [19], which assigns classification based on patterns of textual classifier features found in the text. A similar technique using SVMs has been previously demonstrated[20]. The model of the feature space for the classifier is defined in advance by the user. Users define the names of the intuitive classification features (e.g. "Asthma"), and then also define a model of textual classification features to consider as evidence of an intuitive classification(e.g. [textual]"Asthma Tx"+ [textual] "Asthma Dx") . Any textual features could be associated with the judgment of an intuitive feature.

In this study, the intuitive feature "Asthma", is defined by the occurrence patterns of textual features "Asthma" (Dx) and "Asthma Tx" found in the text. Due to the formative nature of this study and considerations of the NLP competition, the models used for this challenge are oversimplified. The SVM was trained on the occurrence two textual feature classifications to classify the corresponding intuitive classification. Because this is a statistical pattern matching inference mechanism based on imperfect training data, the occurrence of a textual classification of "Y" or "N" does not in all conceivable cases indicate a corresponding intuitive classification of the same. However, ideally (and intuitively) the SVM would detect an association in the training set between a textual classification of "Y" and the correct intuitive classification of "Y". A similar production type system would likely incorporate more than the two features for each intuitive classification used in this study. For example, disease symptoms as well as medical procedures are also plausible features to consider adding to the model but were not used in this study.

The use case for the intuitive classifier begins when a user determines that a particular judgment is appropriate, yet no specific textual evidence supports this. In this case, the user selects the intuitive classification without explicitly annotating any evidence.

The intuitive classifier is trained on texts that 1) already have textual classifications as well as intuitive classifications and 2) are also flagged as having been reviewed and “confirmed”. The intuitive classifier makes judgments (classifications) on all other texts in the corpus. Consequently, intuitive judgments of “Y” and “N” are made for every message in the test set. For this challenge, the training set’s textual features were annotated from the rule-based process described above, and the intuitive designations that were provided were assigned to these same 300 training messages. The intuitive classifier was trained using the results of the text-based classifications from the 300 training messages and the intuitive judgments, and then tested using the results of text-based classifications from the test set.

Evaluation

The system was evaluated using precision, recall, and F-Measure for both intuitive judgments and textual judgments. Precision is the percent of classified texts that are correctly classified. Recall is the fraction of true classifications that were classified correctly by the system. The F measure is the weighted harmonic mean of the two, or F = (2*P*R) /(P+R). Macro averaging gives additional weight to rare classifications by giving each type of classification an equal weighting in the metric, regardless of how comparably rare it is. The primary and secondary evaluation metrics were macro-averaged F-measure and micro-averaged F-measure respectively.

As mentioned, the challenge annotations consist of multiple classification options for intuitive and textual judgments, ("Y", "N", and "Q") and ("Y", "N", "Q", and "U") respectively, where “Q” is “questionable” and “U” is unmentioned. Because this system makes only binary (“Y”, “N”) assignments for intuitive judgments and (“Y”, “N”, and “U”) for textual judgments, an additional evaluation against altered ground truth set is made for discussion. The test sets were evaluated against the “as is” ground truth data (manually annotated by experts) and also with (“Q”) records omitted from the ground truth data.

RESULTS

Precision, recall, and F-measure were calculated for each disease as well as the average calculated across all sixteen diseases. Macro F-measure for specific diseases had a range from . 48 (Obesity Textual) to .98 (Gout Intuitive). Micro averaged F-Measures were notably higher due to less of a penalty for missing the ‘Questionable’ class. F-,measures ranged from .89 (Hypercholesterolemia Intuitive and CAD Textual) to four .99. However, disregarding the 'Questionable ' class negatively affects the Macro-averaged precision calculation to a lesser significance. This is a result of removing the slightly weighted, yet perfect precision of the 'Q' records in the macro-averaged calculation.

Table 1 and Table 2 show the averaged results of the competition runs alongside comparable systems. The ‘as-is’ macro averaged recall and F-measure was dramatically lower than the other scores due to the system’s inability to assign a “Q” judgment. As shown in the “-Q” column, removing the “Q” records from the ground truth improved these values. Note, any document that is incorrectly classified in one class is also missing in the other, therefore micro-averaged precision will equal micro -average recall in this design.

DISCUSSION

This system ranked in the top ten in both the intuitive and textual tasks: 8/28 and 10/28 respectively. Although the system score was categorically penalized by its binary output in Macro level metrics, the system's performance was still comparable to that of other systems participating in the challenge. This suggests that our system, while simple in concept and execution, may still perform at a level that is sufficient for health related text-mining tasks.

Perhaps more important than system-centered performance metrics are the future implications of the system design. This study is a preliminary exploration of valuable contextual insights related to user-centered and interactive text mining systems. Although imperfect, the use of intuitive and textual metaphors for the classifiers appeared to increase the acceptability and familiarity of the system classifiers to users during the design and training processes. Biomedical NLP systems are scarce in practice, and increased attention to user-centered methodologies may help alleviate some barriers to adoption.

Although no formal user-centered evaluation has been performed at this time, the core design objectives (domain expert supervision, transparency of function, and simplicity of use) and subsequent engineering choices were reviewed and approved by stakeholders during the development process. Due to these design objectives, there are important considerations to implementing this system in a real context.

First, the system knowledge base is de novo, that is all of the textual evidence indicating features, negation, and referents arise from within the text without using any external sources of data. A possible advantage of this is there are no external libraries or data sources required to purchase or maintain, and the quality of external data sources is not a concern. Yet, when compared to today's trend of unsupervised learning and large complex knowledge bases, our design appears to be labor intense and limited in the total number of classes possible. However, our system design has the potential to efficiently answer specific questions posed by an expert regarding a reasonable sized corpus. Moreover, NLP systems are currently absent in clinical practice, so there is little-to-no scientific evidence regarding what types of system designs would actually be more effective in real world applications.

Second, this approach requires the input of subject matter experts to train the system. Although the effort required to annotate, train and maintain the knowledge base is non-trivial, requiring experts to train and maintain the system will likely build confidence in the resulting data output. This may be especially true if the end users are also subject experts. Furthermore, attention should be given to ensure the process of maintain the knowledge base is as efficient as possible. This system uses an iterative training /review process, in that the user views the annotated results of all existing rules when viewing the unclassified text. Using this method, the annotation process becomes more of a 'review and accept' process that may save time when compared to annotating the text from scratch.

Third, there are potentially negative considerations of the design related to the many contexts found within discharge summaries. For example, a user may annotate a string of characters as being indicative of a particular feature within a text without realizing that the same characters may have alternate meanings in other texts with different contexts. Immediately after a user annotates a new string of textual evidence, the knowledge base is updated and all texts containing the new evidence are automatically classified according to the updated rules. The interactive design that causes this may also alleviate the risk of these unintended misclassifications. In this case, the user will likely catch the misclassification when reviewing classified documents and update the knowledge base accordingly.

The benefits of this interactive and user-centered methodology have yet to be firmly established. However, the authors envision this type of system being most successful in an environment where 1) user confidence is critical, 2) the judgments may be highly subjective or change over time, 3) there is dedicated time for experts to train and review the knowledge base, and 4) There is a relatively low number of potential classifications (this study used 32 textual and 16 intuitive classifications) .

Experts who maintain the knowledge base as well as interpret and use the data output are potentially good candidates for this type of expert system. For example, a hospital may be interested in using discharge summary analysis to support ongoing quality assurance efforts. However, the tool is generally not intended to operate as standalone high-throughput text mining system. On the contrary, it is intended to be used as an expert's workbench: to analyze and operate in an interactive manner over large sets of text documents. However, the components could easily be adapted for high-throughput applications. For example, the i2b2 challenge is a batch style competition that may not typically be conducive to an expert workbench design system. However, our system performed relatively well in this setting for at least two reasons. First, our interpretation of the challenge 'questions' was quite practical: to answer a question over a large corpus: "Which of these patients are obese and what comorbidities do they have?" Secondly, the interactive nature of the system is found exclusively in training the knowledge base. Once trained, the knowledge base can be applied 'batch style' to any corpus.

Due to the dependency of intuitive features on textual classifications, any errors in textual classifications may propagate into the intuitive classifications as well and compound misclassifications there. However, this does not appear to be a significant problem in our study. In some cases the Intuitive classifier even outperformed the textual classifier. This may be explained by the additional information (disease treatments) used within the intuitive classifier.

The consequences of having such an informal approach to training and maintaining the knowledge base may lead to inconsistencies and difficulty in precisely evaluating the optimal performance of a system. For example, in a 'real world' setting, the output would be a combination of user-classified and machine classified data. However, the convenience and transparency provided to expert users who would like to "see for themselves" what the system is doing and why, may also add to the user's confidence in the system output.

Limitations of Evaluation

Our study has potential limitations related to the system design and the evaluation itself. The system may perform differently using discharge summaries from different computerized medical record systems, or when run on different samples of patients. The system would also likely perform differently when trained differently (e.g. by other users, different training sets, etc.). As for the user interface design, the system 'review for correctness' training approach may inadvertently encourage users to skip over sections or otherwise not thoroughly read the texts. Furthermore, the system's interactive feedback may give users a false confidence although rare events may still be missed.

While the system is designed to be as intuitive and transparent to users as possible, there is still a learning curve associated with annotating the texts and building the knowledge base. Initially, user inexperience may result in lower performance. In addition, the system was designed in collaboration with epidemiologists as key users for text classification tasks that may not translate easily into other contexts. Although we postulate that users will be more likely to trust results from a highly interactive system such as this, it is not known if this type of system would be scalable for general use.

Conclusion

This study illustrates the performance of this formative system at classifying discharge summaries in a controlled environment. The results of this evaluation suggest that this system is comparable to other state of the art text classification systems in terms of output metrics. These findings are useful to the continued development of the system and are a crucial step towards evaluating the value added of the system within the context of real-world problems. Furthermore, the intrinsic and value added potential of a simplified system may outweigh the costs and complexity drawbacks of a more complex system.

'Simplified' workbench style systems such as this one that are based on a fully supervised, interactive learning knowledge base have potential application throughout health services. The ability to analyze large sets of unstructured text documents in a flexible and intuitive manner may give unprecedented access to previously difficult information sources. For example, this system has the potential to analyze discharge summaries over time as an ongoing quality assurance process. It could also provide researchers a lens through which to answer questions pertaining to online health behavior found in message boards, blogs, etc. as well as social networks. However, these potential applications, while very promising, will require further investigation. Additional research is also needed to clarify potential advantages and disadvantages of the interactive and user-centered design philosophy versus other NLP application approaches.

Acknowledgments

This work was supported by the National Library of Medicine training grant #T15LM007442. Special thanks to CJ Inman and Quynh Bui for their help reviewing the classification evidence and to Anne Diekma for reviewing this manuscript.

References

1. Cohen AM, Hersh WR. A survey of current work in biomedical text mining. Brief Bioinform 2005;6 (1):57–71. [PubMed: 15826357]
2. Riloff E, Lehnert W. Information extraction as a basis for high-precision text classification. ACM Transactions on Information Systems (TOIS) 1994;12(3)
3. Jin-Dong K, Tomoko O, Jun'ichi T. Corpus annotation for mining biomedical events from literature. BMC Bioinformatics 2008;9(10)
4. Turner AM, et al. Modeling public health interventions for improved access to the gray literature. J Med Libr Assoc 2005;93(4):487–494. [PubMed: 16239945]
5. Savova G, et al. Mayo clinic NLP system for patient smoking status identification. J Am Med Inform Assoc 2008;15(1)
6. Chapman WW, et al. Classifying free-text triage chief complaints into syndromic categories with natural language processing. Artificial Intelligence in Medicine 2005;33(1):31–40. [PubMed: 15617980]
7. Hripcsak G, et al. Use of Natural Language Processing to Translate Clinical Information from a Database of 889,921 Chest Radiographic Reports. Radiology 2002;224(1):157–163. [PubMed: 12091676]
8. Friedman C, et al. Automated encoding of clinical documents based on natural language processing. Journal of the American Medical Informatics Association 2004;11(5):392–402. [PubMed: 15187068]
9. Zweigenbaum P, et al. Frontiers of biomedical text mining: current progress. Brief Bioinform 2007:bbm045.
10. Zweigenbaum P, et al. New frontiers in biomedical text mining. In. In: (2007). Proc Pac Symp Biocomput 12 2007:205–208.
11. Xiang Z, Zheng W, He Y. BBP: Brucella genome annotation with literature mining and curation. BMC Bioinformatics 2006;7(347)

12. Demaine J, et al. LitMiner: integration of library services within a bio-informatics application. Biomedical Digital Libraries 2006;3(1):11. [PubMed: 17052341]
13. Karamanis N, et al. Integrating natural language processing with Flybase curation. Pac Symp Biocomputing 2007;12:245–256.
14. Manaris BZ, Slator BM. Interactive Natural Language Processing: Building on Success. Computer 1996;Volume 29(7)
15. DeShazo JP, TA. Workshop on Challenges in Natural Language Processing for Clinical Data. Washington DC: 2008. Hands-on NLP:An interactive and user-centered system to classify discharge summaries for obesity and related co-morbidities.
16. Wicentowski R, Sydes MR. Using Implicit Information to Identify Smoking Status in Smoke-blind Medical Discharge Summaries. J Am Med Inform Assoc 2008;15(1):29–31. [PubMed: 17947620]
17. Gilbert EH, et al. Chart Reviews In Emergency Medicine Research: Where Are The Methods? Annals of Emergency Medicine 1996;27(3)
18. Humphries KH, et al. Co-morbidity data in outcomes research Are clinical data derived from administrative databases a reliable alternative to chart review? Journal of Clinical Epidemiology 2000;53(4):343–349. [PubMed: 10785564]
19. Chang, C-C.; Chih-Jen Lin. LIBSVM : a library for support vector machines. 2001 [cited; Available from: Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm
20. Clark C, et al. Identifying Smokers with a Medical Extraction System. J Am Med Inform Assoc 2008;15(1):36–39. [PubMed: 17947619]
21. Uzuner O. Recognizing Obesity and Co-morbidities in Sparse Data. J Am Med Inform Assoc 2009:M3115.

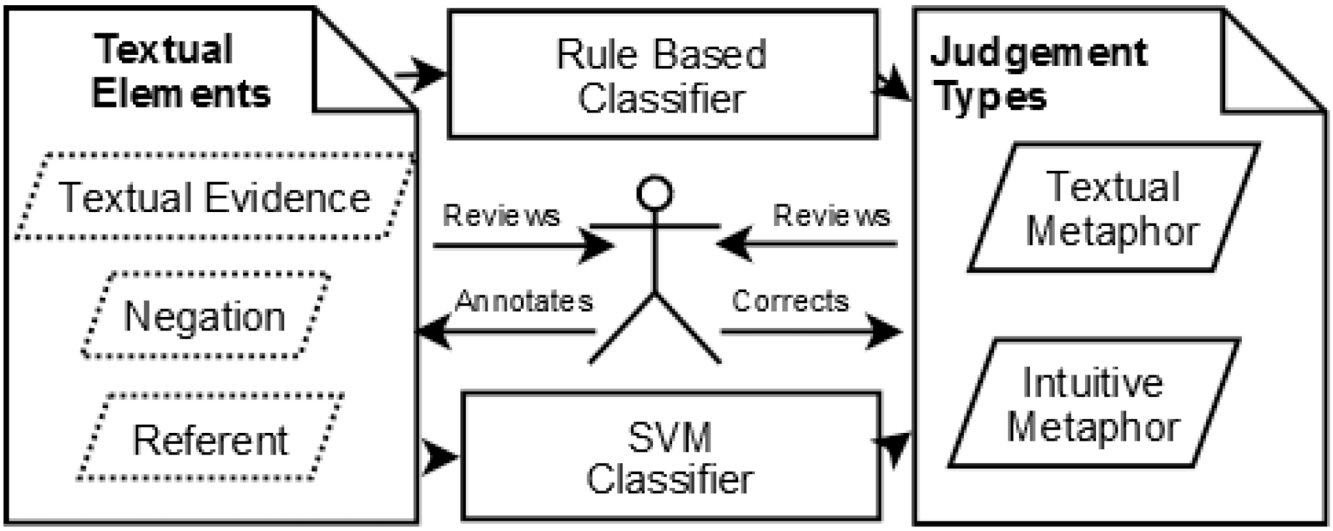

Figure 1.
A fully supervised and interactive classification system using two classifiers. Users interact between textual annotations and classifications made by the system.

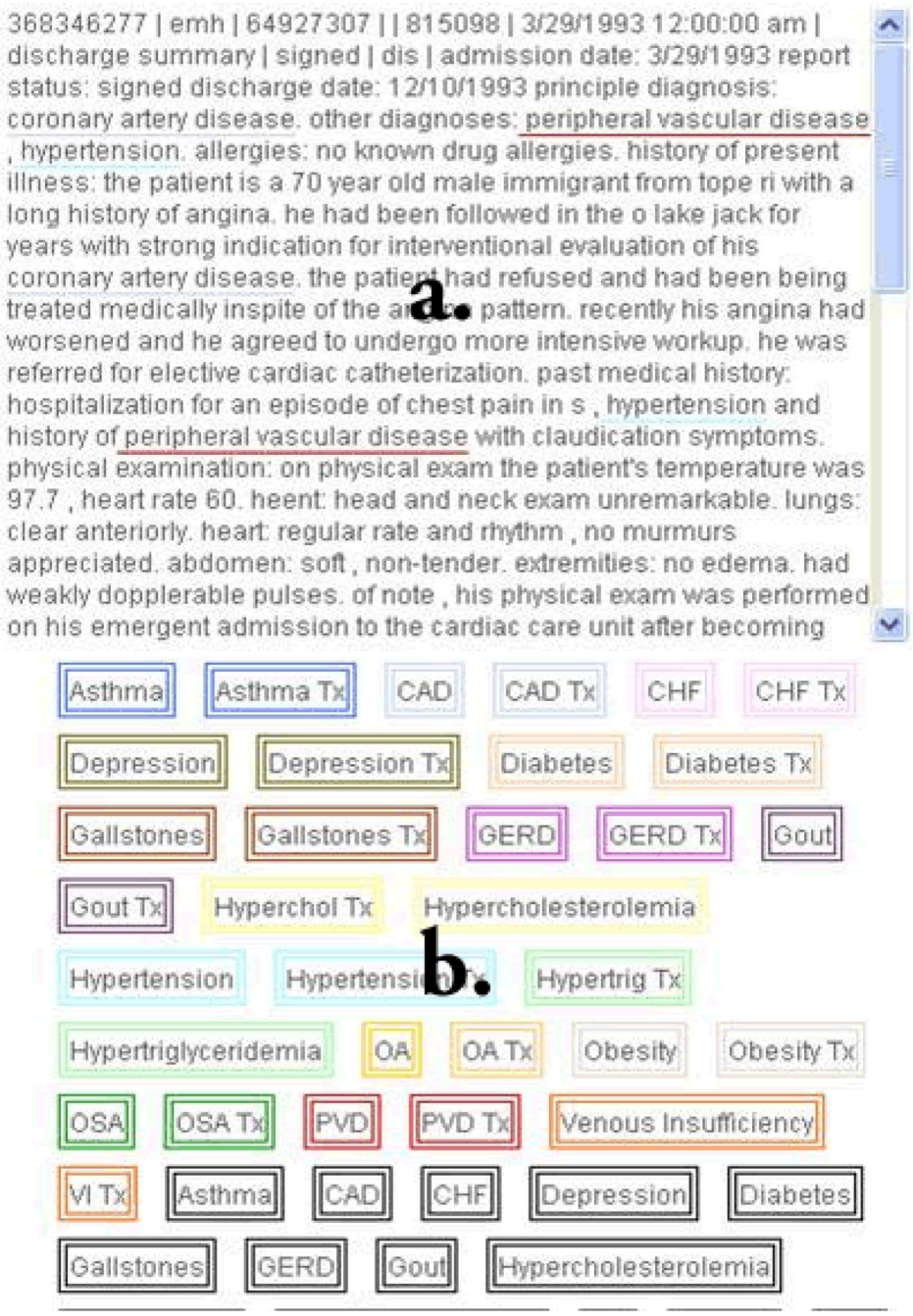

Figure 2.
Cutouts training screen showing underlined evidence for textual judgments (Figure 1 a.) and color coordinated buttons displaying potential features. When training the classifier, users see their textual annotations immediately applied to the knowledge base.

Table 1

System performance compared with 1)the average of the top ten finalists, and 2) the best score in each category from the textual competition of the 2008 *Obesity Challenge*. The tested system does not support "Questionable" (Q) as a classification. Output was evaluated both as-is ('as-is') and also with the 'Q' results omitted from the ground truth (−-Q'). For more details, see text.

		Textual Competition					
	'Q'	Micro Avg. Precision	Macro Avg. Precision	Micro Avg. Recall	Macro Avg. Recall	Micro Avg. F-Measure	Macro Avg. F-Measure
Our System Run	**-Q**	.966	.809	.966	.832	.966	.820
	as-is	.964	.855	.964	.624	.964	.614
Avg. Top 10 Finalists*	**as-is**	.969	.805	.969	.715	.969	.734
Overall Best Scores*	**as-is**	.977	.855	.977	.805	.977	.805

*Compiled from Uzuner 2009[21]

Table 2

System performance compared with 1)the average of the top ten finalists, and 2) the best score in each category from the intuitive competition of the 2008 *Obesity Challenge*. The tested system does not support "Questionable" (Q) as a classification. Output was evaluated both as-is ('as-is') and also with the 'Q' results omitted from the ground truth ('−Q'). For more details, see text.

		Intuitive Competition					
	'Q'*	Micro Avg. Precision	Macro Avg. Precision	Micro Avg. Recall	Macro Avg. Recall	Micro Avg. F-Measure	Macro Avg. F-Measure
Our System Run	**-Q**	.954	.960	.954	.932	.954	.945
	as-is	.952	.972	.952	.622	.952	.629
Avg. Top 10 Finalists*	**as-is**	.957	.694	.957	.638	.957	.645
Overall Best Scores*	**as-is**	.965	.972	.965	.659	.965	.675

*Compiled from Uzuner 2009[21]