Leveraging Rich Annotations to Improve Learning of Medical Concepts from Clinical Free Text

Abstract

Information extraction from clinical free text is one of the key elements in medical informatics research. In this paper we propose a general framework to improve learning-based information extraction systems with the help of rich annotations *(i.e., annotators provide the medical assertion as well as evidences that support the assertion). A special graphical interface was developed to facilitate the annotation process, and we show how to implement this framework with a state-of-the-art context-based question answering system. Empirical studies demonstrate that with about 10% longer annotation time, we can significantly improve the accuracy of the system. An approach to provide supporting evidence for test documents is also briefly discussed with promising preliminary results.*

1. Introduction

Extracting key and actionable information from the Electronic Medical Record (EMR) is critical in improving the quality of care, driving the process efficiencies, and reducing cost in the healthcare industry. Despite the increasing emphasis on collecting key information in structured fields of EMRs, a great amount of information is still hidden in the clinical free text. Information extraction from this type of data has been a hot topic in the last decade, ranging from rule-based expert systems [1,2] to learning-based natural language processing (NLP) tools [3-6]. The learning-based tools are often preferred as they provide higher level of flexibility and can often deal with a variety of data types.

Many of the information extraction tasks involve learning of certain medical concepts from the clinical free text, or learning to answer certain clinical questions about the patient. For instance, hospitals in the United States are required by CMS (Centers for Medicare & Medicaid Services) to submit answers of certain quality related questions (called *quality measures*) after patient discharge [7]. The answers and the corresponding evidences are often found in the free text medical records (e.g., discharge summary) of the patient. Example questions include "*Was the patient given aspirin within 24 hours of admission?*" and "*Did the adult patient smoke cigarettes anytime during the year prior to hospital arrival?*" Most of the questions are yes/no questions, and it requires significant efforts to manually look through available text documents, even if the search is computer assisted using simple tools. The human abstractor has to find the evidence and ascertain it is correct and complete. A decision cannot be made by looking at evidence found at one place (e.g., one clinical note) as there may be conflicting or supporting evidence at other places.

Traditional learning-based tools tackle this type of information extraction tasks as (binary) classification problems. The conventional workflow involves a) collecting a set of clinical documents, b) annotating these records by clinical experts (1 if the answer is yes and 0 otherwise), and c) training a supervised (or potentially semi-supervised) system (or classifier) using the documents and labels. The system can then automatically label new documents as yes or no for the specific question.

The binary labels from the annotators are often called *basic annotations*, since they are the minimal annotations that a learning system would need. In clinical practice we can often obtain *rich annotations*, i.e., the clinical experts not only give the binary label for each example, but also highlight the reason (also called *evidence* or *rationales*) that the label is induced from. The simplest form of rich annotations is the part of text that leads to the yes/no label. For instance, if the example contains the sentence "*the patient has no history of alcohol abuse, and does not smoke*", the annotator will label it as "no" for the aforementioned smoking question, and also highlight the evidence *does not smoke* since this part of the sentence leads to the "no" label. Providing this additional evidence adds minimally to the annotation effort, since the annotators would need to read the whole free text regardless, and highlighting the relevant part of the text would be simple if an easy-to-use graphical user interface is provided, such as selecting a contiguous piece of text using the mouse. We will investigate this in more detail in Section 2.

In this paper, we show how to use the rich annotations to improve learning of medical concepts in a learning-based system. The proposed algorithm is a principled approach to incorporate information from the highlighted evidences into the learning process, and the trained classifier is able to use information from both the original free text and the rich annotations. At the testing time, the classifier can be applied without any annotation. The proposed approach is general and should work with most learning-based systems. In our implementation we choose a state-of-the-art system[6] and show how using rich annotations can significantly improve the classification performance. We also briefly discussed how to provide relevant evidence for testing examples, since for practical systems the users would prefer to see not only the label but also the rationale behind the label.

The rest of the paper is organized as follows. Section 2 describes the annotation process that we used in obtaining the rich annotations, and Section 3 presents the mathematical formulation of how to learn from both the class labels and rich annotations. A detailed example implementation is presented in Section 4, followed by empirical studies in Section 5. Section 6 discusses how to highlight evidence for testing examples and presents some preliminary results. Section 7 concludes the paper.

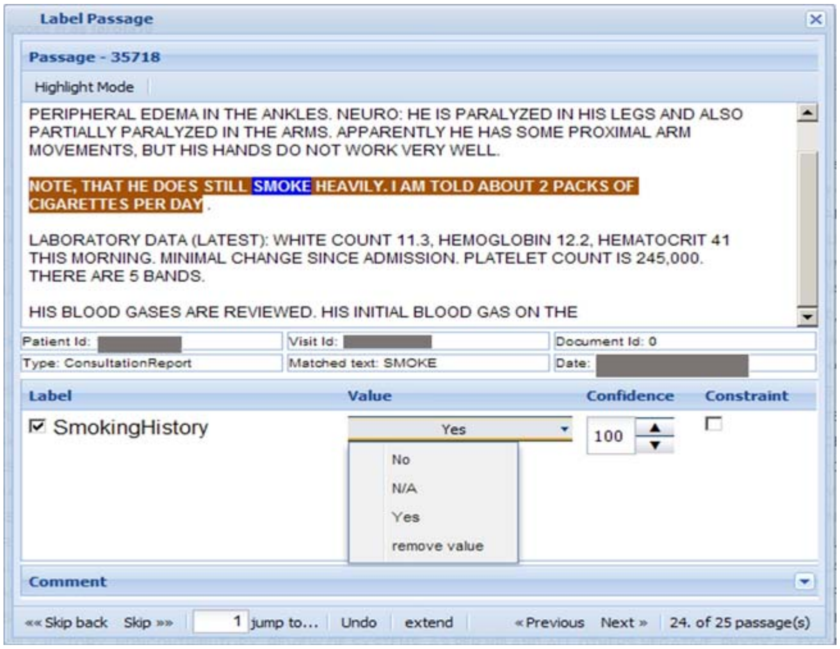
2. The Annotation Process

To get the rich annotations, we show the training example (document or passage) to the annotators, ask them to label it as yes or no (or not applicable) according to the clinical guidelines, and also to highlight the evidence that supports the label. We developed a GUI for this labeling task as shown in Figure 1. The annotator selects the Yes/No/NA label and highlights the portion of text that they consider evidence. In cases, where no evidence can be found or the passage is not applicable to the question, the highlighting step can be skipped. To further simplify the annotation process, we have made the following assumptions:

1) The evidence for each example contains only one continuous portion of the text
2) Examples with Yes label will only have positive evidence highlighted in the text (similarly negative in case of a No)

The first assumption makes highlighting easier, as the annotators do not need to highlight individual words that are relevant, but simply drag their mouse over the whole phrase/sentence. The second assumption automatically links the evidence to the assigned label, so annotation for the example is complete just after relevant texts are highlighted.

With this annotation process, providing rich annotations does not add significantly to the effort for the annotators, as they can perform this while reading the text. In our experiments the annotators spent on average 1.5 hours to annotate 100 examples (of about 200 words each) without providing any rationale. With the rationale, the time changed marginally to 1.6 hours for the same 100 examples which is about 10% longer. In our experience, this additional effort was acceptable to annotators across the board.



Label Passage

Passage - 35718

PERIPHERAL EDEMA IN THE ANKLES. NEURO: HE IS PARALYZED IN HIS LEGS AND ALSO PARTIALLY PARALYZED IN THE ARMS. APPARENTLY HE HAS SOME PROXIMAL ARM MOVEMENTS, BUT HIS HANDS DO NOT WORK VERY WELL.

NOTE, THAT HE DOES STILL SMOKE HEAVILY. I AM TOLD ABOUT 2 PACKS OF CIGARETTES PER DAY .

LABORATORY DATA (LATEST): WHITE COUNT 11.3, HEMOGLOBIN 12.2, HEMATOCRIT 41 THIS MORNING. MINIMAL CHANGE SINCE ADMISSION. PLATELET COUNT IS 245,000. THERE ARE 5 BANDS.

HIS BLOOD GASES ARE REVIEWED. HIS INITIAL BLOOD GAS ON THE

Patient Id: [redacted]	Visit Id: [redacted]	Document Id: 0
Type: ConsultationReport	Matched text: SMOKE	Date: [redacted]

Label	Value	Confidence	Constraint
☑ SmokingHistory	Yes	100	□

Figure 1. GUI for the annotation tool. Annotators can provide label value, confidence of the label (optional), and highlight evidences. See Section 5 for more details. *PHI is redacted*.

3. The Formulation

In standard linear binary classification, each training text is labeled as 1 for yes and 0 for no. Let x_i denote the features for an example text i computed from a dictionary of dimension d. Let $\mathbf{X} = \{x_1,\ldots,x_N\}$ and $\mathbf{Y} = \{y_1,\ldots,y_N\}$ denote the N training examples and their labels, respectively. In the training phase, we learn the model parameter $w \in \Re^d$ by minimizing a cost function C between $\mathbf{X}$ and $\mathbf{Y}$ plus a regularization term on w, which can be denoted in the general form as:

$$w^* = \arg\min_{w} \sum_{i=1}^{N} C(y_i, w^{\mathrm{T}} x_i) + \lambda \cdot g(w) \quad \text{with possible constraints on } w.$$

Here $g(w) \geq 0$ denote the regularization term on w (such as $||w||^2$), and $\lambda \geq 0$ is the regularization parameter.

When the rich annotations are available, let $\mathbf{R} = \{r_1,\ldots,r_N\}$ denote the highlighted evidences, where r_i denote the word sequence of highlighted evidence for example i. The objective is to learn the weight vector w such that the cost function C between $\mathbf{X}$ and $\mathbf{Y}$, conditioned on the rich annotations $\mathbf{R}$, is minimized (with regularization).

Intuitively, the highlighted evidences provide additional insight to the assigned class label. Since each evidence r_i is simply a sequence of words, let us assume that additional features z_i of dimension s can be induced from the annotations for each example i. With this feature augmentation we can formulate the learning problem as:

$$(w^*, v^*) = \arg\min_{w,v} \sum_{i=1}^{N} C(y_i, w^{\mathrm{T}} x_i + v^{\mathrm{T}} z_i) + \lambda_1 \cdot g(w) + \lambda_2 \cdot g(v) \quad \text{with possible constraints on } w \text{ and } v,$$

where $v \in \Re^s$ is the weight vector for the evidence-induced feature z_i, and the regularization term involves both w and v (one can also assume a different regularization term for w and v). Then one can use the same solver as the standard binary classification to solve this optimization problem. One example solver will be discussed in detail in the next section.

One important requirement of the additional features z_i is that *they should not embed the fact that these features are from the highlighted evidences*, i.e., these features can also be obtained for an example without any highlighted evidence. This is critical, as at the testing time there will be no evidence available and one should still be able to apply the learned model as:

$$y^* = \mathrm{sgn}(w^{*\mathrm{T}}\, x^* + v^{*\mathrm{T}}\, z^*),$$

where x^* and z^* are the feature vectors for the testing example, induced from the original dictionary and the training rich annotations, respectively.

The proposed approach adds information from the rich annotations as additional features, which differs from related work [8,9] that incorporate rich annotations as constraints or (generative) observations (primarily for non-medical classification problems). The advantage of generating new features is that we can choose any learning-based system with these additional features, which offers great flexibility.

In the next section we will show some approaches for extracting features z_i from the rich annotations, and explain how this works with a state-of-the-art learning-based system for learning medical concepts from clinical free text.

4. The Implementation

We provide some details on how the proposed framework can be implemented in practice. In the following we focus on bag-of-words (BoW) features, and assume that the learning-based system uses (certain transformation of) these features to train the classifier. Note that the framework is general and there are other implementation options (such as n-gram approaches) that we do not cover in this paper.

4.1 Feature Extraction from Rich Annotations

The rich annotations provide insight to the assigned class label, and since each rich annotation r_i is a word sequence, we can build the augmented dictionary for z_i by selecting the most informative words from the annotations $\mathbf{R}$. This approach is similar to the TF-IDF approach (used in Information Retrieval) of weighing each search term. More specifically, let $\mathbf{\Omega}$ be the set of words that at least occur once in the annotations $\mathbf{R}$, and let u_i be the set of words that was *not* in the annotation r_j for each example i (if r_i is empty then u_i contains all the words in the example). Then we define the *informative score* of each word $t \in \Omega$ as:

$$s(t) = \sum_{i=1}^{N} \mathrm{I}(t \in r_i) \Big/ \sum_{i=1}^{N} \mathrm{I}(t \in u_i) \, .$$

The numerator is the count of the number of occurrences of term t inside the annotations, and the denominator is that of term t outside the annotations. Intuitively, a word which occurs more frequently in annotations will have a higher score, and they should probably help the learning-based system to distinguish positive from negative examples. Hence we rank all the terms in Ω based on this score and select the top words from the list to construct z_i. In our experiments, we select the top 30 words. Note that common stopwords (such as *for* and *the*) will get a low score ($\approx$ 1) and will probably not rank at the top of the list.

This approach of extracting features from rich annotations meets the aforementioned requirement that the features do not embed the fact that they come from the annotations. In other words, they just look like additional (word) features to the learning-based system, and we can easily construct the z vector for a testing example by, e.g., looking at the presence/absence of the specific word.

4.2 Learning Classifiers using Features from Rich Annotations

With an augmented dictionary with word features from the rich annotations, many standard learning-based systems can be used to solve the optimization problem proposed in Section 3. In this subsection we explain how this can be done with the approach from Rosales et al. [6] and show how the proposed method can improve the state-of-the-art solutions for learning medical concepts.

Figure 2 shows the major steps of the approach and visualizes how rich annotations can be incorporated into the system. The training examples are passages (i.e., portion of the document identified by a relevant text, marked as orange). The feature representation in this approach goes beyond BoW and contains distance-based features in the context around the relevant text (marked as light blue). Each word in the nearby context will generate exactly one feature (based on the distance to the relevant text). Some meta-data features, e.g., type of document, date of document, are also used in learning (marked as dark blue). A hyper-plane-based classifier is learned using this feature representation, and the optimization problem is as follows:

$$w^* = \arg\min_{w} \sum_{i=1}^{N} C_{svm}(y_i, w^{\mathrm{T}} x_i) + \lambda \cdot \|w\|_1 \qquad s.t. \quad \sum_{i=1}^{N} \left[y_i \cdot w^{\mathrm{T}} x_i + C_{svm}(y_i, w^{\mathrm{T}} x_i) \right] \ge 1 \, .$$

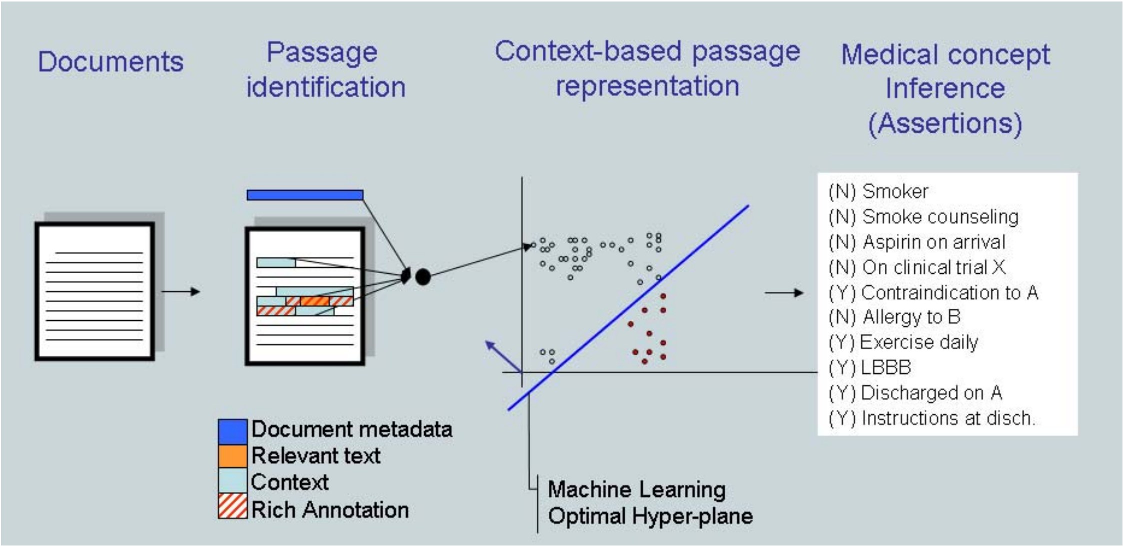


Figure 2. Illustration of how rich annotations can be incorporated into the learning-based system[6].

Here C_{svm} is the hinge loss [10] (used in standard support vector machines), and $||w||_1$ denotes the 1-norm of vector w. This l_1 regularization often leads to a sparse model and lower computational complexity [10].

When the rich annotations are present (marked as red stripes), they provide additional (more informative) words to the dictionary, and they will also be used in computing the distance-based features. The new optimization problem can be defined as follows:

$$(w^*, v^*) = \arg\min_{w,v} \sum_{i=1}^{N} C_{svm}(y_i, w^{\mathrm{T}} x_i + v^{\mathrm{T}} z_i) + \lambda \cdot \big\|[w;v]\big\|_1,$$

$$s.t. \quad \sum_{i=1}^{N} \left[y_i \cdot (w^{\mathrm{T}} x_i + v^{\mathrm{T}} z_i) + C_{svm}(y_i, w^{\mathrm{T}} x_i + v^{\mathrm{T}} z_i) \right] \geq 1$$

Here we consider the 1-norm regularization on the concatenation of vector *w* and *v* to simplify the formulation. The regularizations of *w* and *v* share the same regularization parameter λ, and this can be relaxed if more flexibility is needed. The standard learning approach presented in [6] can be applied to solve this problem.

5. Empirical Studies

Experiments were performed using actual EMR data from various medium/large-size hospitals (names undisclosed due to privacy agreements). We designed our experiments to work at the passage level, similar to the approach taken by [6]. With the help of expert medical personnel (such as expert chart abstractors), we concentrated on gathering information about various medical concepts. These concepts were chosen primarily due to their prevalence in quality reporting [7].

We built 5 datasets, one for each concept. The questions and the number of passages are shown in Table 1 (refer to [7] for the detailed definitions of these questions). These passages were obtained from a set of ~10 million sentences by searching, in each case, for a few keywords related to the concept of interest and provided by the clinical experts. A random subset of the matching sentences were labeled by the expert and saved using the GUI shown in Figure 1. As expected, for all the concepts, keywords were only useful at a first level retrieval. Not surprisingly, a mixture of completely irrelevant, affirmative, or negative sentences were obtained as a first pass; e.g., not all patients with passages containing the keyword *smoke* are actual smokers or even have a history of smoking.

In our problem, the expert labeled each chosen passage with the labels T=True, F=False or N/A=Not Applicable to indicate the following: (1) T → the concept is present and affirmative in the sentence; (2) F → the concept is absent or it is present but negated in the passage; (3) N/A → the passage is not applicable to the concept in question (note that this could also be achieved by simply skipping the passage). In addition to the actual label, the annotator highlighted (by selecting the text with the mouse) the reason (evidence) on which the label was being based.

5.1 Experimental Methodology

For the experiments we used a passage size of 200 tokens, centered at the concept-specific words provided by the clinical experts. For each dataset, we first divided it into two subsets, one held out for testing only (30%) and one used for training (70%). From the training subset, a portion was assigned for actual training (75%) and another portion for cross-validation (25%). The method requires one tunable parameter, namely λ. These settings were the same for all datasets. A total of 5 folds were performed where the above subsets were always randomized. For the proposed approach with rich annotations, we added top 30 words to the dictionary based on the informative score, and then followed the same training methodology.

Table 1. Accuracy and area under the ROC curve (AUC) of various classification schemes for medical concepts.

		BoW		w/o annotations		w/ annotations	
Medical concepts / questions	*N*	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy
ST Elevation Assessed	506	0.54(0.02)	0.56(0.03)	0.87(0.03)	0.82(0.01)	**0.96(0.01)**	**0.90(0.02)**
Assessed for Rehabilitation	278	0.52(0.00)	**0.78(0.04)**	0.72(0.03)	0.74(0.05)	**0.85(0.07)**	0.76(0.03)
VTE Present on Arrival	636	0.62(0.00)	0.68(0.01)	0.90(0.00)	0.83(0.02)	**0.95(0.02)**	**0.89(0.03)**
Smoking History	320	0.49(0.01)	0.50(0.00)	0.72(0.01)	0.61(0.02)	**0.83(0.04)**	**0.80(0.01)**
Joint (*e.g. knee*) Revision	303	0.61(0.01)	0.51(0.00)	0.87(0.00)	0.75(0.03)	**0.94(0.06)**	**0.80(0.05)**

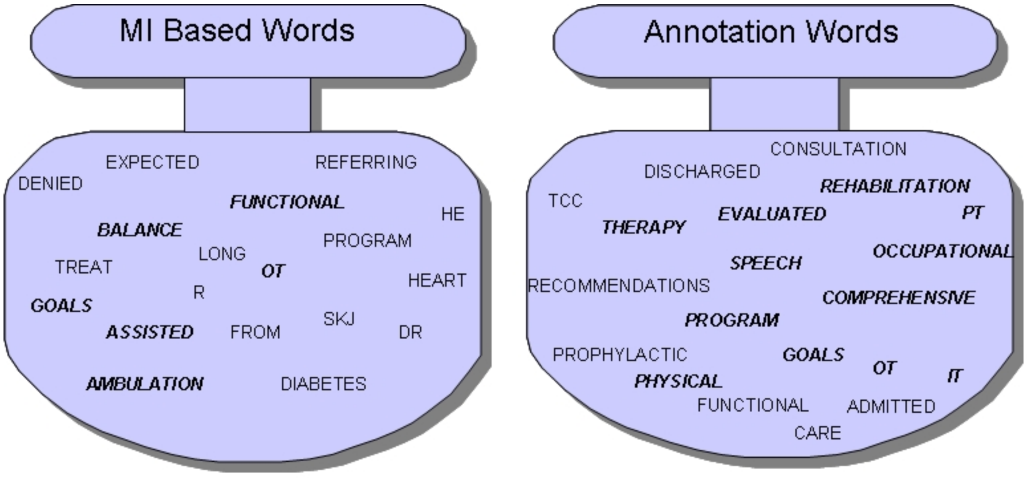


Figure 3. Top 20 words ranked using the MI based (left) and annotation based (right) method for the *Assessed for Rehabilitation* question. Words considered clinically useful are shown in bold and italic.

The learning-based system in [6] was experimented with two different settings: using manually constructed dictionary only, and using an automated query expansion based on mutual information (MI). In the latter approach, MI is used to rank terms specific to the concept of interest given the label value [11], and top ranked terms are added to the dictionary. In Figure 3 we compare the top ranked words based on the MI criteria and on the rich annotations for the Assessed for Rehabilitation question. Clearly the words based on the annotations are more clinically relevant to the question. We observed similar results for the other questions.

5.2 Results

For each concept we compared three competing approaches: (1) A model employing the BoW feature representation and trained using the same large-margin based approach; (2) A model using the context based feature representation, but without using the rich annotations – this is the best approach in [6] where automated query expansion is used to augment the dictionary; (3) A model using our full approach, which utilizes the introduced context sensitive text representation, the automated dictionary construction steps, and the rich annotations. The results are summarized in Table 1. We measure the performance using both the accuracy (defined as the proportion of passages correctly classified) and the AUC (Area Under the Received Operating Characteristic Curve) for all the approaches. For each entry in the table, the standard deviation is also included in parenthesis. Bold numbers show the best performance across the three competing approaches. For cross-validation (selecting the best tuning parameter), we used the AUC as target performance measure.

Table 1 clearly shows that leveraging the rich annotations improved the classification performance for almost all the questions. The improvement is statistically significant for all the questions and both evaluation metrics, except for the accuracy on the *Assessed for Rehabilitation* question in which BoW was slightly better.

6. Highlighting Evidence for Testing Examples

In the previous sections, we discussed how to leverage the available rich annotations from clinical experts to improve the learning system. However, the system is designed to only output the class label for a test example, not the evidence behind the label. For instance, simply highlighting the words used in feature calculation is not a comprehensive solution. This is a critical limitation for clinical applications, as users would often prefer to see the rationale behind the machine generated label (e.g., before making a final decision). For instance for the CMS measure questions, even if an automated system is in place, most hospitals would require an abstractor to work with such a system and confirm the answers before submitting to CMS. It would be hard for the abstractor to do so if only the yes/no answer is provided.

In this section we briefly describe an approach which is able to generate supporting evidences when the system is in use. At the core we develop a probabilistic sequencing model for the (positive and negative) annotations, and at testing time it is able to output, at each token level, a probability score of it being part of the positive or negative evidence. This can be seen as an augmented version of the well known conditional random fields (CRF) models [12].

6.1 Learning to Annotate

Experts annotate clinical documents by reading through the document, locating evidences for the medical concept of interest, and then concluding a final yes/no answer. This is also how an inexperienced person learns to annotate. We try to mimic this learning experience and target at generating the correct evidence as the first step. The class label can be easily induced from these evidences as a second step.

We follow the same annotation process to get the rich annotations on the training examples. Since we know the label value of each example, the rich annotations are automatically assigned positive (if the label value is 1) or negative (if the label value is 0). We then break down each training example into a sequence of word tokens, and assign label "Y", "N" or "N/A" to each token depending on the class label and associated rich annotations. All tokens in a positive annotation get label "Y", whereas all tokens in a negative annotation get label "N". All the other tokens (i.e., not in an annotation) get label "N/A". After this step, the sequence of the token labels determines the rationale in each example.

We build a probabilistic sequencing model that takes these positive and negative annotations as input and learns how to generate an annotation (positive or negative) for an unseen example. Let x_t be the token sequence for one example document, y_t be the token level label sequence (each label in y_t is one of "Y", "N" and "N/A"), and r be the word sequence highlighted from annotation. We model the probability of y_t given x_t conditioned on r as follows:

$$\Pr(y_t \mid x_t, r) \propto \exp\left(\sum_j \mu_j f_j(x_t, y_t) + \sum_k \eta_k g_k(x_t, y_t, r) \right).$$

Here f and g are *feature functions* that contribute to the conditional probability, as used in the CRF language. Each feature has a weight parameter quantifying how much it contributes to the overall conditional probability. We make a distinction between f and g such that f denotes the widely used natural language features (e.g., *n*-grams, part-of-speech tagging) in CRF modeling, and g denotes the specific features that involve the annotation r (which will be discussed in the next subsection). Model fitting involves finding all the feature function weights μ and η, which can be done efficiently via large scale numerical optimization tools such as the L-BFGS. For more details on learning and inference under such models, please refer to the literature [12].

6.2 Features from Rich Annotations

Besides the standard contextual feature functions, we exploit additional features derived from the rich annotations to help improve the probabilistic sequencing model. They are defined under the assumption that each example document is a passage (passages are explained in detail in Section 5). Some of the features used were:

- Dictionary containing the top ranked words from annotations, as described in Section 4.1.
- Regular expression pattern match – if a token matched a part of the relevant text (identified by clinical experts along with the annotations and used for obtaining passages), it would potentially weigh more as there are more chances that it contributes to the evidence.
- Negations features on the matched word in the passage context.

6.3 Preliminary Results

As a preliminary evaluation, the proposed approach is applied on the smoking question (*Did the adult patient smoke cigarettes anytime during the year prior to hospital arrival?*) using the data introduced in Section 5. 124 passages out of the total 320 passages were used for training the system, and 96 passages (corresponding to 53 patients) were used for testing the learned model. There were no overlapping of patients between the training and testing data set.

Along with the additional feature functions from the rich annotations, we used standard unigram, bigram and part-of-speech tagging features as standard feature functions. After we obtained the predicted evidence (token-level class labels) for a test passage, we further induced the passage-level class label (yes/no) by looking at the evidence. Note that at testing time one passage might output many evidences, positive and/or negative, depending on the context. We took a simple rule that if the number of tokens that are labeled positive ("Y") is larger than the number of tokens that are labeled negative (“N”), the passage is assigned 1 (yes) label. Otherwise we assigned 0 (no) label for the passage. This simple rule partially mimics the thinking that an expert would have to label at the passage level.

Since we have the rich annotations for all the 320 passages, we present results at: 1) the word/token level output by looking at how many tokens were correctly labeled; 2) the passage level output by comparing the yes/no predictions with the expert labels; and 3) the patient level output by grouping all passages for a particular patient together (for the smoking question, if one passage says yes, the patient is labeled as yes).

At the word/token level, out of around 200 words per passage, the learned probabilistic sequencing model made an error on 75 words in average. This includes *over-annotating* – more words were labeled than what are necessary, *under-annotating* – not enough words were labeled to make the evidence, and *wrong-annotating* – e.g., the words were labeled as positive evidence but they were actually negative or not evidence at all. For practical applications the first two types of errors are probably acceptable given that the evidence is almost correctly highlighted, and the last type is a critical error as it flips the class label for the passage.

At the passage level, the confusion matrix is shown in Table 2. The sequencing-based model has accuracy of 86/96 = 90%, which is better than the classification based methods shown in Table 1. The sensitivity and specificity are 85% and 93%, respectively, which seems to be better as well. This shows that by correctly identifying the evidences for the test passages, we will be able to get a better passage-level classification performance. It also shows that most of the word/token level mismatches were probably over-annotating or under-annotating, since the wrong-annotating happened for only 10 passages out of 96 passages.

Table 2. Confusion matrix for passage-level classification using the probabilistic sequencing model.

	Predicted Positive Passage Label	Predicted Negative Passage Label
Actual Positive Passage Label	33	6
Actual Negative Passage Label	4	53

At the patient level, out of the 53 patients in the test set the number of patients classified correctly was 44 (accuracy is 83%), showing a solid performance for this difficult medical concept. Comparing this with the passage-level results, we see that the passage-level performance does not always lead to the same patient-level result. The reason is that correctly classifying more passages for the same patient does not necessarily improve the patient-level result. We want to emphasize that the system will output not only the class label for the tested passages / patient, but also the evidence that the label is induced from. So in practice the system would help the human users to make the final decision effectively and efficiently (even in case of wrong-annotating since the users will easily figure this out).

7. Conclusion

With the ever growing regulations such as quality reporting and meaningful use (partly due to the HITECH act of ARRA), the need for the learning-based NLP systems that are able to answer questions mined from clinical data is growing rapidly. At the same time, for such systems to gain acceptance, they need to be more accurate and include knowledge and reasoning from clinical experts in a principled fashion. In the proposed system we established that the use of rich annotations yields better results with only marginal increase in the effort in annotation. We also proposed a principled and generic approach of augmenting learning-based systems with these annotations. Not only does our system improve on accuracies, it also allows us to highlight evidences in text that were important to the classifier to make the final decision. This is especially important in domains like healthcare, where in addition to the final outcome the users want to understand what reasons were considered when that decision was made. This is something existing classification systems were unable to provide.

Our future work in this direction is to combine the sequencing based probabilistic models with the classification models such that richer, more informative features from the sequencing model can be used inside the classification model, and the output of the classification model can be re-fed into the sequencing model to improve highlighting. Combining them into one unified system would provide the most benefit in working with rich annotations.

Another future work is to relax the assumption that positive and negative annotations are not allowed to co-exist in the same passage. That would require more effort from the clinical experts to annotate the passages, but should be necessary for learning of some medical concepts.

Reference

[1] D. Aronow and K. Coltin. *Information technology applications in quality assurance and quality improvement, part II*. Joint Commission Journal on Quality Improvement, 10:465–478, 1993.

[2] Shusaku Tsumoto. *Automated extraction of medical expert system rules from clinical databases based on rough set theory*. Information Sciences, Volume 112, Issues 1-4, December 1998, Pages 67-84

[3] C. Friedman and G. Hripcsak. *Natural language processing and its future in medicine: Can computers make sense out of natural language text*. Academic Medicine, 74(8):890-895, 1999.

[4] X. Zhou, H. Han, I. Chankai, A. Prestrud, and A. Brooks. *Approaches to text mining for clinical medical records*. ACM Symposium on Applied Computing, pages 235-239, 2006.

[5] A. Roberts, R. Gaizauskas, and M. Hepple. *Extracting clinical relationships from patient narratives*. BioNLP 2008: Curr. Trends in Biomedical NLP, 2008.

[6] R. Rosales, F. Farooq, B. Krishnapuram, S. Yu and G. Fung, *Automated Identification of Medical Concepts and Assertions in Medical Text*, Proceedings of AMIA 2010.

[7] CMS / Joint Commission Quality Measures. http://www.jointcommission.org/PerformanceMeasurement/PerformanceMeasurement/Current+NHQM+Manual.htm

[8] O. Zaidan, J. Eisner, and C. Piatko. *Using "annotator rationales" to improve machine learning for text categorization*. In Proceedings of NAACL HLT, pages 260–267, April 2007.

[9] O. Zaidan and J. Eisner. *Modeling annotators: A generative approach to learning from annotator rationales*. In Proceedings of EMNLP, pages 31–40, October 2008.

[10] C. Cortes and V. Vapnik, *Support-Vector Networks*, Machine Learning, 20, 1995.

[11] T. Cover and J. Thomas. *Elements of Information Theory*. Wiley Interscience, Hoboken, 1991.

[12] John Lafferty and Andrew McCallum and Fernando Pereira, *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data*, Proceedings of the International Conference on Machine Learning (ICML-2001), 2001