Automatically Highlighting Relevant Text for Cooperative Human/Computer Document Classification

Abstract

We consider the domain of biomedical information retrieval, where we wish to process the (potentially copious) clinical notes in an electronic medical record to answer a question such as, "Has this patient been treated for breast cancer?" The answer to such a question is often buried deep in a very long document, and thus notoriously difficult to answer accurately in a fully automated fashion. We propose a probabilistic graphical model called a *word-label regression model* that can learn how to specific highlight passages in the text that a human expert should examine to determine the document label.

1 Introduction

Some document classification problems are not amenable to a fully automatic solution. Consider a large database of documents, each chronicling a patient's treatment history (that is, an "Electronic Medical Record", or EMR). A record contains (potentially) hundreds of "clinical notes", which are electronic notes written by physicians. The goal is to use those notes to identify those patients who at some time have had a particular medical condition, such as breast cancer (Stanfill et al., 2010).

The problem is exceedingly difficult. Because breast cancer is uncommon (incidence less than 1%), the classes are highly unbalanced, which is a classic problem (Japkowicz, 2000); also, obtaining an unbiased set of positive training examples is expensive. Further, determining the answer is akin to searching for the proverbial needle-in-a-haystack. Often only a few words in a thousand-line note are indicative of cancer. Finally, the distinctions between positive and negative samples are so subtle that two medical doctors will disagree as to whether the patient described had breast cancer or not. Even using a a domain-specific text-processing pipeline (D'Avolio et al., 2010) (which utilizes medical ontologies and "understands" the difference between patients and family members) we have found it difficult to build a high-recall classifier ($> 90\%$) that has precision exceeding 10%. Were a 50,000 record sample classified automatically, 5,000 records would be returned as "possible" positives, with only 500 being true positives. At fifteen minutes per record, combing through this set would require 1,250 hours of expert time.

With this in mind, our goal is to give the human expert some automated assistance, and reduce the per-EMR examination time. We seek to develop a software supporting the following workflow:

1. The software takes as input a corpus, where each document has been labeled using an error-prone learner. In practice, this "learner" might return the answer to a query such as "Does the document contain one of the terms *breast cancer*, *DCIS*, or *ductal carcinoma in situ*?"

2. If an expert is available, a few of the documents might be processed by a human, who has underlined those passages (contiguous blocks of words) in the text are most likely associated with a positive document label.

3. The software learns how to identify key passages associated with a positive label.

4. A set of candidate positive documents have their key passages underlined; the documents are given to a human expert for a final labeling.

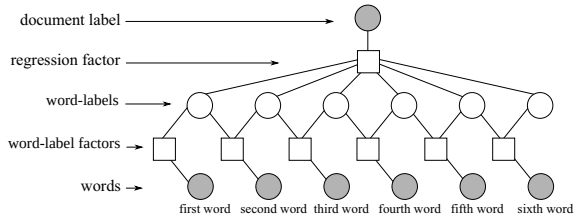

Figure 1: The WLR model. Boxes are factors, shaded circles parameters, and empty circles latent variables.

The hope is that by underlining a small subset of the document, the human expert doing the final labeling can be directed to the most important passages in the text, and it will be possible to greatly speed the final processing of the corpus. The main question we consider in this paper is: How can we develop a software to support this process?

Existing Work. The currently-existing model that is the closest match for solving this problem is the now-ubiquitous conditional random field (CRF) (Lafferty et al., 2001). One could imagine defining a special label that corresponds to a word that should be underlined, then learning a CRF that is able to associate such a label with the words in a test document.

However, there are two reasons that using a CRF is not appropriate here:

1. The learning problem is supervised *at the document level*. A document that is labeled $+1$ by the weak learner should have, relatively speaking, a lot of underlined words.

2. The problem is (mostly) unsupervised *at the word level*. In the most general case, we have no training data at the word level (that is, we have no examples of underlined words during training and must infer the underlining using only the document-level labels). Even in the semi-supervised case, we will have training examples for only one type of label (the one associated with an underlined word).

The Word-Label Regression Model. To handle this, propose a simple factor-based model called the *word-label regression model*, or WLR model, that differs from a CRF in (at least) two important ways:

1. It is supervised at the document level.

2. It is Bayesian, which makes it easy to incorporate our prior expectations (such as the fact that underlined words tend to sequentially follow one another) in a principled fashion, via the use of appropriate priors.

In this paper, we describe the WLR model, as well as how to perform full-Bayesian inference for the WLR model using an efficient Gibbs sampler. We develop a dynamic programming algorithm that finds the most likely underlining for a particular document, given a learned WLR model. We also show experimentally the utility of the method.

2 The WLR Model

2.1 Overview

A simple WLR model is depicted in Figure 1. There is an unseen label associated with each word in the document, as well as visible label describing the document as a whole. The model is discriminative, in that a document's words and the document-label are taken as input parameters. A factor connects each word-label with its associated word, as well as the word-label of the previous word. Another key aspect of the model is the presence of a "regression" factor, which connects the set of word-labels present in the document with the document's label via a logistic regression function. This factor measures whether the set of word-labels present in the document are in-keeping with the document's label.

2.2 Initial Formulation

Let $D = d_1, d_2, \ldots, d_N$ be a document corpus. Each document d is a sequence of N_d words (tokens) and the total number of unique words, i.e. the dictionary size, is A. A latent word-label, which represents the context within which that word appears, is associated with each word. There are K possible labels. Assuming that we are using a WLR model to learn to underline the important passages in a document, one of those K labels will be designated the "underlined label", and a word in the document is underlined if and only if it is assigned that label.

The association of word-labels to words in a particular document can be represented with a simple factor graph, as shown in Figure 2. The vector $\mathbf{w}_d = \langle w_1, w_2, ..., w_{N_d} \rangle$ contains the words in the document and the latent vector $\mathbf{l}_d = \langle l_1, l_2, ..., l_{N_d} \rangle$

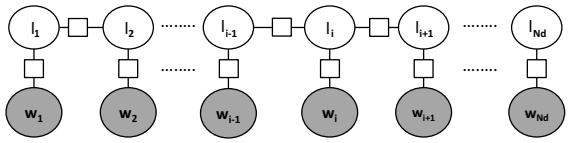

Figure 2: Factor graph for the initial formulation.

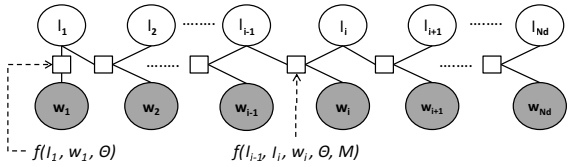

Figure 3: No-normalization factor graph.

is the word-labels. Note that since the model is not generative with respect to the document words, $\mathbf{w}_d$ is viewed as a constant parameter rather than a variable. Assuming there are K possible word-labels, we define $\mathbf{M}$ to be a $K \times K$ matrix where entry $m_{i,j}$ is the model's affinity for transitioning from word-label i to word-label j. A larger $m_{i,j}$ means it will be more likely that the transition is taken. Moreover, we define Θ to be a $K \times A$ matrix where entry $\theta_{i,w}$ is the model's affinity for associating word-label i with word w. Since we apply a Bayesian approach, each $m_{i,j}$ and $\theta_{i,w}$ is viewed as a latent variable, with an appropriately chosen prior.

Then, the joint probability of $\mathbf{l}_d$ given $\mathbf{w}_d$ is:

$$P(\mathbf{l}_d|\mathbf{w}_d, \Theta, \mathbf{M}) = \frac{\theta_{l_1,w_1} \times \prod_{i=2}^{N_d} m_{l_{i-1},l_i} \times \theta_{l_i,w_i}}{Z_d}$$

Z_d is a normalization term:

$$Z_d = \sum_{l'_1=1}^{K} \sum_{l'_2=1}^{K} \cdots \sum_{l'_{N_d}=1}^{K} \theta_{l'_1,w_1} \times \prod_{i=2}^{N_d} m_{l'_{i-1},l'_i} \times \theta_{l'_i,w_i}$$

Alternative Model. During Bayesian inference—we will employ a Gibbs sampler to perform inference—whenever we need to evaluate $P(m_{i,j}|.)$ or $P(\theta_{i,w}|.)$ for a candidate $m_{i,j}$ or $\theta_{i,w}$ value, we need to re-evaluate Z. This is problematic, because evaluating Z requires time proportional to the size of the corpus. This renders inference impractical.

A slight modification addresses this. This alternative graph is presented in Figure 3. Specifically:

$$f(l_1, w_1, \Theta) = \frac{\theta_{l_1,w_1}}{\sum_{l'_1=1}^{K} \theta_{l'_1,w_1}} \text{ and, if } i \geq 2,$$

$$f(l_{i-1}, l_i, w_i, \Theta, \mathbf{M}) = \frac{m_{l_{i-1},l_i} \times \theta_{l_i,w_i}}{\sum_{l'_i=1}^{K} m_{l_{i-1},l'_i} \times \theta_{l'_i,w_i}}$$

So $P(\mathbf{l}_d|\mathbf{w}_d, \Theta, \mathbf{M}) =$

$$\frac{f(l_1, w_1, \Theta) \prod_{i=2}^{N_d} f(l_{i-1}, l_i, w_i, \Theta, \mathbf{M})}{Z_d}$$

That is, while we still incorporate the label-to-label and word-to-label affinities, these affinities are grouped and normalized so as to guarantee that:

$$\sum_{l'_1=1}^{K} f(l'_1, w_1, \Theta) = 1 \text{ and}$$

$$\sum_{l'_i=1}^{K} f(l_{i-1}, l'_i, w_i, \Theta, \mathbf{M}) = 1 \text{ if } i \geq 2$$

As a result, the term Z_d becomes:

$$Z_d = \sum_{l'_1=1}^{K} f(l'_1, w_1, \Theta) \sum_{l'_2=1}^{K} f(l'_1, l'_2, w_2, \Theta, \mathbf{M}) \ldots$$
$$\sum_{l'_{N_d}=1}^{K} f(l'_{N_d-1}, l'_{N_d}, w_{N_d}, \Theta, \mathbf{M}) = 1$$

Thus, we can ignore the denominator entirely since it always evaluates to one. As we will discuss in detail in the next section of the paper, this alternative factorization will permit an efficient MCMC sampling algorithm that requires only $O(A)$ (or $O(K)$) time to evaluate the conditional likelihood of a candidate $m_{i,j}$ (or $\theta_{i,m}$) value, respectively.

Is a Markovian model problematic? The concern with forcing local normalization on sequential models of this type is that the labeling of words becomes Markovian; this is known to be problematic for this type of sequential model. For example, consider the bigram "ductal carcinoma" (referring to the presence of abnormal cells in a milk duct), which should almost always be underlined when searching for patients who have had breast cancer. Let the words "ductal" and "carcinoma" be w_1 and w_2, respectively. Imagine that the word "ductal" is (erroneously) not given the correct label c, which would have marked it as an underlined word. Under Markovian normalization, $\Pr(l_2|l_1 \neq c \wedge w_2)$

must always be at least $\frac{1}{K}$ for *some* value of l_2, given that the probabilities over all possible labels on "carcinoma" must sum to one. Hence, "carcinoma" is limited in how much it can punish a bad decision at "ductal". In contrast, an un-normalized model give a tiny weight to *all* factors of the form $f(l_2, l_1 \neq c$, "carcinoma", $\Theta, \mathbf{M})$, penalizing $l_1 \neq c$ at "carcinoma" arbitrarily.

This can be sidestepped in the WLR model. Since the model is Bayesian, over-fitting is not a concern, and we can increase K (the number of labels) to reduce $\frac{1}{K}$ as desired, allowing for an arbitrary punishment at "carcinoma" in our example, at the cost of slower inference. In a sense, Markovian normalization allows a user to explore a trade-off between computational efficiency and model quality.

2.3 Regression

The WLR model is supervised in the sense that the individual word-labels chosen must be in-keeping with the document label L_d. To force this, a factor f_R corresponding to a logistic regression over the word-labels is added to the model, as depicted in Figure 1. f_R is defined as follows. Let $p(\mathbf{l}_d, i) = \frac{\sum_j I(l_j=i)}{N_d}$ (I in this expression is the identity function, returning one if the argument evaluates to true, and zero otherwise), be the proportion of word-label i in the document d. Given a vector of weights $\Phi = \langle \phi_1, \phi_2, ..., \phi_K \rangle$, then $f_R(\mathbf{t}_d, L_d, \Phi) =$

$$\begin{cases} \frac{1}{1+\exp(-\sum_{i=1}^{K} \phi_k \times p(\mathbf{l}_d, i))} \text{ if } L_d = +1 \\ \frac{1}{1+\exp(\sum_{i=1}^{K} \phi_k \times p(\mathbf{l}_d, i))} \text{ if } L_d = -1 \end{cases}$$

Note that this is equivalent to a generative view where we assume that the label L_d was produced via a trial over a Bernoulli variable where the probability that $L_d = +1$ is $\frac{1}{1+\exp(-\sum_{i=1}^{K} \phi_i \times p(\mathbf{l}_d, i))}$.

2.4 Complete Formulation

Before combining the sequential model and the regression to form the final PDF, we define priors over $\mathbf{M}$, Θ, and Φ (Γ denotes the Gamma distribution):

$$\begin{aligned} m_{i,j} &\sim \Gamma(k_1, \beta_1) \\ \theta_{i,w} &\sim \Gamma(k_2, \beta_2) \\ \phi_i &\sim \text{Laplace}(0, b) \end{aligned}$$

Here k_1, θ_1, k_2, θ_2, b are given constants. Putting it all together, we have:

$$P(\mathbf{l}_d | \mathbf{w}_d, L_d, \Theta, \mathbf{M}, \Phi) = f_R(\mathbf{l}_d, L_d, \Phi) \times f(l_1, w_1, \Theta) \prod_{i=2}^{N_d} f(l_{i-1}, l_i, w_i, \Theta, \mathbf{M})$$

And so the posterior PDF for the entire dataset is:

$$P(\{\mathbf{l}_d\}_{d=1}^N, \Theta, \mathbf{M}, \Phi | \{(\mathbf{w}_d, L_d)\}_{d=1}^N) = \prod_{i,j} \Gamma(m_{i,j} | k_1, \beta_1) \times \prod_{i,j} \Gamma(l_{i,j} | k_2, \beta_2) \times \prod_i \text{Laplace}(\phi_i | 0, b) \times \prod_d P(\mathbf{l}_d | \mathbf{w}_d, L_d, \Theta, \mathbf{M}, \Phi)$$

3 Inference

3.1 Efficient Gibbs Sampling

We employ a Gibbs sampler for inference. This requires that we be able to derive the conditional posterior distributions for each $\theta_{i,w}$, for each ϕ_i, for each $m_{i,j}$, and for each l_i in each $\mathbf{l}_d$.

Given $P(\{\mathbf{l}_d\}_{d=1}^N, \Theta, \mathbf{M}, \Phi | .)$ from the previous section, deriving the required distributions is quite mechanical, and for brevity it is not covered in the paper. However, we discuss a few of the key implementation details.

Updating l_i in $\mathbf{l}_d$. Given a candidate value l_i' for l_i, let $\mathbf{l}_d'$ denote $\mathbf{l}_d$ after it has been updated so that $l_i = l_i'$. To re-sample l_i, we must sample from a categorical distribution where

$$P(l_i = l_i' | .) = f(l_{i-1}, l_i', w_i, \Theta, \mathbf{M}) f_R(\mathbf{l}_d', L_d, \Phi).$$

Evaluating the first factor is straightforward, but a naive evaluation of the second would make a pass through the document d and require $O(N_d)$ time. This can be reduced to constant time by recording, for each document, the current sum $\sum_{i=1}^{K} \phi_k \times p(\mathbf{l}_d, i)$. Then when evaluating $P(\mathbf{l}_d = \mathbf{l}_d' | .)$ for a candidate value of l_i', we subtract $\frac{\phi_{l_i}}{N_d}$ from the sum, and add $\frac{\phi_{l_i'}}{N_d}$, to re-evaluate the sum for the candidate.

Updating $m_{i,j}$. There are several ways to update $m_{i,j}$; the most straightforward is to use a rejection sampler (Robert and Casella, 2010). As above, given a candidate value $m_{i,j}'$ for $m_{i,j}$, let

$\mathbf{M}'$ denote $\mathbf{M}$ updated to incorporate $m'_{i,j}$. Using a rejection sampler for $m_{i,j}$ requires evaluating $P(m_{i,j} = m'_{i,j}|.)$ efficiently. It is easy to show that this value is proportional to $\Gamma(m'_{i,j}|k_1, \beta_1) \prod_d P(\mathbf{l}_d|\mathbf{w}_d, L_d, \Theta, \mathbf{M}', \Phi)$. Note that $P(\mathbf{l}_d|\mathbf{w}_d, L_d, \Theta, \mathbf{M}', \Phi)$ itself is a product of N_d factors, but the value of the kth factor in this expression is constant with respect to $m_{i,j}$ unless $l_{k-1} = i$ and $l_k = j$. We can take advantage of this to greatly speed evaluation of $P(m_{i,j} = m'_{i,j}|.)$. For each (i, j, w) triple, we count the number of times that word-label i transitioned to word-label j, where word-label j was associated with word w. Let $c_{i,j,w}$ denote this count. Then $P(m_{i,j} = m'_{i,j}|.)$ can be evaluated in only $O(A)$ time as proportional to:

$$\Gamma(m'_{i,j}|.) \prod_{w=1}^{A} f(i, j, w, \Theta, \mathbf{M}')^{c_{i,j,w}}.$$

The numerator of each of the K factors in the above expression clearly takes $O(1)$ time to evaluate, but if evaluated naively, the denominator of each factor will take $O(K)$ time, resulting in an overall $O(K \times A)$ cost. However, all affinities out of word-label i (that is, all $m_{i,*}$) have the same set of denominators. Thus, they can all be evaluated once and saved for a particular i in $O(K \times A)$ time, then maintained incrementally as each $m_{i,j}$ is updated. After the initial evaluation, each denominator need only be looked up in a table. This amortizes the cost down to $O(1)$ for evaluating the denominator in a factor, and the cost is then $O(A)$ for evaluating $P(m_{i,j} = m'_{i,j}|.)$. Also note that $O(A)$ is only an upper bound; since most of words only occur a few times in the corpus, the majority of the $c_{i,j,w}$ values will be zero and most evaluations will be far faster.

Updating $\theta_{i,w}$. An almost identical strategy can be used to update each $\theta_{i,w}$ using $O(K)$ time for each evaluation of $P(\theta_{i,w} = \theta'_{i,w}|.)$. The difference compared to the above case is that here, the kth factor in the formula for $P(\mathbf{l}_d|.)$ is constant with respect to $\theta_{i,w}$ unless $l_k = i$ and $w_k = w$. We can again take advantage of this by collecting the various $c_{i,j,w}$ values in a single pass over the corpus. At the same time, we also count the number of times in the corpus that word-label i and word w begin a document; this is denoted as $c_{i,w}$. Then $P(\theta_{i,w} = \theta'_{i,w}|.) \propto$

$$\Gamma(\theta'_{i,w}|.) f(i, w, \Theta')^{c_{i,w}} \prod_{j=1}^{K} f(j, i, w, \Theta', \mathbf{M})^{c_{i,j,w}}.$$

As above, a naive evaluation of the denominator in each factor would take $O(K)$ time. But again, the K denominators for each factor are identical for all word-label affinities involving word w. Again, these K denominators can be evaluated once in $O(K^2)$ time, and re-used for updating each of the A different $\theta_{i,w}$ variables. This amortizes the cost down to $O(K/A) \approx O(1)$ per evaluation of the denominator, leading to a $O(K)$ cost for each i, w, pair.

3.2 Incorporating Training Data

If word-level training data are available, a word-label 1 is arbitrarily declared to be the word-label that will be highlighted. To incorporate training data into MCMC sampling, all expert-underlined words are assigned label 1 and are never updated. Further, word-label 1 can never be assigned to those words that were not highlighted by the expert. Finally, when the regression coefficients are initialized (and when they are updated during MCMC sampling), we impose the constraint that ϕ_1 must be no smaller than ϕ_i for $i \in \{2...K\}$. This ensures that word-label 1 is the label that is most strongly associated with a $+1$ document label. This is easily handled during Gibbs sampling by truncating the conditional posterior distribution for ϕ_1 on the lower side at $\max_{i>1}\{\phi_i\}$, and truncating the conditional posterior for all other ϕ_i on the upper side at ϕ_1.

3.3 Choosing Appropriate Priors

The extent to which large blocks of contiguous words have the same word-label (and hence are highlighted together) is greatly influenced by the priors chosen for the individual entries in $\mathbf{M}$. A prior that tends to assign a relatively high affinity to a word-label-to-itself transition ensures that the trained model highlights longer passages. This should be seen as a "feature" of the model, providing a principled tuning mechanism, which is important in the absence of word-level training data.

4 Document Highlighting

Producing an actual highlighting is tricky. During MCMC inference, M, Θ and Φ quickly become stable after a short burnin. However, the actual word-labels are constantly in a state of flux. Labellings drift from one state to another, holes in word-label sequences open up and then close again, and no one labeling is particularly suitable.

The logical solution is to first run the MCMC sampler for a sufficiently long burnin period, average $\mathbf{M}$, Θ and Φ over the last few iterations, and use those parameters as input to a maximum likelihood (ML) computation, which is far more stable. This is relatively easy to do efficiently using a Viterbi-like dynamic programming (DP) algorithm (AJ, 1967), with the caveat that the ML computation does not take into account the regression factor.

Let α denote the number of words in document d to highlight, and let U denote the set of word-labels to highlight—this will typically include those word-labels with the largest ϕ_i regression coefficients. Let $F[i, j, a]$ to be the maximum likelihood obtainable by highlighting a of the first i words in the document, subject to the constraint that $l_i = j$. In the DP formulation, we define the base cases for $F[]$:

- $F[1, j, 1] = f(j, w_1, \Theta)$ if $j \in U$, else 0.
- $F[1, j, 0] = f(j, w_1, \Theta)$ if $j \notin U$, else 0.
- $F[1, j, a] = 0$ if $a \geq 2$.

The first two cases simply apply the factor from Section 2 appropriately. The third case reflects the fact that we cannot highlight more than one of the first one words in the document. Then the recurrence relation on $F[]$ is defined as $F[i, j, a] =$

- $\max_k\{F[i-1, k, a-1] \times f(k, j, w_i, \Theta, \mathbf{M})\}$ if $j \in U$
- $\max_k\{F[i-1, k, a] \times f(k, j, w_i, \Theta, \mathbf{M})\}$ if $j \notin U$

The first case applies when we are highlighting word i; here the best solution must be constructed by highlighting $a-1$ of the first $i-1$ words. The second applies when we do not highlight word i; here we construct the best solution by highlighting a of the first i words. The optimal highlighting itself is easily obtained by tracing the path used to obtain the ML value (located at $\max_{j \in U}\{F[N_d, j, \alpha]\}$) back through the $F[.]$ array.

	Shared words	Only MD$_A$	Only MD$_B$
Doc A	216	7	42
Doc B	386	10	236
Doc C	348	17	272
Doc D	221	29	304
Doc E	589	46	689

Table 1: Difference in highlighting between human experts.

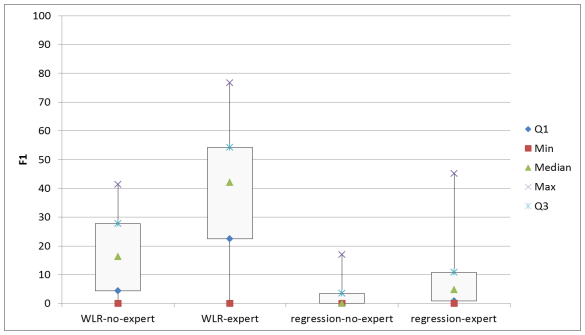

Figure 4: Results for experiment (a). The goal is to predict the underlining of MD$_B$. The y-axis is the per-document F1. The boxes show the 25th to 75th F1 percentiles over all 50 test documents.

5 Experimental Evaluation

5.1 Study Goals

Our ultimate goal is to assess the utility of our model for producing useful highlightings, particularly in our stated, biomedical application domain.

The best way to do this would have undoubtedly been to recruit several medical doctors, and then ask them to label a number of documents, both with and without the aide of the automated highlighting provided by the WLR model. The accuracy and speed with which the doctors labeled the documents with and without the WLR labeling would be examined to decide if it was helpful.

Unfortunately, running such a study with the primary goal of evaluating the WLR model is not feasible.[1] As such, we do the next best thing, and evaluate whether the WLR model is able to produce high-

[1]The problem is obtaining IRB (institutional review board)

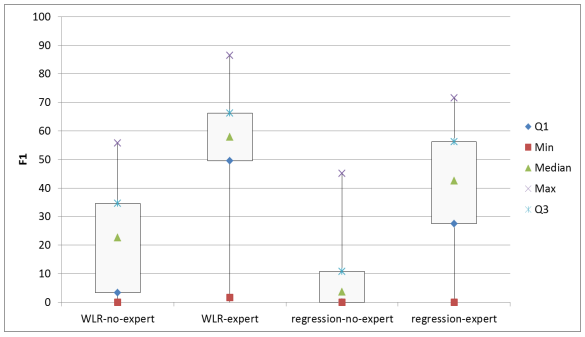

Figure 5: Results on 50 real test documents for experiment (b). MD_A was the target.

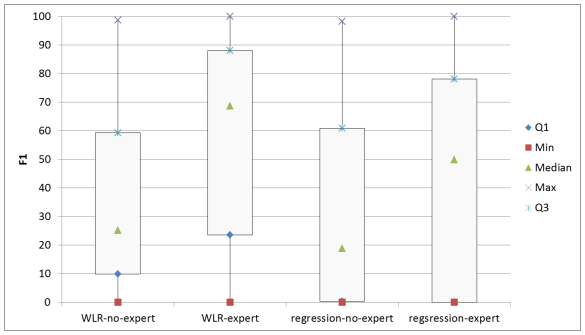

Figure 6: Results on 20-newsgroups (experiment (c)).

lightings that closely match highlightings produced by domain experts (in particular, we study the highlightings produced by the last two authors of this paper, who are medical doctors). If one assumes that an expert-supplied highlighting would likely be useful for another expert who is attempting to label the document, then the ability of the WLR model to closely match an expert-supplied highlighting is a reasonable proxy for a direct measurement.

5.2 Experimental Methodology

We evaluated our methods on two data sets.

Biomedical Data Set. We obtained a set of 1000 outpatient EMRs. The set consists of 500 sets of textual clinical notes obtained from EMRs associated with patients who have been billed for breast cancer (that is, where the EMR is labeled with ICD-9-CM codes 174.* (ICD 9 Codes,)) and 500 notes describing patients that have not been billed for breast cancer. We labeled the notes of the former patients as $+1$. Patients who were not billed for breast cancer or a related condition were assumed to have no history of breast cancer, and are labeled -1. The labellings are thus somewhat noisy, since (for example) former cancer patients may be treated for another condition. The notes varied greatly in length, from a few hundred to a few tens of thousand words.

We then selected two sets of 50 patients at random from those that had been billed for breast cancer, and the last two authors of the paper (hereafter called MD_A and MD_B) highlighted the documents, noting any text that they felt would be important to an expert who was reading the document, trying to determine if the patient had breast cancer. 5 of the 50 documents for each doctor were shared, so 95 documents in all were underlined. This took both MD_A and MD_B the better part of a week.

Synthetic Data Set. We also used the ubiquitous 20-newsgroups data set to create a synthetic experiment. Since we needed to simulate a situation where there was interesting text embedded within a document that needed to be underlined, we did the following. To create a $+1$ labeled synthetic document, we first sampled a document "length" l from a Poisson(10) distribution. We then selected l documents at random from the 20-newsgroups data set, subject to the constraint that the ith of the l documents had a 25% chance of being from soc.religion.christian, and a 75% chance of being from talk.*. All of the l sampled documents are then concatenated together to form the synthetic document. Furthermore, the portion of the document that came from soc.religion.christian was highlighted. In this way, we create a document where 25% of the text is from soc.religion.christian, and should be highlighted. To create a -1 document, none of the text comes from soc.religion.christian; that is, every synthetic document is composed entirely of newsgroup articles from talk.* that have been concatenated together.

Methods Tested. It is challenging to put together

approval. The IRB is tasked with performing a risk-reward analysis on all such research. Since EMRs contain copious amounts of personal health information and are generally impossible to de-identify, there is a risk of disclosure associated with presenting a large number of them to a third party—in this case, the doctors whose labeling acumen we would study. Without a biomedical study that directly benefits from undertaking this risk (for example, a clinical trial that needs to locate a large number of patients suffering from a particular ailment), obtaining IRB approval would be an uphill battle.

a strawman constructed out of off-the-shelf components to test the WLR model against, without having to solve some difficult problems in the process. After a lot of thought, we came up with two options.

We term the first option "regression-no-expert". In this case, we learn a regression model that does not take into account any example underlinings, only the overall document label. We pre-process each document so that it is represented as a bag-of-words TF-IDF vector, then train an L2 regularized logistic regression model. In order to perform an underlining during testing, the TF-IDF vector for each sentence in the test document is computed, and its dot product with the vector of regression coefficients is computed. All of the sentences are then sorted (from large to small) on the resulting weight, and those sentences with the highest weight are underlined (assuming that a target number of words are known, we can keep adding underlined sentences until the target number of words have been underlined).

We term the second option "regression-expert". In this case, we learn a model that makes use of example underlinings. We assume a set of training documents that have been underlined by an expert. We pre-preprocess each sentence in the set of documents that the expert has underlined so that it is represented as a bag-of-words TF-IDF vector, and then train an L2 regularized logistic regression model on all of those *sentences*; all underlined sentences are given a $+1$ label, and those that are not underlined are given a -1 label. During testing, we again sort the sentences based upon the likelihood that they should be underlined, and given a target number of words, we can keep adding underlined sentences until the target is reached. This option is similar to some methods applied in automatic document summarization (see the related work section).

In addition to these two methods, we also test two WLR-based methods. In the first, the WLR model is trained without access to any example underlining, and only with access to the document label. 20 different latent labels are used. We call this "WLR-no-expert". In the second, the WLR model is trained with access to example underlinings. 20 different latent labels are used. We call this "WLR-expert".

Experiments Run. On the real data, we performed two sub-experiments, (a) and (b). In (a), the goal was to predict the underlining performed by MD_A—so 945 documents (900 of which had only document-wide labels, 45 of which had MD_B's underlining available to WLR-expert and regression-expert) were used as training with the goal of predicting MD_A's 50 underlinings. In (b), things were reversed and the goal was to predict the underlining performed by MD_B.

On the synthetic data, we ran experiment (c). In this experiment, 1000 synthetic documents were used for training, and 1000 synthetic documents were used for testing. Each of the 1000 training documents had an underlining available to WLR-expert and regression-expert during model training.

In each experiment, it is necessary to choose how many words to highlight. To allow for a meaningful and easy-to-interpret comparison, the number of words highlighted is always set at the "correct" number (the number or words chosen by MD_A or MD_B); see Footnote 2 for more discussion. Choosing this number automatically is a problem for future work.

5.3 Quantitative Results

First, to give an idea of how much variance there might be a high-quality highlighting, in Table 1, we compare, for each of the five shared documents, the highlightings performed by MD_A and MD_B. In fact there is quite a lot of agreement between MD_A and MD_B. In each case, more than 90% of the words underlined by MD_A were also highlighted by MD_B, though MD_B tended highlight a super-set of MD_A's words, highlighting about twice as many.

Second, the degree of agreement between the automatic highlighting and the highlighting of MD_B (experiment (a)) is presented in the boxplot of Figure 4. The y-axis of the plot is the per-document F1; the box depicts the 25th to 75th percentiles of the F1 observed over all test documents.[2] Since the number

[2]The reader may wonder why we did not consider AUC. Typically, a classifier will assign each item a probability of being labeled $+1$. By decreasing the threshold below which a $+1$ label is assigned, more objects will be classified positively, resulting in higher recall, and lower precision. AUC quantifies this trade-off. However, this trade-off does not exist in our setting since the labellings of words within a document are *not independent*. One could decide to underline more words and have the effect of actually *lowering* recall and/or *increasing* precision, because a completely different set of words will be underlined at the higher number. Hence AUC is not applicable, at

of automatically underlined words is constrained to be equal to the number underlined by the expert, the precision, the recall, and the F1 are all the same.

Results, obtained using the model to predict MD_A's underlining (experiment (b)) is given in Figure 5. The plot for the 20-newsgroups data (experiment (c)) is given in Figure 6.

5.4 Discussion Of Quantitative Results

In every case, the WLR model was dominant. The WLR model with word-label training data had a recall/precision/F1 that was 20 to 40 points better than the logistic regression strawman. Without word-label training data, the gap was 10-20 points. On the breast cancer labeling task, the WLR-expert option reached a recall/precision/F1 of between 0.4 and 0.6, which we feel is quite high given that the amount of the document to be highlighted was typically just a few percent, and that training and testing were done on the highlightings of different doctors.

One interesting finding was that it was much more difficult to accurately match MD_B's highlighting than MD_A's. This is likely explained by looking at Table 1. MD_A was generally much more judicious with his highlighting, suggesting that it was more precise. One may reasonably deduce that this makes it much easier to mimic.

5.5 Qualitative Results

To give the reader some idea of what the documents and highlightings actually look like, we give two of them in Figures 7 and 8. The human in this case was MD_A, and MD_B's highlightings were used to train the WLR model. Figure 7 shows a case where the accuracy was high (75%), and Figure 8 a case where it was low (35%). In the former case, a huge amount of text was highlighted, and so only a portion of each highlighting is shown. In the latter case, the highlighting of each method is shown in its entirety. One thing that is striking from these examples is that all four highlightings seem more or less defensible. Take Figure 7. The expert chose to highlight a passage regarding family history, while the WLR model choose to replace this highlighted passage with a note about a right breast mass. Either choice seems reasonable.

least in the classic sense.

Human Expert Highlighting

...Ms — is a —year-old female who presents for further evaluation of a right breast mass. The patient first noted the lesion in her right breast on —. Approximately 3 months prior to that... **[long text]** ...and a biopsy was recommended. The patient was referred for further management. The patient denies any change in the lesion since she has noticed it and has no history of any pain or nipple discharge and no history of any lesions on the left by the patient report and normal mammogram 3 months prior. PAST MEDICAL HISTORY: Asthma. PAST SURGICAL HISTORY: Tubal ligation, cholecystectomy, and rotator cuff repair. FAMILY HISTORY: She had a grandmother who died of breast cancer in her early —'s as well as maternal great aunt who died of breast cancer. No family history of... **[long text]** On the patient history form, again, she noted a right breast mass, which has not change in size since she has noticed it. The rest of the review of systems is negative per the patient's history... **[long text]** BREASTS: She has no supraclavicular or infraclavicular or axillary adenopathy. On inspection, she has no erythema or dimpling. No nipple retraction. On palpation of her right breast she a 3 x 4-cm mass located at the 6 o'clock position 6 cm from the nipple. It is firm and consistent with a malignancy. It is moveable, but I believe it is fixed to the muscle. She has no palpable masses in the left breast. IMPRESSION: ...

WLR Highlighting

...Ms — is a —year-old female who presents for further evaluation of a right breast mass. The patient first noted the lesion in her right breast on —. Approximately 3 months prior to that... **[long text]** ...and a biopsy was recommended. The patient was referred for further management. The patient denies any change in the lesion since she has noticed it and has no history of any pain or nipple discharge and no history of any lesions on the left by the patient report and normal mammogram 3 months prior. PAST MEDICAL HISTORY: Asthma. PAST SURGICAL HISTORY: Tubal ligation, cholecystectomy, and rotator cuff repair. FAMILY HISTORY: She had a grandmother who died of breast cancer in her early —'s as well as maternal great aunt who died of breast cancer. No family history of... **[long text]** On the patient history form, again, she noted a right breast mass, which has not change in size since she has noticed it. The rest of the review of systems is negative per the patient's history... **[long text]** BREASTS: She has no supraclavicular or infraclavicular or axillary adenopathy. On inspection, she has no erythema or dimpling. No nipple retraction. On palpation of her right breast she a 3 x 4-cm mass located at the 6 o'clock position 6 cm from the nipple. It is firm and consistent with a malignancy. It is moveable, but I believe it is fixed to the muscle. She has no palpable masses in the left breast. IMPRESSION: ...

Figure 7: Comparison of expert and WLR highlightings.

6 Related Work

Extracting important sections in the text is long-studied. An overview can be found in (Das and Mar-

Human Expert Highlighting

...PRIMARY MEDICAL ILLNESS: This is a —year-old female patient originally from who comes to the clinic for followup and evaluation. The patient is seen in the absence of Dr. —. She has a past medical history of breast cancer, triple negative, status post mastectomy and adjuvant chemotherapy in —. The patient has remained free of disease and living a normal life. The patient offered no significant complaints in regards to her breast cancer past medical history. She... **[long text]** The lungs were clear to auscultation. There was no pain on percussion of the spine or the costophrenic angles. The mastectomy scar is clean with no evidence of recurrence, and the contralateral breast is negative with no palpable masses. No palpable axillary adenopathy...

WLR Highlighting

...PRIMARY MEDICAL ILLNESS: This is a —year-old female patient originally from who comes to the clinic for followup and evaluation. The patient is seen in the absence of Dr. —. She has a past medical history of breast cancer, triple negative, status post mastectomy and adjuvant chemotherapy in —. The patient has remained free of disease and living a normal life. The patient offered no significant complaints in regards to her breast cancer past medical history. She... **[long text]** The lungs were clear to auscultation. There was no pain on percussion of the spine or the costophrenic angles. The mastectomy scar is clean with no evidence of recurrence, and the contralateral breast is negative with no palpable masses. No palpable axillary adenopathy...

Figure 8: Comparison of expert and WLR highlightings.

tins, 2007). Most approaches measure section importance based on human-supplied, global features such as sentence length, sentence position, existence of a proper name or adjective, and so on. Another long-studied problem is identifying shifts in document topics (Hearst, 1997; Blei and Moreno, 2001; Beeferman et al., 1999), though in contrast to the WLR model, these approaches are typically unsupervised.

There have been only a few Bayesian variants of the CRF proposed in the literature, such a the aptly-named Bayesian CRF (Qi et al., 2005), which employs power Expectation Propagation (Minka and Lafferty, 2002) to approximate the posterior distribution. A significant difference between our approach and CRF-based work is that the WLR model is supervised. Supervised latent topic models have previously been proposed (Zhu et al., 2009; Wang et al., 2011; Blei and McAuliffe, 2007), but these models are non-sequential and utilize a bag-of-words view of a document. There has been work on semi-supervised CRFs (Mahdaviani and Choudhury, 2007; Mann and McCallum, 2008), but these do not take into account document-level information.

The CRF variant most closely related is the Conditional Topic Random Field (CTRF) (Zhu and Xing, 2010) model, where a latent topic assignment is defined with a Conditional Random Field. In contrast to purely generative models to capture correlations between topic assignments (such as structured models with Markov properties (Gruber et al., 2007; Verbeek and Triggs, 2007; Wang et al., 2009b) or latent permutations (Chen et al., 2009)) the CTRF model provides a conditional scheme for incorporating word-level features. That said, the CTRF model is still generative in the sense that the words in the document are generated by the model.

Numerous proposals have been developed to address the issue of normalization in document and topic models. Most apply approximation techniques (Murray and Ghahramani, 2004; Qi et al., 2005; Welling and Parise, 2006; Wang et al., 2009a; Zhu and Xing, 2010). A recent paper proposed the conditional topic coding (Zhu et al., 2011), which is a non-probabilistic formulation of the CTRF model that is not subject to strict normalization constraints.

Finally, we mention the large body of work on automatic text summarization (ATS). The References section of the paper includes a few relevant methods (Shen et al., 2007; Wang et al., 2008; Goldstein et al., 2000). However, there is an key difference between the vast majority of the ATS methods and the problem we consider: in ATS the goal is to summarize the entire document, rather than attempting to find those (potentially) few discriminative passages.

7 Conclusions

In this paper we tackled the problem of automatically identifying document passages that are most relevant to a user-defined label. We proposed the word-label regression model for this purpose, as well as associated learning algorithms and a Viterbi-style dynamic programming algorithm to perform the actual highlighting. Our preliminary experimental study using electric medical records shows the promise of the proposed method.

References

Viterbi AJ. 1967. Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269.

Doug Beeferman, Adam L. Berger, and John D. Lafferty. 1999. Statistical models for text segmentation. *Machine Learning*, 34(1-3):177–210.

David M. Blei and Jon D. McAuliffe. 2007. Supervised topic models. In *NIPS*.

David M. Blei and Pedro J. Moreno. 2001. Topic segmentation with an aspect hidden markov model. In *SIGIR '01*, pages 343–348, New York, NY, USA. ACM.

Harr Chen, S.R.K. Branavan, Regina Barzilay, and David R. Karger. 2009. Content Modeling Using Latent Permutations. *Journal of Artificial Intelligence Research*, 36:129–163.

Dipanjan Das and Andr F. T. Martins. 2007. A Survey on Automatic Text Summarization.

L.W. D'Avolio, T.M. Nguyen, W.R. Farwell, Y. Chen, F. Fitzmeyer, O.M. Harris, and L.D. Fiore. 2010. Evaluation of a generalizable approach to clinical information retrieval using the automated retrieval console (arc). *Journal of the American Medical Informatics Association*, 17(4):375–382.

Jade Goldstein, Vibhu Mittal, Jaime Carbonell, and Mark Kantrowitz. 2000. Multi-document summarization by sentence extraction. In *Proceedings of the 2000 NAACL-ANLPWorkshop on Automatic summarization-Volume 4*, pages 40–48. Association for Computational Linguistics.

Amit Gruber, Michal Rosen-Zvi, and Yair Weiss. 2007. Hidden Topic Markov Models. In *AISTATS*.

Marti A. Hearst. 1997. Texttiling: segmenting text into multi-paragraph subtopic passages. *Comput. Linguist.*, 23:33–64.

ICD 9 Codes. Available at: http://icd9cm.chrisendres.com/.

N. Japkowicz. 2000. The class imbalance problem: Significance and strategies. In *Proceedings of the 2000 International Conference on Artificial Intelligence (ICAI2000)*, volume 1, pages 111–117.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *ICML*.

Maryam Mahdaviani and Tanzeem Choudhury. 2007. Fast and scalable training of semi-supervised crfs with application to activity recognition. In *NIPS*.

Gideon S. Mann and Andrew McCallum. 2008. Generalized expectation criteria for semi-supervised learning of conditional random fields. In *ACL*, pages 870–878.

Tom Minka and John Lafferty. 2002. Expectation Propagation for the Generative Aspect Model. In *UAI*.

Iain Murray and Zoubin Ghahramani. 2004. Bayesian Learning in Undirected Graphical Models: Approximate MCMC Algorithms. In *UAI*.

Yuan (Alan) Qi, Martin Szummer, and Thomas P. Minka. 2005. Bayesian Conditional Random Fields. In *AISTATS*.

Christian P. Robert and George Casella. 2010. *Monte Carlo Statistical Methods*. Springer.

Dou Shen, Jian-Tao Sun, Hua Li, Qiang Yang, and Zheng Chen. 2007. Document summarization using conditional random fields. In *Proceedings of IJCAI*, volume 7, pages 2862–2867.

M.H. Stanfill, M. Williams, S.H. Fenton, R.A. Jenders, and W.R. Hersh. 2010. A systematic literature review of automated clinical coding and classification systems. *Journal of the American Medical Informatics Association*, 17(6):646–651.

Jakob Verbeek and Bill Triggs. 2007. Region Classification with Markov Field Aspect Models. In *CVPR*.

Dingding Wang, Tao Li, Shenghuo Zhu, and Chris Ding. 2008. Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 307–314. ACM.

Chong Wang, David Blei, and Li Fei-Fei. 2009a. Simultaneous image classification and annotation. In *CVPR*.

Chong Wang, Bo Thiesson, Christopher Meek, and David Blei. 2009b. Markov Topic Models. In *AISTATS*.

Hongning Wang, Yue Lu, and ChengXiang Zhai. 2011. Latent Aspect Rating Analysis without Aspect Keyword Supervision. In *KDD*.

Max Welling and Sridevi Parise. 2006. Bayesian Random Fields: The Bethe-Laplace Approximation. In *UAI*.

Jun Zhu and Eric P. Xing. 2010. Conditional Topic Random Fields. In *ICML*.

Jun Zhu, Amr Ahmed, and Eric P. Xing. 2009. Medlda: Maximum margin supervised topic models for regression and classification. In *ICML*.

Jun Zhu, Ni Lao, Ning Chen, and Eric P. Xing. 2011. Conditional Topical Coding: an Efficient Topic Model Conditioned on Rich Features. In *KDD*.