Wilbur and Kim *BMC Bioinformatics* 2011, **12**(Suppl 3):S5
http://www.biomedcentral.com/1471-2105/12/S3/S5

BMC Bioinformatics

RESEARCH Open Access

Improving a gold standard: treating human relevance judgments of MEDLINE document pairs

W John Wilbur*, Won Kim

From Machine Learning for Biomedical Literature Analysis and Text Retrieval in the International Conference for Machine Learning and Applications 2010
Washington, DC, USA. 12-14 December 2010

Abstract

Given prior human judgments of the condition of an object it is possible to use these judgments to make a maximal likelihood estimate of what future human judgments of the condition of that object will be. However, if one has a reasonably large collection of similar objects and the prior human judgments of a number of judges regarding the condition of each object in the collection, then it is possible to make predictions of future human judgments for the whole collection that are superior to the simple maximal likelihood estimate for each object in isolation. This is possible because the multiple judgments over the collection allow an analysis to determine the relative value of a judge as compared with the other judges in the group and this value can be used to augment or diminish a particular judge's influence in predicting future judgments. Here we study and compare five different methods for making such improved predictions and show that each is superior to simple maximal likelihood estimates.

Introduction

Human relevance judgments of a document in answer to a query are important as a means of evaluating the performance of a search engine and as a source of training data for machine learning methods to improve search engine performance [1,2]. Because human judgments are difficult, time consuming and expensive to obtain, it is important to extract as much advantage or information from human judgments as possible. If one is fortunate enough to have multiple judgments for the same query-document pair, the question arises as to how these multiple answers can best be used. It is the purpose of this paper to argue that ideally all available data should be used. It is not uncommon that relevance judgments are made on an ordinal scale consisting of {0,1,2,...,k} categories of relevance where k is as large as four[3,4]. We will not concern ourselves here with why a particular application might benefit from judgments on a scale with k greater than 2, but will simply assume that if this is important then it is important to predict the relevance of documents on this same scale. We propose that all available judgment data should be used to produce the most accurate assignment of probabilities to the different relevance categories for a document in answer to a query. The meaning of these probabilities must be the probabilities that these categories would be assigned by some new unseen judge (or user). Such probabilities will then provide optimal training data for improving system performance. But this leads to the important question, how shall we measure the quality of the probabilities produced from the human judgments? Our answer is to leave out one judge's judgments and measure the quality of the predicted probabilities by how well they predict the held out judge's judgments.

Before we proceed further with our discussion it is important to point out a distinction between what we are doing and work that has been done on a related problem. There are many examples of classification problems for which the true class of any object definitely exists. For example a patient either is or is not fit to undergo anaesthesia [5], a certain number of volcanoes are present in a given region of the surface of venus or

* Correspondence: wilbur@ncbi.nlm.nih.gov
National Center for Biotechnology Information, National Library of Medicine, National Institute of Health, Bethesda, MD 20894, USA
Full list of author information is available at the end of the article

that many are not present [6], a mushroom is either known to be edible or not known to be edible [7] , etc. For such data where it is known that there is a ground truth it makes sense to study models of the labeling process that incorporate an estimate of the reliability of labelers and an estimate of the ground truth for a task. Several such models have been developed and applied to a variety of data [5,7-14]. Such models are generally tested on how well they predict the ground truth which is known independently of the labeling process and labels being studied. This situation is fundamentally different than the problem we are interested in. Our data consist of multiple judgments of relevance of a query to a document and we consider each of these judgments to be legitimate and valuable. Judgments of relevance are generally understood to be highly subjective and their diversity represents different interests and insights of the judges [15-22]. Search services and online merchants are interested in what interests their customers and how to predict this interest and to the question of interest there is generally no one correct answer. Thus we make no assumption regarding correctness, but only seek how best to predict what some new searcher will find relevant.

Given the goal of producing probabilities of relevance categories from multiple human judgments, the next question is what are the options to do this? Clearly the simplest and most obvious approach is to compute the maximal likelihood estimates of class probabilities for each document. As an example suppose we have a document d retrieved by a query q and we require judgments to be made from the set of numbers {0,1,2,3,4} where 0 means clearly irrelevant and 4 means clearly relevant and the other options represent grades between these two extremes. Suppose we have ten prior human judgments {2,3,1,2,4,2,3,2,2,0}. Then the maximal likelihood predictions for future human judgments are

$p_0 = 1/10$, $p_1 = 1/10$, $p_2 = 5/10$, $p_3 = 2/10$, and $p_4 = 1/10$

and are proportional to the number of times each different judgment was seen in the past. Based on these predictions it seems much more likely that some future judge will assign a label of 2 than a label of 4 to the question of d 's relevance to q. The maximal likelihood approach treats all the judges as of equal value, i.e., we have assumed that all make judgments that are equally predictive of what a future judge would do. However, there is already in the data {2,3,1,2,4,2,3,2,2,0} a hint that some judges might be more valuable than others. There is a consensus in the data that 2 may be more likely as the value of a future judgment than other values. Thus a judge who chose the value 2 may be more useful than a judge who chose a different value. Of course we cannot really rate the usefulness of judges based on their judgment of a single object. But with judgments over a reasonable sized collection of objects it becomes quite feasible to rate judges for usefulness. To put this approach into practice, methods must be designed which account for the predictive value of judges.

As far as we have been able to ascertain, little work has been done in this area. Yu and colleagues [23,24] proposed a method to estimate the hidden intrinsic values of a set of objects that have been evaluated by a group of judges. They argue that the intrinsic value of an object judged by a group of judges is a suitably weighted average over the judgments of those judges where the weights represent the rating power of the judges. The intrinsic values determined in this way are interpreted as the consensus values of the group and each individual judge j's mean squared deviation σ_j^2 from the consensus values over the set of objects represents the reputation of the judge. They propose that the weight for judge j should be proportional to $1/\sigma_j$ and by normalizing one obtains the weights. Beginning with uniform weights one may calculate intrinsic values and then a more refined set of weights. This procedure may be iterated to convergence. They then suggest using the final intrinsic value for an object as the mean of a Gaussian distribution representing the distribution predictive of future judgments. This requires determining the variance of this predictive distribution, but this can be done using held out data. We evaluate this approach and compare it with our proposed methods and find that it performs well.

One of our approaches is related to the method proposed by Yu, et al. [24] in that we assume there are weights that represent the value of the individual judges. However, our approach differs from theirs in several respects. First, we are dealing with a small discrete set of possible judgments (five in number). In this setting it is convenient to combine prior judgments in a weighted manner to approximate a distribution predictive of future judgments. Instead of obtaining the weights by some iterative procedure we take a machine learning approach and learn optimal weights based on predicting held out data from the training set. We obtain our best results with this approach.

Our second approach is to treat the problem of predicting future judgments as a multiclass (five classes) classification problem. It is then natural to apply the maximum entropy classifier to this problem as it readily allows the computation of probabilities for multiple classes. In this approach the features are judgments of the training set judges and each training judge takes a turn at being held out to provide labels used to learn the weights for the features derived from the other training judges. When the training is completed the learned weights are then suitable for prediction. While

this method works well it seems to be somewhat less reliable than the other methods.

The paper is organized as follows. Section 2 describes the judgment data we study and how it was obtained. Section 3 presents the six different methods of predicting future judgments that we tested. The results are in section 4 and the discussion of these results comprises section 5. Section 6 presents conclusions and possible directions for future work.

Judgement data

The data that we study are human judgments of relevance between a query document q and a second document d where both documents were extracted from approximately a million MEDLINE documents dealing with aspects of molecular biology [25]. There are one hundred q 's that were selected at random and for each q a generic cosine retrieval algorithm [26,27] was used to find the top 50 documents d in relation to q. The resulting set of 5,000 query-document pairs will be denoted here by *DP*. The human judge was asked to judge for each pair in *DP* whether they would want to read d if they had to write the paper represented by q. They were asked to make their judgments on a scale of 0-4 where 0 means the document is clearly not relevant; 1, the document has a 0.25 probability of relevance to writing the query document; 2, a 0.50 probability of relevance; 3, a 0.75 probability of relevance; 4, the document is certainly relevant to the query-writing task [28]. Initially, a panel of seven judges trained in the area of molecular biology was hired to judge the set *DP*. Multiple judges were asked to perform the task because of the known variability in human judgments [18,29]. Later, because of questions raised by the work of the first panel [25,28], a panel of six untrained judges was hired to judge the 5000 query-document pairs of *DP*. One of the interesting findings coming from the work of the second panel was that while the untrained judges on average did not perform as well as the trained judges, some of the untrained judges were competitive and the pooled results of the untrained judges were almost as good as the pooled results for the trained judges and better than any single trained judge. Here we study the full set of thirteen judges who have judged *DP*.

Let us define notation for our study as:

J: set of 13 judges where J = {0,1,2,3,4,5,6,7,8,9, 10,11,12}.

dp: a query-document pair.

DP: set of 5,000 query-document pairs.

C : set of possible judgment values, i.e., C = {0,1,2,3,4}.

ζ_k^{dp} : judgment value of the query-document pair $dp \in DP$ made by the judge k.

$\Xi^{dp}(J)$:set of judgment values of the query-document pair $dp \in DP$ made by the judge set J, i.e., $\Xi^{dp}(J) = \left\{\zeta_k^{dp}\right\}_{k\in J}$.

$\Xi(J)$:set of judgment values of all 5,000 query-document pairs made by the judge set J, i.e., $\Xi(J) = \{\Xi^{dp}(J)\}_{dp\in DP}$.

Methods

Our approach is to consider a method $M(\Phi)$ that depends on a set of parameters Φ and that can be applied to a set of judgments $\Xi(J)$ to make predictions about the judgments of an as yet unseen judge k who has also judged the members of *DP*. We require these predictions to be in the form of probability distributions $P_{dp}(c|M(\Phi), \Xi(J))$ where $c \in C$.

We can then evaluate the performance of $M(\Phi)$ by the log probability that it assigns to k's judgments

$$S(k) = \sum_{dp\in DP} \log\left[P_{dp}\left(\zeta_k^{dp} \mid M(\Phi), \Xi(J)\right)\right] \quad (2)$$

Because our data is limited to 13 judges on the set *DP*, we follow a standard leave-one-out cross validation scheme for training and testing. We remove a judge k from the set J and denote the remaining set by $J_k = J - \{k\}$.

(Hereafter, any judges marked as subscripts on the set J are to be understood as removed from the set J. For example, the set $J_{k,i}$ means that the judges k and i have been removed from the set J.)

Now there is a particular problem in applying a method $M(\Phi)$ to data such as $\Xi(J_k)$. The problem is how to choose Φ to achieve good performance and yet avoid overtraining. If we choose Φ according to

$$\Phi^* = \arg\max_{\Phi}\left\{\sum_{dp\in DP} \log\left[P_{dp}\left(\zeta_k^{dp} \mid M(\Phi), \Xi(J_k)\right)\right]\right\} \quad (4)$$

there is a serious risk of overtraining. In order to overcome this issue, we apply a cross inductive learning process to get optimal parameters Φ^* as follows: Given a test judge k let us exclude one more judge $i \neq k$ from the set J. Cycling through all 12 judges $i \in J_k$, then the induction process is to find the optimal Φ according to

$$\Phi_k^* = \arg\max_{\Phi}\left\{\sum_{i\in J_k}\sum_{dp\in DP} \log\left[P_{dp}\left(\zeta_k^{dp} \mid M(\Phi), \Xi(J_{k,i})\right)\right]\right\}. \quad (5)$$

Then the optimal parameters Φ_k^* obtained from (5) may be applied to training on $\Xi(J_k)$ and the success $S(k)$ of the method $M(\Phi)$ is given by

$$S(k) = \sum_{dp\in DP} \log\left[P_{dp}\left(\zeta_i^{dp} \mid M\left(\Phi_i^*\right), \Xi(J_k)\right)\right] \quad (6)$$

To perform the cross validation we compute $S(k)$ in (6) over all judges $k \in J$ and average the results

$$Ave = \frac{1}{|J|}\sum_{k\in J} S(k) \quad (7)$$

and use Ave as the measure of overall performance for the method M in this study. We consider six different methods to predict a probability distribution over the possible judgment values of a query document pair $dp \in DP$. Each method is applied to induce the associated optimal parameters Φ^* according to (5) and then evaluated according to (6) and (7).

We proceed to a description of the individual methods. Here we present the basic ideas of the methods. The mathematical details can be found online at: http://www.ncbi.nlm.nih.gov/CBBresearch/Wilbur/IRET/research/methods.pdf

Method M_1: direct probability estimation

We take as our estimate of the probability of a given relevance category and a given query-document pair the fraction of the training judges that assigned that category to that pair, i.e., we take the maximal likelihood estimate of the probability of that pair based on the training judges. However, we must modify this estimate slightly to avoid predicting zero for any category because the test judge may have chosen a category that no training judge chose. We do this by mixing in a small fraction τ of the training judge probabilities of choosing the categories over the whole set of query-document pairs. We optimize the choice of τ by holding out from the training each training judge in turn and choosing the single τ that gives the best overall average of predictions over all such experiments.

Method M_2: direct probability estimation with weighting parameters

It is not optimal to put each judge on an equal footing for his class label judgments of query- document pairs as the previous method M_1 does since the predictive value of judgments will differ among judges. To deal with this we assign an arbitrary positive weight to each judge and instead of counting as in the previous method to obtain probabilities we add the weights of judges to obtain probabilities. Thus if three training judges chose category $c \in C$ for a given query-document pair dp we add the weights for the three judges and divide by the sum of the weights for all the training judges to obtain the probability assigned to c for dp. We also smooth in the same way as for the previous method and for the same reason. In fact we use the value of τ determined in M_1. Finally we optimize the choice of the weights by leaving each training judge out in turn and predicting his/her judgments based on the weights and optimizing their choice base on the whole set of such experiments at once. (The choice of τ here can be arbitrary since the weights will always adjust themselves to produce the same optimal results.)

Method M_3: correlation matrix with weighting parameters

If a given judge j assigns a category c to a query-document pair dp we can examine all the instances dp' when this judge assigned the category c. Based on all these instances we can come up with probabilities

$p(c'$ assigned by any judge $\neq j | c$ assigned by $j)$. This matrix of probabilities should have predictive value and may capture aspects not captured by the previous methods. Thus if for a particular dp if j has judged c we will use the distribution

$p(c'$ assigned by any judge $\neq j | c$ assigned by $j)$ as part of our prediction for dp. If a different judge j' has assigned c' we also want

$p(c''$ assigned by any judge $\neq j' | c'$ assigned by $j')$ to contribute to our prediction. Thus we take a weighted average over all these distributions to obtain our prediction and we smooth as before for the same reason. For each judge and each category there is assigned a weight and these same weights are used whenever the corresponding distribution is used in the predictions. Thus there are more weights here than in the previous method. The approach to optimization is the same as before.

Method M_{23}: combining the methods M_2 and M_3

We can combine the methods M_2 and M_3 defining the probabilities as a mixture of weighted terms coming from each method plus the smoothing. The optimization is then performed over all the weights at once.

Method M_4: intrinsic judgments from a weighted average

Yu et. al. devised the method whereby a community judgment can be obtained from a suitably weighted average over judgments for any given item. Given the judge set $J_{k,i}$ and a query- document pair $dp \in DP$, one defines the weighted average of judgments by

$$\mu_{k,i}^{dp} = \sum_{j\in J_{k,i}} r_j \zeta_j^{dp}. \quad (8)$$

Here the numbers r_j are a nonnegative normalized set of weights and are designed to reflect the importance of each judge's judgments. A judge's predictive capability is reflected in the average quadratic error in her judging history on all query-document pairs in DP:

$$e_j^2 = \frac{1}{|DP|}\sum_{dp\in DP}\left(\zeta_j^{dp} - \mu_k^{dp}\right)^2 \text{ for any } j \in J_{k,i}. \quad (9)$$

Then the weights may be defined by

$$r_j = \frac{1 / e_j^{2\beta}}{\sum_{j' \in J_{k,i}} 1 / e_{j'}^{2\beta}}. \quad (10)$$

While $\beta = 1$ in (10) gives the optimal weighting for statistical estimation [24,30], we use $\beta = 0.5$ for better numerical stability [24].

Starting with uniform weighting, the algorithm iterates eqs. (8), (9), and (10) to convergence to a solution. Once the solution has been obtained and we have the intrinsic class value $\bar{\mu}_{k,i}^{dp}$ for each dp, $\bar{\mu}_{k,i}^{dp}$ is taken as the mean of a Gaussian distribution which is used to predict the judgments of a test judge. There is one parameter and that is the σ for the predictive distributions and this is taken to be the same number for all dp. The value of σ is optimized as in the previous methods by optimizing the predictions for all $i \in J_k$ simultaneously. Once σ is determined, then the method is evaluated by its predictions for the judge k and the evaluation is completed by averaging over all such $k \in J$.

Method M_5: maximum entropy classifier

For details of the Maximum Entropy classifier we refer the reader to Berger, Pietra, and Pietra [31]. Here data points to be classified correspond to the query-document pairs $dp \in DP$. In order to apply a maximum entropy classifier we need to define a class label and features for each instance. The basic approach is to use one judge to supply the label for a pair dp and let other judges paired with their judgments on dp serve as the features. The same pair dp can serve repeatedly as an instance with each judge in turn supplying the label and the other judges and their judgments supplying the features. Since the labels are treated as true it is not crucial that they remain connected to the judges that produced them. But for the features it is crucial that they are pairs consisting of the judgment and the judge who produced that judgment. In this way when the features are weighted the weights reflect the predictive value of the judges that are involved. The scheme that we use is straightforward but a little complicated by several levels of held out judges. First we hold out judge k for testing leaving the set of judges J_k for training. Then we hold out judge i for determining the regularization parameter for the Maximum Entropy classifier leaving judges $J_{k,i}$ for training. Finally, we leave out judge j for labeling the instances coming from all the pairs in DP and use the judges remaining in $J_{k,i,j}$ to provide the features for each such instance. When we have created instances from all of DP for each $j \in J_{k,i}$ we train the classifier over all these instances together and then evaluate performance at predicting judge i 's labels for different values of the regularization parameter. We choose as optimal that value of the regularization parameter that gives the best average performance at prediction over all the $i \in J_k$ at once. When this regularization parameter is determined we use it and repeat the training on all instances coming from all $j \in J_k$ and test the prediction of k's labels. By repeating this for all $k \in J$ and averaging the results we measure the method's performance.

One parameter optimizations

The foregoing methods rigorously avoid overtraining in choosing the optimal parameter set Φ_k^* by equation (5). This clearly has advantages. On the other hand for methods where Φ_k involves only a single parameter, it is reasonable to consider the optimization of that single parameter for performance on the test data. This means optimization of (7) by choice of a single parameter value for all k. We have done this for the methods M_1, M_4, and M_5. The optimal parameters for these methods are given in Table 1.

Results

For a baseline performance of the predicted probability of human judgments for query document pairs in the set DP, we assume the uniform distribution where all pairs receive the probability 1/5 for all relevance categories. The measure (7) for this baseline method is

$$Ave_{random} = 5000 \cdot \log\left(\frac{1}{5}\right) = -8047.19 \quad (11)$$

We applied each of the methods $M_1 - M_5$ to induce the optimal parameters in (6) and the results are shown in Table 2. We also applied the parameters given in Table 1 for methods M_1, M_4, and M_5 and the results are shown in Table 3. Overall, one can observe that the performances of all methods are almost always better than the random level on each judge. The major exception is judge 0 where almost all methods make predictions that are less accurate than random predictions. Judge 12 is also challenging to predict and about half the predictions are worse than random. In a comparison of different methods we see that among the methods based on a rigorous determination of Φ_k^* method M_{23} performs best based on the average log probability measure. M_4 is a close second. Using the same measure for the single parameter optimizations in Table 3, the method M_5 performs best. If one considers the

Table 1 Optimal parameters associated with the methods M_1, M_4 and M_5 accurate to two digits.

τ^*:M_1	σ^*:M_4	λ^*:M_5
0.63	1.8	0.022

Table 2 Log of Probability Measures for all the methods using rigorous Φ_k^* values. The best performance in each row is marked with an asterisk

Judge	M_1	M_2	M_3	M_{23}	M_4	M_5
0	-8897	-8884	-8384	-8704	-8501	-8202*
1	-7103	-7085	-7006	-6843	-6940	-6690*
2	-6900	-6884	-6889	-6687	-6701	-6371*
3	-6806	-6729	-6699	-6493	-6734	-6192*
4	-7694	-7637	-7501*	-7560	-8121	-9350
5	-7131	-7045	-6912	-6872*	-7259	-7514
6	-7044	-6993	-6884	-6814*	-7026	-7237
7	-7110	-7149	-7035	-6876	-6557	-6446*
8	-7354	-7521	-7374	-7266	-6559*	-6838
9	-7122	-7040	-7100	-6911*	-7004	-7125
10	-8032	-8128	-7862	-7881	-7576	-7545*
11	-7281	-7123	-7071	-7008*	-7450	-7593
12	-8153	-8305	-8044	-8047	-7694*	-8056
Ave	-7433	-7425	-7289	-7228	-7240	-7320

predictions for individual judges the method M_5 achieves the best result more than any of the other methods in both Tables. However, the differences between methods do not achieve statistical significance by the sign test.

While the results in Tables 2 and 3 provide a useful performance gauge, they do not allow meaningful statistical testing of the differences seen. To allow statistical testing we consider the 5,000 predicted probabilities for the judgments of each test judge by the different methods under the same circumstances as those used to obtain the results in Tables 2 and 3.

To compare methods M_4 and method M_5 on how well they predict the judgments of judge 0 we examine judge 0's assigned label for each $dp \in DP$ and consider the difference in the probability it receives from method M_4 and the probability it receives from M_5. If this difference is positive it favors method M_4, but if negative method M_5. Of course the magnitude of the difference is also important as a large magnitude is more important than a small magnitude. This leads us to apply the Wilcoxon signed rank test [32] to determine the significance of differences. For the conditions of Table 2 we find method M_4 makes the better prediction for judge 0's judgments in 2,197 cases and M_5 for 2,803 cases and there are no ties. We then apply the signed rank test to see that the likelihood of the observed differences happening by chance if the two methods were equally good at making such predictions would be a probability of 5.12×10^{-63}. This indicates there is a very significant difference in the ability of the two methods in predicting this judge's judgments over the 5,000 query-document pairs.

Table 3 Log of Probability Measures for test set optimized single parameters. The best performance in each row is marked with an asterisk.

Judge	M_1	M_4	M_5
0	-8902	-8425	-7805*
1	-7087	-6925	-6833*
2	-6872	-6641*	-6674
3	-6760	-6675	-6462*
4	-7694*	-8068	-7703
5	-7121	-7259	-6942*
6	-7032	-7015	-6913*
7	-7094	-6482*	-6836
8	-7352	-6487*	-7199
9	-7113	-6992	-6874*
10	-8041	-7576	-7442*
11	-7275	-7450	-6909*
12	-8160	-7694*	-7784
Ave	-7423	-7207	-7106

Discussion

It is evident from the results of Table 2 that each of the six different methods of predicting relevance judgments for the unseen judge are far better than random, i.e., the -8047.19 given in (11). The method M_1 which takes the simplest approach of making the maximal likelihood estimate under the assumption that all the judges are of equal value in making predictions for what an unknown judge would judge gives the poorest result. Improved results come from making estimates of how to weight individual judges in combining their judgments. When these weights are learned by the iterative method of Yu, et al. [24] we see that the result is very good. When the weights are learned from the training judges using a held out judge, methods M_2, M_3, and M_{23}, we see our best result in M_{23}. The methods M_2 and M_3 each represent only a part of the solution and to get the best result both methods have to be combined in M_{23}. The method M_5, based on the maximum entropy method, comes in fourth in the competition based on the summary figure of -7320 for the average log probability of the judgments computed over all judges. From one point of view this summary figure is a little deceptive in that M_5 actually obtained the best score on six of the judges and this is a greater number of best scores than even the method M_{23} which achieved the overall best average. An examination of the scores for different judges shows that M_5 would have done much better had it not done very poorly predicting the judgments for judge 4. Analysis for judge 4 shows the algorithm attempts to use a regularization parameter λ_4^* that is much too small and hence overtrains and makes poor predictions. This problem led us to ask what performance would be if optimization were done to produce a single optimal λ^* for

all test data at once as given in Table 1. As seen in Table 3, one obtains improved overall performance. Of course there is a small risk of overtraining. The same single parameter optimization for method M_4 essentially does not work. The reason for this failure is not clear. For M_1 the single optimization just involves the smoothing parameter and has little effect, but does not degrade performance. The methods M_2, M_3, and M_{23} all involve multiple parameters and have a higher risk of overtraining and hence are not included in this analysis.

While the log probability of the judgments of an unseen judge averaged over all judges in turn seems like a reasonable way to rate overall performance, it does not provide a method to determine whether an observed difference between methods has statistical significance. In order to compute such significance values we have resorted to examining the difference in the probabilities assigned by two methods to a judge's judgments over the whole set *DP*. We can apply the Wilcoxon signed rank test to this data to ascertain statistical significance in a comparison of two methods for each judge. Such data is contained in Table 4 and Table 5 and is based on the same calculations reported in Table 2 and Table 3, respectively. The results are interesting in that they show that method M_5 is superior in the comparison of the rigorous approaches reported in Table 2 except for its performance in predicting judge 4's judgments. The data in Table 5 do not support any conclusion regarding the comparison of methods M_4 and M_5.

Table 4 In order to measure which method best predicts the individual class values made by a test judge between two methods, we apply the signed rank test. We also count query document pairs where the predicted probability of the class value is bigger for each method (and also ties). An asterisk marks the better result when the difference has a p-value less than 0.05 by the signed rank test. The optimal parameters are obtained through the rigorous induction method as in Table 2.

Judge	M_{23}*vs* M_4			M_{23}*vs* M_5			M_4*vs* M_5		
	M_{23}	M_4	=	M_{23}	M_5	=	M_4	M_5	=
0	2326	2674*	0	1763	3237*	0	2197	2803*	0
1	2576*	2423	1	1741	3259*	0	1808	3192*	0
2	2336*	2664	0	1580	3420*	0	1892	3108*	0
3	2637*	2363	0	1616	3384*	0	1592	3408*	0
4	3130*	1870	0	2817*	2183	0	2788	2212	0
5	2955*	2045	0	2463	2537*	0	2341	2659*	0
6	2692*	2308	0	2302	2698*	0	2301	2699*	0
7	1829	3171*	0	1504	3496*	0	1972	3028*	0
8	1398	3602*	0	1504	3496*	0	2313	2687*	0
9	2449*	2551	0	1964	3036*	0	2024	2976*	0
10	1970	3030*	0	1689	3311*	0	2337	2663*	0
11	3035*	1965	0	2199	2801*	0	2096	2904*	0
12	1965	3035*	0	1915	3085*	0	2452	2548*	0
Total	31298	33701	1	25057	39943	0	28113	36887	0

Table 5 In order to measure which method best predicts the individual class values made by a test judge between two methods, we apply the signed rank test. We also count query document pairs where the predicted probability of the class value is bigger for each method (and also ties). An asterisk marks th better result when the difference has a p-value less than 0.05 by the signed rank test. The optimal parameters are the single parameter optimizations of Table 1.

Judge	M_4*vs* M_5		
	M_4	M_5	=
0	1992	3008*	0
1	2546	2454*	0
2	2864*	2136	0
3	2598	2402*	0
4	2148	2851*	1
5	2247	2753*	0
6	2527	2473*	0
7	3392*	1608	0
8	3798*	1202	0
9	2676	2324*	0
10	2802*	2198	0
11	2084	2916*	0
12	2938*	2062	0
Total	34612	30387	1

A natural question that may have occurred to the reader is why not apply some of the techniques used in studying noisy classification labels [5,6,9,10,13,33] to our problem. Indeed this might be an interesting thing to try. However, there are two reasons we have not done it. First, all these models involve latent values of annotator reliability and true labels and are more complicated in concept and/or in application than the methods we use. Second and more important, all these models are constructed to predict the reliability of the judges and the true labels for items and are evaluated on how well they predict these true labels. Since we do not have true labels and true labels are philosophically inconsistent with our data and how it was obtained, we would not be able to evaluate such an application of these models to our data except in how well they could predict judgments of held out judges. But it is not readily apparent how such predictions could or should be derived in these approaches. Therefore we consider this a question beyond the scope of our current investigation.

Conclusions and future work

We have studied basically three methods of predicting human judgments from known human judgments. We find that method M_{23} gives the best overall predictions

for all judges. On the other hand the maximum entropy method M_5 gave the best results on twelve of the thirteen judges. However, it failed badly on one judge. As a result we conclude that M_5 is usually the best method, but is subject to occasional large errors. It is possible that such large errors could be prevented by setting a lower limit for the regularization parameter which is followed regardless of training. The method M_4 we regard as somewhere between M_{23} and M_5 in that it does not give quite as good results as M_{23} on the one hand and did not experience the large error seen with M_5 on the other hand.

Several directions for further investigation are suggested by our results. First, it is possible that the method M_4 of Yu, et al. [24] could be improved by taking their same basic approach, but determining the weights for individual judges as those that are optimal for predicting held out data. Determining the weights using their iterative algorithm works well, but there is no theoretical reason why that approach should be optimal for the purpose of making the desired predictions. Second, it may be useful to explore the connections of our methods with methods for fusing multiple classifiers [34-36] as both problems have solutions involving weighting the individual members to be combined or involving second stage machine learning to learn how to combine individuals. On the other hand the problems are distinct because combining human judgments employs no gold standard and attempts to predict what an unknown member typical of the group would do, whereas the classification problem generally works with a gold standard set of training data. Third, in a real application the predictive value of judges could be used to control the judgment process so that if less predictive judges judge material, then more such judgments are needed to obtain a certain level of assurance regarding the predictive value achieved. This suggests an active learning scenario in which not only the entities to be judged, but the judges, are controlled for maximum efficiency much as has been done for the classification problems [8,13,14].

Acknowledgment

This research was supported by the Intramural Research Program of the NIH, National Library of Medicine.

This article has been published as part of *BMC Bioinformatics* Volume 12 Supplement 3, 2011: Machine Learning for Biomedical Literature Analysis and Text Retrieval. The full contents of the supplement are available online at http://www.biomedcentral.com/1471-2105/12?issue=S3.

Authors' contributions

WJW designed and guided the project, and wrote most of this paper. WK participated in the design of the study, and carried out the experimentation of the machine learning methods and helped to draft the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Published: 9 June 2011

References

1. Baeza-Yates R, Ribeiro-Neto B: **Modern Information Retrieval.** Harlow, England: Addison-Wesley Longman Ltd.; 1999.
2. Manning CD, Raghavan P, Schütze H: **Introduction to Information Retrieval.** Cambridge, England: Cambridge University Press; 2009.
3. Grady C, Lease M: **Crowdsourcing Document Relevance Assessment with Mechanical Turk.** *NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk* Los Angeles, California. Association for Computational Linguistics; 2010, 172-179.
4. Wilbur WJ: **The knowledge in multiple human relevance judgments.** *ACM Transactions on Information Systems* 1998, **16**:101-126.
5. Dawid AP, Skene AM: **Maximum likelihood estimation of observer error-rates using the EM algorithm.** *Applied Statistics* 1979, **28**:20-28.
6. Smyth P, Fayyad U, Burl M, Perona P, Baldi P: **Inferring ground truth from subjective labelling of venus images.** California Institute of Technology; 1995.
7. Sheng VS, Provost F, Ipeirotis PG: **Get another label? improving data quality and data mining using multiple, noisy labelers.** *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* Las Vegas, Nevada, USA: ACM; 2008.
8. Donmez P, Carbonell JG, Schneider J: **Efficiently learning the accuracy of labeling sources for selective sampling.** *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* Paris, France: ACM; 2009.
9. Raykar VC, Yu S, Zhao LH, Valadez GH, Florin C, Bogoni L, Moy L: **Learning From Crowds.** *Journal of Machine Learning Research* 2010, **11**:1297-1322.
10. Rzhetsky A, Shatkay H, Wilbur WJ: **How to get the most out of your curation effort.** *PLoS Comput Biol* 2009, **5**:e1000391.
11. Smyth P: **Bounds on the mean classification error rate of multiple experts.** *Pattern Recogn Lett* 1996, **17**:1253-1257.
12. Snow R, O'Connor B, Jurafsky D, Ng AY: **Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks.** *Proceedings of the Conference on Empirical Methods in Natural Language Processing* Honolulu, Hawaii: Association for Computational Linguistics; 2008.
13. Welinder P, Perona P: **Online crowdsourcing: rating annotators and obtaining cost-effective labels.** *Workshop on Advancing Computer Vision with Humans in the Loop at CVPR'10* 2010.
14. Whitehill J, Ruvolo P, Wu T, Bergsma J, Movellan J: **Whose vote should count more: optimal integration of labels from labelers of unknown expertise.** *Advances in Neural Information Processing Systems* 2009, 2035-2043.
15. Burgin R: **Variations In Relevance Judgments and the Evaluation Of Retrieval Performance.** *Information Processing & Management* 1992, **28**:619-627.
16. Harter SP: **Psychological relevance and information science.** *Journal of the American Society of Information Science* 1992, **43**:602-615.
17. Saracevic T: **Individual differences in organizing, searching, and retrieving information.** In *Proceedings of the 54th Annual ASIS Meeting.* Washington; D. C. Learned Information, Inc.;Griffiths J-M 1991:82-86.
18. Saracevic T: **Relevance: A review of and a framework of the thinking on the notion in information science.** *Journal of the American Society for Information Science* 1975, **26**:321-343.
19. Schamber L, Eisenberg MB, Nilan MS: **A re-examination of relevance: Toward a dynamic, situational definition.** *Information Processing & Management* 1990, **26**:755-776.
20. Schamber L: **Relevance and Information Behavior.** In *Annual Review of Information Science and Technology. Volume 29.* Medford, New Jersey: Learned Information, Inc.;Williams ME 1994:3-48.
21. Harter SP: **Variations in relevance assessments and the measurement of retrieval effectiveness.** *Journal of the American Society for Information Science* 1996, **47**:37-49.
22. Froehlich TJ: **Relevance reconsidered-towards an agenda for the 21st century: introduction to special topic issue on relevance research.** *Journal of the American Society for Information Science* 1994, **45**:124-134.
23. Laureti P, Moret L, Zhang Y-C, Yu Y-K: **Information filtering via iterative refinement.** *Europhysics Letters* 2006.

24. Yu Y-K, Zhang Y-C, Laureti P, Moret L: **Decoding information from noisy, redundant, and intentionally distorted sources.** *Physica A* 2006, **371**:732-744.
25. Wilbur WJ: **Human subjectivity and performance limits in document retrieval.** *Information Processing & Management* 1996, **32**:515-527.
26. Lucarella D: **A document retrieval system based on nearest neighbor searching.** *Journal of Information Science* 1988, **14**:25-33.
27. Salton G: **Automatic Text Processing.** Reading, Massachusetts: Addison-Wesley Publishing Company; 1989.
28. Wilbur WJ: **A comparison of group and individual performance among subject experts and untrained workers at the document retrieval task.** *Journal of the American Society for Information Science* 1998, **49**:517-529.
29. Swanson DR: **Historical note: Information retrieval and the future of an illusion.** *Journal of the American Society for Information Science* 1988, **39**:92-98.
30. Chi-Hoon Lee RG, Shaojum Wang: **Using Query-Specific Variance Estimates to Combine Bayesain Classfiers.** *Proceedings of the 23rd international conference on Machine learning* 2006, **148**:529-536.
31. Berger AL, Pietra SAD, Pietra VJD: **A maximum entropy approach to natural language processing.** *Computational Linguistics* 1996, **22**:39-71.
32. Larson HJ: **Introduction to Probability Theory and Statistical Inference.** New York: John Wiley & Sons;, 3 1982.
33. Whitehill J, Ruvolo P, Wu T, Bergsma J, Movellan J: **Whose vote should count more: optimal integration of labels from labelers of unknown expertise.** *Proceedings of the 2009 Neural Information Processing Systems (NIPS) Conference* 2009.
34. Al-Ani A, Deriche M: **A new technique for combining multiple classifiers using the Dempster-Shafer theory of evidence.** *Journal of Artificial Intelligence Research* 2002, **17**:333-361.
35. Ho TK: **Multiple classifier combination: lessons and next steps.** In *Hybrid methods in pattern recognition. Volume 47.* Singapore: World Scientific Pub. Co., Ptc. Ltd.;Bunke H, Kandel A 2002, Machine Perception Artificial Intelligence.
36. Kittler J: **Combining classifiers: a theoretical framework.** *Pattern Analysis & Applications* 1998, **1**:18-27.

doi:10.1186/1471-2105-12-S3-S5
Cite this article as: Wilbur and Kim: Improving a gold standard: treating human relevance judgments of MEDLINE document pairs. *BMC Bioinformatics* 2011 **12**(Suppl 3):S5.