

# BioCreative 2. Gene Mention Task

**John Wilbur**

wilbur@ncbi.nlm.nih.gov

**Larry Smith**

lsmith@ncbi.nlm.nih.gov

**Lorrie Tanabe**

tanabe@ncbi.nlm.nih.gov

National Center for Biotechnology Information, Bethesda, Maryland

## Abstract

There were 21 participants in the BioCreative 2 Gene Mention Task, with a highest F-score of 87.21. We discuss the statistical significance of these results, and estimate how these systems would perform on alternate corpora. We also demonstrate that by combining the results from all submissions, an F-score of 90.66 is feasible, and furthermore, that the best result makes use of the lowest scoring submissions.

## 1 Introduction

Finding gene names in scientific text is both important and difficult. It is important because it is needed for tasks such as document retrieval, information extraction, summarization, and automated text mining, reasoning, and discovery. Technically, finding gene names in text is a kind of named entity recognition similar to the tasks of finding person names and company names in newspaper text [2]. However, finding gene names may be significantly harder for several reasons:

1. There are millions of gene names used.
2. New names are created continuously.
3. Authors usually do not use proposed standardized names, which means that the name used depends on preference.
4. Gene names naturally co-occur with other types, such as cell names, that have similar morphology, and even similar context.
5. Expert readers may disagree on which parts of text correspond to a gene name.
6. Unlike companies and individuals, genes are not defined unambiguously. A gene may refer to a specified sequence of DNA base pairs, but that sequence may vary in nonspecific ways, as a result of polymorphism, multiple alleles, translocation, and cross-species analogues.

All of these things make gene name finding a unique and persistent problem. An alternative approach to finding gene names in text, is to decide upon the actual genes that are referenced in a sentence. This is the goal of the gene normalization task [10]. While success in gene normalization to some degree eliminates the need to find explicit gene mentions, it will probably never be the case that gene normalization is more easily achieved. Therefore, the need for finding gene mentions will probably continue into the future.

## 1.1 Task Description

BioCreative is called a “challenge evaluation” (competition or contest), in which participants are given well defined text-mining or information extraction tasks in the biological domain. Participants are given a common training corpus, and a period of time to develop systems to carry out the task. At a specified time, the participants are then given a test corpus, and a very short period of time in which to apply their systems and return the results to the organizers for evaluation. All submissions are then evaluated according to numerical criteria, specified in advance. The results are then returned to the participants and subsequently made public in a workshop and coordinated publication. The first challenge was carried out in 2003 (with a workshop in 2004) and consisted of a gene mention task, a gene normalization task and a functional annotation task. The current challenge took place in 2006 and the workshop is taking place in 2007. There were three tasks in “BioCreative 2”, called the gene mention (GM), gene normalization (GN) and protein-protein interaction (PPI) tasks. This paper summarizes the performance of the participants in the gene mention task, and also suggests a prospective view of the task.

The BioCreative 2 Gene Mention task builds on the similar task from BioCreative 1. The training corpus for the current task consists mainly of the training and testing corpora from the previous task, and the testing corpus for the current task consists of an additional 5,000 sentences that were held “in reserve” from the previous task. In the time since the previous challenge, the corpus has been reviewed for consistency using a combined automated and manual process. In the previous task, participants were asked to identify gene mentions by giving a range of tokens in the pretokenized sentences of the corpus. In the current corpus, tokenization is not provided, instead participants are asked to identify gene mentions by giving the start and end characters in each sentence. As before, the training set consists of a set of sentences, and to each sentence a set of gene mentions. Each “official” gene mention in a sentence may optionally have alternate boundaries that are judged by human annotators to be essentially equivalent references.

Every substring identified by a run is considered either a true positive or a false positive. If the string matches a gene or alternate in the humanly annotated corpus, it is counted as a true positive with the exception that only one true positive is permitted per gene given in the corpus. If a gene that is given in the corpus does not match any strings nominated by a run, and none of the allowed alternates are matched by a run, then the gene is counted as a false negative. A run is scored by counting the true positives ( $TP$ ), false positives ( $FP$ ), and false negatives ( $FN$ ). Let  $T = TP + FN$  denote the total number of genes in the corpus, and let  $P = TP + FP$  denote the total number of nominated gene mentions by a run. The evaluation is based on the performance measures  $p$  (precision),  $r$  (recall), and their harmonic average  $F$

$$p = \frac{TP}{P} \quad r = \frac{TP}{T} \quad F = \left( \frac{p^{-1} + r^{-1}}{2} \right)^{-1} = \frac{TP}{(T + P)/2}$$

Different applications may favor a different weighting between precision and recall, but this is beyond the scope of our analysis. We assume this simple form of F-score in all of our analysis.

Despite being called a “challenge evaluation”, competition, or contest, there are several reasons to view the results differently. As is pointed out repeatedly in the TREC workshop [8], the “absolute value of effectiveness measure is not meaningful”, that is, the scores provided are not meaningful outside of the context of the challenge. The F-score is a specific metric, not without controversy, and the value achieved on the corpora of the challenge is no guarantee of performance on other corpora. We will demonstrate how it may be possible to estimate the performance on alternative corpora, but there is no way to determine the accuracy of these estimates. All performance measures have a natural statistical variation, even within the narrow confines of the corpora defined for this task. We will estimate the statistical significance of pairwise comparisons. Finally, runs that score below the median may still give valuable insights into the task, and we will provide some evidence that

this is the case. In short, this competition is not a horse race, but a scientific forum in which the state-of-the-art is advanced through comparison and sharing of ideas.

## 1.2 Corpus Preparation

In 2003, as part of a project to improve on the AbGene tagger [6], a corpus of 20,000 sentences was selected and annotated for training and testing purposes. As described in [6], a Bayesian classifier was developed to recognize documents that are likely to contain gene names, and it was found that the precision and recall of the tagger was much better for high scoring documents. With this motivation, 10,000 sentences from high scoring documents and 10,000 sentences from low scoring documents were selected and combined to form the 20,000 sentence corpus. The corpus was further subdivided into *train*, *test*, *round1*, and *round2* sets of 5,000 sentences, each of which contained equal numbers of high scoring and low scoring sentences. The *train* and *test* sets were provided as the training set in BioCreative 1, and the *round1* set was used as the final evaluation. With some modifications, the *train*, *test*, and *round1* sets were provided as the training set in BioCreative 2, and the *round2* set was used as the final evaluation.

For BioCreative 2, the entire corpus of 20,000 sentences and approximately 44,500 GENE and ALTGENE annotations, was converted to the MedTag database format [5]. To do this, the original sentence in MEDLINE was located (though a few had been removed from MEDLINE and were replaced with sentences existing at the time). The bibliographic information for each sentence was also determined. The token specifications of all previous annotations were changed to character specifications. And because annotations were no longer limited to preset token boundaries, it was necessary to manually review every annotation to confirm or relocate the annotation boundaries. For example, it became possible to annotate a gene that is hyphenated to another word, the combination of which is not a gene mention.

To improve the consistency of annotation, approximately 1,500 strings (containing 2 or more characters) were found that were annotated as GENE or ALTGENE in one sentence and unannotated in another sentence. These strings occurred in approximately 13,500 mentions, of which 4,300 were GENE annotations, 2,200 were ALTGENE annotations, and 7,000 were unannotated. All of these cases were manually reviewed for accuracy and several corrections were made.

## 2 Summary of Submitted Runs

The BioCreative 1 gene mention task had 15 participants and each was allowed to submit up to 4 runs, categorized as either closed (no additional lexical resources), or open (no restriction). The BioCreative 2 gene mention task had 21 participants and each team was allowed to submit up to 3 runs. There were no restrictions placed on the submissions. The highest achieved F-score for the BioCreative 1 gene mention task was 82.2 while in the current challenge the highest achieved F-score was 87.2. For the purposes of presenting results, and all further analysis in this paper, only one submission from each of the 21 teams with the highest F-score was considered.

The precision, recall, and F-score for each team, in rank order based on F-score, is shown in Table 1. To compute significance, bootstrap resampling was used on the test corpus. For 10,000 trials, a random sample of 5,000 sentences was selected *with replacement* from the test corpus, and the precision, recall, and F-score was computed using these sentences for each of the 21 submissions. For each pair of submissions, say *A* and *B*, the proportion of times in these 10,000 trials that the F-score of *A* exceeded the F-score of *B* was noted, and we label that pair statistically significant if this proportion is greater than 95%. Significant differences are shown in Table 1. One can see that the top 3 F-scores did not have statistically significant differences. Also, the top 6 F-scores are all statistically significant compared to the remaining scores, and so on. Every pair of F-scores ( $\times 100$ ) that differed by approximately 1.23 or more was significant, and every pair of F-scores that

rank	BioCreative					MEDLINE			Trans. Factors		
	<i>p</i>	<i>r</i>	<i>F</i>	<i>signif</i>	<i>% alt</i>	<i>p</i>	<i>r</i>	<i>F</i>	<i>p</i>	<i>r</i>	<i>F</i>
1	88.48	85.97	87.21	4-21	32.48	80.06	83.62	81.80	90.15	86.57	88.32
2	89.30	84.49	86.83	6-21	14.02	83.21	81.14	82.16	90.52	85.31	87.84
3	84.93	88.28	86.57	6-21	14.08	76.53	85.22	80.64	86.77	89.02	87.88
4	87.27	85.41	86.33	7-21	31.77	79.93	82.78	81.33	88.80	85.96	87.36
5	85.77	86.80	86.28	7-21	16.67	73.53	83.84	78.35	88.45	87.47	87.96
6	82.71	89.32	85.89	7-21	16.02	69.78	87.85	77.78	85.72	89.66	87.65
7	86.97	82.55	84.70	8-21	14.83	78.62	78.88	78.75	88.75	83.43	86.01
8	84.35	81.39	82.85	10-21	14.57	74.42	77.30	75.83	86.60	82.40	84.45
9	86.28	79.66	82.84	10-21	14.55	79.33	75.97	77.61	87.77	80.48	83.97
10	85.22	78.44	81.69	12-21	33.02	75.82	74.73	75.27	87.16	79.29	83.04
11	85.54	76.83	80.95	12-21	19.76	75.64	74.07	74.84	87.73	77.48	82.29
12	72.95	88.49	79.97	14-21	16.82	50.75	88.31	64.46	79.26	88.46	83.61
13	92.67	68.91	79.05	15-21	19.73	89.88	64.39	75.02	93.25	70.10	80.04
14	88.83	69.70	78.11	16-21	37.05	82.44	64.39	72.30	90.05	71.20	79.52
15	80.46	73.61	76.88	17-21	20.43	71.92	70.85	71.38	82.10	74.08	77.89
16	82.28	71.08	76.27	18-21	16.80	73.40	67.33	70.23	84.26	71.95	77.62
17	84.32	68.57	75.63	18-21	34.02	80.40	64.61	71.64	85.01	69.39	76.41
18	71.68	62.33	66.68	19-21	28.23	54.16	61.33	57.52	75.99	62.33	68.49
19	65.83	61.55	63.62	20, 21	27.23	49.98	55.95	52.79	69.39	62.78	65.92
20	60.56	64.11	62.29	21	31.71	39.30	62.45	48.24	66.98	64.54	65.74
21	50.09	46.12	48.02	-	28.46	36.71	43.86	39.97	53.44	46.42	49.68

Table 1: In the left panel, under “BioCreative”, the precision, recall, and F-score for the best submitted run from each of 21 participants, sorted by F-score. Each team has an F-score that has a statistically significant comparison ( $p < 0.05$ ) with the teams indicated in the *signif* column. The column labeled *% alt* is the percentage of true positives in the submission that matched an ALTGENE annotation. The panels under “MEDLINE” and “Trans. Factor” are the precision, recall and F-score after reweighting sentences for MEDLINE and a “human blood transcription factors” query, respectively (see text).

differed by approximately 0.35 or less was insignificant.

Table 1 also shows the alternates in each run as a percentage of the corresponding true positives, which varies from about 15 to 30%. It is interesting to observe that the number of alternates in a run is not predictive of the score, as the top 3 runs represented both extremes. Nevertheless, there was an overall negative correlation of -0.40, and it could be hypothesized that methods which were less effective at learning the boundaries of the primary gene mentions were still able to get close enough to match alternatives, resulting in a higher representation of alternates among their true positives.

If the percentage of alternates among true positives is denoted by  $\alpha$ , then the F-score that obtains after omitting the alternates is

$$F^* = F \frac{1 - \alpha}{1 - F\alpha/2} = F(1 - \alpha)(1 + F\alpha/2 + (F\alpha/2)^2 + \dots).$$

With  $\alpha$  in the observed range, removing the alternates reduces the F-score by a multiple ranging from 1/2 to 3/4 of  $\alpha$ , with the greater reduction occurring for lower scoring runs.

Along with each submitted run, each team was required to submit answers to a list of questions which are paraphrased in Figure 1. We read the submitted answers and developed a list of mentioned features, which are shown in Table 2. That table also shows the total number of features

*How would you summarize your overall approach?*  
*List training data that you used in addition to the training data provided.*  
*List machine learning techniques used.*  
*List NLP techniques used.*  
*List Bio-NLP techniques used.*  
*List external lexical resources used, such as dictionaries and ontologies.*

Figure 1: Paraphrase of the questionnaire required from each participant in the Gene Mention task.

<b>Techniques</b>	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	Tot
SVM	*		*											*								3
CRF		*	*	*		*	*	*	*				*		*	*			*			11
Merge		*	*					*			*	*										5
Online learning					*																	1
n-gram							*															1
MaxEnt										*												1
HMM												*										1
Manual rules																	*			*		2
Case based																		*				1
C4.5																					*	1
<b>NLP</b>	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	
POS tagger	*	*	*		*	*	*		*	*	*		*	*	*							12
NP chunker	*						*								*							3
Paren matching	*					*																2
Stemming		*	*				*		*								*					5
Bidirectional		*	*																			2
Abbreviations				*		*	*															3
LSA				*																		1
Character based						*						*										2
Tokenization							*									*						2
Parser									*								*				*	3
<b>Systems</b>	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	
Mallet		*	*	*		*										*						5
GENIA tagger				*		*							*		*							4
LingPipe						*					*											2
Abner											*					*						2
TnT											*											1
MedPost													*	*								2
<b>Data</b>	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	
Medline	*			*	*	*																4
Mesh	*				*																	2
LocusLink	*																					1
SwissProt	*																					1
HUGO		*		*	*								*									4
Other lists				*	*			*											*			4
ALTGENE				*						*												2
AbGene List				*																		1
Biothesaurus						*																1
UMLS						*																1
RefSeq							*															1
MedPost							*															1
EntrezGene													*				*			*		3
Genia															*		*					1
Uniprot																	*					1
WordNet																				*		1
Totals	8	7	7	10	6	11	8	4	4	3	5	3	6	3	4	5	5	1	2	3	2	

Table 2: Features mentioned in system questionnaires, as interpreted by the authors, for the best run from each team. Column headings are the F-score rank. The last column is the number of teams that mention a feature, and the last row is the number of features mentioned by each team.

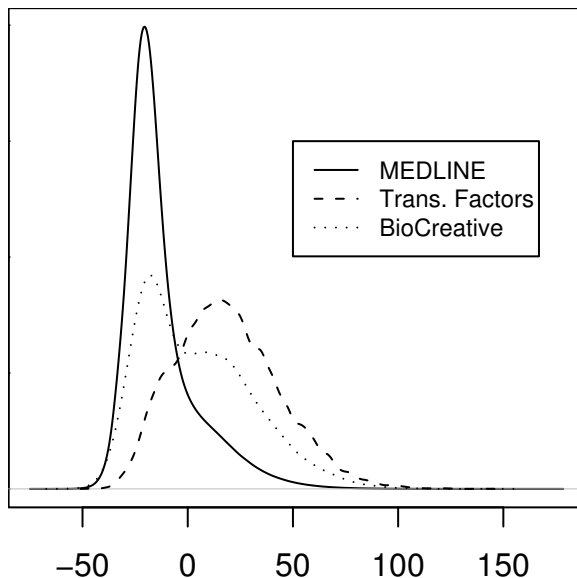


Figure 2: The distributions of gene indicator scores for the sentences of MEDLINE, the sentences in the query “Human Blood Transcription Factors” and the 5,000 sentences comprising the BioCreative 2 evaluation set. The distribution of scores for the BioCreative 2 sentences reflects the origin of the corpus as an equal mixture of high scoring and low scoring sentences.

mentioned by each team, and the total number of teams mentioning each feature (the rows are sorted in each group according to the highest ranking team mentioning the feature). Because this questionnaire was not a controlled study, it is not possible to draw definite conclusions as to the relative effectiveness of the techniques. However, it is interesting to note that the number of features mentioned by a team has a significant correlation of  $-0.757$  with the rank of the submission.

### 3 Estimated Performance on Alternate Corpora

As noted in Section 1.2, the corpus provided for training and testing was selected from MEDLINE so as to equally represent sentences likely to contain gene names and sentences not likely to contain gene names. Because of this selection bias, the performance measures obtained in this evaluation do not directly predict the system performance in any other situation. Nevertheless, by weighting the sentences appropriately, it is possible to estimate a system’s performance on corpora with a different distribution of sentences.

Based on the selection bias in the original 20,000 sentences, we used an updated system for scoring sentences to indicate whether they are likely to contain a gene name, and we used this score to obtain weights for alternative corpora. Suppose it is desired to estimate the performance of a system on a given alternate corpus. If  $f_t$  is an estimate of the probability density function for the scores in the test set, and  $f_a$  is the probability density function for the scores in the alternate corpus, then sentence number  $i$  from the test set with score  $s_i$  should have weight  $w(i) = f_a(s_i)/f_t(s_i)$  in the alternate corpus. Then, if sentence number  $i$  contains  $TP(i)$  true positive gene annotations,  $FP(i)$  false positives and  $FN(i)$  false negatives, then the weighted performance on the alternate collection is computed using

$$TP' = \sum_i w(i)TP(i) \quad FP' = \sum_i w(i)FP(i) \quad FN' = \sum_i w(i)FN(i)$$

To estimate the densities  $f_t$  and  $f_a$  we computed the scores for all of the sentences of the col-

lections and then applied the *density* function using the R 2.2.1 statistical program (with spline interpolation).

We carried out this weighting for random sentences selected from MEDLINE and for sentences selected as the result of a query for human blood transcription factors. The distribution of the gene score for the sentences in the BioCreative test corpus is shown in Figure 2, along with the distribution of scores of random sentences from MEDLINE and the sentences from the PubMed query

```
"Transcription Factors" [MeSH]  
AND "Blood Cells" [MeSH]  
AND Humans [MeSH]
```

which returns 9,003 abstracts. One can see that the BioCreative test corpus has a greater representation of high scoring sentences than the MEDLINE corpus, as does the human blood transcription factor corpus. The computed precision, recall, and F-scores for each team for the alternate corpora are shown in Table 1. Whether a system performs better or worse on a collection roughly depends on the difference in its performance on high scoring and low scoring sentences. Note that the estimated F-scores for random MEDLINE is generally lower than the scores on the BioCreative 2 evaluation, while the estimated F-scores for the human blood transcription factor set are generally higher.

## 4 Combined Performance

We wanted to know if it is possible to improve on the best scores obtained in this workshop. To do this, we used machine learning to predict gene mentions using all of the the submitted runs as feature data.

In order to simulate what might result if all of the methods were combined into a single system, we extracted features from the submitted runs. By holding out 25 sentences at a time, and training on the remaining 4,975 sentences, we could apply the result to the held out set and then merge all of the results to obtain a single "fusion" run for all 5,000 sentences.

For each candidate, which is defined by a particular start and end offset within a sentence, the features described in Figure 3 were generated. We used two different machine learning techniques with this feature data, boosted decision trees, and conditional random fields.

For boosted decision trees, the training set consisted of all candidates whose starting and ending offsets coincided with a nominated string from at least one team (but the starting and ending offsets need not both be nominated by the same team). Each character of a candidate was also required to overlap a nominated string from at least one team. This meant that every candidate had at least one "nom" feature from Figure 3. Each candidate was further marked as a "positive" depending on whether it appeared exactly as a gene or alternate gene mention, and all other candidates were marked as a "negative". A boosted decision tree algorithm [4, 1] was applied to this data set (holding out 25 sentences at a time, as mentioned above) to learn which candidate is a positive. Each tree was allowed to have a depth of 5 and boosting was repeated 1,000 times. The induced set of decision trees was applied to the held-out set of 25 sentences to obtain gene mentions for them. Where gene mentions overlap, only the gene mention with the highest score is retained, so that the final result does not contain any overlapping gene mentions. We repeated this training using only nomination features, only word features, and combined nomination and word features. The results are shown in Table 3, and the nomination features combined with words performed best with an F-score of 90.50. As this is 3.29 greater than the highest F-score obtained by an individual team, the difference is statistically significant.

We also used conditional random fields (with gaussian prior) to learn gene mention [3]. Each sentence was tokenized and each token was marked as being positive or negative depending on

$not(T)$	Team $T$ did not nominate any gene mention that overlaps with this candidate.
$nom(T)$	Team $T$ nominated a gene mention that overlaps with this candidate.
$nom_s(T, S)$	Team $T$ nominated a gene mention that overlaps with this candidate, and that starts before ( $S = -1$ ), starts after ( $S = 1$ ) or coincides with the start of this candidate ( $S = 0$ ).
$nome(T, E)$	Team $T$ nominated a gene mention that overlaps with this candidate, and that ends before ( $E = -1$ ), ends after ( $E = 1$ ) or coincides with the end of this candidate ( $E = 0$ ).
$nom(T, S, E)$	Team $T$ nominated a gene mention with $S$ and $E$ as above.
$nom_s(S)$	Some team nominated a gene mention with $S$ as above.
$nome(E)$	Some team nominated a gene mention with $E$ as above.
$word(W)$	Word $W$ occurs in the candidate.
$firstword(W)$	Word $W$ is the first word of this candidate.
$lastword(W)$	Word $W$ is the last word of this candidate.
$context(P, W)$	Word $W$ at position $P$ relative to this candidate. The possible values for $P$ are $-2, -1, 1, 2$ .

Figure 3: The features generated for each candidate gene mention, based on the submitted runs.

whether it was part of an annotated gene (alternates were not used in this approach). The features described in Figure 3 were generated for each token, in which, for the purposes of generating features, each token is treated as a candidate. By holding out 25 sentences at a time, the CRF was trained on the remaining 4,975 sentences (the gaussian prior defined in [3] was taken to be  $1/2\sigma^2 = 300$ ). The trained CRF was then applied to tag the 25 sentences, and any sequence of consecutive positive labels were combined into a single gene mention. The results from each set of 25 sentences were combined to form a single run. The result, shown in Table 3 was an F-score of 90.66. This is slightly higher than the result obtained using boosted decision trees (with nomination and word features), but the difference is not statistically significant.

A question of interest to us is whether the alternate annotations could be used in machine learning to improve performance in the gene mention task. There were some teams that did train

Exp	Method	$p$	$r$	$F$	$signif$	% alt
A	CRF noalt, nom and word	92.55	88.85	90.66	1-21, C-F	13.62
B	BDT nom and word	92.21	88.85	90.50	1-21, C-F	25.67
C	BDT nom and word, top 10 teams	91.18	87.68	89.40	1-21, E, F	23.37
D	BDT nom only	90.92	87.73	89.29	1-21, E, F	25.42
E	BDT noalt, nom and word	92.42	81.65	86.70	7-21, F	9.58
F	BDT word only	71.65	61.87	66.40	19-21	37.07

Table 3: The precision, recall, and F-score of machine learning experiments to learn gene mentions using the data extracted from all submitted runs as features. Method column: BDT = boosted decision trees, CRF = conditional random fields, nom = all nomination features, word = words of candidate, noalt = alternate gene data not used. The column  $signif$  indicates the ranks of runs for which there was a significant difference, and the letters indicate the machine learning experiments for which there was a significant difference. The column % alt gives the percentage of alternate gene mentions among the resulting true positives.

with alternates, but the data from individual runs is not sufficient to settle the issue. Given that the boosted decision tree result, which uses alternates, is about the same as the conditional random field result, we might conclude that training with alternates does not make the task significantly easier. We therefore trained with boosted decision trees, marking candidates as positive only if they appear as gene annotations, *i.e.* ignoring alternates. The result was an F-score of 86.70, which is a statistically significant difference with the result 90.50 obtained by training in the same way with alternates positive. Training with alternates generated true positives that contained 25.67% alternates, while training without alternates generated true positives containing only 9.58% alternates.

We believed that the results from the lowest scoring teams, if used appropriately, could contribute useful information towards identifying gene mentions. To test the hypothesis, we trained with boosted decision trees using word features plus all nomination features from teams ranked 1 through 10 only. The result gave an F-score of 89.40, which is significantly lower than the 90.50 obtained when features from teams with ranks 11 through 21 were included. This confirms the importance of results from teams with lower individual performance. We note, for example, that the lowest ranking team obtained 8 true positives that were not obtained by any other run.

## 5 Discussion

The submission data can be used as a source for exploring the consistency and accuracy of corpus annotations. There were no false positives common to all 21 submissions, but there were 2 that were common to 17 submissions, for the names *GH* and *FAK*, both of which should have been annotated as true. There are more of these false positives with less than 17 common submissions that deserve further review. We also found 34 gene mentions that were false negatives in all 21 submissions, but all of these were found to be correctly annotated in the corpus. Mentions with a high false negative rate may be clues to difficult or under-represented gene mentions, and studying these may give some guidance to future systems developers. As much as we would like to increase the representation of “difficult” gene mentions, this may be infeasible because it is likely that they obey a Zipf-like distribution: there are as many uniquely difficult gene mentions as there are common and easy ones.

It can be argued that the difficulty experienced by human annotators in reaching mutual agreement directly limits the performance of automated systems, and this can be influenced by the clarity of the annotation guidelines. It has been pointed out that the guidelines for annotating genes are surprisingly short and simple given the complex guidelines for annotating named entities in news wires [2]. However, a gene is a scientific concept, and it is only reasonable to rely on domain experts to recognize and annotate gene mentions. Thus, the gene annotation guidelines can be conveyed by reference to a body of knowledge shared by individuals with experience and training in molecular biology, and it is not feasible to give a complete specification for gene annotation that does not rely on this extensive background knowledge. Nevertheless, we believe that some improvement could be achieved by documenting current annotation decisions for difficult and ambiguous gene mentions.

The highest F-score obtained on the BioCreative 2 evaluation is 87.21, and we have shown that by combining the efforts of all systems it is possible to achieve an F-score of 90.66, a significant improvement. This proves that future systems should be able to achieve improved performance. Though this F-score is only relevant to the BioCreative 2 test corpus, it is feasible, as illustrated here, to estimate the performance of future systems on the current corpus, and thus to measure the improvement in future systems’ performance. We are also optimistic that, through a combination of refining the corpus for annotation consistency and improving systems design through collaboration, even greater improvements in performance are achievable.

## Acknowledgments

This research was supported by the Intramural Research Program of the NIH, NLM, and NCBI.

## References

- [1] Carreras, X and Marquez, L. (2001) *Boosting trees for anti-spam email filtering*. In RANLP2001. Tzigov Chark, Bulgaria.
- [2] Hirschman, L and Chinchor, N. (1997) *Muc-7 named entity task definition*. In Proceedings of the 7th Message Understanding Conference.
- [3] McCallum, A. (2003) *Efficiently inducing features of conditional random fields*. In Nineteenth Conference on Uncertainty in Artificial Intelligence (UAI03).
- [4] Schapire, RE and Singer, Y. (1999) *Improved boosting algorithms using confidence-rated predictions*. Machine Learning, 1999. 37(3): p. 297-336.
- [5] Smith LH, Tanabe L, Rindfleisch T and Wilbur WJ. (2005) *MedTag: A Collection of Biomedical Annotations*. In Proceedings of the ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics, pp. 32-37. Detroit, June 2005.
- [6] Tanabe, L, and Wilbur, WJ. (2002) *Tagging Gene and Protein Names in Biomedical Text*. Bioinformatics, 18:1124-1132, 2002.
- [7] Tanabe L, Xie N, Thom LH, Matten W, Wilbur WJ. (2005) *GENETAG: A Tagged Corpus for Gene/Protein Named Entity Recognition*. BMC Bioinformatics 6(Suppl 1):S3.
- [8] Voorhees, EM. (2005) *Overview of TREC 2005*. NIST Special Publication 500-266.
- [9] Yeh, AS, Morgan, A, Colosimo, M, Hirschman, L. (2005) *BioCreAtIvE task 1A: gene mention finding evaluation*. BMC Bioinformatics 6(Suppl 1):S2.
- [10] *BioCreative 2: Gene Normalization Task*. These proceedings.