A Sequential Algorithm for Training Text Classifiers: Corrigendum and Additional Data

David D. Lewis
AT&T Bell Laboratories
Murray Hill, NJ 07974
USA
lewis@research.att.com

1 Introduction

At ACM SIGIR '94, I compared the effectiveness of ***uncertainty sampling*** with that of random sampling and relevance sampling in choosing training data for a text categorization data set [1]. (Relevance sampling is the application of relevance feedback [3] to producing a training sample.)

I have discovered a bug in my experimental software which caused the relevance sampling results reported in the SIGIR '94 paper to be incorrect. (The uncertainty sampling and random sampling results in that paper were correct.) I have since fixed the bug and rerun the experiments. This note presents the corrected results, along with additional data supporting the original claim that uncertainty sampling has an advantage over relevance sampling in most training situations.

2 Methods

The SIGIR '94 experiment, and the experiments reported here, proceeded as follows. (See the original paper [1] for full details.) The experimental variable was the method used to choose training samples for a text categorization problem. Uncertainty sampling, relevance sampling, and random sampling were tested.

The data, a set of 371,454 titles of AP newswire items, was divided randomly into a training set of 319,463 titles and a test set of 51,991 titles. Ten categories of titles were defined based on the keyword slug line of the AP items, and are listed in Table 1 (identical to Table 1 of the original paper).

Each experimental run involved selecting a sample from the 319,463 training titles, looking up the class label for each item in the sample (i.e. looking up whether each item belonged to the category of interest), and using the labeled sample to train a binary classifier by statistical methods. The binary classifier was then run on each of 51,991 test titles. Its ability to correctly decide whether test titles belonged to the category of interest was measured using the effectiveness measure $F_{\beta=1} = 2PR/(P+R)$, where P is precision and R is recall.

The samples used were built up incrementally. The initial sample in all cases was a randomly selected set of three titles which were category members. The sampling method being tested then built up larger samples by adding additional titles to these three in different ways. The effectiveness of classifiers trained on samples of several sizes was measured. Ten runs were made for each category, using ten different randomly selected initial sets of three category members. The initial sets were controlled (held constant) across variations in sampling method. For random sampling twenty runs were done, two with each of the ten initial sets of three category members.

	Training		Test	
Category	Number	Freq.	Number	Freq.
tickertalk	208	0.0007	40	0.0008
boxoffice	314	0.0010	42	0.0008
bonds	470	0.0015	60	0.0012
nielsens	511	0.0016	87	0.0017
burma	510	0.0016	93	0.0018
dukakis	642	0.0020	107	0.0021
ireland	780	0.0024	117	0.0023
quayle	786	0.0025	133	0.0026
budget	1176	0.0037	197	0.0038
hostages	1560	0.0049	228	0.0044

Table 1: The ten categories used in the SIGIR-94 paper, with number of occurrences and frequency of occurrence on training and test sets.

	3 + 996 uncer.		3 + 7 rand.		3 + 996 rel. (buggy)		3 + 996 rel. (correct)		3 + 319,460 full	
Category	mean	SD	mean	SD	mean	SD	mean	SD	mean	SD
tickertalk	.033	(.031)	.018	(.023)	.023	(.039)	.071	(.046)	.047	(.001)
boxoffice	.700	(.041)	.222	(.172)	.481	(.053)	.714	(.029)	.647	(.023)
bonds	.636	(.034)	.146	(.134)	.541	(.069)	.629	(.019)	.509	(.020)
nielsens	.801	(.016)	.291	(.218)	.567	(.132)	.762	(.031)	.741	(.022)
burma	.653	(.035)	.032	(.033)	.201	(.057)	.691	(.032)	.464	(.023)
dukakis	.136	(.046)	.101	(.075)	.035	(.021)	.119	(.038)	.163	(.015)
ireland	.416	(.041)	.050	(.033)	.170	(.038)	.448	(.030)	.288	(.030)
quayle	.386	(.040)	.081	(.064)	.140	(.072)	.395	(.026)	.493	(.009)
budget	.290	(.039)	.058	(.046)	.141	(.029)	.310	(.030)	.235	(.005)
hostages	.477	(.021)	.068	(.042)	.177	(.039)	.436	(.021)	.498	(.003)

Table 2: Mean and standard deviation of $F_{\beta=1}$ for training on initial 3 examples combined with each of 996 uncertainty selected examples, 7 random examples, 996 relevance selected examples (both buggy SIGIR-94 results and new corrected results), or 319,460 remaining examples. Means are over ten runs for uncertainty and relevance sampling, and over twenty runs for random and full sampling.

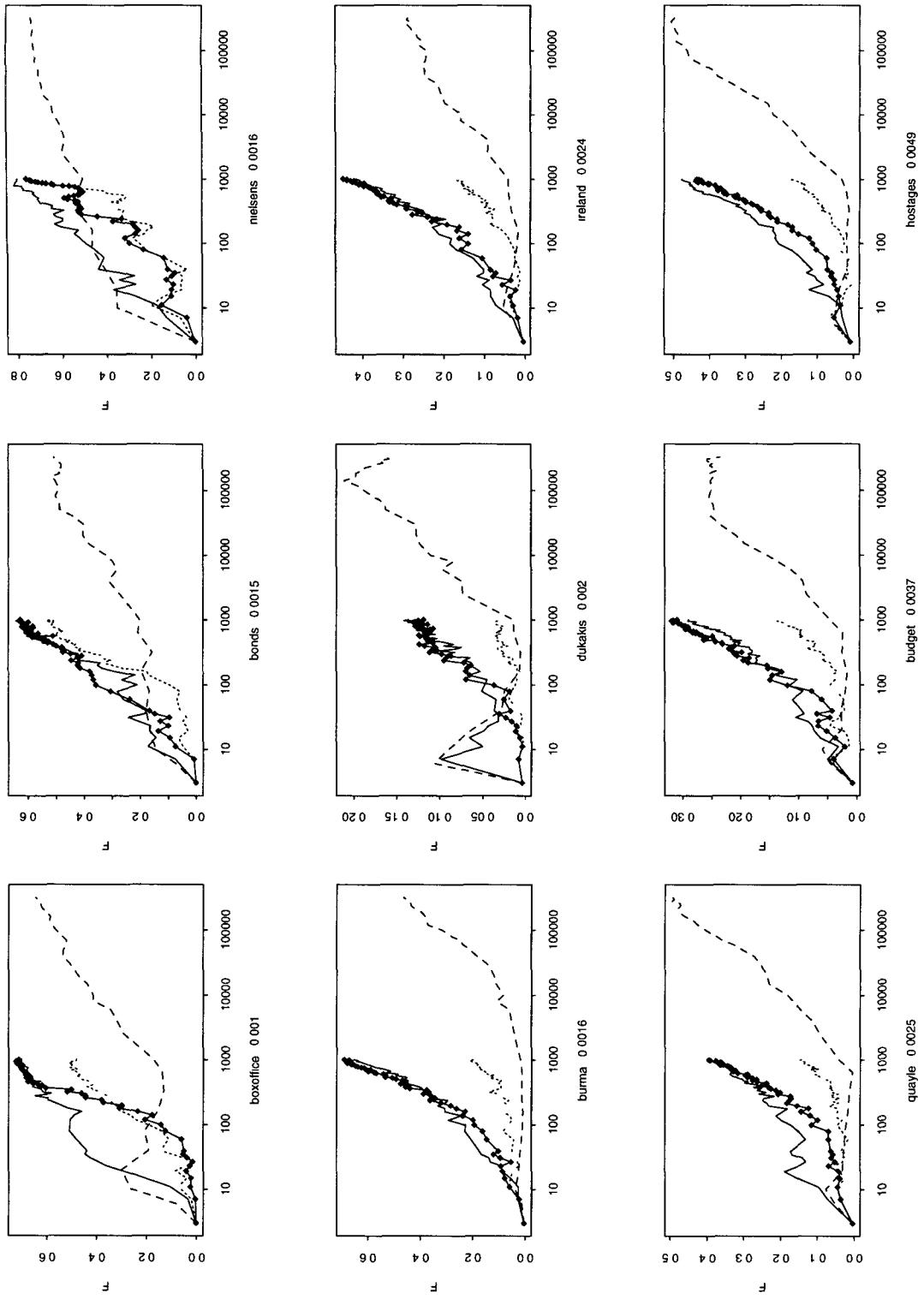

Figure 1: Mean $F_{\beta=1}$ values for text classifiers trained using uncertainty sampling (solid line), relevance sampling (buggy: dotted line, corrected: solid plus diamonds), and random sampling (dashed line). Means are over ten runs for uncertainty and relevance sampling, and over twenty runs for random sampling. Results are shown for nine categories. Frequency of category on training set is shown after category name.

	3 + 36 uncer.		3 + 36 rel.		3 + 96 uncer.		3 + 96 rel.	
Category	mean	SD	mean	SD	mean	SD	mean	SD
tickertalk	.001	(.002)	.007	(.022)	.017	(.031)	.011	(.024)
boxoffice	.437	(.156)	.049	(.073)	.506	(.053)	.143	(.162)
bonds	.178	(.174)	.167	(.127)	.212	(.140)	.360	(.141)
nielsens	.441	(.198)	.129	(.192)	.495	(.178)	.298	(.243)
burma	.176	(.119)	.103	(.089)	.227	(.131)	.194	(.086)
dukakis	.036	(.022)	.017	(.023)	.048	(.048)	.037	(.026)
ireland	.102	(.066)	.085	(.044)	.184	(.064)	.139	(.041)
quayle	.173	(.118)	.063	(.054)	.171	(.108)	.116	(.096)
budget	.095	(.101)	.042	(.059)	.116	(.034)	.120	(.037)
hostages	.114	(.056)	.073	(.059)	.206	(.074)	.112	(.059)

Table 3: Mean and standard deviation of $F_{\beta=1}$ for training on initial 3 examples combined with samples of size 39 and 99 chosen by uncertainty sampling or relevance sampling. Means are over ten runs.

3 Results

Table 2 of this note repeats the data from Table 2 of the SIGIR-94 paper, but now includes both the buggy original relevance sampling results and the new corrected relevance sampling results. Similarly, Figure 1 of this note corresponds to Figure 2 of the SIGIR-94 paper, but includes both the original and corrected relevance sampling results.

The corrected relevance feedback results are substantially higher than the buggy ones, reducing the difference observed between uncertainty sampling and relevance sampling. For the larger sample sizes the effectiveness of the two becomes very similar. In retrospect this is not surprising, given the low frequency of the categories used. The total number of positive examples in the training data ranged from 208 to 1560 for the ten categories. Therefore training samples of several hundred titles chosen by either method are likely to contain many of the same category members, leading to similar effectiveness of the trained classifiers.

This suggests that any advantage of uncertainty sampling over relevance sampling would be more apparent when the sample size is relatively small in comparison to the number of positive examples available. This can be shown in two ways: by decreasing the sample size, or increasing the number of positive examples.

The first approach can be taken simply by examining the leftmost portion of the graphs in Figure 1, where uncertainty sampling dominates the corrected relevance sampling for most categories when the sample size is below 100. Table 3 presents the corresponding numeric data for two specific sample sizes, 39 and 99. (The particular sample sizes shown are those convenient given the iterative procedure by which the samples were built up and data collected.) For samples of size 39, uncertainty sampling is more effective than relevance sampling on nine out of ten categories, and for samples of size 99, it is better on eight out of ten categories.

The advantage of uncertainty sampling can also be shown by keeping our sample size constant while increasing the number of category members in the training set. Table 4 presents summary data on ten additional AP categories chosen to have 1800 to 9000 positive training instances, about a factor of five larger than those in the first set. I replicated the SIGIR-94 experiments (using the corrected relevance sampling) on this second set of ten categories. Table 5 is analogous to Table 2 but shows mean F-values for the ten new categories. With these higher frequency categories uncertainty sampling is strikingly superior to relevance sampling, even with sample sizes as large as 999.

Figure 2 graphs the results for nine of the Set 2 categories (omitting the category *aparts*, for which all

	Training		Test	
Category	Number	Freq.	Number	Freq.
aparts	1835	0.0057	287	0.0055
dollargold	2219	0.0069	376	0.0072
yugoslavia	2708	0.0085	491	0.0094
weatherpageweather	2800	0.0088	432	0.0083
german	3397	0.0106	568	0.0109
britain/british	4753	0.0149	804	0.0155
israel	5442	0.0170	862	0.0166
bush	6057	0.0190	934	0.0180
japan	6445	0.0202	1023	0.0197
gulf	8637	0.0270	1403	0.0270

Table 4: The Set 2 categories, with number of occurrences and frequency of occurrence on training and test sets.

	3 + 996 uncer.		3 + 7 rand.		3 + 996 rel.		3 + 319,460 full	
Category	mean	SD	mean	SD	mean	SD	mean	SD
aparts	.038	(.017)	.035	(.019)	.041	(.019)	.018	(.007)
britain/british	.364	(.012)	.024	(.038)	.196	(.096)	.415	(.001)
bush	.120	(.012)	.079	(.073)	.103	(.015)	.244	(.002)
dollargold	.977	(.002)	.702	(.185)	.515	(.269)	.930	(.002)
german	.320	(.017)	.043	(.052)	.295	(.018)	.457	(.002)
gulf	.258	(.026)	.033	(.030)	.204	(.023)	.415	(.001)
israel	.437	(.023)	.022	(.033)	.280	(.093)	.537	(.002)
japan	.579	(.015)	.108	(.080)	.214	(.190)	.580	(.001)
weatherpageweather	.893	(.011)	.335	(.124)	.207	(.136)	.861	(.002)
yugoslavia	.569	(.028)	.027	(.024)	.357	(.051)	.600	(.002)

Table 5: Results for Set 2 categories: Mean and standard deviation of $F_{\beta=1}$ for training on initial 3 examples combined with each of 996 uncertainty selected examples, 7 random examples, 996 relevance selected examples, or 319,460 remaining examples. Means are over ten runs for uncertainty and relevance sampling, and over twenty runs for random and full sampling.

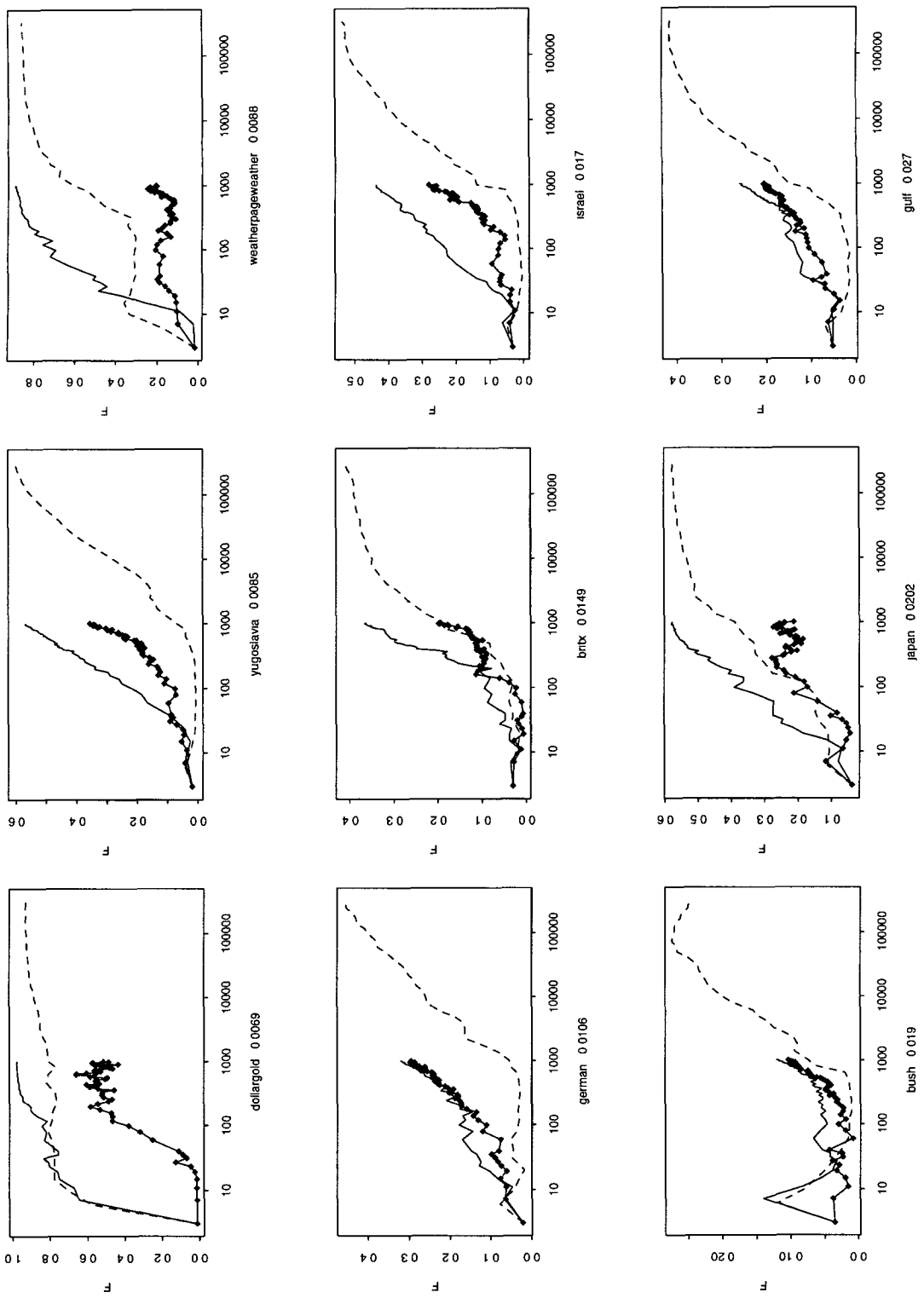

Figure 2: Results for Set 2 categories: Mean $F_{\beta=1}$ values for text classifiers trained using uncertainty sampling (solid line), corrected relevance sampling (solid plus diamonds), and random sampling (dashed line). Means are over ten runs for uncertainty and relevance sampling, and over twenty runs for random sampling. Results are shown for nine categories. Frequency of category on training set is shown after category name.

methods did poorly), using the same format as Figure 1. For these higher frequency categories, uncertainty sampling clearly dominates relevance sampling over the entire range of sample sizes considered.

4 Summary

Uncertainty sampling and relevance sampling are two of a range of possible active exemplar selection [2] approaches to choosing training data. Both are far superior to random sampling when limitations of time or money mean that only a fraction of a data set can be labeled for training classification rules.

Our corrected results show less difference between uncertainty sampling and random sampling than originally appeared to be the case. However uncertainty sampling has a large advantage when the size of the training sample to be produced is small in comparison with the number of positive examples in the unlabeled training data. This is a common situation in practice, since data sets are constantly growing while the time available for labeling data is shrinking. Uncertainty sampling and other active learning methods are still at an early stage of development, so further improvements are likely.

Acknowledgments

I thank Doug McIlroy for useful comments on this note.

References

[1] David D. Lewis and William A. Gale. A sequential algorithm for training text classifiers. In W. Bruce Croft and C. J. van Rijsbergen, editors, *SIGIR 94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 3–12, London, 1994. Springer-Verlag.

[2] Mark Plutowski and Halbert White. Selecting concise training sets from clean data. *IEEE Transactions on Neural Networks*, 4(2):305–318, March 1993.

[3] Gerard Salton and Chris Buckley. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41(4):288–297, 1990.