In Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005),
November 2005, Gaithersburg, Maryland.

Fusion of knowledge-intensive and statistical approaches for retrieving and annotating textual genomics documents

Alan R. Aronson,[a] Dina Demner-Fushman,[a,b] Susanne M. Humphrey,[a] Jimmy Lin,[a,b] Hongfang Liu,[c] Patrick Ruch,[a] Miguel E. Ruiz,[a] Lawrence H. Smith,[a] Lorraine K. Tanabe,[a] W. John Wilbur,[a]

[a]National Library of Medicine, Bethesda, Maryland
{alaronson, ddemner, shumphrey, larsmith, tanabe, wilbur}@mail.nih.gov;
patrick.ruch@sim.hcuge.ch; meruiz@buffalo.edu

[b]University of Maryland, College Park, Maryland
jimmylin@umd.edu
[c]University of Maryland, Baltimore County, Maryland
hfliu@umbc.edu

Abstract

This paper represents a continuation of research into the retrieval and annotation of textual genomics documents (both MEDLINE® citations and full text articles) for the purpose of satisfying biologists' real information needs. The overall approach taken here for both the ad hoc retrieval and categorization tasks within the TREC genomics track in 2005 was one combining the results of several NLP, statistical and ML methods, using a fusion method for ad hoc retrieval and ensemble methods for categorization. The results show that fusion approaches can improve the final outcome for the ad hoc and the categorization tasks, but that care must be taken in order to take advantage of the strengths of the constituent methods.

Keywords: Genomics; MEDLINE/PubMed; MeSH; Information Retrieval; Vector Space Models; Statistical Natural Language Processing; Machine Learning; Thematic Analysis.

1 Introduction

In the first two years of the TREC genomics track, the NLM/UMd team's best systems performed adequately in both the ad hoc retrieval and categorization tasks, but did not lead by a wide margin in the first year and suffered from over-training in the second year. In 2005 we participated in both the ad hoc retrieval and categorization tasks of the genomics track and strove to overcome the above weaknesses by combining the results of various NLP, statistical and ML methods to achieve our final results. For ad hoc retrieval, we used a fusion approach; for categorization, ensemble methods.

Section 2 of this paper describes our efforts on the ad hoc retrieval task, section 3 discusses the categorization task, and section 4 contains some conclusions about the work.

2 Ad hoc Retrieval Task

For the ad hoc retrieval task we combined the retrieval results of four systems (Smart, InQuery, easyIR and Theme) each of which is known to perform well for some IR tasks. Our fusion approach for ad hoc retrieval consisted of normalizing the scores from each system on a query by query basis and then using these normalized scores to compute a new combined score for the union of all results returned by all four systems. The top 1,000 results were selected based on the combined score.

2.1 Basic ad hoc approaches

2.1.1 Query expansion

All systems that were used in the ad hoc retrieval task experimented with various query expansions: gene name expansion, MeSH profile expansions by template, and disease-name expansions. In addition some systems indexed and searched specific MEDLINE document fields, e.g., MeSH headings.

The gene names were identified using ABGene (Tanabe and Wilbur, 2002) and then expanded with their synonyms. Four methods of synonym identification were developed:

1. MeSH-based expansion – Pattern matching against MeSH regular descriptor files and MeSH supplementary records file.
2. Entrez Gene database synonym lookup.
3. Popularity expansion – Synonyms for the expansion were selected based on their popularity defined as the number of different databases that contain this synonym, and ambiguity defined as the number of different UniRef50 groups that include this synonym

or its textual variants as symbols, names, or synonyms. Only synonyms with high popularity and low ambiguity were used for expansion.

4. Thesaurus-based expansion – A set of online resources, such as UniProt/SwissProtKB, GPSDB and other well known databases was used to build our thesaurus.

Diseases were identified using MetaMap (Aronson, 2001) and expanded using all textual strings with the same UMLS unique concept identifier.

As part of our strategy for the ad hoc retrieval topics, templates were expanded using very broad PubMed searches. Table 1 contains the names of template components and their Boolean intersection or union, and Table 2 contains the template components and the corresponding actual PubMed searches which should be substituted for the component names. For example, Template1 consists of the intersection of INVESTIGATIVE_TECHNIQUES and METHOD_OR_PROTOCOL which, according to its components, translates into the following actual PubMed search: investigative techniques[mh] AND (methods[sh] or isolation and purification[sh]).

Template1	INVESTIGATIVE_TECHNIQUES and METHOD_OR_PROTOCOL
Template2	GENES and DISEASE
Template3	GENES and BIOLOGICAL_PROCESS
Template4	GENES and (DISEASE or BIOLOGICAL_PROCESS)
Template5	GENES and MUTATIONS and (DISEASE or BIOLOGICAL_PROCESS)

Table 1. Templates

2.1.2 Systems

Smart. We used the Smart system created by Salton (1971) and his collaborators. Our version of Smart has been modified to add modern weighting schemes and to handle 11 European languages. In this work we used a simple stemmer that only removes plurals and a stopword list that was reviewed to avoid discarding terms that could have potential meaning in the genomics domain. Documents and queries were indexed using 4 ctypes:

ctype 1: words in title and abstract (using a stop list and a simple stemmer that only stems plurals)

ctype 2: Substances and terms found in the RN field in the MEDLINE record

INVESTIGATIVE_TECHNIQUES
investigative techniques[mh]

METHOD_OR_PROTOCOL
methods[sh] OR isolation and purification[sh]

GENES
proteins[mh] OR enzymes[mh] OR peptide hormones[mh] OR intercellular signaling peptides and proteins[mh] OR intracellular signaling peptides and proteins[mh] OR genes[mh] OR genetics[sh] OR genetic processes[mh] OR genetic phenomena[mh] OR genetic structures[mh] OR immunogenetics[mh]

DISEASE
disease category[mh] OR mental disorders[mh] OR abnormalities[sh] OR injuries[sh]

BIOLOGICAL_PROCESS
cell physiology[mh] OR genetic processes[mh] OR biochemical phenomena[mh] OR metabolism[mh] OR metabolism[sh:noexp] OR enzymology[sh] OR biosynthesis[sh] OR immunology[sh] OR physiology[sh:noexp] OR cytology[sh] OR chemistry[sh:noexp] OR antagonists and inhibitors[sh] OR genetics[sh] OR physiopathology[sh] OR deficiency[sh]

MUTATIONS
mutation[mh] OR mutagenesis[mh]

Table 2. Template components and PubMed searches

ctype 3: MeSH terms (represented by single words and word bigrams)

ctype 4: word bigrams from the title and abstract.

The similarity between query and documents was computed using a linear combination of the scores of each ctype with weights 7, 1, 2 and 1 for each of the ctypes described above.

Queries were expanded before retrieval by adding gene-names synonyms extracted from MeSH (MeSH-based gene expansion). The weighting scheme used in this run was *atn.ann*. Both the weighting scheme and weights of each ctype were selected by maximizing the MAP over the 10 training topics. More details can be found in (Ruiz, 2005).

InQuery. We indexed the document collection without stemming. Queries were performed using the InQuery sum operator and expanding only disease

names. The details of our use of the InQuery system are described in (Lin et al., 2005)

easyIR. Significant variance in the document length in MEDLINE motivated our evaluation of the effectiveness of a statistical weighting model based on a pivoted normalization factor (Singhal et al. 1996, Fujita 2004). MEDLINE document length seems to be the result of a two-Gaussian mixture with a maximum at 204 and 32 (tokens). We used the ten training topics to select the best statistical parameters and the best normalization and expansion methods. The best weighting was obtained using a slightly modified dtu.dtn formula (Singhal 2001, Ruch et al. 2004), with slope = 13 and using a slightly modified Porter stemmer (in particular, 'a' was removed from the stop words and '-' was not considered a separator).

Gene and protein names can be highly variable, and their recognition is far from trivial and could result in some inappropriate expansion due to lexical ambiguities. From a comprehensive set of experiments including thesaurus-based gene name expansion and some very conservative and minimal approaches (Ruch et al. 2005, Abdou et al. 2005), it is worth observing that thesaurus-based gene expansion seems rather ineffective for MEDLINE retrieval.

In addition to gene-name expansion, we evaluated the impact of expanding other types of entities: chemicals (calcium), diseases (cancer), species (rats), and body parts (spleen). Results are reported in Table 3.

Baseline (slope = 13)	0.1751
Expanding chemicals, diseases, species and body parts and removing documents not containing the species	0.1775

Table 3. The result of expanding entities other than gene names

Theme. In this approach, queries were performed on the current MEDLINE database indexed on MeSH terms, single word terms and two word phrases in titles, abstracts. The final results were intersected with the TREC test set.

Probabilistic Method for Query Expansion. Given a set of documents, a query can be expanded with the nearest neighbors algorithm (Wilbur and Coffee, 1994) using concepts extracted from the documents (Kim and Wilbur, 2005). The Bayesian weights of the high weight terms in these documents were used to rescore all of the documents in MEDLINE.

Each topic was handled by combining multiple queries and query expansions. The result of each query, query expansion, or nearest neighbor was that each document in MEDLINE was scored with a log odds score. The log odds scores were normalized and converted to probabilities using the formula $p = 1 / (1 + exp(a\ z + b)$ where z is the document score and a and b are computed so that the original number of documents have probability 0.9 or greater and no more than 10 times the original number of documents has probability greater than 0.5. The probabilities were combined using standard fuzzy logic formulas: $p_{AND} = p1 * p2$ and $p_{OR} = p1 + p2 - p1 * p2$. This allowed expanded sets to be combined with unexpanded sets (with document probability 0 or 1). When query expansion was applied to a fuzzy set, it was first converted to a set by thresholding.

To prepare the query, topic fields were extracted using separate patterns for each of the five templates.

Template 1 Nearest Neighbors and Boolean Logic. For template 1 queries, the entire text of the ad hoc query was used as a document for nearest neighbor retrieval (Wilbur and Coffee, 1994). The top 100 of these documents were then used for query expansion and then intersected with the documents of the test collection. This result was ANDed with a query expansion of a "gene and experiment" query (synonyms of the word gene and experiment also appear in this query).

Template 2-5 Synonym Lookup, Nearest Neighbors, Boolean Logic and Fuzzy Logic. For templates 2-5, gene names were expanded using synonyms from the Entrez Gene database[1]. Each of the alternatives was queried, and the results were combined into a single set which was not query-expanded. The remaining fields were analyzed and broken into single or double word terms and the query was expanded and combined. Finally, the query expanded fuzzy set was ANDed with the gene set and the final result was intersected with the test collection.

Example Query. As an example, query 118 is "Provide information about the role of the gene Transforming growth factor-beta1 (TGF-beta1) in the disease Cerebral Amyloid Angiopathy (CAA)." In the following expression, Query(...) denotes the result of an unexpanded query, and Expand(...) the result of expanding a query. The operations AND and OR are fuzzy set operations.

[1]http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene

```
1  (Query("transforming growth"
2           AND "growth factor"
3           AND "factor beta1")
4   OR
5   Query("tgf beta1"))
6  AND
7  Expand(Query("cerebral amyloid"
8                AND "amyloid angiopathy")
9    OR (Query("cerebral amyloid")
10         AND
11       Expand("angiopathies"))
12   OR Expand("caa"))
```

Lines 1-5 are the standard query for the gene, which has two synonyms, "transforming growth factor beta1" and "tgf beta1". The first synonym is grouped into 3 overlapping double word terms. The result of this gene portion of the query is ANDed with the disease portion from lines 7-12. The disease has 3 synonyms, "cerebral amyloid angiopathy", "cerebral amyloid angiopathies" and "caa". The system grouped the first synonym into 2 overlapping double word terms. The second synonym was obtained from UMLS. For this synonym, the query on overlapping pairs was empty and in this case all terms after the first one are expanded before ANDing. The third term was (correctly) recognized as an alternate name.

2.2 Fusion of ad hoc approaches

Our initial experiments indicated that the combination of all four systems resulted in a significant improvement compared to any of the four systems individually.

The final fusion run consisted of results from ***Smart*** using pre-retrieval expansion of gene names, results from ***InQuery*** using expansion of diseases, a retrieval run using the ***easyIR*** system with pivoted length normalization, slope=10, and disease expansion, and a retrieval run using the ***Theme*** probabilistic retrieval system. Combination of the scores was performed by normalizing each of the scores of individual systems on a query by query basis and adding them as proposed by Fox and Shaw (1994) and confirmed later by (Savoy, 2004).

$$rsv_{fusion}(i) = \sum_{s \in S} \lambda_s \frac{rsv_s(i) - \min_s}{\max_s - \min_s}$$

where $rsv_{fusion}(i)$ represents the final score of document i, S is the set of systems participating in the fusion, $\min_s$ and $\max_s$ are the minimum and maximum scores reported by system s. λ_s is a weighting factor that can be used to favor the most effective retrieval system participating in the fusion. Note that all these values are computed on a query by query basis.

A second ad hoc run was generated by selecting the best fusion runs for each of the five templates of the topics. For this purpose we evaluated several types of combinations of the four systems on the 10 training topics.

2.3 Ad hoc results

The results on the training topics showed that the fusion runs performed above each of the single systems. Table 4 shows the summary of our results on the training set. The first row represents our Baseline system which is a simple ***Smart*** run without any query expansion. The following four rows correspond to the performance of each individual system. The remaining rows represent different combinations of systems. Note that in this table the last two rows are second order fusions which are equivalent to assigning a higher weight to the ***Theme*** system. The second column shows the difference in performance with respect to the baseline. The results on the training set showed the fusion approach significantly improved performance with respect to the baseline system. For example, the fusion run labeled (TS)-(TI)-(TE) combined first the ***Theme*** system with each of the other three and then combined the tree resulting runs into a single run. This is equivalent to assign a weight 3 to the ***Theme*** and 1 to each of the other systems.

	MAP	Diff-baseline
Baseline	0.1713	
Theme (T)	0.2554	49%
Smart-gene-exp (S)	0.17	-1%
InQuery-disease (I)	0.1394	-19%
easyIR (slope=10) (E)	0.1710	0%
I-E	0.1801	5%
T-E	0.2847	66%
T-I	0.2648	55%
S-E	0.1968	15%
S-I	0.1852	8%
S-T	0.264	54%
S-T-I	0.2569	50%
S-T-E	0.2794	63%
S-T-I	0.2569	50%
S-I-E	0.1957	14%
T-I-E	0.2495	46%
S-T-I-E	0.2569	50%
(TS)-(TI)-(TE)	0.3021	76%
T-(TS)-(TI)-(TE)	0.3148	84%

Table 4. Performance on ad hoc training topics

Template	Best fusion Run	MAP
Methods	S-I-E	0.3297
Gene-disease	S-T-I-E	0.4446
Gene-Biological-Process	T-(TS)-(TI)-(TE)	0.2292
Gene-Function-Disease	T-(TS)-(TI)-(TE)	0.3858
Mutation-Gene-Function	S-T	0.4728
	Overall Performance	**0.3232**

Table 5. Performance by template (training queries)

We submitted two official runs. NLMfusionA corresponds to our second level run (TS)-(TI)-(TE). We debated on whether to submit as our second run a more conservative fusion approach that assumes equal weight to each system. However, we decided to explore whether tuning the fusion to each of the templates would yield a better approach. Our second official run was produced by selecting the best fusion run for each template. Table 5 shows the best performance run for each template and the corresponding performance of the template specific fusion.

Table 6 shows the official results as well as the unofficial results of our fusion runs in the test set. For comparison purposes we will use the ***Smart*** and the ***InQuery*** baseline (using the topics with no modifications) runs as a baseline since their performance is pretty close to the median system. Both of our official runs show results that are slightly above the median system. However, the difference with respect to the baseline system is not statistically significant. The best of our runs corresponds to a fusion run that weights equally all four systems (S-T-I-E). This run shows a significant improvement above the baseline (21%). This result shows that although individually each of the original systems does not perform significantly above the median system, the fusion approach can actually yield a significant improvement. Individually, the run produced with pivoted normalization (***easyIR***) performed slightly better than the other systems and was not affected by overfitting which was observed for some of the other systems. Pivoted length normalization seems effective for retrieval in MEDLINE.

	map	**deviation from baseline**	**bpref**
Unofficial runs			
Smart (S)	0.2262		0.2254
easyIR (E)	0.2373	5%	0.2546
Theme (T)	0.1777	-21%	0.1761
InQuery (I)	0.1729	-24%	0.1738
InQuery (basic)	0.2237	-1%	0.2266
S-E	0.2473	9%	0.2382
S-T	0.2432	8%	0.2420
S-I	0.2185	-3%	0.2158
T-E	0.2439	8%	0.2430
T-I	0.2120	-6%	0.2191
I-E	0.2311	2%	0.2290
S-T-E	0.2621	16%	0.2536
S-T-I	0.2512	11%	0.2474
S-I-E	0.2567	13%	0.2480
T-I-E	0.2589	14%	0.2591
S-T-I-E	**0.2736**	**21%**	**0.2680**
T-(TS)-(TI)-(TE)	0.2406	6%	0.2443
Official runs			
Template NLMfusionB *	0.2453	8%	0.2351
(TS)-(TI)-(TE) NLMfusionA *	0.2479	10%	0.2499

Table 6. Performance on test queries

A query by query analysis of the performance of both official runs shows that our best run is above the median on 31 queries and achieves the best score once. Our second official run, which uses the template optimization, performs above the median on 34 queries. The best unofficial run performs above the median system on 36 queries (See Figures 1-3).

With the exception of the first two templates (information about methods and protocols, and roles of genes in diseases) the difference between the results of two types of fusion for individual topics is in the second decimal point. For most of the topics in the first two templates, selecting a combination specifically for the template was beneficial with a larger effect observed for the second template. Although the template based optimization did improve the performance on several queries, it did

not achieve the best possible performance on each template. We believe that this is due to the fact that we tuned the template optimization with only two queries per template.

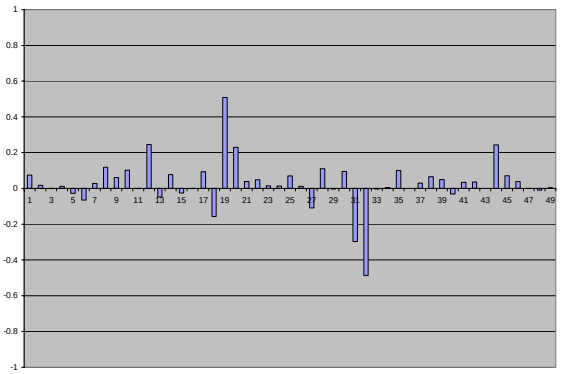

Figure 1. Comparison of NLMfusionA against the median

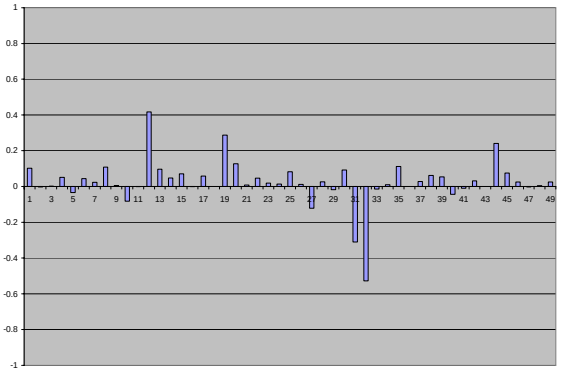

Figure 2. Comparison of NLMfusionB against the median

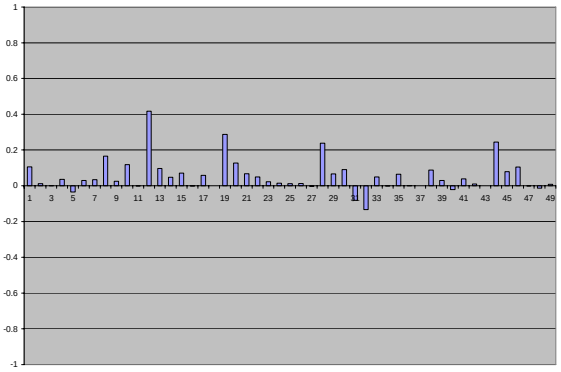

Figure 3. Comparison of best fusion run against the median

In addition to the official MAP measure that is determined by the ranks of the relevant documents in the result set and makes no distinction between documents explicitly judged as not relevant and the documents that are unjudged, the trec-eval package provides a preference-based measure, bpref, that depends on the number of judged non relevant documents retrieved before the relevant ones (Buckley and Voorhees, 2004). Table 6 reports our results for both measures. As expected, there is an excellent correlation between the measures (Kendall $\tau = 0.9131$) for submitted runs. Please note that the significant rank swaps reported in our notebook paper were caused by a problem in the bpref calculation from the originally disseminated judgments.

3 Categorization Task

The categorization task required triage of scientific articles for four types of information: Alleles of mutant phenotypes (task "A"), Embryologic gene expression (task "E"), GO annotation (task "G"), and Tumor biology (task "T"). For the categorization task, we used four machine learning methods (k-NN, SVM, NBL and Theme Detection). With the exception of k-NN, each of the machine learning classifiers was tuned on the training set using 5-fold cross validation.

3.1 Machine learning approaches

Machine learning for text categorization requires transforming each document into a feature representation (usually a feature vector) where features are usually words or word stems in the document. In addition to word or word stems in free text, we also explored other features that could be extracted from online resources.

Several supervised learning algorithms have been adapted for text categorization: Naïve Bayes learning (NBL) (Yang and Liu, 1999), neural networks (Wiener, 1995), instance-based learning (Iwayama and Takunaga, 1995), and Support vector machines (SVM) (Joachims, 1998). Yang and Liu (1999) provide an overview and a comparative study of different learning algorithms. We applied two of these machine learning algorithms in addition to k-NN and Theme Generation approaches.

k-NN. We used the easyIR engine, as described in the ad hoc retrieval task to compute the similarity between the document to be categorized and the training instances. We did not perform any feature selection on the indexing units; and we did not use the full-text of the articles, only the RN, MeSH, Title and Abstract fields of MEDLINE records. The

distance metrics defined by the engine showed some effectiveness for the ad hoc retrieval task and was applied to task 2.

To tune the k-NN system (mainly the k parameter), the training set was divided into two data sets: the first subset (10%) was used to evaluate the system, the second subset (90%) was used to tune the k-NN. Other classifiers were trained using a more powerful cross-validation method. Final runs were computed using all the available data: k was respectively set to 5, 30, 1 and 1 for the "A", "E", "G" and "T" tasks.

SVM and NBL. For each category, we built multiple classifiers using SVM and NBL on five different feature representations: MeSH, abstracts/titles, methods/figures, discussion/conclusion/results, and all available text including abstract and full text. We then used a pooling strategy followed by a voting scheme where the parameters were tuned using the training set.

NBL (Duda, 1973) is widely used in machine learning due to its efficiency and its ability to combine evidence from a large number of features. An NBL classifier chooses the category with the highest conditional probability for a given feature vector; while the computation of conditional probabilities is based on the Naïve Bayes assumption: the presence of one feature is independent of another when conditioned on the category variable. The training of the naïve Bayes classifier consisted of estimating the prior probabilities for different categories as well as the probabilities of each category for each feature.

The SVM method is a supervised learning algorithm proposed by Vladimir Vapnik and his co-workers (Vapnik, 1998). For a binary classification task with classes {+1, –1}, given a training set with n class-labeled instances, (x1, y1), (x2, y2), ..., (xi, yi), …, (xn, yn), where xi is a feature vector for the ith instance and yi indicates the class, an SVM classifier learns a linear decision rule, which is represented using a hyper-plane. The tag of an unlabelled instance x is determined by the side of the hyperplane on which x lies. The purpose of training the SVM is to find a hyper-plane that has the maximum margin to separate the two classes.

The SVM and the NBL methods were combined to obtain runs NLM1 and NLM2. We extracted MEDLINE citations information for both the training set and the test set. Besides that, we used a simple parser which parsed the full text so that each paragraph was associated with a section header such as (DISCUSSION, CONCLUSION etc).

For NLM1, we used SVM trained on the "AbstractText_ArticleTitle_Mesh Heading" features of the training set. A ranked list of the test set documents was then used in the ensemble methods to obtain combined results.

For NLM2 we then obtained ten classifiers (fp, ml), where ml is chosen from NBL and SVM, and fp has five values:

AbstractText_ArticleTitle, MeshHeading, RESULTS_RESULTSDISCUSSION_DISCUSSION_CONCLUSION, FIGURE_MATERIALMETHODS_PROCEDURES, ALLTEXT.

Except for MeshHeading that used the complete MeSH Headings as values, all other features used stemmed words in the free text as features. All classifiers were trained using the training set and obtained a ranked list of the test set. We then used threshold values to select top ranked documents and the final result, NLM2, was obtained by voting of the ten classifiers. For A and G, an instance was included in the final result if at least seven out of the ten classifiers had it in the top ranked list. For E and T, an instance was included in the final result if at least one out of the ten classifiers had it in the top ranked list.

Theme Detection. We applied an EM algorithm to generate themes for each category. Theme detection is based on a novel approach for discovering themes within text (Wilbur 2002). Theme detection was done on full text features using TexTool (Aronson et al., 2004). This year we added MeSH terms, and optimized for each category both 1) the number of theme terms, and 2) the score cutoff. We generated themes using 90% of the training data, and tested using 10%. For the final submission, we trained four themes (A, G, T, E) on 100% of the training data.

The full text sections, MeSH, Results, Discussion and Materials/Methods, contained useful terminology for categorization (see Table 7). For category A, the theme method recalled 99% of the true positives, and for category T, 100% of the true positives. This indicates that the theme method may work well for automated annotation, if the goal is to retain all true positives. The tradeoff for this high recall is low precision (see Figure 4).

T Top 20/100 Theme Terms		
Score	**Term**	**Section**
61.45	mice	Text
52.98	tumors	Results
49.37	tumors	Discussion
48.72	mice	Materials and Methods

46.01	paraffin	Materials and Methods
45.82	tumorigenesis	Results
45.25	eosin	Materials and Methods
44.45	hematoxylin	Materials and Methods
44.02	tumorigenesis	Discussion
42.10	southern	Materials and Methods
39.18	tumorigenesis	Text
38.07	tumors	Text
37.54	tumors	Materials and Methods
36.19	histological	Results
35.99	sections	Materials and Methods
35.53	tumor	Discussion
35.22	tumor	Results
35.09	mice	Results
34.13	tumor development	Discussion
33.25	mice, knockout	MeSH

Table 7. NCBI Theme for T

3.2 Fusion of categorization approaches

On the training data, our classifiers performed equally according to the utility measures (see Table 8.) However important differences in document selection were also found, therefore we decided to combine our results.

A-KNN Normalized Utility	0.6920
A-NCBI Normalized Utility	0.8824
A-SVM Normalized Utility	0.8824
E-KNN Normalized Utility	0.7552
E-NCBI Normalized Utility	0.7465
E-SVM Normalized Utility	0.6302
G-KNN Normalized Utility	0.5706
G-NCBI Normalized Utility	0.5338
G-SVM Normalized Utility	0.5377
T-KNN Normalized Utility	0.7532
T-NCBI Normalized Utility	0.9773
T-SVM Normalized Utility	0.7403

Table 8. Performance on training data

We opted for a voting model in which each of the three classifiers was given the same weight so that every document that was provided by at least two systems was selected. Except for task "T", where the NCBI alone obtained the best results (0.9773), the voting model outperformed other classifiers with respect to utility measures (see Table 9).

A Normalized Utility	0.9014
E Normalized Utility	0.8403
G Normalized Utility	0.5938
T Normalized Utility	0.9708

Table 9. Fusion Results (training)

Stacking was the second method for constructing ensembles of heterogeneous classifiers that we evaluated for the categorization task. We used stacking with probability distributions and multi-response linear regression that has been shown to perform best (Ting and Witten, 1999). During training, for all but the allele classification task, the coefficient assigned to one of the systems was so large that the contribution of the other two systems was insignificant. For the allele task, the coefficients were determined to be 2/3 for the ncbi classifier, 1/3 for svm and 0 for k-NN. This motivated our submission of the combined run for this task and selecting the second-best performing classifier for each of the remaining tasks. Due to a misunderstanding in the NLM coordination team that assumed the best runs for these tasks were submitted separately, some of our best runs were not submitted. These runs are labeled NCBIx in Figure 4.

Another combination that performed well in training was the simple voting scheme described above. The benefits displayed by the simple voting scheme during training were not confirmed in the test classification task; however the coefficients determined using stacking were stable and resulted in our best submission that also outperforms both base contributing systems (see Figure 4)

Of importance in understanding the relative subperformance of the k-NN classifier, we must observe that the tool was validated only on 10% of the available data, while the other classifiers were tuned using cross-validation. In addition, these results suggest that for such a task, a feature weighting and selection based on full-text articles might be more effective than simple MEDLINE records.

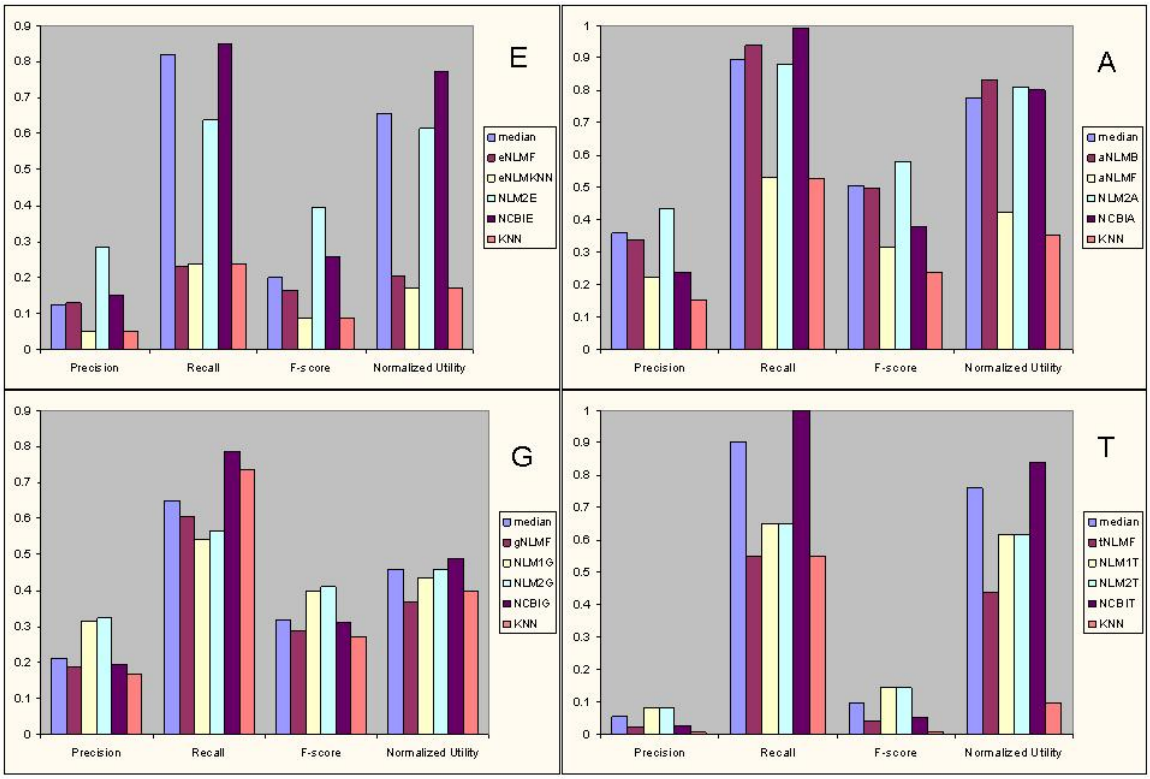

Figure 4. Classification task results. (xNLMF = voting; NLM1x=SVM; NLM2x= NBL&SVM; aNLMB=Fusion; NCBIx=theme queries; KNN=K-nearest neighbor)

4 Conclusions

Our results on the ad hoc retrieval task show consistent performance improvement. Our two official runs perform above the median even though some of the original runs used in the fusion were not significantly better than the median. We also found that using a more conservative weighting of the contribution of each system is a safer approach to improve retrieval performance.

The results of the query expansion techniques and template specific retrieval are inconclusive and require further investigation.

Given the excellent correlation of the MAP and bpref observed in the submitted runs we would like to suggest bpref as a second official evaluation measure (in addition to the mean average precision) for the genomics track. Due to bpref's known stability with respect to incomplete relevance judgments and the fact that relevant documents' scores are independent of the rank of other relevant documents when measured using bpref (Buckley and Voorhees, 2004), this will ensure usefulness of the genomics track collection to systems that did not participate in the evaluation.

Our results on the categorization task show that full text features provide useful information.

Stacking is a good strategy for fusing categorization results. The insights provided by this method during training were confirmed in testing: the optimal combination of the individual runs resulted in our best fusion run that significantly improved over both contributing base systems. On the other hand, the inability of this method to combine systems for tasks other than "A" was probably indicative of the poor performance of ensembles of our classifiers for these tasks.

If systems cannot be combined well on a training set using numerical coefficients (see above), a simple voting procedure is unlikely to perform well. The poor performance by our simple voting system can largely be attributed to the loss of high ranking unique documents due to our 2/3 voting requirement.

Voting strategies should take into account the compatibility of the combined systems, including the types of features used. For example, our NB/SVM and Theme Generation methods used full text features, but k-NN did not. Thus, our 2/3 voting requirement had the undesired effect of negating the contribution of full text features when NB/SVM and Theme Generation results did not agree.

Acknowledgements and Affiliations

We would like to acknowledge the significant contribution to this research of three recent visiting faculty members to NLM: Patrick Ruch (University Hospital of Geneva), Miguel E. Ruiz (State University of New York at Buffalo) and Jimmy Lin (University of Maryland).

References

Abdou S., Ruch P., Savoy J. (2005) General vs. Specific Blind Query Expansion for Biomedical Searches. In *Proceedings of the Fourteen Text Retrieval Conference TREC 2005*. Gaithersburg, MD.

Aronson A.R. (2001) "Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program." *Proc AMIA Symp*., 17-21.

Aronson A.R., Demner D., Humphrey S.M., Ide N.C., Kim W., Liu H., Loane R.R., Mork J.G., Smith L.H., Tanabe L.K., Wilbur, W.J. and Xie N. (2004) "Knowledge-intensive and statistical approaches to the retrieval and annotation of genomics MEDLNE citations." *The Thirteenth Text Retrieval Conference, TREC-2004*, Gaithersburg, MD.

Buckley C. and Voorhees E.M. (2004) Retrieval evaluation with incomplete information. *SIGIR 2004*: 25-32.

Duda R. and Hart P. (1973) *Pattern Classification and Scene Analysis*. John Wiley and Sons, NY.

Fox E.A. and Shaw J.A. (1994). Combination of multiple searches. In *Proceedings TREC-2*, (pp. 243-249). Gaithersburg: NIST Publication #500-215.

Fujita S. (2004) "Revisiting Again Document Length Hypotheses: TREC-2004 Genomics Track Experiments at Patolis." *The Thirteenth Text Retrieval Conference, TREC-2004*, Gaithersburg, MD.

Iwayama M. and Tokunaga T. (1995) Cluster-based text categorization: a comparison of category search strategies, *SIGIR 1995*, 273-281.

Joachims T. (1998) Text categorization with support vector machines: learning with many relevant features, *ECML 1998*, 137-142.

Kim W. and Wilbur W.J. (2005) A Strategy for Assigning New Concepts in the MEDLINE Database. *Proc AMIA Symp*., to appear.

Lin J., Abels E., Demner-Fushman D., Oard D.W., Wu P., Wu Y. (2005) A Menagerie of Tracks at Maryland: HARD, Enterpise, QA, and Genomics, Oh My! In *Proceedings of the Fourteen Text Retrieval Conference TREC 2005*. Gaithersburg, MD.

Ruch, P., Chichester, C., Cohen, G., Ehrler, F., Fabry, P., Marty, J., Muller, H. and Geissbuhler, A. (2004) "Report on the TREC 2004 Experiment: Genomics Track." *The Thirteenth Text Retrieval Conference, TREC-2004*, Gaithersburg, MD.

Ruch P., Ehrler F., Abdou S., Savoy J. (2005) Report on the TREC 2005 Experiment: Genomics Track. In *Proceedings of the Fourteen Text Retrieval Conference TREC 2005*. Gaithersburg, MD.

Ruiz, M.E. (2005) Experiments on Genomics ad hoc Retrieval. In *Proceedings of the Fourteen Text Retrieval Conference TREC 2005*. Gaithersburg, MD.

Salton, G (1971) *The SMART Retrieval System - Experiments in automatic Document Processing*. NJ: Prentice Hall.

Singhal A., Buckley C. and Mitra, M. (1996) Pivoted document length normalization. *SIGIR 1996*, 21-29.

Singhal A. (2001) Modern information retrieval: A brief overview. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 24(4):35--43, 2001.

Tanabe, L. and Wilbur, W.J. (2002) Tagging gene and protein names in biomedical text. *Bioinformatics*, Aug 2002; 18: 1124 – 1132.

Ting K.M. and Witten I.H. (1999) Issues in Stacked Generalization. *Journal of Artificial Intelligence Research*. 1999; 10:271-289.

Vapnik V. (1998) *Statistical Learning Theory.* John Wiley and Sons, NY.

Wiener E.D., Pedersen J.O. and Weigend A.S. (1995) A neural network approach to topic spotting, *SDAIR 1995*, 317-332.

Wilbur W.J. and Leona Coffee L. (1994) The effectiveness of document neighboring in search enhancement, *Information Processing & Management*, 30(2):253-266, 1994.

Wilbur, W.J. (2002) "A thematic analysis of the AIDS literature." *Pac Symp Biocomput*.: 386-97.

Yang Y. and Liu X. (1999) A re-examination of text categorization methods, *SIGIR 1999*, 42-49.