Proceedings of the SIGIR 2004 Workshop on Search and Discovery in Bioinformatics,
Sheffield, UK, July 2004.

Gene Terms and English Words: An Ambiguous Mix

Aditya Kumar Sehgal[+], Padmini Srinivasan[!*+], Olivier Bodenreider[@]

[+]Department of Computer Science
[!]School of Library and Information Science
[*]Department of Management Sciences
The University of Iowa
Iowa City, IA 52242

[@]National Library of Medicine
Bethesda, MD 20892

sehgal@cs.uiowa.edu, padmini-srinivasan@uiowa.edu, olivier@nlm.nih.gov

ABSTRACT

Continuing technical advances have made it possible for large-scale genetic analysis experiments where data for thousands of genes can be produced at a time. This has led to a burgeoning need for information on genes and the proteins they encode. The exceedingly large amount of biomedical information available today makes it very difficult for someone to completely follow the literature. This requirement is a major motivating factor in the development of automatic text processing techniques that enable easier and more efficient analysis of relevant information.

Recognizing gene terms in biomedical text is crucially important for applications in information retrieval and the extraction of higher level information. There are however many challenges associated with this task. One difficult aspect is negotiating the various kinds of ambiguity in gene and protein nomenclature. In this research we look at one of the most challenging kinds in which gene terms are also common English words. For example, TRAP, ART, ACT, are all gene symbols that also have English meanings. This kind of ambiguity makes retrieval of relevant information more difficult. We describe IR-based ranking methods applied to document sets retrieved for ambiguous gene terms in LocusLink and present our results. We find that using summary and product information from LocusLink records in addition to the gene term performs the best in terms of re-ranking the retrieved documents.

1. INTRODUCTION

Consider the term BAD, the official symbol for a human gene that encodes a member of the BCL-2 family of proteins, which regulate programmed cell death. The authors of a document that refers to this gene by BAD [PMID: 11878929[1]] are also likely to include the same term as a search term when looking for other papers on this gene. However, this term retrieves 14,815 MEDLINE documents when searched via the PubMed interface[2]. These include many documents retrieved because of the general English language meaning of BAD as for example in 'bad news' [PMID: 10085998] and 'bad prediction' [PMID: 10149445]. Unfortunately, using the gene symbol aliases for this term (BBC2 and BCL2L8) does not solve the ambiguity of the first term and its effect on retrieval. Deciding to drop this ambiguous term from the query, may be too risky as the consequent loss of relevant documents may be too great. In the case of BAD, there are 14 relevant documents[3] (out of 15) that have only the symbol BAD and none of its alias symbols. As another example, the gene symbol ACHE retrieves 45,961 documents. Again it is quite likely that a portion of the retrieved subset is on 'body aches' [PMID: 12118459, 11837753]. This challenge in gene term ambiguity is indeed well acknowledged in the literature [3, 8, 17, 20, 5]. Several approaches have been used to disambiguate the individual occurrences in documents of a string that has more than one meaning including one that points to a gene. For example, in [16] the authors use a combination of automatically generated and manually generated linguistic rules to identify gene and protein names in MEDLINE articles.

Most of the prior disambiguation research has targeted the decision: Is this observed instance of an ambiguous term referring to the gene or not? In contrast, we consider the problem of retrieval with these ambiguous terms more directly. That is, given a set of documents retrieved by an ambiguous gene string, we investigate strategies that may be used to filter through only those documents that are about that gene. Unfortunately, it is not sufficient to state that one may add a context indicating word such as 'gene' to the search syntax. Searching on 'BAD gene' for example, actually misses 10 relevant documents out of the 15 identified in the LocusLink database for that gene. Searching on 'ACHE gene' misses 16 out of 18 relevant documents. In general the ef-

SIGIR'04 Workshop on Search and Discovery in Bioinformatics, July 29, 2004, Sheffield, UK.

[1]PubMed Document Identifier
[2]http://www.ncbi.nlm.nih.gov/entrez/query.fcgi
[3]as identified in the LocusLink database.

fectiveness of adding qualifiers such as 'gene' from a *context* - indicating lexicon is bounded not only by lexicon coverage but also by the precision in meaning of these context terms. Such a lexicon must moreover, be appropriate for all ambiguous gene terms, unless one commits to building gene specific context lexicons.

The problem of retrieval and filtering strategies for ambiguous gene terms may appear to be a simple restatement of the problem of disambiguating occurrences of these same terms in documents. However, we suggest that there are key characteristics that differentiate the two. For example, we may tailor the retrieval and filtering strategies depending on the size of the retrieved set. If it is reasonable to assume that larger retrieved sets are more likely to include false positive documents, then we may apply more stringent filtering methods. This level of stringency may be relaxed with decreasing set sizes. Viewing the problem as a retrieval challenge also allows one to explore many standard information retrieval (IR) techniques such as the use of ranking queries to re-rank the retrieved set [14], pseudo relevance feedback, true relevance feedback, etc. These methods have been proven to be successful in many IR problem contexts. However, a systematic study of these IR methods for retrieval using ambiguous gene terms is absent. If successful these methods will offer a valuable alternative to the many linguistic [16, 19] and machine learning [2, 1, 3] instance-based disambiguation techniques observed in the literature.

In this paper we explore a variety of filtering methods that are applied to document sets retrieved by the ambiguous gene terms. Each filtering method is essentially a *ranking* strategy using a secondary ranking query to re-order the retrieved set. Our objective is to rank relevant documents higher than the non-relevant ones. We assess the quality of ranking using different IR measures. Secondary queries are automatically built from different resources. In most cases these are tailored to the individual gene while in some cases these are generic, i.e., the same for all genes. Each type of ranking query implies a different set of assumptions on the information available apriori about the gene.

2. RANKING METHODS

As mentioned before our approach is to take the set of documents retrieved by the ambiguous gene term and re-order them using a secondary query (henceforth called ranking query). Our goal is to rank the relevant documents higher than the non relevant ones. The key question here is: how are these ranking queries to be derived? Ideally, we would like an expert to describe how a typical relevant document for a particular gene is likely to be written. More practically we wish to know about the key words and phrases that are likely to appear in such a document. Since it is generally not possible to contact all the experts needed for the large number of genes, we can simulate this by looking at human gene annotations. In particular we can use manually curated and compiled resources such as LocusLink[4], HUGO[5], GenBank[6] to extract descriptions of genes and use these as ranking queries. The implied assumption is that some annotation is available for the gene. We also explore generic ranking strategies that do not make this assumption. These will be explained later.

Table 1 below shows the different queries for two example gene terms.

GAB	TRAP
Baseline1	
GAB	TRAP
Baseline2	
GAB gene genetics genome oncogene	TRAP gene genetics genome oncogene
Summary	
GAB The protein encoded by this gene is a plasma glycoprotein of unknown function. The protein shows sequence similarity to the variable regions of some immunoglobulin supergene family member proteins.	TRAP Acid phosphatase 5 is an iron containing glycoprotein which catalyzes the conversion of orthophosphoric monoester to alcohol and orthophosphate. ACP5 is the most basic of the acid phosphatases and is the only form not inhibited by L(+)-tartrate.
Summary+Product	
GAB alpha 1B-glycoprotein The protein encoded by this gene is a plasma glycoprotein of unknown function. The protein shows sequence similarity to the variable regions of some immunoglobulin supergene family member proteins.	TRAP tartrate resistant acid phosphatase 5 precursor Acid phosphatase 5 is an iron containing glycoprotein which catalyzes the conversion of orthophosphoric monoester to alcohol and orthophosphate. ACP5 is the most basic of the acid phosphatases and is the only form not inhibited by L(+)-tartrate.
Summary+Product+Name	
alpha 1B-glycoprotein A1BG A1B ABG GAB alpha-1-B glycoprotein The protein encoded by this gene is a plasma glycoprotein of unknown function. The protein shows sequence similarity to the variable regions of some immunoglobulin supergene family member proteins.	tartrate resistant acid phosphatase 5 precursor alpha-1-B glycoprotein serine (or cysteine) proteinase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 3 alanyl-tRNA synthetase ABO blood group (transferase A, alpha 1-3-N-acetylgalactosaminyltransferase; transferase B, alpha 1-3-galactosyltransferase) acyl-Coenzyme A dehydrogenase, C-2 to C-3 short chain acetyl-Coenzyme A acetyltransferase 1 (acetoacetyl Coenzyme A thiolase) acetylcholinesterase (YT blood group) ACP5 TRAP acid phosphatase 5, tartrate resistant Acid phosphatase 5 is an iron containing glycoprotein which catalyzes the conversion of orthophosphoric monoester to alcohol and orthophosphate. ACP5 is the most basic of the acid phosphatases and is the only form not inhibited by L(+)-tartrate.
Product	
GAB alpha 1B-glycoprotein	TRAP tartrate resistant acid phosphatase 5 precursor

Table 1: Example Ranking Queries

In this research we use only the LocusLink database as a source for gene descriptions. Each LocusLink record contains a wide variety of information. We select particular fields to explore different types of ranking queries. We avoid using fields such as GO (the Gene Ontology annotation field)[7] since these often directly point to PubMed documents and may seem unfair. We also do not use the GRIF (Gene Reference Into Function) field, again because these include direct pointers into PubMed. Also the component sentences

[4]http://www.ncbi.nlm.nih.gov/LocusLink/
[5]http://www.gene.ucl.ac.uk/hugo/
[6]http://www.ncbi.nlm.nih.gov/Genbank/
[7]http://www.geneontology.org/

may actually come directly from the documents [4], which is likely to make the retrieval task simple. Instead, we use the GO and GRIF fields as a source for gold standard relevant documents and use them to evaluate our methods as explained later. We use 6 ranking queries all of which include the ambiguous gene term; these are:

1. Baseline1: Here the ambiguous gene term alone is used to re-rank the documents.
2. Baseline2: A simple query, gene term + 'gene, genetics, oncogene, genome', is used to re-rank the documents.
3. Summary: This a multi-sentence field that describes for example, the gene's function, its structure, the protein it produces, and details regarding its expression as well as associated phenotype information. This information is obtained from the SUMMARY field in a LocusLink record. Note that this summary field is generated using data from the Reference Sequence(RefSeq) collection and other databases such as OMIM, Proteome, and Protein Reviews on the Web(PROW)[7].
4. Product: This is the name of the protein encoded by the gene. This information is available under the PRODUCT, PREFERRED_PRODUCT, and ALIAS_PROT fields in a LocusLink record.
5. Summary+Product: This is a combination of the previous two queries.
6. Summary+Product+Name: This is a combination of the summary, product, gene name (OFFICIAL_GENE_NAME), official symbol (OFFICIAL_ SYMBOL), and alias symbols (ALIAS_SYMBOL) for the gene.

3. EXPERIMENTAL DETAILS

3.1 Datasets

Since this is exploratory work we use only the LocusLink database as a source for gene terms. Moreover, we limit our attention to the human gene terms, from the OFFICIAL_GENE_NAME, OFFICIAL_SYMBOL, ALIAS_SYMBOL, PREFERRED_PRODUCT, and ALIAS_PROT fields, listed in LocusLink. A preliminary study of these gene terms indicate that there are different kinds of ambiguities. Often a single gene is represented by multiple gene terms. This phenomenon is known as *synonymy* and is the most common source of ambiguity in gene terms. For example, the gene AGRP is also known by its aliases ART, AGRP, and ASIP2. Sometimes a single gene term can refer to multiple genes. This condition is known *homonymy* and is less common but more difficult to resolve than synonymy. For example, AFP is the official symbol of a gene that encodes alpha-fetoprotein and it is also an alias symbol of another gene that encodes a protein member of the tripartite motif (TRIM) family. As mentioned before, many gene terms are also ambiguous because they have a meaning in the English language. Weeber et al.[17] consider the first two kinds of ambiguities in LocusLink and use an abbreviation expansion based approach to resolve them and create a thesaurus of disambiguated gene terms. In [9] Morgan et al. describe the affect of the third kind of ambiguity on the performance of their system. They mention that the precision of their system goes down mainly due to the presence of common English words such as *if*, *to*, etc., as positive instances (of gene symbols) in their training set. To get around this problem they plan to filter out those gene symbols that are also common English words, from their data.

It may be that each variety of ambiguity requires a different strategy for document retrieval and filtering. Thus we begin by focusing on one type of ambiguity where, in addition to representing a gene, the string also represents a general English language concept. Examples include TRAP, GRAIL, RAGE, APEX and MAT. In our pool there are 1,051 gene terms that also have an English meaning as determined by a look up in the WordNet database[8]. However, 28 of these had to be eliminated for various reasons. Some, for example, were eliminated because they retrieved 0 PubMed documents (eg. the gene symbol IN) others because there were no documents in LocusLink corresponding to these genes (eg. COP). (As explained later, we use the documents associated with the gene entry in LocusLink as defining the gold standard set of documents to retrieve). All experiments reported here are based on the remaining 1,023 genes. It should be mentioned that over 99% of the gene terms in our pool are single words. Given the low percentage of multi-word gene terms (less than 1%), we do not treat them differently in this research. Consequently each word in a multi-word gene term is considered independently in constructing the ranking queries. Further research on ambiguous multi-word gene terms will be done in the future.

It is important to observe that the unique set of documents retrieved jointly by the starting set of 1,051 strings is close to 3 million! This fact underlines the importance of having effective filtering strategies to funnel through the retrieved documents retaining just those that are relevant. Ranking the relevant documents above the non relevant ones will certainly be of immense help to users.

3.2 Evaluation Strategies

3.2.1 Gold Standards

The re-ordered documents are assessed using a gold standard collection of relevant documents for each gene that is also identified from the LocusLink database. In particular the PMID, GRIF fields identify MEDLINE documents. The GO field may also contain pointers to MEDLINE documents. Documents in the GO field are identified by curators of annotation databases for the various model organisms. These provide supporting evidence for a GO annotation applied to a gene. Documents in the GRIF field are identified by MEDLINE indexers. These documents contain information about the function of the gene. Since these gene to document connections are made by trained individuals, we are confident that these documents are 'relevant' to the gene.

Unfortunately the documents as identified above may not be the *complete* set of relevant documents for the gene. In fact because of the time it takes for human annotations, it is quite possible that relevant documents are missing from these PMID-GO-GRIF sets. Thus we a generate a second, expanded gold standard set. In particular, we apply a neighborhood method to expand the pool of relevant documents for each gene. This method was designed and tested for MEDLINE by Wilbur et. al [18] and is the basis for the 'Related Articles' function that is available in PubMed. For

[8]http://www.cogsci.princeton.edu/~wn/

each document in the PMID-GO-GRIF set, which we now refer to as a *seed* document, we get its neighbors using the ELink tool[9] provided by PubMed. In order to maintain confidence in the expanded pool of neighboring documents, we apply a majority rule. Specifically a document D is accepted as a relevant document for a gene G, if D is retrieved as a neighbor for the majority of the seed documents of G. Thus if G has 5 seed documents identified from PMID-GO-GRIF, then at least 3 seeds should have D as a neighbor before D is considered a relevant document. In some cases, the majority rule results in 0 documents selected. In such cases the majority rule is relaxed to 33%, 25%, and then 20% till a minimum number of relevant documents are added. For our collection of 1,023 genes, the majority rule was used successfully for 896 genes (87.5%). In each case a minimum criteria of adding 5 documents was satisfied. By relaxing it to a 33% rule, a 25% rule, and a 20% rule an additional 108 (10.5%), 10 (1%), and 9 (1%) genes respectively satisfied the criteria.

In all experiments that follow, we present results using both gold standard sets.

3.2.2 Measures

A standard measure for evaluating ranking effectiveness is AP[4] or average precision. Given a ranked set of documents for a particular gene, we first calculate precision at each position where a relevant document appears. The average of these precision scores is the average precision. This serves a user who is interested in retrieving all relevant documents.

Since AP is sensitive to the rank of each relevant document including the last ranked one, it is not appropriate for a user who is satisfied with just a few relevant documents. Thus we also compute precision in the top 10 ranked documents (Top10P).

3.3 Ranking System

We use the SMART system[11] to rank each retrieved set. In essence each retrieved document is indexed using SMART's *atc* weighting scheme after removing stopwords and stemming the rest. This means that each stem is weighted using the augmented term frequency and inverse document frequency. Also the vector of weights are normalized for length. Weights were calculated within the context of a set of retrieved documents. The same weighting scheme is used for each ranking query. SMART compares each document with the query and computes a similarity score in [0,1]. The documents ranked by similarity score are evaluated using AP and Top10P.

4. RESULTS

Figures 1 and 2 summarize the relative performance of the different strategies in AP score for the two gold standard relevance sets. Each figure shows scores of the four ranking strategies and for the 2 baseline strategies. For each strategy, the X axis represents the genes sorted by AP score. Thus for example, give the top ranked 200 genes for the Baseline1 strategy the minimum AP score is 0.1. Whereas, given the top ranked 200 genes for the Summary+Product strategy the minimum AP score is above 0.9. In fact for all ranking strategies, except for Baseline2, the minimum AP score for the top 200 is at least 0.625. These represent at a minimum an over 5-fold increase in performance from Baseline1 and over 1-fold increase from Baseline2. For the top 600 (i.e., more than 1/2 the genes in our pool) genes the minimum increases are almost 20-fold and over 3-fold from Baseline1 and Baseline2 respectively.

The best ranking strategy for both gold standards is Summary+Product. This strategy achieves the best AP score of 1 for 142 (14%) genes. It is at least 0.5 for 59% of the genes. We observe that a ranking query on Product alone is quite effective as well. Performances are generally better using the expanded gold standard than the basic gold standard. What is important is that the relative ordering of strategies stays the same. Interestingly adding Names to Summary+Product decreases performance considerably. This could be due to ambiguity in the added Names since this includes other symbols as well as the name(s) for the gene.

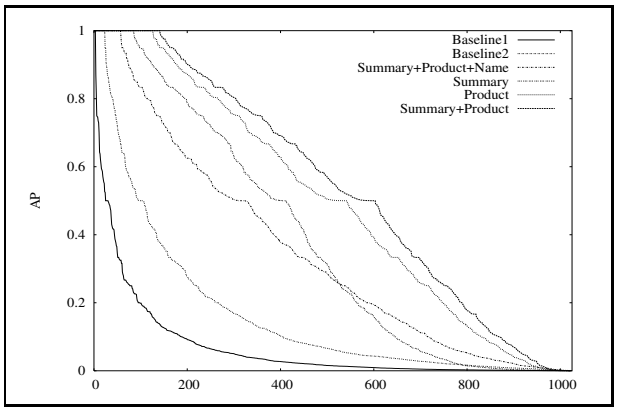

Figure 1: AP: Average Precision Scores (Gold Standard: Basic)

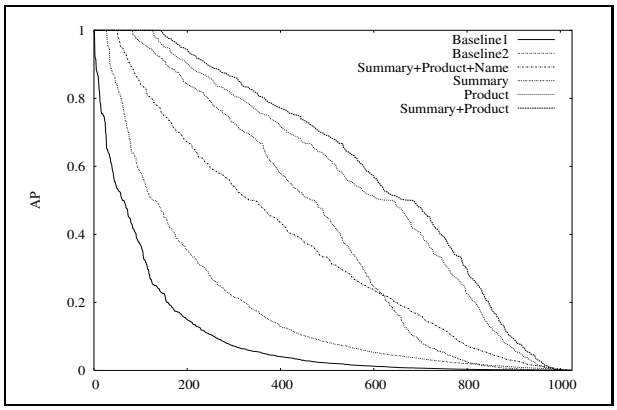

Figure 2: AP: Average Precision Scores (Gold Standard: Expanded)

Figures 3 and 4 summarize relative performance in Top10P. We see that Baseline1 and Baseline2 have 938 (92%) and 760 (74%) genes respectively in the 0.1 Top10P group. Thus for at least 74% of the genes only 1 of the top 10 ranked documents is relevant. In contrast, for the best ranking strategy (Summary+Product), only 351 genes (34%) are in the 0.1 group while 259 (25%) achieve at least 0.5 Top10P. Product alone is also a good strategy with at least 236 (23%)

[9]http://eutils.ncbi.nlm.nih.gov/entrez/query/static/elink_help.html

achieving at least 0.5 Top10P. The relative effectiveness of the 4 ranking queries stays generally the same for both gold standard sets. Again Baseline2 appears much better than Baseline1. For example, there are at least 150 fewer genes in the 0.1 Top10P bin for Baseline1. Given that Baseline2 is much better than Baseline1 in both AP and Top10P scores, henceforth we drop Baseline1 from our analysis.

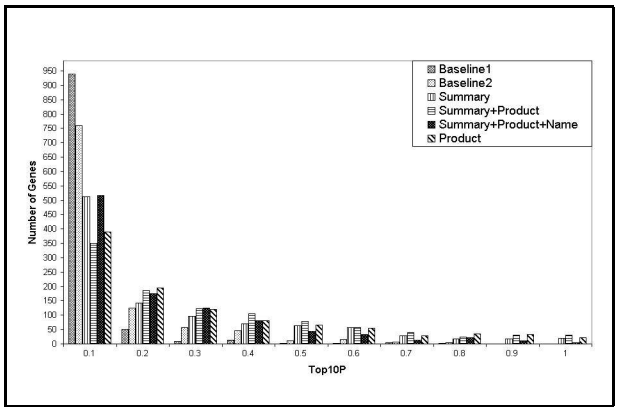

Figure 3: Top10P: Precision at Top 10 Rank (Gold Standard: Basic)

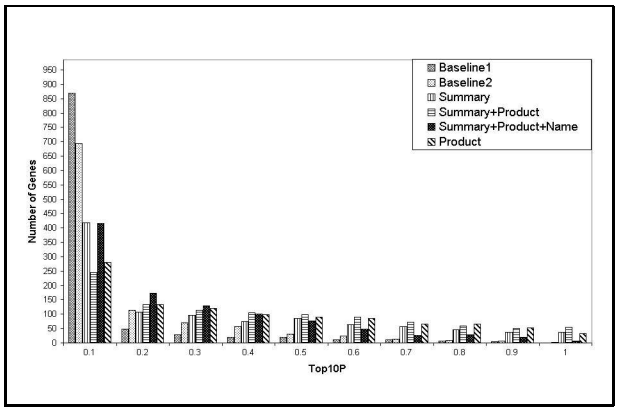

Figure 4: Top10P: Precision at Top 10 Rank (Gold Standard: Expanded)

Figures 5 and 6 compare performance at the specific gene level. The X axis (at Y = 0) represents the baseline AP performance for each gene. The genes are organized in bins ordered by ascending AP scores. For example, the first bin (closest to the origin) contains all genes whose Baseline2 AP is in the [0.0, 0.1] range. The second bin has genes with Baseline2 AP scores in the (0.1, 0.2] range. The row of the numbers along the X axis (in square brackets) shows the number of genes in each bin. The last row of numbers (in parenthesis) is the average number of documents retrieved in each bin when their gene terms are used to search PubMed. Thus the majority of the gene terms (618, 60%) fall into the first bin in AP score and retrieve on average 23,485 documents. This highlights the ambiguity problem related to these gene terms. Given a particular ranking strategy and a specific bin, we calculate the mean difference in AP score between the strategy and the baseline for genes of that bin. This is plotted as a bar on the graph. Therefore, bars below the Y = 0 line indicate performance that is worse than the baseline.

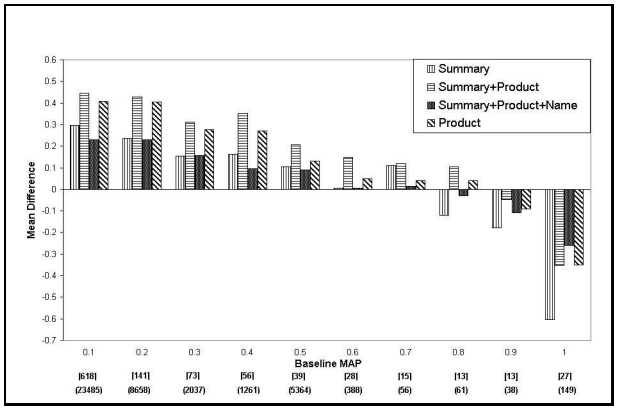

Figure 5: Difference in AP score between Ranking Queries and Baseline2 (Gold Standard: Basic)

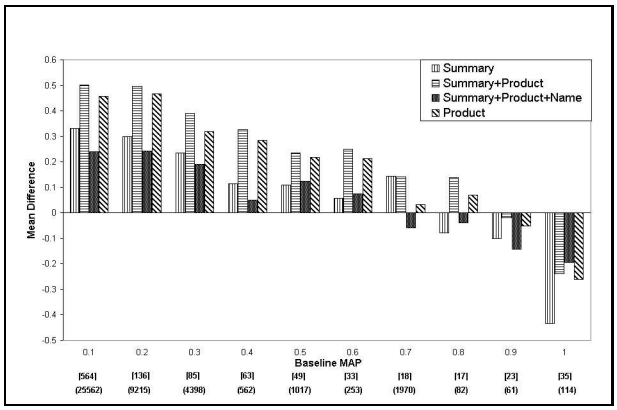

Figure 6: Difference in AP score between Ranking Queries and Baseline2 (Gold Standard: Expanded)

Looking at figure 5 we see that Summary+Product is the best strategy producing a 0.3 to 0.45 increase in AP for the first 4 bins. The Product strategy is second best with score increases from 0.3 to 0.4 for the same bins. Note that these bins contain 87% (888) of the genes. For the highest 3 bins (0.7 and higher) it becomes difficult to improve performance using any of the strategies. Note that there are only 53 (5%) genes that fall into these bins. In these cases the best strategy would be to do nothing. The challenge is in being able to predict these cases. The data suggest that it may be fruitful to explore heuristics based on the number of retrieved documents. These will be studied in future research.

When examining the plot for the extended gold standard set we observe that the order of strategies does not change. The improvements are higher with for example, an increase of 0.34 to 0.5 for the Summary+Product strategy in the first 4 bins.

Analyzing differences in Top10P scores compared to Baseline2 more clearly identifies a winner in the Summary+Product strategy. In Figures 7 and 8, displaying these differences, this strategy is the best for both gold standard sets and performance is hurt only on the 10 genes in the highest 2 bins using the expanded gold standard.

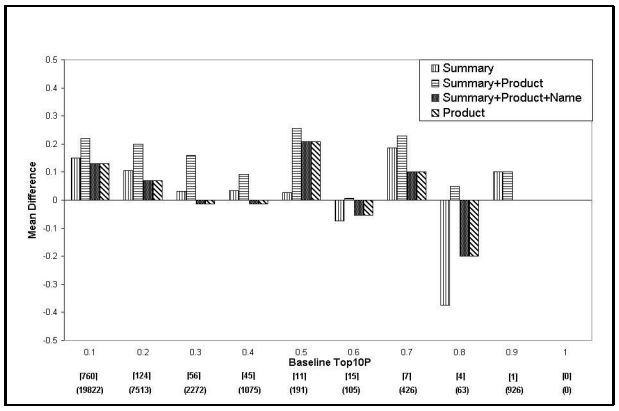

Figure 7: Difference in Top10P score between Ranking Queries and Baseline2 (Gold Standard: Basic)

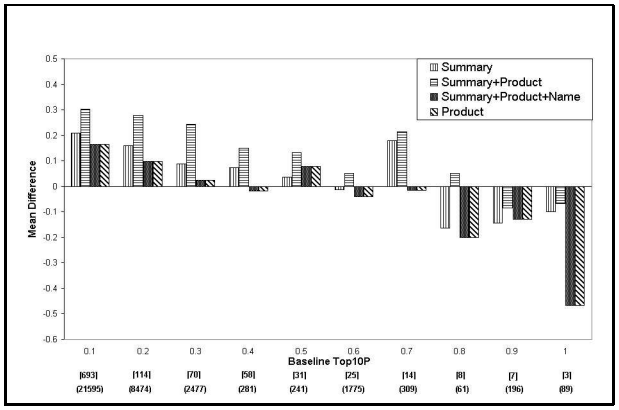

Figure 8: Difference in Top10P score between Ranking Queries and Baseline2 (Gold Standard: Expanded)

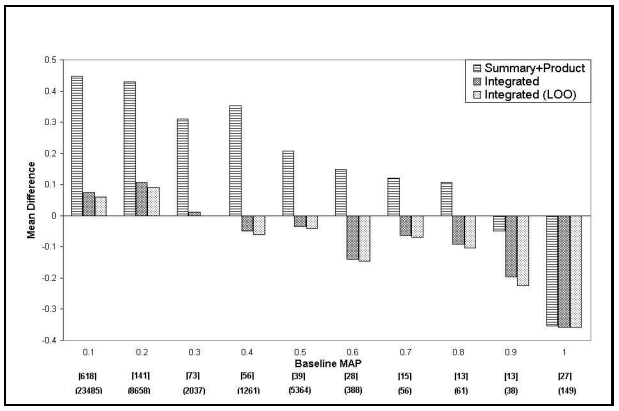

Figure 9: Difference in AP score, Summary+Product, LOO and Integrated Ranking Strategies (Gold Standard: Basic)

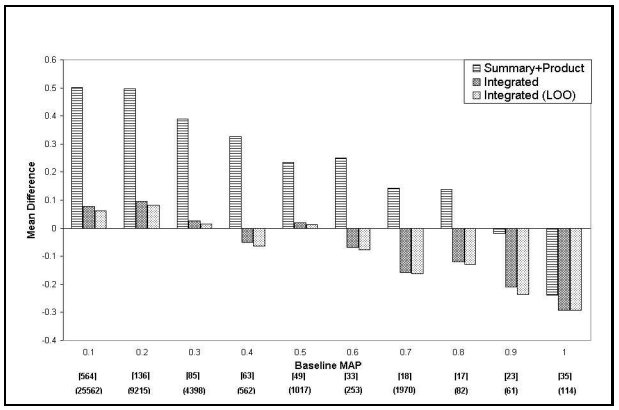

Figure 10: Difference in AP score, Summary+Product, LOO and Integrated Ranking Strategies (Gold Standard: Expanded)

All strategies thus far have assumed that the gene term being searched has an entry in LocusLink. Stated differently, these are 'known' gene terms with summaries, product names, etc. provided. The question asked is: what can we do if we do not have these descriptions - for instance, when the gene term is quite new? We tried a strategy where we combined the summary and product fields for all gene terms except the one being queried (which we refer to as G). We then used this combined query to re-rank the documents for G. We did this in turn for each gene in the pool. Figures 9 and 10 show these results. As expected, this strategy, denoted LOO for leave one gene out, fared poorly when compared to the gene specific Summary+Product strategy. However, it should be noted that for the first two bins containing 74% of the genes for the basic and 68% for the expanded sets, AP increased by 0.05 to 0.1. This represents for example, a 50% improvement for the second bin using the basic gold standard set. Thus even these crude strategies can be useful with genes that are most ambiguous.

Finally we tried a strategy which simply measured the value of combining all summaries and products without discarding any genes' information. Results for this strategy, known as Integrated, are also shown in Figures 9 and 10. Although the performance is slightly better than LOO for some bins, overall this strategy falls short of the one which is tailored to the specific gene.

5. RELATED RESEARCH

Recognizing gene terms in biomedical text is important for applications in information filtering, retrieval and extraction of higher level information such as functional relationships between genes [13, 15], interaction between genes and gene products [12], etc. The task is made difficult because different nomenclature schemes are followed for different organisms. This difficulty is further compounded by the presence of various kinds of ambiguities in gene terms, such as *synonymy*, *homonymy*, and English words.

Different kinds of methods, viz. machine learning based, statistical, linguistic, and rule-based, have been proposed for automatically identifying gene terms in text. Hatzivassiloglou et al. [3] used a machine learning based approach to identify gene and protein terms in text and were able to obtain accuracy rates upto 85%, using over 9 million words to train and test their system. Morgan et al. [8] describe a hidden markov model based approach that is trained on automatically generated data using existing resources available for Drosophilla (Flybase). Evaluating on 86 abstracts they obtain an F-score of 0.75. In [19] Yu et al. describe an

algorithm to identify gene-related pairs of abbreviations and long forms and use these to identify gene symbols and names in MEDLINE text. Using 50 MEDLINE documents to test their system, they obtained recall and precision values of 0.73 and 0.93 respectively. Linguistic approaches mainly based on part-of-speech tagging have also been applied in this problem domain. The TREC genomics track[10] is a task-based competition organized every year with problems in the genomics domain. The primary task of the 2003 TREC genomics track was to retrieve, for each gene in a given set of genes, those documents that focused on specific properties of the gene, such as its function, etc. The best performance (measured by AP) was achieved by Kayaalp et al. [6] from a research group at NLM (National Library of Medicine). They achieved an average AP score of 0.4165. It may be observed that our goal, which is focused entirely on a specific kind of ambiguity, while related to the 2003 TREC genomics goal is also quite distinct.

6. CONCLUSIONS

In summary, we see that it is possible to derive ranking queries from a resource such as LocusLink to refine the documents retrieved by an ambiguous gene term. The best ranking query, across both AP and Top10P measures and for both gold standard sets, is one that is derived from the summary and product fields of the LocusLink record. Interestingly, a close contender is a query derived from the product fields alone. Baseline2 performance suggests that at a minimum a query composed of the gene term combined with 'genetics, oncogene, genome, gene' may be used to rank the documents. Compared with this baseline our best ranking queries yield excellent returns especially for the most difficult subset of genes. For the first bin of genes in Figures 6 and 7, the average increase in AP scores is 0.455 to 0.5. For the same bins in Figures 7 and 8, Top10P scores increase 0.23 to 0.3 units. Performance for the high score bins is sometimes worse than the baseline. As mentioned before, it may be that heuristics based on the number of retrieved documents will allow us to identify conditions under which we should not apply a ranking strategy. We also find that a naive ranking query which assumes that no information is available for a particular gene also succeeds in improving AP score by at least 50% when the baseline score is 0.2 or less. This result suggests that we may be able to improve search results even for new genes or when searching newly coined names for existing genes.

In future research we plan to use these ranking strategies to create positive and negative examples of documents for each gene. These may then be used to train classifiers that might, hopefully, further improve upon the effectiveness of the ranking strategies explored in this research. This will also lead us quite naturally to the next step which is filtering, where we will make a retrieval decision on each ranked document.

7. ACKNOWLEDGMENTS

Padmini Srinivasan acknowledges NSF Grant No. IIS-0312356, which partly funded this research.

8. REFERENCES

[10]http://medir.ohsu.edu/~genomics/

[1] Bickel, S., Brefeld, U., Faulstich, L., Hakenberg, J., Leser, U., Plake, C. and Scheffer, T. A Support Vector Machine Classifier for Gene Name Recognition. EMBO Workshop: A critical assessment of text mining methods in molecular biology, Granada, Spain, March 2004.

[2] Collier, N.H., Nobata, C. and Tshjii, J. Extracting the names of genes and gene products with a hidden markov model. In Proceedings of the 18th international Conference on Computational Linguistics (COLING'2000), pg. 201-7, 2000.

[3] Hatzivassiloglou, V., Pablo A. and Rzhetsky, D. Disambiguating proteins, genes, and RNA in text: A machine learning approach. ISMB (Supplement of Bioinformatics): pg. 97-106, 2001.

[4] Hersh, W. and Bhupatriraju, R.T. TREC Genomics Track Overview. In Notebook of the TREC-2003, pg. 148-157, 2003.

[5] Koike, A. and Takagi, T. Gene/Protein/Family Name Recognition in Biomedical Literature. HLT Biolink 2004.

[6] Kayaalp, M., Aronson, A.R., et al. Methods for accurate retrieval of MEDLINE citations in functional genomics. The Twelfth Text REtrieval Conference: TREC 2003, Gaithersburg, MD. National Institute for Standards & Technology, 2003.

[7] Maglott, D. LocusLink: A Directory of Genes. The NCBI Handbook (Part 3: Chapter 19), 2003.

[8] Morgan, A., Hirschman, L., Yeh, A. and Colosimo, M. Gene Name Extraction Using FlyBase Resources. In Proceedings of ACL 2003 Workshop on Natural Language Processing in Biomedicine, Sapporo, Japan, 2003.

[9] Morgan, A., Hirschman, L. and Yeh, A. Gene Name Finding Using FlyBase. Poster at The Pacific Symposium for Biocomputing, Kaua'i, 2003.

[10] Nobata C., Collier N.H. and Tsujii J. Automatic term identification and classification in biology texts. In Proceedings of the Natural Language Pacific Rim Symposium (NLPRS'99), 1999.

[11] Salton, G., ed. The Smart Retrieval System-Experiments in Automatic Document Processing. Prentice Hall Inc. Englewood Cliffs NJ, 1971.

[12] Sekimizu T., Park H.S. and Tsujii J. Identifying the Interaction between Genes and gene Products Based Frequently Seen verbs in Medline Abstracts. Genome Informatics, 9:62-71, 1998.

[13] Shatkay, H., Edwards, S., Wilbur, W.J. and Boguski, M. Genes, Themes and Microarrays Using Information Retrieval for Large-Scale Gene Analysis. In Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology, pg. 317-328, 2000.

[14] Singhal, A., Mitra, M. and Buckley, C. Learning

Routing Queries in a Query Zone. ACM SIGIR'97, pg. 25-32, 1997.

[15] Stephens M., Palakal M., Mukhopadhyay S. and Raje R. Detecting Gene Relations from MEDLINE Abstracts. Pacific Symposium on Biocomputing 2001, pg. 483-95, 2001.

[16] Tanabe, L. and Wilbur W.J. Tagging gene and protein names in full text articles. ACL Workshop on Natural Language Processing in the Biomedical Domain: 9-13, 2002.

[17] Weeber, M., Schijvenaars, R.J.A., van Mulligen, E.M., Mons, B., Jelier, R., van der Eijk, C.C. and Kors, J.A. Ambiguity of Human Gene Symbols in LocusLink and MEDLINE. Creating an Inventory and a Disambiguation Test Collection. In Proceedings of AMIA Symposium: pg. 704-708, 2003.

[18] Wilbur, W.J. and Coffee, L. The Effectiveness of Document Neighboring in Search Enhancement. Information Processing & Management. 30(2):253-266, 1994.

[19] Yu, H., Hatzivassiloglou, V., Rzhetsky, A. and Wilbur, W.J. Automatically identifying gene/protein terms in MEDLINE abstracts. Journal of Biomedical Informatics, 35(5-6):322-30, 2002.

[20] Yu, H. and Agichtein, E. Extracting Synonymous Gene and Protein Terms from Biological Literature, Journal of Bioinformatics Vol 19(Suppl): pg. i340-i349, 2003.