

# AZuRE, a Scalable System for Automated Term Disambiguation of Gene and Protein Names

Raf M. Podowski  
AstraZeneca R&D Boston  
and Karolinska Institutet  
rpodowski@cmmt.ubc.ca

John G. Cleary  
Reel Two, Ltd. and  
University of Waikato  
jcleary@reeltwo.com

Nicholas T. Goncharoff  
Reel Two, Inc.  
nicko@reeltwo.com

Gregory Amoutzias  
AstraZeneca  
gregory.amoutzias@astrazeneca.com

William S. Hayes  
AstraZeneca R&D Boston  
william.s.hayes@astrazeneca.com

## Abstract

*Researchers, hindered by a lack of standard gene and protein-naming conventions, endure long, sometimes fruitless, literature searches. A system is described which is able to automatically assign gene names to their LocusLink ID (LLID) in previously unseen MEDLINE abstracts. The system is based on supervised learning and builds a model for each LLID. The training sets for all LLIDs are extracted automatically from MEDLINE references in the LocusLink and SwissProt databases. A validation was done of the performance for all 20,546 human genes with LLIDs. Of these, 7,344 produced good quality models ( $F$ -measure  $> 0.7$ , nearly 60% of which were  $> 0.9$ ) and 13,202 did not, mainly due to insufficient numbers of known document references. A hand validation of MEDLINE documents for a set of 66 genes agreed well with the system's internal accuracy assessment. It is concluded that it is possible to achieve high quality gene disambiguation using scaleable automated techniques.*

## 1. Introduction

Biological researchers are constantly hindered in their work by a lack of standard naming conventions for genes and proteins. Near-frivolous choices of gene synonyms result in gene names like "IT" "midget", or "ER". These inherently ambiguous names cannot be effectively filtered by current search tools, nearly all of which are based on keyword queries. As a result, researchers must endure long, and sometimes fruitless, searches for literature about genes or proteins. Automated disambiguation of gene

and protein names could significantly help improve access to biological literature and increase the efficiency of text analytics in the biomedical domain.

We present a system, called AZuRE, for performing automated term disambiguation that can easily scale to tens of thousands of unique gene and protein names. AZuRE uses a combination of machine learning and natural language processing technologies to identify abstracts relevant to specific genes and return these results as a ranked list.

Over 20,000 human genes have been identified in LocusLink and over 100,000 different names have been used to refer to them. A gene disambiguation system that is truly useful to a wide range of researchers must address some key, heretofore unsolved, challenges:

- It must scale to cover tens of thousands of genes and proteins per organism;
- It must be able to automatically generate training data with minimal human intervention;
- It must be able to make use of varying quantities of training data;
- It must be able to make use of low-quality training data that hasn't been annotated or enhanced with meta-data;
- It mustn't rely on a comprehensive list of all possible gene and protein synonyms, since updating such a list is impractical.

AZuRE addresses these gaps using a supervised learning system. This article presents test results that show AZuRE is capable of accurately distinguishing between highly ambiguous gene terms, as well as between synonymous gene and non-gene terms.

## 2. Background

### 2.1. Problem overview

The absence of an automated approach for resolving ambiguity between gene synonyms is a key problem [6,7]. Further, text analytics in the biomedical domain are dependent upon good gene name tagging and disambiguation. Natural Language Processing in particular is dependent upon term disambiguation, which has been called the “great open problem” of natural language lexical analysis [13]. In the biomedical domain, gene and protein name disambiguation is essential for providing quality protein-protein interactions, disease associations, and other complex biomedical analysis. This problem can also have a substantial impact on the efficiency of information retrieval methods, such as biomedical thesauri [2] or molecular pathway identification [4].

Disambiguation tasks fall into two basic categories: determining if a term refers to a gene or gene product (does “PI” refer to “glutathione transferase” or “Permeability Index”); and identifying the true meaning of a synonymous gene name or abbreviation (does “PI” refer to “glutathione transferase” or “alpha-1-antitrypsin”). Both of these problems often elude keyword searches.

### 2.2. Previous work

Natural language researchers began focusing on automated approaches to term disambiguation in the late 1980s and early 1990s. Yarowsky [16] used statistical models built from entries in Roget’s Thesaurus to assign sense to ambiguous words in text, using a Bayesian model to weight the importance of words related to the targeted ambiguous term. Gale, Church and Yarowsky [3] outlined an approach that used the 50 words preceding and following the target term to define a context for that term’s sense. In developing a method for general word sense disambiguation using unsupervised learning, Yarowsky [17] took a document classification approach to solving the problem of general term disambiguation. He also showed in this study that generic English language terms often have only one sense per co-location with neighboring words.

Around the year 2000, computational linguists and computational biologists began to look at term disambiguation in the biomedical domain. A number of researchers [1,2,14] have proposed solutions that involve manually crafted rules to help natural language processing and information retrieval

systems correctly process ambiguous synonyms. These rules are often combined with supervised learning methods (in which systems are provided with human-curated training data) and in some cases unsupervised learning methods (also often referred to as “clustering”). Recent work by Yu and Agichtein [18] compared four different approaches to solving the disambiguation problem – manual rules, fully supervised learning, partially supervised learning and unsupervised. The manual method is then combined with several of the machine learning approaches to yield a system capable of extracting synonymous genes and proteins from biomedical literature. Liu et al. [6] also explore a partially supervised learning approach based on disambiguation rules defined in the Unified Medical Language System. In the case of both papers, results are promising, but the systems require a pre-existing set of hand-crafted corpora, raising questions about scaling up to a level where a significant portion of human genes and proteins can be covered. Hatzivassiloglou, Duboue and Rzhetsky[4] apply machine learning to the problem of gene, protein and RNA in text, showing that accuracy levels, as defined by F-measure, of nearly 85% can be attained for classifying terms as belonging to the class of gene or protein. Note, however, that the problem they have tackled is simpler than the one reported here, which seeks to identify the specific gene referred to.

## 3. Methods

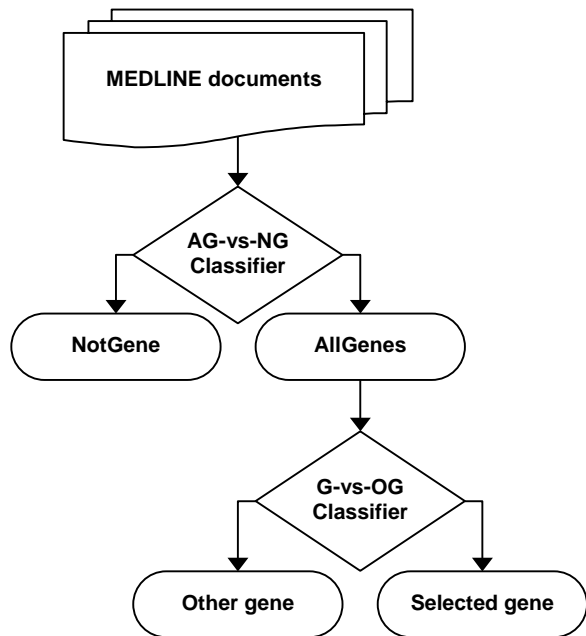
### 3.1. Data collection

Individual genes included in AZuRE are defined by the LocusLink (LL) [5,11] human gene set. Gene names, symbols and synonyms were collected from LL and SwissProt (SP) [10] databases. The system is designed to query and recognize gene or gene product context in MEDLINE abstracts.

### 3.2. Disambiguation Strategy

The disambiguation process, shown in Figure 1, outlines the steps followed to decide if a document contains a reference to a particular LLID. To begin, a gene and the corpus of documents to be searched are selected (e.g. MEDLINE). The first step of the disambiguation process begins with the documents being classified by a model that decides if a document genuinely refers to genes or gene products. This is called the “AllGenes vs NotGene” (“AG-vs-NG”) model. If the document is classified as a NotGene, it is rejected and not considered further. The next step is a classification of each retained

document with a model that is specific to the selected LLID. This is called a “Gene vs OtherGene” (“G-vs-OG”) model. If the model classifies the document as “Gene” then it is accepted as having a reference to the LLID.



**Figure 1. Disambiguation Scheme**

Section 3.4 describes the process of building the “AG-vs-NG” model and the 20,546 “G-vs-OG” models. Section 3.5 describes how the accuracy of these models was validated.

### 3.3. Machine Learning Classification System

The classification models were all built using Reel Two’s proprietary Classification System [12]. This uses a modified form of the Naïve Bayes algorithm, the Weighted Confidence Learner (WCL) [9] to generate a score for each document, as well as a threshold that determines whether the document is counted as being in the category or out of it. This threshold can be modified to vary the precision and recall. Normally, it is set at the break-even point where the precision, recall and F-measure are all equal. The Classification System also has the ability to do Leave-One-Out (LOO) validation on its training sets. This gives an accurate measure of how well the system can be expected to perform on unseen data and is used later for validation.

No manual tuning or setting of parameters was done for the individual models. The WCL algorithm uses all the words in an abstract as the context for a decision. Because the underlying algorithm is

sufficiently robust, performance is not significantly improved by altering the feature sets it generates. For example, there is no pruning of the word set, nor is a context window around target terms used. The models were adapted for MEDLINE by using a specialized tokenizer that dealt well with the complex biomedical terms in MEDLINE abstracts. Some tuning was also done to account for the fact that some abstracts are very short, often consisting of nothing more than the title of the paper.

### 3.4. “AG-vs-NG” model

The “AG-vs-NG” classifier is an initial screen designed to eliminate documents that are clearly not about genes. For example, this filter should eliminate documents where the symbol “AR” refers to the gas argon, autoregressive model, acrosome reaction, etc., and retain any referring to the genes androgen receptor, aldose reductase, amphiregulin or any other gene-related content.

**3.4.1. Training set selection.** A pool of training documents for the AG-category was obtained by searching MEDLINE for articles containing one or more terms suggestive of gene or gene product context. Using a query composed of the terms “gene”, “genes”, “cDNA” and “mRNA”, we obtained 672,675 documents containing one or more of the terms in the title or abstract.

The NG-category training set document pool comprised MEDLINE documents that had at least 500 characters of text and did not contain terms from a stoplist. This stoplist included terms such as: “gene”, “protein”, “cDNA”, “mRNA”, “kinase”, “receptor”, “amino acid”, “encode”, “subunit”, “express”, “pathway”, “repress”, “inhibit”, “transcript”, “oncogene” and “oncoprotein” as well as plurals and other variants. This gave a final pool of 4.5 million documents. The AG and NG category document pools were then used to obtain random subsets for the final classifier training sets (see Sections 3.4.3 and Section 3.5).

**3.4.2. Training set bias.** When selecting training documents, a bias will be introduced if the proportions of positive and negative examples in the categories do not represent the “real world” data distribution. The result of this bias is that suboptimal decisions are made about which categories the document belongs to. The classification system chooses category membership at the “break-even point” where the number of False Positives (FP) is predicted to equal the number of False Negatives (FN).

There are two possible remedies for this problem. Both require an estimation of the correct “real-world” document ratios. The first method involves changing the prediction threshold. This threshold will differ from the default break-even point because of the training bias. An alternative and preferred method is to use the correct document ratios in the training sets. This will improve the accuracy of the result since at the break-even region for the threshold there will be roughly equal numbers of FP and FN documents from each category, unlike the first method. Thus the noise or error in the region will be minimized. We adopted the latter approach.

**3.4.3. Model parameters.** The “AG-vs-NG” model was prepared with 175,000 training documents. Based on an analysis of real-world data distribution, a training data ratio of 1:2.5 was selected, resulting in 50,000 AG and 125,000 NG training documents. This ratio selection was based on keyword screening of random sets of 10,000 MEDLINE documents with abstracts (45% of all MEDLINE records do not contain an abstract) and an independent classification of the same set with a number of AG vs. NG models. To validate this estimate, groups of 10,000 randomly sampled MEDLINE articles were categorized by the “AG-vs-NG” model several times, each time varying the model bias slightly. Despite the changes in the bias, the AG:NG ratios only ranged between 1:2.41 and 1:2.68. Because the bias change had little effect on the data distribution, it was decided a final ratio of 1:2.5 is reasonable.

### 3.5. Individual gene models: “G-vs-OG”

Individual “G-vs-OG” models are intended to recognize and prioritize documents with context matching a specific gene, recognizing and eliminating those documents with context matching that of other, ambiguously named genes or non-gene entities that evaded the “AG-vs-NG” filter. There are 20,546 “G-vs-OG” models, one for each human LLID. The classifier is expected to realize that an abbreviation such as “ER” referring to “Endoplasmic Reticulum” is not, for example, the desired target “Estrogen Receptor”, even when it occurs in a gene-context abstract.

**3.5.1. Training set selection.** A number of possible training set sources were considered. Information linking MEDLINE documents to genes through keyword searching, MeSH [8], SP and LL databases was evaluated. MeSH linkage turned out to be too non-specific, with a large fraction of documents containing no actual mention of a given gene in the

title or abstract. (MeSH headings appear to better define the contents of a full text document.) Gene name-based keyword searching, even with phrase searching capabilities, cannot be relied on to automatically supply a quality set for more than a handful of genes. Therefore, we chose to use the SP and LL MEDLINE references to compile training set documents for the initial gene models. Documents referencing more than 20 individual genes were excluded, as they most often represent large scale sequencing projects, rather than discussing individual genes. Finally, functional description texts for each gene from LL and SP were added to each gene’s positive category (G) training set.

For the negative category (OG) in the “G-vs-OG” model, we relied on a combination of random AG-category documents (excluding any overlaps with G documents) and documents with known name or symbol ambiguities to the gene in the G category. For example, the androgen receptor gene (LLID 367) is frequently referred to by the symbol “AR”. “AR” is also known to refer to the “aldose reductase” (LLID 231) and “amphiregulin” (LLID 374) genes. The OG-category training set would then include SP and LL references to the latter two genes.

**3.5.2. Training set bias.** To accurately set the “real-world” proportions of documents in the G and OG categories an estimate is needed of the proportions of documents that are: associated with the gene; associated with other known genes, and; those associated with unlisted genes or with terms that are not genetic. Because of the large numbers of models that must be built it is impossible to require human intervention for this step.

There is also the possibility that some documents do not refer to any of the synonymous genes, but might instead refer to an unknown gene or to another non-genetic meaning (for example, “ER” can refer to “Endoplasmic Reticulum” or “Emergency Room”). The AG-vs-NG filter eliminates most documents where a gene name synonym refers to a non-gene subject and no other gene. To account for all the above possibilities, the OG-category training set included an equal number of documents to those in the G-category, but chosen at random from the AG-category set (excluding any that were already in the G-category). It has yet to be confirmed that this is the correct number of such documents to include. This is discussed further in the results section.

To recapitulate, the training sets were chosen as follows. For the G-category, the training set documents comprised gene-specific MEDLINE references from SP and LL databases. For the OG-category, a set of documents equal in number to the

G-category set were chosen at random from the complete AG-category document pool (taking care to exclude any that already occurred in G). This OG-category training set also included documents with SP and LL references to known, ambiguously named genes. The preceding assumptions give a ratio of 1:1 +  $x$  for the G:OG categories, where  $x$  is determined by the number of other synonymous genes included in category OG.

The effect of a bias between the training sets and the “real-world” proportions becomes smaller the more accurate the classifier is. Consider for example a classifier that is perfect, that is, the true and false instances are completely separated by the threshold point generated by the classifier. The break-even point will be set between the lowest-scoring true instance and the highest-scoring false instance. This will work regardless of the real world data distribution and thus for a perfect classifier, bias does not matter. So one way of compensating for the crude initial estimates used here is to accumulate more data and to improve the performance of the classifier. Methods for doing this are discussed in Section 5.

**3.5.3. Model parameters.** AZuRE requires a specialized model for each gene. Thus the training set size and composition differs for each gene model.

### 3.6. Validation document set collection and markup

Validation of our approach was done for a total of 66 genes (Tables 2 and 3). Twenty genes were selected based on the inherent ambiguity of at least one of their commonly used symbols in regard to other genes as well as to non-gene acronyms and English words (see the second column of Table 2). For each of these 20 LLIDs, a subset of MEDLINE documents containing the selected ambiguous symbol was collected (excluding any that were used to construct the training sets). Each of these was marked up by hand as one of the following: having the ambiguous symbol referencing a specific gene in human LL; another gene not in human LL or having unexpected usage based on LL gene symbol information; or as referring to a non-gene entity. The documents were then categorized with the appropriate “G-vs-OG” classifiers. In the case of the ambiguous symbol “AR”, the validation was performed for three individual genes based on one set of appropriately marked up documents. Performance of each model was evaluated against the LOO validation (see Section 4.2).

Models for 46 members of the human Nuclear Receptor (NR) gene family (Table 3) were similarly evaluated, with an additional screening by the “AG-vs-NG” model. For each NR gene, MEDLINE documents matching any one of the LL-defined gene symbols were selected and processed with the disambiguation algorithm (Figure 1). Accuracy of each gene model was estimated by human examination of all - or in some case the first several hundred top-scoring - documents in the G and OG categories.

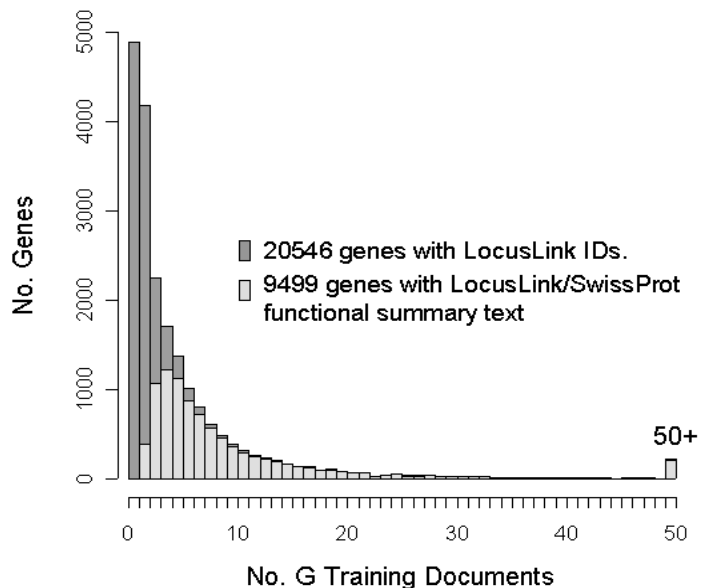
**Table 1. AG vs. NG model performance.**

| Model Category  | Automatic |         | Partially Curated |         |
|-----------------|-----------|---------|-------------------|---------|
|                 | allgenes  | notgene | allgenes          | Notgene |
| Total documents | 50000     | 125000  | 49493             | 125108  |
| F-Measure       | 0.990     | 0.922   | 0.996             | 0.924   |
| FP = FN         | 519       | 9689    | 185               | 9490    |

## 4. Results and Validation

### 4.1. “AG-vs-NG” model

The automatically generated “AG-vs-NG” model’s performance is summarized in Table 1. The AG-category accuracy is 99.0% at the break-even point, based on a LOO validation. Partial hand curation of the model increased the LOO validation accuracy to 99.6%. Any document with a below-threshold score in the AG category is treated as having a non-gene context regardless of the NG category score. It is important to note that the



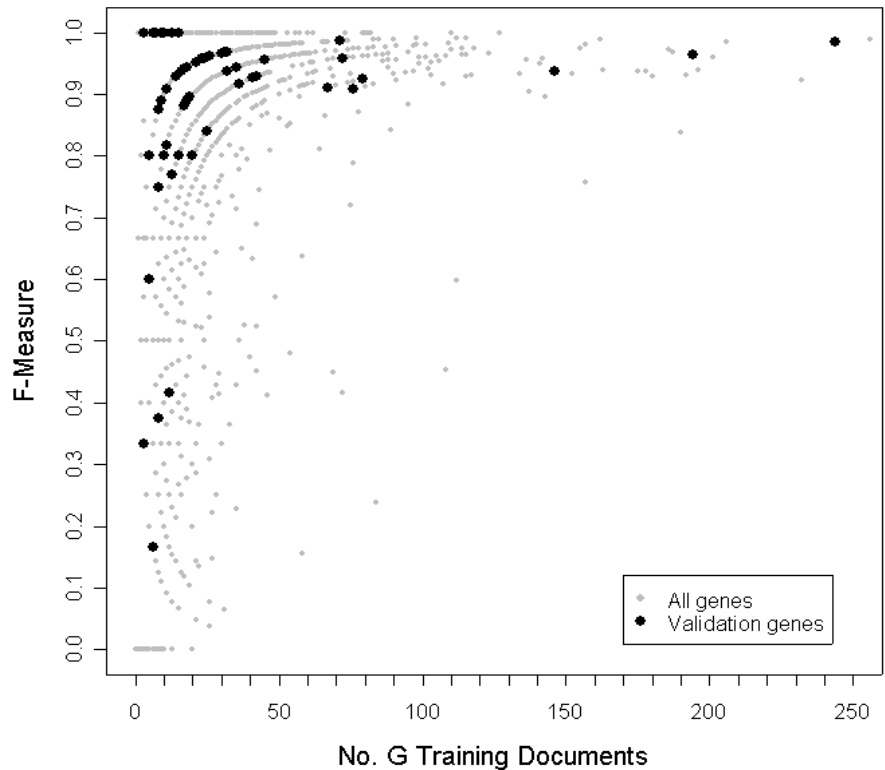
**Figure 2. Number of Genes vs. Number of Training Documents**

majority of NG documents in the False Negative (FN) group are not classified as AG, but as “other”, that is not fitting either the AG or NG categories.

#### 4.2. “G-vs-OG” model

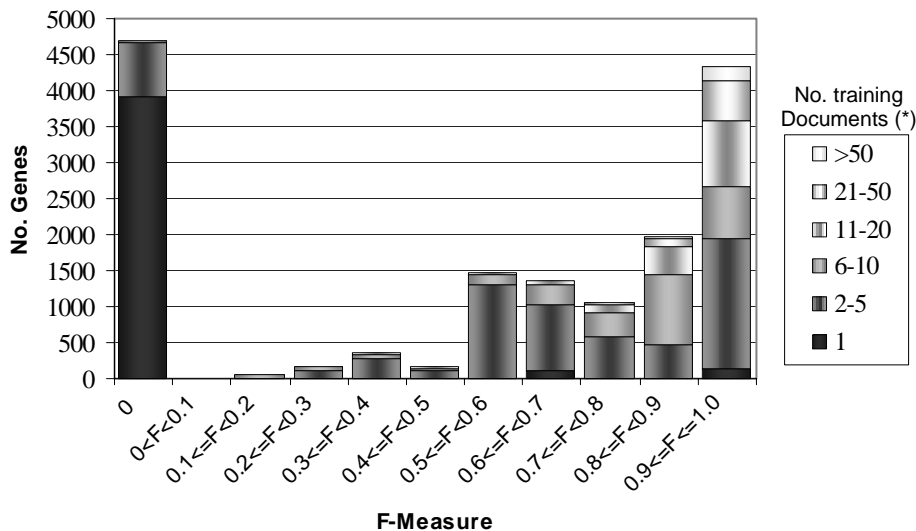
Models for 20,546 LocusLink human genes were generated automatically. Of these, 4,879 genes contained no references in the LL or SP databases and thus offered no training data to create models. Figure 2 shows the distribution of the training documents for all the genes, comparing the number of genes (LLIDs) against the number of G-category training documents. The distribution is heavily skewed toward low numbers of documents.

Despite this, Figures 3 and 4 indicate it is possible to get consistent, good model performance for genes with sufficient training examples; usually five documents



**Figure 3. Number of Genes vs. Number of Training Documents. The patterned lines result from integer values of FP and FN.**

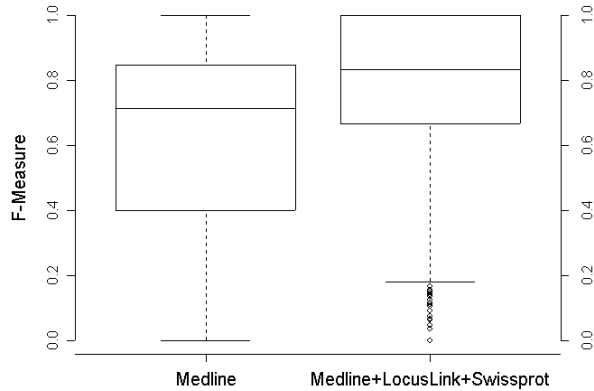
were sufficient. Figure 3 plots the LOO F-measure for all 20,564 gene models against the number of G-category training documents. The bulk of the gene models have F-measures above 70% and very few models with more than 20 training documents fall below 70%. The 66 hand-validated genes are also highlighted, and as well seem representative of the overall distribution.



(\*) The number of training documents represents MEDLINE documents only, and does not reflect the addition of LocusLink summaries for 6528 genes nor the SwissProt functional description text for 7119 genes to the model training sets.

**Figure 4. Effect of number of training documents on gene context predictive accuracy.**

The chart in Figure 4 shows a different view of the same data, with bars broken down by the number of documents in the G training set. It can be seen that almost all the poorer performing gene models have small numbers of training documents. In addition, more than 84% of gene models with more than five training documents have



**Figure 5. Effect of LocusLink and SwissProt functional text on Gene model performance.**

an accuracy of greater than 70%. For models with more than 10 training documents, this increases to 91% of the models.

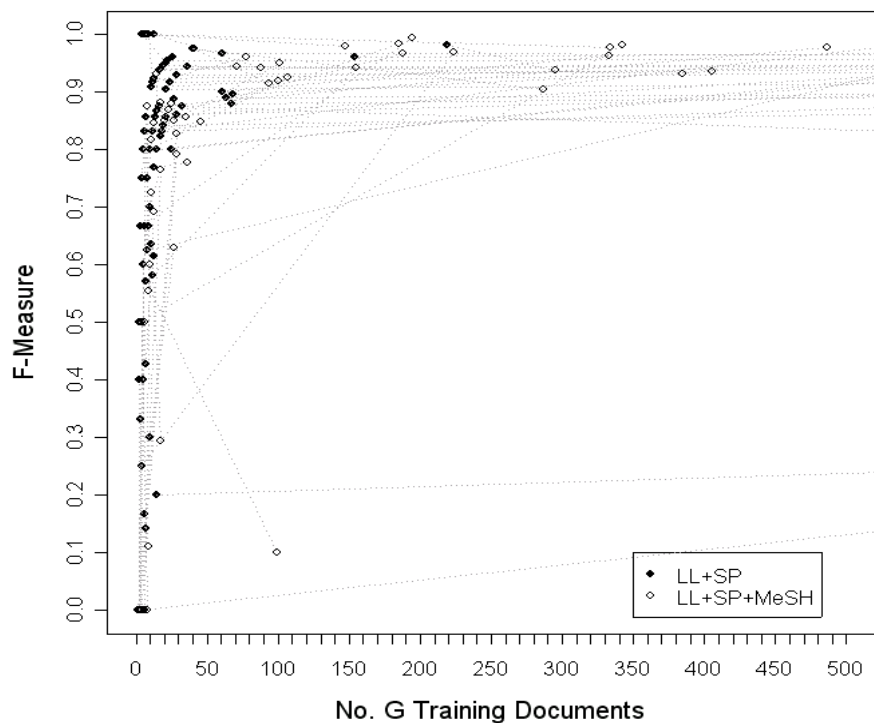
Model accuracy was greatly improved by adding functional description texts from the LL and SP databases to the training data. There were 6,528 genes with such information in LL and 7,119 in SP. Overall, 9,499 genes benefited from one or both sources. As shown in Figure 5, the median F-Measure for the 9,499 genes improved from .714 to .833, while the middle 50% of the data improved from a range of 0.4-0.846 to 0.667-1.0.

To ensure the training set contained documents that were indeed about genes, all 61,727 LL and SP gene reference documents were categorized with the “AG-vs-NG” model. The classification system predicted 58,241 (94.4%) documents as AG, while the partially hand-curated model (described in Section 4.1) indicated 57,631 (93.4%).

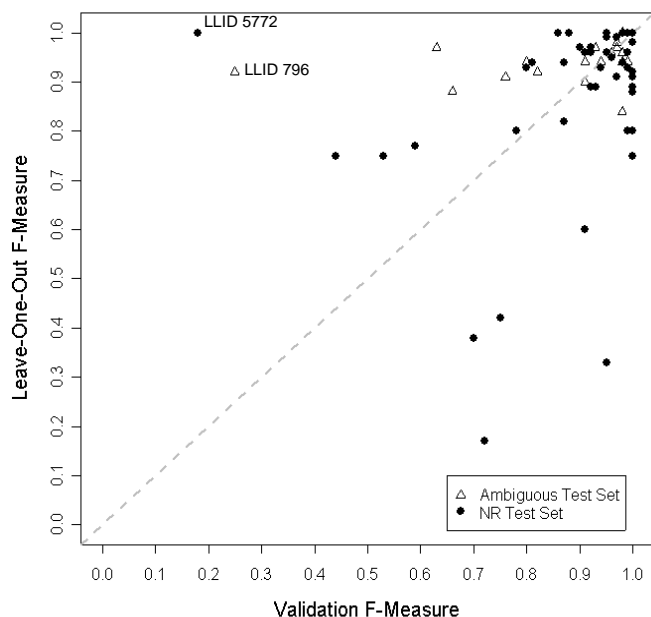
The addition of MeSH gene references to LL and SP increases the total gene model training set size from 61,727 to 1,090,641 individual documents. However, adding MeSH references did not increase, and sometimes decreased, accuracy. Figure 6 shows lines joining the LOO F-measures of the training sets with and without the MeSH-linked references. The majority of the lines trend horizontally or downward showing that MeSH-linked references have a

negligible effect on the model performance for most genes with 10 or more SP and LL training documents, and can cause an occasional significant drop in accuracy. The decrease in accuracy is often due to non-gene context of the document or lack of references to the specific gene in the abstract. Training data based purely on MeSH references thus does not seem to be useful.

Initial model predictive performance, based on a LOO method, was compared to a real-life predictive performance for a set of 20 genes with highly ambiguous gene symbols (Table 2) and 46 genes from the human NR gene family (Table 3). F-Measure comparisons between the two methods are shown in Figure 7 and the last two columns of Tables 2 and 3. Points below the diagonal in Figure 7 indicate models where human validation showed higher performance than that suggested by the LOO evaluation. Points above the diagonal reflect instances where the classification system’s LOO estimate is more optimistic than the results obtained by human validation. The general consistency between the LOO F-measures and the validated F-measures gives some confidence that the performance on the bulk of genes is actually reflected in the results in Figures 3 and 4. Tables 2 and 3 also show that the ratios of G:OG in the training data differ markedly from the actual ratios in the



**Figure 6. Impact of MeSH references on model performance.**



**Figure 7. Comparison of G vs. OG model performance.**

validation sets. Despite this, the models still yield good performance.

One outlier in the ambiguous gene symbol data set is LLID 796 (symbol CT), which shows drastically worse performance on documents marked

up as LLID 796 due to a large False Positive set. This is due to frequent use of the gene name “calcitonin-related polypeptide alpha” in non-gene, clinical context in a large fraction of documents. Strategies for overcoming this problem are presented in Section 5. Another poorly performing model is for the NR gene, NCOA5 (LLID 5772). This gene has known ambiguous gene symbols while there are only three training documents for this model - additional training documents are required.

## 5. Discussion

The problems resulting from ambiguous gene and protein names have caused enormous difficulties in biomedical text mining as well as simple text searches for gene-related information. The algorithms presented here provide a scalable system for disambiguating gene and protein names for a variety of purposes. One can use it for improving text searches against the literature by tagging all potential gene names in the literature with their canonical forms. Much more accurate NLP systems for gene and protein relation extraction will be possible given accurate disambiguation.

The results show that when more than 20 abstracts per gene are available for training,

**Table 2. Validation performance on 20 genes selected for high ambiguity.**

| LocusLink ID | Ambiguous Gene Symbol | Model Training Documents |     | Marked up Validation Documents |       | Model Predictions |      |    |      | Validation Performance |      |             | Leave-One-Out F-Measure |
|--------------|-----------------------|--------------------------|-----|--------------------------------|-------|-------------------|------|----|------|------------------------|------|-------------|-------------------------|
|              |                       | G                        | OG  | G                              | Other | TP                | FP   | FN | TN   | Prec                   | Rec  | F-Measure   |                         |
| 12           | ACT                   | 30                       | 40  | 163                            | 35    | 146               | 6    | 17 | 29   | 0.96                   | 0.90 | 0.93        | <b>0.97</b>             |
| 54           | TRAP                  | 18                       | 139 | 120                            | 6     | 106               | 0    | 14 | 6    | 1.00                   | 0.88 | 0.94        | <b>0.94</b>             |
| 231          | AR                    | 19                       | 278 | 254                            | 2306  | 215               | 5    | 39 | 2301 | 0.98                   | 0.85 | <b>0.91</b> | 0.90                    |
| 367          | AR                    | 244                      | 276 | 1700                           | 859   | 1615              | 25   | 85 | 834  | 0.98                   | 0.95 | 0.97        | <b>0.98</b>             |
| 374          | AR                    | 16                       | 275 | 99                             | 2461  | 98                | 49   | 1  | 2412 | 0.67                   | 0.99 | 0.80        | <b>0.94</b>             |
| 434          | ASP                   | 8                        | 72  | 21                             | 118   | 21                | 22   | 0  | 96   | 0.49                   | 1.00 | 0.66        | <b>0.88</b>             |
| 718          | ASP                   | 31                       | 67  | 62                             | 77    | 33                | 10   | 29 | 67   | 0.77                   | 0.53 | 0.63        | <b>0.97</b>             |
| 796          | CT                    | 36                       | 48  | 172                            | 1069  | 170               | 1019 | 2  | 50   | 0.14                   | 0.99 | 0.25        | <b>0.92</b>             |
| 847          | CAT                   | 24                       | 27  | 282                            | 318   | 279               | 21   | 3  | 297  | 0.93                   | 0.99 | <b>0.96</b> | 0.96                    |
| 948          | FAT                   | 35                       | 61  | 56                             | 38    | 55                | 0    | 1  | 38   | 1.00                   | 0.98 | <b>0.99</b> | 0.94                    |
| 1356         | CP                    | 26                       | 25  | 313                            | 24    | 304               | 3    | 9  | 21   | 0.99                   | 0.97 | <b>0.98</b> | 0.96                    |
| 1890         | TP                    | 25                       | 59  | 216                            | 102   | 212               | 5    | 4  | 97   | 0.98                   | 0.98 | <b>0.98</b> | 0.84                    |
| 2099         | ER                    | 194                      | 196 | 468                            | 7     | 455               | 5    | 13 | 2    | 0.99                   | 0.97 | <b>0.98</b> | 0.96                    |
| 2950         | PI                    | 79                       | 145 | 149                            | 153   | 141               | 53   | 8  | 100  | 0.73                   | 0.95 | 0.82        | <b>0.92</b>             |
| 3240         | HP                    | 18                       | 30  | 512                            | 4     | 497               | 3    | 15 | 1    | 0.99                   | 0.97 | <b>0.98</b> | 0.94                    |
| 4860         | NP                    | 18                       | 41  | 43                             | 25    | 36                | 0    | 7  | 25   | 1.00                   | 0.84 | 0.91        | <b>0.94</b>             |
| 5241         | PR                    | 45                       | 48  | 438                            | 26    | 427               | 10   | 11 | 16   | 0.98                   | 0.97 | <b>0.98</b> | 0.96                    |
| 5265         | PI                    | 67                       | 145 | 100                            | 202   | 65                | 6    | 35 | 196  | 0.92                   | 0.65 | 0.76        | <b>0.91</b>             |
| 6476         | SI                    | 7                        | 20  | 117                            | 16    | 117               | 4    | 0  | 12   | 0.97                   | 1.00 | 0.98        | <b>1.00</b>             |
| 7298         | TS                    | 32                       | 56  | 137                            | 12    | 131               | 1    | 6  | 11   | 0.99                   | 0.96 | <b>0.97</b> | 0.97                    |

**Table 3. Validation performance on 46 Nuclear Receptor genes.**

| LocusLink ID | Official Gene Symbol | Model Training Documents |     | Marked up Validation Documents |       | Model Predictions |     |    |     | Validation Performance |      |             | Leave-One-Out |
|--------------|----------------------|--------------------------|-----|--------------------------------|-------|-------------------|-----|----|-----|------------------------|------|-------------|---------------|
|              |                      | G                        | OG  | G                              | Other | TP                | FP  | FN | TN  | Prec                   | Rec  | F-Measure   | F-Measure     |
| 190          | NR0B1                | 25                       | 24  | 225                            | 75    | 196               | 10  | 29 | 65  | 0.87                   | 0.95 | 0.91        | <b>0.96</b>   |
| 367          | AR                   | 244                      | 276 | 102                            | 100   | 102               | 0   | 0  | 100 | 1.00                   | 1.00 | <b>1.00</b> | 0.98          |
| 2063         | NR2F6                | 3                        | 3   | 28                             | 4     | 28                | 3   | 0  | 1   | 1.00                   | 0.90 | <b>0.95</b> | 0.33          |
| 2099         | ESR1                 | 194                      | 196 | 45                             | 102   | 40                | 2   | 5  | 100 | 0.89                   | 0.95 | 0.92        | <b>0.96</b>   |
| 2100         | ESR2                 | 72                       | 72  | 100                            | 96    | 99                | 1   | 1  | 95  | 0.99                   | 0.99 | <b>0.99</b> | 0.96          |
| 2101         | ESRRA                | 15                       | 14  | 32                             | 0     | 32                | 0   | 0  | 0   | 1.00                   | 1.00 | <b>1.00</b> | 0.80          |
| 2103         | ESRRB                | 6                        | 4   | 16                             | 0     | 16                | 0   | 0  | 0   | 1.00                   | 1.00 | <b>1.00</b> | 1.00          |
| 2104         | ESRRG                | 8                        | 12  | 5                              | 0     | 5                 | 0   | 0  | 0   | 1.00                   | 1.00 | <b>1.00</b> | 0.75          |
| 2494         | NR5A2                | 11                       | 11  | 54                             | 159   | 54                | 16  | 0  | 143 | 1.00                   | 0.77 | <b>0.87</b> | 0.82          |
| 2516         | NR5A1                | 18                       | 24  | 117                            | 82    | 100               | 0   | 17 | 82  | 0.86                   | 1.00 | <b>0.92</b> | 0.89          |
| 2649         | NR6A1                | 13                       | 12  | 43                             | 18    | 43                | 2   | 0  | 16  | 1.00                   | 0.96 | 0.98        | <b>1.00</b>   |
| 2908         | NR3C1                | 71                       | 80  | 104                            | 118   | 97                | 3   | 7  | 115 | 0.93                   | 0.97 | 0.95        | <b>0.99</b>   |
| 3164         | NR4A1                | 17                       | 39  | 99                             | 97    | 99                | 0   | 0  | 97  | 1.00                   | 1.00 | <b>1.00</b> | 0.88          |
| 3172         | HNF4A                | 41                       | 42  | 90                             | 110   | 89                | 11  | 1  | 99  | 0.99                   | 0.89 | <b>0.94</b> | 0.93          |
| 3174         | HNF4G                | 5                        | 4   | 2                              | 0     | 2                 | 0   | 0  | 0   | 1.00                   | 1.00 | <b>1.00</b> | 0.80          |
| 4306         | NR3C2                | 23                       | 22  | 94                             | 106   | 94                | 6   | 0  | 100 | 1.00                   | 0.94 | 0.97        | <b>0.99</b>   |
| 4929         | NR4A2                | 21                       | 35  | 130                            | 27    | 119               | 0   | 11 | 27  | 0.92                   | 1.00 | <b>0.96</b> | 0.95          |
| 5241         | PGR                  | 45                       | 48  | 182                            | 112   | 180               | 16  | 2  | 96  | 0.99                   | 0.92 | 0.95        | <b>0.96</b>   |
| 5465         | PPARA                | 42                       | 43  | 104                            | 43    | 82                | 18  | 22 | 25  | 0.79                   | 0.82 | 0.80        | <b>0.93</b>   |
| 5467         | PPARD                | 17                       | 17  | 80                             | 21    | 79                | 2   | 1  | 19  | 0.99                   | 0.98 | <b>0.98</b> | 0.94          |
| 5468         | PPARG                | 146                      | 155 | 146                            | 0     | 100               | 0   | 46 | 0   | 0.69                   | 1.00 | 0.81        | <b>0.94</b>   |
| 5914         | RARA                 | 31                       | 35  | 111                            | 87    | 94                | 4   | 17 | 83  | 0.85                   | 0.96 | 0.90        | <b>0.97</b>   |
| 5915         | RARB                 | 36                       | 45  | 101                            | 99    | 100               | 0   | 1  | 99  | 0.99                   | 1.00 | <b>1.00</b> | 0.92          |
| 6095         | RORA                 | 9                        | 19  | 47                             | 34    | 42                | 1   | 5  | 33  | 0.89                   | 0.98 | <b>0.93</b> | 0.89          |
| 6096         | RORB                 | 9                        | 9   | 20                             | 2     | 18                | 0   | 2  | 2   | 0.90                   | 1.00 | 0.95        | <b>1.00</b>   |
| 6097         | RORC                 | 8                        | 8   | 8                              | 39    | 8                 | 14  | 0  | 25  | 1.00                   | 0.36 | 0.53        | <b>0.75</b>   |
| 6256         | RXRA                 | 30                       | 32  | 116                            | 0     | 99                | 0   | 17 | 0   | 0.85                   | 1.00 | 0.92        | <b>0.97</b>   |
| 6257         | RXRB                 | 14                       | 14  | 102                            | 0     | 99                | 0   | 3  | 0   | 0.97                   | 1.00 | <b>0.99</b> | 0.93          |
| 6258         | RXRG                 | 5                        | 5   | 94                             | 0     | 92                | 0   | 2  | 0   | 0.98                   | 1.00 | <b>0.99</b> | 0.80          |
| 7025         | NR2F1                | 12                       | 18  | 86                             | 0     | 52                | 0   | 34 | 0   | 0.61                   | 1.00 | <b>0.75</b> | 0.42          |
| 7026         | NR2F2                | 6                        | 59  | 43                             | 45    | 24                | 0   | 19 | 45  | 0.56                   | 1.00 | <b>0.72</b> | 0.17          |
| 7067         | THRA                 | 20                       | 27  | 148                            | 54    | 97                | 3   | 51 | 51  | 0.66                   | 0.97 | 0.78        | <b>0.80</b>   |
| 7068         | THRB                 | 32                       | 32  | 128                            | 82    | 99                | 0   | 29 | 82  | 0.77                   | 1.00 | 0.87        | <b>0.94</b>   |
| 7101         | NR2E1                | 5                        | 38  | 37                             | 33    | 35                | 5   | 2  | 28  | 0.95                   | 0.88 | <b>0.91</b> | 0.60          |
| 7181         | NR2C1                | 7                        | 38  | 50                             | 49    | 47                | 13  | 3  | 36  | 0.94                   | 0.78 | 0.86        | <b>1.00</b>   |
| 7182         | NR2C2                | 10                       | 31  | 33                             | 101   | 33                | 1   | 0  | 100 | 1.00                   | 0.97 | <b>0.99</b> | 0.80          |
| 7376         | NR1H2                | 8                        | 22  | 25                             | 158   | 24                | 60  | 1  | 98  | 0.96                   | 0.29 | 0.44        | <b>0.75</b>   |
| 7421         | VDR                  | 76                       | 99  | 101                            | 99    | 100               | 0   | 1  | 99  | 0.99                   | 1.00 | <b>1.00</b> | 0.91          |
| 8431         | NR0B2                | 2                        | 2   | 52                             | 172   | 52                | 73  | 0  | 99  | 1.00                   | 0.42 | 0.59        | <b>0.77</b>   |
| 8856         | NR1I2                | 18                       | 72  | 100                            | 100   | 100               | 0   | 0  | 100 | 1.00                   | 1.00 | <b>1.00</b> | 0.89          |
| 9572         | NR1D1                | 8                        | 27  | 35                             | 45    | 19                | 0   | 16 | 45  | 0.54                   | 1.00 | <b>0.70</b> | 0.38          |
| 9970         | NR1I3                | 17                       | 72  | 99                             | 101   | 99                | 1   | 0  | 100 | 1.00                   | 0.99 | <b>1.00</b> | 0.88          |
| 9971         | NR1H4                | 15                       | 72  | 98                             | 102   | 98                | 2   | 0  | 100 | 1.00                   | 0.98 | 0.99        | <b>1.00</b>   |
| 10002        | NR2E3                | 10                       | 10  | 14                             | 102   | 14                | 4   | 0  | 98  | 1.00                   | 0.78 | 0.88        | <b>1.00</b>   |
| 10062        | NR1H3                | 11                       | 12  | 107                            | 2     | 100               | 0   | 7  | 2   | 0.94                   | 1.00 | <b>0.97</b> | 0.91          |
| 57727        | NCOA5                | 3                        | 10  | 12                             | 180   | 12                | 108 | 0  | 72  | 1.00                   | 0.10 | 0.18        | <b>1.00</b>   |

accuracy of the system is mostly over 90%. Even with 5 documents we usually get significant enrichment of the search results. The system can easily be altered dynamically to provide greater

precision or recall by altering the thresholds associated with the gene disambiguation models.

The next step (in process) is a developing a central, publicly available web service to allow

researchers to access this system when searching the literature for specific genes or proteins. The users of the public system will be able to provide performance feedback and additional training data if the gene of interest has too little training data to yield accurate disambiguation results, or if the existing training data displays an unexpected bias (such as that found in LLID 796 as documented in Section 4.2). Although models will initially have small numbers of training documents, training can be quickly bootstrapped as users submit feedback on initial predictions. In this way, additional training data can be collected in a scaleable manner based on distributed feedback/ annotation. Further, genes of specific interest such as pharmaceutically relevant genes (GPCR's, NHR's, Kinases, etc.) can be enhanced in an organized way based on their family membership or if the gene shows a low F-measure.

## Acknowledgements

We thank Ian Dix for providing the MeSH to LLID linkage map and Julia Kozlovsky for help in validation of model predictions.

## References

- [1] A. R. Aronson, "Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program", *Proc. AMIA Symp*, 2001, pp. 17-21.
- [2] A. R. Aronson, "Ambiguity in the UMLS Metathesaurus", National Library of Medicine, 2001.
- [3] W. A. Gale, K. W. Church, D. Yarowsky, "A Method for Disambiguating Word Senses in a Large Corpus", *Computers and the Humanities*, 26, 1993, pp. 415-439.
- [4] V. Hatzivassiloglou, P. A. Dubou'e, A. Rzhetsky, "Disambiguating Proteins, Genes, and RNA in Text: A Machine Learning Approach", *Bioinformatics*, 1, 2001, pp. 1-10.
- [5] "LocusLink Database", 2003. [ftp://ftp.ncbi.nlm.nih.gov/refseq/LocusLink/LL\\_tmpl.gz](ftp://ftp.ncbi.nlm.nih.gov/refseq/LocusLink/LL_tmpl.gz) (accessed December 2004).
- [6] H. Liu, S. B. Johnson, C. Friedman, "Automatic Resolution of Ambiguous Terms Based on Machine

Learning and Conceptual Relations in the UMLS", *J Am Med Inform Assoc.*, 9, 2003, pp. 621-636.

[7] J. S. MacNeil, "What Big Pharma Wants", *Genome Technology*, 29, 2003, pp. 31-38.

[8] "Medical Subject Headings", National Library of Medicine, 1998. <http://www.nlm.nih.gov/mesh/meshhome.html>

[9] T. Mitchell, "Machine Learning", McGraw-Hill, 1997.

[10] C. O'Donovan, M. J. Martin, A. Gattiker, E. Gasteiger, A. Bairoch, R. Apweiler, "High-quality protein knowledge resource: SWISS-PROT and TrEMBL", *Brief. Bioinform.*, 3, 2002, pp. 275-284.

[11] K. D. Pruitt, D. R. Maglott, "RefSeq and LocusLink: NCBI gene-centered resources", *Nucleic Acids Res*, 29(1), 2001, pp. 137-140.

[12] "Reel Two Classification System", Reel Two Inc. San Francisco, CA, 2001-2004. <http://www.reeltwo.com>.

[13] P. Resnik, D. Yarowsky, "Distinguishing Systems and Distinguishing Senses: New Evaluation Methods for Word Sense Disambiguation", *Natural Language Engineering*, Cambridge University Press. 5 (3), 2000, pp. 113-133.

[14] T. Rindflesch, L. Tanabe, J. Weinstein, and L. Hunter,, "EDGAR: extraction of drugs, genes and relations from the biomedical literature", *Proceedings of the Pacific Symposium on Biocomputing*, 5, 2000, pp. 514-525.

[15] "Unified Medical Language System Metathesaurus", 2003. <http://lhncbc.nlm.nih.gov/csb/CSBPages/UMLProject.html>

[16] D. Yarowsky, "Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora", *Proceedings of the 14th International Conference on Computational Linguistics*, 2000, pp. 454-460.

[17] D. Yarowsky, "Unsupervised Word Sense Disambiguation Rivaling Supervised Methods", *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, Cambridge, MA, 1995, pp. 189-196.

[18] H. Yu, E. Agichtein, "Extracting synonymous gene and protein terms from biological literature", *Bioinformatics*, 19, Suppl. 1, 2003, pp. i340-i349.