Combining Multiple Evidence for Gene Symbol Disambiguation

Hua Xu
Dept. of Biomedical Informatics, Columbia University
622 W 168th St. NY, USA
hux7002@dbmi.columbia.edu

Jung-Wei Fan
Dept. of Biomedical Informatics, Columbia University
622 W 168th St. NY, USA
fan@dbmi.columbia.edu

Carol Friedman
Dept. of Biomedical Informatics, Columbia University
622 W 168th St. NY, USA
friedman@dbmi.columbia.edu

Abstract

Gene names and symbols are important biomedical entities, but are highly ambiguous. This ambiguity affects the performance of both information extraction and information retrieval systems in the biomedical domain. Existing knowledge sources contain different types of information about genes and could be used to disambiguate gene symbols. In this paper, we applied an information retrieval (IR) based method for human gene symbol disambiguation and studied different methods to combine various types of information from available knowledge sources. Results showed that a combination of evidence usually improved performance. The combination method using coefficients obtained from a logistic regression model reached the highest precision of 92.2% on a testing set of ambiguous human gene symbols.

1 Introduction

In the past decade, biomedical discoveries and publications have increased exponentially due to high-throughput technologies such as automated genomic sequencing, and therefore, it is impossible for researchers to keep up-to-date with the most recent knowledge by manually reading the literature. Therefore, automated text mining tools, such as information retrieval and information extraction systems, have received great amounts of interest (Erhardt et al., 2006; Krallinger and Valencia, 2005). Biomedical entity recognition is a first crucial step for text mining tools in this domain, but is a very challenging task, partially due to the ambiguity (one name referring to different entities) of names in the biomedical field.

Genes are among the most important biological entities for understanding biological functions and processes, but gene names and symbols are highly ambiguous. Chen et al. (2005) obtained gene information from 21 organisms and found that ambiguities within species, across species, with English words and with medical terms were 5.02%, 13.43%, 1.10%, 2.99%, respectively, when both official gene symbols and aliases were considered. When mining MEDLINE abstracts, they found that 85.1% of mouse genes in the articles were ambiguous with other gene names. Recently, Fundel and Zimmer (2006) studied gene/protein nomenclature in 5 public databases. Their results showed that the ambiguity problem was not trivial. The degree of ambiguity also varied among different organisms. Unlike other abbreviations in the literature, which usually are accompanied by their corresponding long forms, many gene symbols occur alone without any mention of their long forms. According to Schuemie et al. (2004), only 30% of gene symbols in abstracts and 18% in full text were accompanied by their corresponding full names, which makes the task of gene symbol normalization much harder.

Gene symbol disambiguation (GSD) is a particular case of word sense disambiguation (WSD), which has been extensively studied in the domain of general English. One type of method for WSD uses established knowledge bases, such as a machine readable dictionary (Lesk, 1986; Harley and Glennon, 1997). Another type of WSD method uses supervised machine learning (ML) technolo-

BioNLP 2007: Biological, translational, and clinical language processing, pages 41–48,
Prague, June 2007.

gies (Bruce and Wiebe, 1994; Lee and Ng, 2002; Liu et al., 2002).

In the biomedical domain, there are many gene related knowledge sources, such as Entrez Gene (Maglott et al., 2005), developed at NCBI (National Center for Biotechnology Information), which have been used for gene symbol disambiguation. Podowski et al. (2004) used MEDLINE references in the LocusLink and SwissProt databases to build Bayesian classifiers for GSD. A validation on MEDLINE documents for a set of 66 human genes showed most accuracies were greater than 90% if there was enough training data (more than 20 abstracts for each gene sense).

More recently, information retrieval (IR) based approaches have been applied to resolve gene ambiguity using existing knowledge sources. Typically, a profile vector for each gene sense is built from available knowledge source(s) and a context vector is derived from the context where the ambiguous gene occurs. Then similarities between the context vector and candidate gene profile vectors are calculated, and the gene corresponding to the gene profile vector that has the highest similarity score to the context vector is selected as the correct sense. Schijvenaars et al. (2005) reported on an IR-based method for human GSD. It utilized information from either Online Mendelian Inheritance in Man (OMIM) annotation or MEDLINE abstracts. The system achieved an accuracy rate of 92.7% on an automatically generated testing set when five abstracts were used for the gene profile. Xu et al. (2007) studied the performance of an IR-based approach for GSD for mouse, fly and yeast organisms when different types of information from different knowledge sources were used. They also used a simple method to combine different types of information and reported that a highest precision of 93.9% was reached for a testing set of mouse genes using multiple types of information.

In the field of IR, it has been shown that combining heterogeneous evidence improves retrieval effectiveness. Studies on combining multiple representations of document content (Katzer et al., 1982), combining results from different queries (Xu and Croft, 1996), different ranking algorithms (Lee, 1995), and different search systems (Lee, 1997) have shown improved performance of retrieval systems. Different methods have also been developed to combine different evidence for IR tasks. The inference-network-based framework, developed by Turtle and Croft (1991), was able to combine different document representations and retrieval algorithms into an overall estimate of the probability of relevance. Fox et al. (1988) extended the vector space model to use sub-vectors to describe different representations derived from documents. An overall similarity between a document and a query is defined as a weighted linear combination of similarities of sub-vectors. A linear regression analysis was used to determine the value of the coefficients.

Though previous related efforts (Schijvenaars et al., 2005, Xu et al., 2007) have explored the use of multiple types of information from different knowledge sources, none have focused on development of formal methods for combining multiple evidence for the GSD problem to optimize performance of an IR-based method. In this study, we adapted various IR-based combination models specifically for the GSD problem. Our motivation for this work is that there are diverse knowledge sources containing different types of information about genes, and the amount of such information is continuously increasing. A primary source containing gene information is MEDLINE articles, which could be linked to specific genes through annotation databases. For example, Entrez Gene contains an annotated file called “gene2pubmed”, which lists the PMIDs (PubMed ID) of articles associated with a particular gene. From related MEDLINE articles, words and different ontological concepts can be obtained and then be used as information associated with a gene. However they could be noisy, because one article could mention multiple genes. Another type of source contains summarized annotation of genes, which are more specific to certain aspects of genes. For example, Entrez Gene contains a file called “gene2go”. This file lists genes and their associated Gene Ontology (GO) (Ashburner et al., 2000) codes, which include concepts related to biological processes, molecular functions, and cellular components of genes. Therefore, methods that are able to efficiently combine the different types of information from the different sources are important to explore for the purpose of improving performance of GSD systems. In this paper, we describe various models for combining different types of information from MEDLINE abstracts for IR-based GSD systems. We also evaluated the combination models using two data sets containing ambiguous human genes.

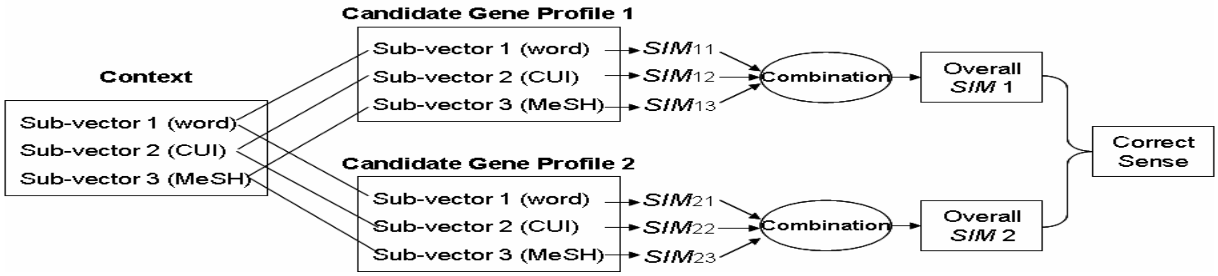

Figure 1 Overview of an IR combination-based gene symbol disambiguation approach using different types of information.

2 Methods

In this paper, we extend the IR vector space model to be capable of combining different types of gene related information in a flexible manner, thus improving the performance of an IR-based GSD system. Figure 1 shows an overview of the IR combination-based approach. We generated three different sub-vectors for the context and three for the profile, so that each sub-vector corresponded to a different type of information. The similarity scores between context and profile were measured for each type of sub-vector and then combined to generate the overall similarity scores to determine the correct sense. We explored five different combination methods using two testing sets.

2.1 Knowledge Sources and Available Information

The “gene2pubmed” file in Entrez Gene was downloaded in January 2006. A profile was then built for each gene using information derived from the related articles. We used the following three types of information: 1) Words in the related MEDLINE articles (title and abstract). This is the simplest type of information about a gene. General English stop words were removed and all other words were stemmed using the Porter stemming algorithm (Porter, 1980). 2) UMLS (Unified Medical Language System) (Bodenreider 2004) CUIs (Concept Unique Identifier), which were obtained from titles and abstracts of MEDLINE articles using an NLP system called MetaMap (Aronson 2001). 3) MeSH (Medical Subject Headings) terms, which are manually annotated by curators based on full-text articles at the National Library of Medicine (NLM) of the United States.

2.2 Document Set and Testing Sets

Using the “gene2pubmed” file, we downloaded the MEDLINE abstracts that were known to be related to human genes. Articles associated with more than 25 genes (as determined by our observation) were excluded, since they mostly discussed high-throughput technologies and provided less valuable information for GSD. This excluded 168 articles and yielded a collection of 116,929 abstracts, which were used to generate gene profiles and one of the test sets. Two test sets were obtained for evaluating the combination methods: testing set 1 was based on the “gene2pubmed” file, and testing set 2 was based on the BioCreAtIvE II evaluation.

Testing set 1 was automatically generated from the 116,929 abstracts, using the following 3 steps:

1) Identifying ambiguous gene symbols in the abstracts. This involved processing the entire collection of abstracts using an NLP system called BioMedLEE (Biomedical Language Extracting and Encoding System) (Lussier et al. 2006), which was shown to identify gene names/symbols with high precision when used in conjunction with GO annotations. When an ambiguous gene was identified in an article, the candidate gene identifiers (GeneID from Entrez Gene) were listed by the NLP system, but not disambiguated. For each ambiguous gene that was detected, a pair was created consisting of the PMID of the article and the gene symbol, so that each pair would be considered a possible testing sample. Repeated gene symbols in the same article were ignored, because we assumed only one sense per gene symbol in the same article. Using this method, 69,111 PMID and ambiguous human gene symbol pairs were identified from the above collection of abstracts.

2) Tagging the correct sense of the ambiguous gene symbols. The list of candidate PMID/gene symbol pairs generated from the articles was then compared with the list of gene identifiers known to be associated with the articles based on "gene2pubmed". If one of the candidate gene senses matched, that gene sense was assumed to be the correct sense. Then the PMID/gene-symbol pair was tagged with that sense and set aside as a testing sample. We identified a pool of 12,289 testing samples, along with the corresponding tagged senses.

3) Selecting testing set 1. We randomly selected 2,000 testing samples from the above pool to form testing set 1.

Testing set 2 was derived using the training and evaluation sets of the BioCreAtIvE II Gene Normalization (GN) task (Morgan 2007). The BioCreAtIvE II GN task involved mapping human gene mentions in MEDLINE abstracts to gene identifiers (Entrez Gene ID), which is a broader task than the GSD task. However, these abstracts were useful for creating a testing set for GSD, because whenever a gene mention mapped to more than one identifier, disambiguation was required. Therefore, it was possible to derive a list of ambiguous gene symbols based on data that was provided by BioCreAtIvE. We combined both manually annotated training (281 abstracts) and evaluation (262 abstracts) sets provided by BioCreAtIvE. Using the same process as described in step 1 of testing set 1, we processed the abstracts and identified 217 occurrences of ambiguous gene symbols from the combined set. Following a similar procedure as was used for step 2 in the testing set 1 (except that the reference standard in this case was the manually annotated results obtained from BioCreAtIvE instead of "gene2pubmed"), we obtained 124 PMID/gene-symbol pairs with the corresponding tagged senses, which formed testing set 2.

Because one article may contain multiple ambiguous gene symbols, a total of 2,048 PMIDs were obtained from both testing sets 1 and 2. Articles with those PMIDs were excluded from the collection of 116,929 abstracts. We used the remaining document set to generate gene profiles, which were used for both testing sets.

2.3 Profile and Context Vectors

For each gene in "gene2pubmed" file, we created a profile. It consisted of three sub-vectors containing *word*, *CUI,* or *MeSH,* respectively, using the information derived from the related MEDLINE abstracts. Similarly, a context vector was also formed for each testing sample, using three sub-vectors containing *word*, *CUI,* or *MeSH*, which were derived from the abstract whose PMID was stated in the testing sample. The tf-idf weighting schema (Salton and Buckley, 1988) was used to assign weights to index terms in the profile and context sub-vectors. Given a document *d*, the Term Frequency (tf) of term *t* is defined as the frequency of *t* occurring in *d*. The Inverse Document Frequency (idf) of term *t* is defined as the logarithm of the number of all documents in the collection divided by the number of documents containing the term *t*. Then term *t* in document *d* is weighted as tf*idf.

2.4 Similarity Measurement

The similarity score between the same type of context and profile sub-vectors were measured as cosine similarity of two vectors. The cosine similarity between two vectors *a* and *b* is defined as the inner product of *a* and *b*, normalized by the length of two vectors. See the formula below:

$$\mathrm{Sim}(a,b) = \text{cosine } \theta = \frac{a \cdot b}{|a||b|} \quad \text{where}$$

$$|a| = \sqrt{a_1^2 + a_2^2 + \ldots + a_n^2} \qquad |b| = \sqrt{b_1^2 + b_2^2 + \ldots + b_n^2}$$

We built three basic classifiers that used only one type of sub-vector: *word*, *CUI,* or *MeSH,* respectively, recorded three individual similarity scores of each sub-vector for each candidate gene of all testing samples. We implemented five methods to combine similarity scores from each basic classifier, which are described as follows:

1) *CombMax* - Each individual similarity score from a basic classifier was normalized by dividing the sum of similarity scores of all candidate genes for that basic classifier. Then the decision made by the classifier with the highest normalized score was selected as the final decision of the combined method.
2) *CombSum* - Each individual similarity score from a basic classifier was normalized by dividing the maximum similarity score of all candidate genes for that basic classifier. The overall similarity score of a candidate gene was considered to be the sum of the normalized similarity scores from all three basic classifiers for that gene. The candidate gene

with the highest overall similarity was selected as the correct sense.

3) *CombSumVote* - The overall similarity score was considered as the similarity score from *CombSum,* multiplied by the number of basic classifiers that voted for that gene as the correct sense.

4) *CombLR* - The overall similarity score was defined as a predicted probability (*P*) of being the correct sense, given the coefficients obtained from a logistic regression model and similarity scores from all three basic classifiers for that gene. The relation between dependent variable (probability of being the correct sense) and independent variables (similarity scores from individual basic classifiers) of the logistic regression model is shown below, where *Cs* (C_{word}, C_{cui}, C_{mesh} and *C*) are the coefficients, and *SIMs* (SIM_{word}, SIM_{cui}, SIM_{mesh}) are the individual similarity scores from the basic classifiers. To obtain the model, we divided 2,000 testing samples into a training set and a testing set, as described in section 2.5. For samples in the training set, the correct gene senses were labeled as "1" and incorrect gene senses were labeled as "0". Then logistic regression was applied, taking the binary labels as the value of the dependent variable and the similarities from the basic classifiers as the independent variables. In testing, coefficients obtained from training were used to predict each candidate gene's probability of being the correct sense for a given ambiguous symbol.

$$P = \frac{e^{Cword*SIMword+Ccui*SIMcui+Cmesh*SIMmesh+C}}{1+e^{Cword*SIMword+Ccui*SIMcui+Cmesh*SIMmesh+C}}$$

5) *CombRank* – Instead of using the similarity scores, we ranked the similarity scores and used the rank to determine the combined output. Following a procedure called Borda count (Black, 1958), the top predicted gene sense was given a ranking score of N-1, the second top was given N-2, and so on, where N is the total number of candidate senses. After each sense was ranked for each basic classifier, the combined ranking score of a candidate gene was determined by the sum of ranking scores from all three basic classifiers. The sense with the highest combined ranking score was selected as the correct sense.

2.5 Experiments and Evaluation

In this study, we measured both precision and coverage of IR-based GSD approaches. Precision was defined as the ratio between the number of correctly disambiguated samples and the number of total testing samples for which the disambiguation method yielded a decision. When a candidate gene had an empty profile or different candidate gene profiles had the same similarity scores (e.g. zero score) with a particular context vector, the disambiguation method was not able to make a decision. Therefore, we also reported on coverage, which was defined as the number of testing samples that could be disambiguated using the profile-based method over the total number of testing samples. We evaluated precision and coverage of different combined methods for gene symbol disambiguation on both testing sets.

Results of three basic classifiers that used a single type of information were reported as well. We also defined a baseline method. It used the majority sense of an ambiguous gene symbol as the correct sense. The majority sense is defined as the gene sense which was associated with the most MEDLINE articles based on the "gene2pubmed" file.

To evaluate the *CombLR*, we used 10-fold cross validation. We divided the sense-tagged testing set into 10 equal partitions, which resulted in 200 testing samples for each partition. When one partition was used for testing, the remaining nine partitions were combined and used for training, which also involved deriving coefficients for each round. To make other combination methods comparable with *CombLR*, we tested the performance of other combination methods on the same partitions as well. Therefore, we had 10 measurements for each combination method. Mean precision and mean coverage were reported for those 10 measurements. For testing set 2, we did not test the *CombLR* method because the set was too small to train a regression model.

We used Friedman's Test (Friedman, 1937) followed by Dunn's Test (Dunn, 1964), which are non-parametric tests, to assess whether there were significant differences in terms of median precision among the different single or combined methods.

3 Results

Results of different combination methods for testing set 1 are shown in Table 1, which contains the mean precision and coverage for 10-fold cross validation, as well as the standard errors in parentheses. All IR-based gene symbol disambiguation approaches showed large improvements when compared to the baseline method. All of the combination methods showed improved performance when compared to results from any run that used a single type of information. Among the five different combination methods, *CombLR* achieved the highest mean precision of 0.922 for testing set 1. *CombSum,* which is a simple combination method, also had a good mean precision of 0.920 on testing set 1. The third Column of Table 1 shows that coverage was in a range of 0.936-0.938.

Run	Precision	Coverage
Baseline	0.707 (0.032)	0.992 (0.005)
Word	0.882 (0.023)	0.937 (0.017)
CUI	0.887 (0.022)	0.938 (0.017)
MeSH	0.900 (0.021)	0.936 (0.017)
CombMax	0.909 (0.020)	0.938 (0.017)
CombSum	0.920 (0.019)	0.937 (0.017)
CombSumVote	0.917(0.019)	0.938 (0.017)
CombLR	**0.922** (0.019)	0.938 (0.017)
CombRank	0.918 (0.020)	0.938 (0.017)

Table 1. Results on testing set 1.

Run	Precision	Coverage
Baseline	0.593	0.991
Word	0.872	0.944
CUI	0.897	0.944
MeSH	0.863	0.944
CombMax	**0.906**	0.944
CombSum	**0.906**	0.944
CombSumVote	0.897	0.944
CombRank	0.889	0.944

Table 2. Results on testing set 2.

We performed Friedman's test followed by Dunn's test on each single run: *word*, *CUI* or *MeSH,* with all combination runs respectively. Friedman tests showed that differences of median precisions among the different methods were statistically significant at α=0.05. Dunn tests showed that combination runs *CombSum*, *CombSumVote, CombLR,* and *CombRank* were statistically significantly better than single runs using *word* or *CUI*. For single run using *MeSH*, combination runs *CombLR* and *CombSum* were statistically significantly better.

The results of different runs on testing set 2 are shown in Table 2. Most combined methods, except *CombRank*, showed improved precision. The highest precision of 0.906 was reached when using *CombSum* and *CombMax* methods. Note that the logistic regression method was not applicable. The coverage for testing set 2 was 0.944 for all of the methods.

4 Discussion

4.1 Why Combine?

As stated in Croft (2002), a Bayesian probabilistic framework could provide the theoretical justification for evidence combination. Additional evidence with smaller errors can reduce the effect of large errors from one piece of evidence and lower the average error.

The idea behind *CombMax* was to use the single classifier that had the most confidence, but it did not seem to improve performance very much because it ignored evidence from the other two basic classifiers. The *CombSum* was a simple combination method, but with reasonable performance, which was also observed by other studies for the IR task (Fox and Shaw, 1994). *CombSumVote* was a variant of *CombSum*. It favors the candidate genes selected by more basic classifiers. In Lee (1997), a similar implementation of *CombSumVote* (named "CombMNZ") also achieved better performance in the IR task. *CombLR*, the combination method trained on a logistic regression model, achieved the best performance in this study. It used a set of coefficients derived from the training data when combining the similarities from individual basic classifiers. Therefore, it could be considered as a more complicated linear combination model than *CombSum*. In situations where training data is not available, *CombSum* or *CombSumVote* would be a good choice. *CombRank* did not perform as well as methods that used similarity scores, probably due to the loss of subtle probability information in the similarity scores. We explored ranking because it was independent of the weighting schema and could be valuable if it performed well.

The typical scenario where combination should help is when a classifier based on one type of information made a wrong prediction, but the other(s), based on different types of information, made the correct predictions. In those cases, the overall prediction may be correct when an appropriate combination method applies. For example, an ambiguous gene symbol *PDK1* (in the article with PMID 10856237), which has two possible gene senses ('GeneID:5163 pyruvate dehydrogenase kinase, isoenzyme 1' and 'GeneID:5170 3-phosphoinositide dependent protein kinase-1'), was incorrectly predicted as 'GeneID: 5163' when only "word" was used. But the classifiers using "CUI" and "MeSH" predicted it correctly. When the *CombSum* method was used to combine the similarity scores from all three classifiers, the correct sense 'GeneID: 5170' was selected. When all three classifiers were incorrect in predicting a testing sample, generally none of the combination methods would help in making the final decision correct. Therefore, there is an upper bound on the performance of the combined system. In our case, we detected that all three classifiers made incorrect predictions for 65 testing samples of the 2,000 samples. Therefore, the upper bound would be 1,935/2,000=96.7%.

The methods for combining different types of information from biomedical knowledge sources described in this study, though targeted to the GSD problem, could be also applicable to other text mining tasks that are based on similarity measurement, such as text categorization, clustering, and the IR task in the biomedical domain.

4.2 Coverage of the Methods

The IR-based gene symbol disambiguation method described in this paper aims to resolve intra-species gene ambiguity. We focused on ambiguous gene symbols within the human species and used articles known to be associated with human genes. Fundel and Zimmer (2006) reported that the degree of ambiguity of the human gene symbols from Entrez Gene was 3.16%–3.32%, which is substantial. However, this is only part of the gene ambiguity problem.

Based on the "gene_info" file downloaded in January 2006 from Entrez Gene, there were a total of 32,852 human genes. Based on the "gene2pubmed" file, 24,170 (73.4%) out of 32,852 human genes have at least one associated MEDLINE article, which indicates that profiles could be generated for at least 73.4% of human genes. On average, there are 9.02 MEDLINE articles associated with a particular human gene. Coverage reported in this study was relatively high because the testing samples were selected from annotated articles as listed in "gene2pubmed", and not randomly from the collection of all MEDLINE abstracts.

4.3 Evaluation Issues

It would be interesting to compare our work with other related work, but that would require use of the same testing set. For example, it is not straightforward to compare our precision result (92.2%) with that (92.7%) reported by Schijvenaars et al. (2005), because they used a testing set that was generated by removing ambiguous genes with less than 6 associated articles for each of their senses, and they did not report on coverage. The data set from the BioCreAtIvE II GN task therefore is a valuable testing set that enables evaluation and comparison of other gene symbol disambiguation methods. From the BioCreAtIvE abstracts, we identified 217 occurrences of ambiguous gene symbols, but only 124 were annotated in the BioCreAtIvE data set. There are a few possible explanations for this. First, the version of the Entrez Gene database used by the NLP system was not the most recent one, so some new genes were not listed as possible candidate senses. The second issue is related to gene families or genes/proteins with multiple sub-units. According to the 'gene_info' file, the gene symbol "IL-1" is a synonym for both "GeneID: 3552 interleukin 1, alpha" and "GeneID: 3553 interleukin 1, beta". Therefore, the NLP system identified it as an ambiguous gene symbol. When annotators in the BioCreAtIvE II task saw a gene family name that was not clearly mapped to a specific gene identifier in Entrez Gene, they may not have added it to the mapped list. In Morgan et al. (2007), it was suggested that mapping gene family mentions might be appropriate for those entities. Testing set 2 was a small set and results from that set might not be statistically meaningful, but it is useful for comparing with others working on the same data set.

In this paper, we focused on the study of improvements in precision of the gene symbol disambiguation system. When combining information from different knowledge sources, coverage may also be increased by benefiting from the cross-coverage of different knowledge sources.

5 Conclusion and Future Work

We applied an IR-based approach for human gene symbol disambiguation, focusing on a study of different methods for combining various types of information from available knowledge sources. Results showed that combination of multiple evidence usually improved the performance of gene symbol disambiguation. The combination method using coefficients obtained from a logistic regression model reached the highest precision of 92.2% on an automatically generated testing set of ambiguous human gene symbols. On a testing set derived from BioCreAtIvE II GN task, the combination method that performed summation of individual similarities reached the highest precision of 90.6%. However, the regression-based method could not be used, because the testing sample was small.

In the future, we will add information that is specifically related to genes, such as GO codes, into the combination model. Meanwhile, we will also study the performance gain in terms of coverage by integrating different knowledge sources.

Acknowledgements

This work was supported in part by Grants R01 LM7659 and R01 LM8635 from the National Library of Medicine, and Grant NSF-IIS-0430743 from the National Science Foundation. We would like to thank Alexander Morgan for providing the evaluation set from the BioCreAtIvE II GN task.

References

Aronson, A. R. 2001. *Proc. AMIA. Symp.*, 17-21.

Ashburner, M. et al. 2000. *Nat Genet*, 25, 25-29.

Black, D. 1958. *Cambridge University Press.*

Bodenreider, O. 2004. *Nucleic Acids Research*, 2004, 32, D267-D270.

Bruce, R. and Wiebe, J. 1994. *Proceedings of ACL 1994*, 139-146.

Chen, L., Liu, H. and Friedman, C. 2005. *Bioinformatics*, 21, 248-256.

Croft, W. 2002. *Advances in Information Retrieval.* Springer Netherlands, Chapter 1, 1-36

Dunn, O. J. 1964. *Technometrics*, 6, 241-252.

Erhardt, R.A., Schneider, R. and Blaschke, C. 2006. *Drug Discov. Today*, 11, 315-325.

Fox, E., Nunn, G., and Lee, W. 1988. *Proceedings of the 11th ACM SIGIR Conference* on Research and Development in Information Retrieval, 291–308.

Fox, E. and Shaw, J. 1994. *Proceedings TREC-2*, 243–252.

Friedman, M. 1937. *Journal of the American Statistical Association*, 32, 675-701.

Fundel, K. and Zimmer, R. 2006. *BMC. Bioinformatics.*, 7: 372.

Harley, A. and Glennon, D. 1997. *Proc. SIGLEX Workshop "Tagging Text With Lexical Semantics"*, 74-78.

Katzer, J., McGill, M., Tessier, J., Frakes,W., and DasGupta, P. 1998. *Information Technology: Research and Development*, 1(4):261–274.

Krallinger, M. and Valencia, A. 2005. *Genome Biol.*, 6, 224.

Lee, J. 1995. *Proceedings of the 18th ACMSIGIR Conference on Research and Development in Information Retrieval*, 180–188.

Lee, J. 1997. *Proceedings of the 20th ACM SIGIR Conference on Research and Development in Information Retrieval*, 267–276.

Lee, Y. K. and Ng, H. T. 2002. *Proc EMNLP 2002,* 41-48.

Lesk, M. 1986. *1986 SIGDOC Conference*, 24-26.

Liu, H., Johnson, S. B. and Friedman, C. 2002. *J. Am. Med. Inform. Assoc.*, 9, 621-636.

Lussier, Y., Borlawsky, T., Rappaport, D., Liu, Y., Friedman, C. 2006. *Pac. Symp. Biocomput.*, 11, 64-75.

Maglott D, Ostell J, Pruitt KD, Tatusova T. 2005. *Nucleic Acids Res.*, 3, D54-D58.

Morgan, A., Wellner, B., Colombe, J. B., Arens, R., Colosimo, M. E., Hirschman L. 2007. *Pacific Symposium on Biocomputing* 12:281-291.

Podowski, R.M., Cleary, J.G., Goncharoff, N.T., Amoutzias, G., Hayes W.S. 2004. *Proc IEEE Comput Syst Bioinform Conf*, *2004*, 415-24.

Porter,M.F. 1980. *Program*, 14, 130-137.

Salton, G. and Buckley, C. 1988. *Information Processing & Management*, 24, 513-523.

Schijvenaars, B.JA. et al. 2005. *BMC. Bioinformatics.*, 6:149.

Schuemie, M.J. et al. 2004. *Bioinformatics*, 20, 2597-2604.

Turtle, H. and Croft, W. 1991. *ACM Transactions on Information Systems*, 9(3):187–222.

Xu, H., Fan, J. W., Hripcsak, G., Mendonça A. E., Markatou, M., Friedman, C. 2007. *Bioinformatics*, doi: 10.1093/bioinformatics/btm056

Xu, J. and Croft,W. 1996. Proceedings of the 19th ACM SIGIR Conference on Research and Development in Information Retrieval, pages 4–11.