Unsupervised gene/protein named entity normalization using automatically extracted dictionaries

Aaron M. Cohen
Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, OR, USA
cohenaa@ohsu.edu

Abstract

Gene and protein named-entity recognition (NER) and normalization is often treated as a two-step process. While the first step, NER, has received considerable attention over the last few years, normalization has received much less attention. We have built a dictionary based gene and protein NER and normalization system that requires no supervised training and no human intervention to build the dictionaries from online genomics resources. We have tested our system on the Genia corpus and the BioCreative Task 1B mouse and yeast corpora and achieved a level of performance comparable to state-of-the-art systems that require supervised learning and manual dictionary creation. Our technique should also work for organisms following similar naming conventions as mouse, such as human. Further evaluation and improvement of gene/protein NER and normalization systems is somewhat hampered by the lack of larger test collections and collections for additional organisms, such as human.

1 Introduction

In the genomics era, the field of biomedical research finds itself in the ironic situation of generating new information more rapidly than ever before, while at the same time individual researchers are having more difficulty getting the specific information they need. This hampers their productivity and efficiency. Text mining has been proposed as a means to assist researchers in handling the current expansion of the biomedical knowledge base (Hirschman *et al.*, 2002). Fundamental tasks in text mining are named entity recognition (NER) and normalization. NER is the identification of text terms referring to items of interest, and normalization is the mapping of these terms to the unique concept to which they refer. Once the concepts of interest are identified, text mining can proceed to extract facts and other relationships of interest that involve these recognized entities. With the current research focus on genomics, identifying genes and proteins in biomedical text has become a fundamental problem in biomedical text mining research (Cohen and Hersh, 2005). The goal of our work here is to explore the potential of using curated genomics databases for dictionary-based NER and normalization. These databases contain a large number of the names, symbols, and synonyms and would likely enable recognition of a wide range of genes on a wide range of literature without corpus-specific training.

Gene and protein NER and normalization can be viewed as a two-step process. The first step, NER, identifies the strings within a sample of text that refer to genes and proteins. The second step, normalization, determines the specific genes and proteins referred to by the text strings.

Many investigators have examined the initial step of gene and protein NER. One of the most successful rules-based approaches to gene and protein NER in biomedical texts has been the AbGene system (Tanabe and Wilbur, 2002), which has been used by several other researchers. After training on hand-tagged sentences from biomedical text, it applies a Brill-style tagger (Brill, 1992) and manually generated post-processing rules. AbGene achieves a precision of 85.7% at a recall of 66.7% (F1 = 75%). Another successful system is GAPSCORE (Chang *et al.*, 2004). It assigns a numeric score to each word in a sentence based on appearance, morphology, and context of the word and then applies a classifier trained on these features. After training on the Yapex corpus (Franzen *et al.*, 2002), the system achieved a precision of 81.5% at a recall of 83.3% for partial matches.

For many applications of text mining, the second step, normalization is as important as the first step. Many biomedical concepts, including genes and proteins, have large numbers of synonymous terms (Yu and Agichtein, 2003, Tuason *et al.*, 2004). Without normalization, different terms for the same concept are treated as distinct items, which can distort statistical and other analysis. Normalization can aggregate references a given gene or protein and can therefore increase the sample size for concepts with common synonyms. However, normalization of gene and protein references has not received as much attention as the NER step.

One recent conference, the BioCreative Critical Assessment for Information Extraction in Biology (Krallinger, 2004), had a challenge task that addressed gene and protein normalization. The task was to identify the specific genes mentioned in a set of abstracts given that the organism of interest was mouse, fly, or yeast. Training and test collections of about 250 abstracts were manually prepared and made available to the participants along with synonym lists. Seven groups participated in this challenge task (Hirschman *et al.*, 2004), with the best F-measures ranging from 92.1% on

Proceedings of the ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics, pages 17–24, Detroit, June 2005. ©2005 Association for Computational Linguistics

yeast to 79.1% on mouse. The overall best performing system used a combination of hand built dictionaries, approximate string matching, and parameter tuning based on the training data, and performed match disambiguation using a collection of biomedical abbreviations combined with approximate string match scoring and preferring concepts with a high count of occurring terms (Hanisch *et al.*, 2004).

One thing that almost all of these systems have in common is that they need to be trained on a text corpus and/or use manually built dictionaries based on the training corpus. Since the training corpus may be a small sample of the total relevant biomedical literature, it is uncertain how the performance of these systems will change over time or when applied to other sources of biomedical text. Also, since new genes and proteins are being described all the time, it is unclear how these systems will handle genes discovered after system training is complete. This is may especially be a problem for normalization.

Dictionary-based approaches to gene and protein NER and normalization that require no training have several advantages over orthographic, lexical, and contextual based approaches. Currently there are few test collections for gene and protein normalization, and they are relatively small (Hirschman *et al.*, 2004). Unsupervised systems therefore may perform more uniformly over different data sets and over time for the near future. Since they are not dependent upon training to discover local orthographic or lexigraphic clues, they can recognize long multi-word names as easily as short forms. Dictionary-based approaches can also normalize gene and protein names, reducing many synonyms and phrases representing the same concept to a single identifier for that gene or protein.

In addition, dictionary-based approaches can make use of the huge amount of information in curated genomics databases. Currently, there is an enormous amount of manual curation activity related to gene and protein function. Several genomics databases contain large amounts of curated gene and protein name symbols as well as full names. Groups such as the Human Genome Organisation (HUGO), Mouse Genome Institute (MGI), UniProt, and the National Center for Biotechnology Information (NCBI) collect and organize information on gene and proteins, much of it from the biomedical literature, including gene names, symbols, and synonyms. Dictionary-based approaches provide a way to make use of this information for gene and protein NER and normalization. As the databases are updated by the curating organization, a NER system based on these databases can automatically incorporate additional new names and symbols. These approaches can also be very fast. Much of the computation can be performed during the construction of the dictionary. This can leave the actual searching for dictionary terms a simple and rapid process.

Tsuruoka and Tsujii recently studied the use of dictionary-based approaches for protein name recognition (Tsuruoka and Tsujii, 2004), although they did not evaluate the normalization performance. They applied a probabilistic term variant generator to expand the dictionary, and a Bayesian contextual filter with a sub-sentence window size to classify the terms in the GENIA corpus as likely to represent protein names. Overall they obtained a precision of 71.1%, at a recall of 62.3% and an F-measure of 66.6%. Tsuruoka and Tsujii did not make use of curated database information, and instead split the GENIA corpus into training and test data sets of 1800 and 200 abstracts respectively, and extracted the tagged protein names from the training set to use as a dictionary. These results compare well to, being a bit below, other non-dictionary based methods applied to the GENIA corpus (Lee *et al.*, 2004, Zhou *et al.*, 2004).

In this work we attempt to answer several questions pertaining to dictionary-based gene/protein NER:

- What curated databases provide the best collection of names and symbols?
- Can simple rules generate sufficient orthographic variants?
- Can common English word lists be used to decrease false positives?
- What is the overall normalization performance of an unsupervised dictionary-based approach?

2 Methods

A dictionary-based NER system starts out with a list, potentially very large, of text strings, called *terms*, which represent concepts of interest. In our system, the terms are organized by *concept*, in this case a unique identifier for the gene or protein. All terms for a given concept are kept together. The combination of terms indexed by concept is similar to a traditional thesaurus, and when used for NER and normalization is usually called a *dictionary*. When a term is found in a sample of text, it is a simple process to map the term to the unique gene or protein that it represents. There are several unique identifiers in use by the gene curation organizations, we chose to use the official symbol as a default, but it is easy to use other database identifiers as needed.

2.1 Building the dictionary

Building the initial dictionary is an essential first step in dictionary-based NER. The dictionaries we used in this study were built automatically from five databases available for download: MGI, Saccharomyces, UniProt (the curated SwissProt portion only), LocusLink, and the Entrez Gene database. For each of these databases, the official symbol, unique identifiers, name, symbol, synonym, and alias fields were extracted. Symbols, synonyms, and aliases corresponding to the same official symbol were combined into a single list. At this stage in dictionary generation, any leading or trailing white space characters are removed. The original capitalization of each term is kept. This will be important in a later step

Like several other investigators (Tanabe and Wilbur, 2002, Chang *et al.*, 2004), we do not discriminate between the names of genes and the proteins that they code for. For many text mining purposes, recognizing a mention of a gene or the coded protein has been treated as equivalent (Cohen and Hersh, 2005). Therefore, combining terms corresponding to the same official symbol is justified, even if one database is composed of genes and the other proteins.

2.2 Generating orthographic variants

Our previous work on gene and protein name synonyms (Cohen *et al.*, 2005) led us to make the observation that many name synonyms are simple orthographic variants of each other, and that most of these variants can be generated with a few simple rules. The next step in dictionary generation is to generate variant terms for each term extracted from the downloaded databases.

Our system uses seven simple rules to generate variants:

(1) If the original term includes internal spaces, these can be replaced by hyphens (e.g., “IL 10” to “IL-10”).

(2) If the original term includes internal hyphens, these can be replaced by spaces (e.g., “mmac-1” to “mmac 1”).

(3) If the original term includes internal spaces or hyphens, these can be removed (e.g., “nf-kappa b” to “nfkappab”).

(4) If the original term ends in a letter followed by single digit, or a letter followed by single digit and then a single letter, a hyphen can be added before the digit (e.g., “NFXL1” to “NFXL-1”).

(5) If the original term ends in a digit, followed by the single letter ‘a’ or ‘b’, we can add a hyphen before the ‘a’ or ‘b’ and also expand ‘a’ to ‘alpha’ and ‘b’ to ‘beta’ (e.g., “epm2b” to “epm2-beta”).

(6) If the original term ends in ‘-1’ or a ‘-2’, replace this ending with the Roman numeral equivalent, ‘-i’ or ‘-ii’ respectively.

(7) For yeast only, if the original term consists of one space-delimited token, append a “p” (see (Cherry, 1995)).

These rules are applied iteratively until no new terms are generated.

2.3 Separating common English words

The next step aids in discriminating mentions of gene and protein names from common English words. The dictionary now contains a large number of terms extracted from the databases along with generated variants. At this point the dictionary is split into two parts. Terms that case-insensitively match a list of common English words are put into the one dictionary, and other terms are put into a separate dictionary.

In practice, this creates a small dictionary of terms easy to confuse with common English words (the *confusion* dictionary) and a much larger dictionary of terms that are not confused with English words (the *main* dictionary). When searching text for gene and protein names, the terms in the smaller dictionary will be handled differently than the terms in the larger dictionary.

For the work presented here, a file of 74,550 common English words was used to filter the terms. This file is available as part of the Moby lexical resource, and is available at (Ward, 2000).

2.4 Screening out the most common English words

Some English words are so common that when they occur they are rarely references to gene and protein names. Our approach includes a list of about 300 English words that is used as a “stop” list. In our system these words are never recognized as gene or protein terms, even if those terms appear in one of the curated databases.

We obtained our list of the 300 most common words in the English language (Carroll *et al.*, 1971). To this list we added a few terms that are commonly found in the biomedical literature that should not be confused with specific gene names. These include “gene”, “genes”, “protein”, “proteins”, ”locus”, ”site”, “alpha”, “beta”, and “as a”.

Terms appearing in this most common word list are removed from both of the dictionaries. The final product of the four preceding steps are two dictionaries, a main dictionary and a confusion dictionary, each which map terms to the unique identifier for the gene/protein symbol corresponding to that term.

2.5 Searching the text

With the two dictionaries complete it is straightforward to search input text for mentions of gene and protein names. While the algorithm can handle practically any size input text, in practice the input will usually be individual sentences or abstracts, and this is the input size to which we have tuned our system.

For speed and accuracy, we first search the input text for the terms within the dictionary, and if a term is found, we then check to ensure that the matching text is bounded by characters that are acceptable delimiters for gene and protein names. In our system this includes white space characters as well as these characters: .,/\\(){}[]=;?*!". Note that our approach does not prohibit these characters from appearing within the name, only that the matching sequence of characters is bounded by these delimiters. Also, the approach does not require tokenization of the input string. We consider this more flexible than delimiter-based tokenization, which would not allow delimiters to appear within the terms.

This method of searching and checking delimiters is applied for every term in both the main and confusion dictionaries with one essential difference. Case-insensitive search is performed on the terms in the main dictionary. Strict case-sensitive search is performed on terms in the confusion dictionary. This requires terms in the confusion dictionary to exactly match the capitalization of the input text. The observation here is that a string like “dark” appearing in biomedical text is most often being used as a normal English word, while a string like “DARK”, is likely being used as a gene name.

Finally, the algorithm examines all matching terms on the input text. Overlaps are resolved with a combination of criteria based on comparing the confidence and length of each recognized entity. In the current implementation, the confidence of the dictionary-based NER is always 1.0, so in practice the system resolves overlap by keeping the entity recognized by the longest overlapping term and discarding any shorter overlapping entities.

2.6 Disambiguation

It has been shown that a large number of gene and protein terms refer to more than one actual concept with over 5% of terms being ambiguous intra-species and 85% being ambiguous with gene names for other organisms (Tuason *et al.*, 2004, Chen *et al.*, 2005). For normalization, occurrences of these ambiguous terms need to be resolved to the correct concept. This is called *disambiguation*.

Various disambiguation approaches have been proposed, including the method of Hanisch previously described, as well as simply ignoring ambiguous terms. Ignoring all ambiguous terms can be wasteful, since context may allow disambiguation to a unique concept. This can helpful for increasing the sample size for further text mining. For example NER and normalization can be performed on abstracts, and further processing (e.g., co-occurrence detection) performed at the sentence level. Our approach to disambiguation makes two assumptions about the biomedical literature. First, ambiguous terms are often synonyms for other, non-ambiguous terms within the same text sample, and second, authors usually explicitly provide sufficient context for readers to resolve ambiguous terms.

For each ambiguous term, we collect the potential normalized concepts. If any of those concepts appears in the text sample using an unambiguous term for that concept, we assign the ambiguous term to the concept with the unambiguous term. If there is more than one concept with an unambiguous term (this occurs infrequently), we select one of these concepts at random. We ignore terms that cannot be resolved in this manner. Notice that this is a general dictionary disambiguation algorithm and does not require any information specific to genes and proteins.

2.7 Optimization

One of the benefits of the dictionary-based approach is that it is simple and amenable to code optimization. In our case we were able to gain almost a thousand-fold speed improvement over brute force searching against every term in the database. We accomplished this using an approach based on indexing the term prefixes, taking each unique sequence of *n* initial term characters as the index for all terms with that initial sequence. In our system we chose an *n* of 6 as a good balance between performance and memory requirements.

Searching for gene and protein terms then becomes an efficient matter of only searching for the terms that correspond to 6 character sequences (prefixed by a delimiter) that actually exist in the input text. This greatly reduces the number of searching operations necessary. While other more complex optimization algorithms are possible, such as organizing the terms character-by-character into an n-way tree, or completely grouping the terms into a complete prefix tree, our approach is simple, very fast, and has modest memory needs.

3 Evaluation

We based our evaluation on two test corpora that have been previous used to evaluate gene and protein NER and normalization. We used the GENIA corpus, version 3.02 (Kim *et al.*, 2003), to evaluate the utility of each online database as a source of terms for gene and protein NER, and we used the BioCreative Task1B mouse and yeast collections to evaluate the performance of our system for normalized gene and protein identification.

The GENIA corpus is a key resource in biomedical text mining, and has been used by many investigators (e.g., (Collier and Takeuchi, 2004, Lee *et al.*, 2004, Tsuruoka and Tsujii, 2004)). However, some system-dependent decisions still need to be made in order to use it as a gold standard for gene and protein NER. First, GENIA marks genes separately from proteins. While the "protein_molecule" attribute appears to be used in a manner that tightly and specifically delimits mentions of proteins, other attributes such as the "DNA_domain_or_region" attribute and the "protein_family_or_group" attribute are used more loosely. "DNA_domain_or_region" can be used to mark a specific gene (e.g., "IL-2 gene", "peri kappa-B site"), sometimes including words such as "gene" and "site". At other times the attribute marks a non-specific gene concept (e.g., "viral gene"). Similar observations are true about the "protein_family_or_group" attributes (e.g., "CD28", "transcription factor"). Clearly when evaluating dictionary-based (possibly as opposed to corpus trained) gene/protein NER, many of the concepts marked with the "DNA_domain_or_region", "protein_family_or_group" and other similar attributes should be treated as correct for the purposes of precision. However, the large number of more generic concepts that these attributes mark should not be included in the calculation of recall.

Because of these issues, here we have used a hybrid technique in order to produce the most meaningful results in choosing a database for wide coverage of gene and protein names and symbols. Entities marked with the "protein_molecule" attribute are included for computation of both precision and recall. The text marked with the DNA and protein family attributes are only used for the computation of precision. This method is different from that applied by others using the GENIA corpus for both training and testing and therefore our NER results here are not directly comparable to prior work using GENIA.

In the first set of experiments we are primarily concerned with evaluating the richness of each database and combination of databases as a source of names for gene and protein NER. Therefore, we use the weak match criteria of Chang et al., to evaluate performance (Chang *et al.*, 2004). The weak match criteria treats any overlap of identified text with the gold standard as a positive.

In the second set of experiments we use the BioCreative mouse and yeast test collections to evaluate the performance of our unsupervised dictionary-based method of gene and protein NER and normalization. For mouse, the more challenging organism, we evaluate the effect of each system feature separately and in combination. We also evaluate the effect of using just the organism-specific database to populate the dictionary, along with the organism-specific database in combination with the richest database determined in the first set of experiments. Table 1 shows information on the databases that were used to generate the dictionaries and the fields taken from each database.

4 Results

Table 2 presents the results of applying our dictionary-based NER to the GENIA 3.02 corpus using the three multi-organism databases individually. The Entrez Gene database performs the best, having both the highest F-measure of 75.5% at a precision of 73.5% and a recall of 77.6%. The LocusLink database is next, and not significantly different in performance (LocusLink is being phased out and replaced with Entrez Gene as of March 2005). The UniProt database performs much worse overall. This is surprising, performing well on precision at 78.5%, but having recall of 59.1%, poorer than we expected for a multi-species database.

Table 1. Databases used to create protein/gene NER dictionaries. Fields marked with an asterisk were used as the unique identifier.

Database & Organism	**Fields used**	**Dictionary Size**
Entrez multi-organism	SYMBOL*, SYNONYMS, DESCRIPTION	59 Mbytes
LocusLink multi-organism	PRODUCT, OFFICIAL_SYMBOL*, PREFERRED_SYMBOL, OFFICIAL_GENE_NAME, PREFERRED_GENE_NAME, PREFERRED_PRODUCT, ALIAS_SYMBOL, ALIAS_PROT	14 Mbytes
MGI mouse only	MGI MARKER ACCESSION ID*, MGI GENE TERM, STATUS	7 Mbytes
UniProt multi-organism	Name*, Synonyms, OrderedLocusNames, ORFNames	5 MBytes
Saccharomyces yeast only	Locus, ORF, SGID*, alias, standard name, feature name	1.5 MBytes

Table 2. Results of creating dictionary from a single database for NER of GENIA genes and proteins.

Dictionary	**Precision**	**Recall**	**F-measure**
Entrez	0.735	0.776	0.755
LocusLink	0.723	0.773	0.747
UniProt	0.785	0.474	0.591

Table 3. Results of creating dictionary from a combination of two databases for NER of GENIA genes and proteins.

Dictionaries	Precision	Recall	F-measure
Entrez	0.735	0.776	0.755
Entrez+UniProt	0.707	0.792	0.747
Entrez+LocusLink	0.734	0.780	0.756

Table 4. Results of using dictionary created from databases for NER and normalization for mouse.

Dictionary	Precision	Recall	F-measure
Entrez/MGI	0.775	0.726	0.750
MGI	0.710	0.535	0.610

Having found that Entrez Gene was the single best online database for dictionary creation, we tried combining it with the other databases. As can be seen from Table 3, this did not result in any meaningful performance improvement.

For the remainder of our experiments we used the BioCreative mouse and yeast test collections and gold standard files to evaluate the performance of our system for gene/protein NER and normalization. The gold standard required the unique identifiers to be MGI or SGD accession numbers. To accomplish this, we performed a join between the Entrez database and the MGI (or Saccharomyces) database using a mapping identifier between the MGI (or SGI) database entries and the Entrez Gene ids while extracting dictionary terms.

Table 4 shows the results of using the joined Entrez/MGI dictionary for mouse NER and normalization compared to using the dictionary created from the MGI database alone. Using the MGI database alone has much worse recall than using the dictionary created with a combination of Entrez and MGI databases, with recall falling almost 20%. Restricting the dictionary to the MGI database also results in a 6.5% decrease in precision.

Table 5 shows the results of individually removing each of the three main dictionary pre-processing features and the disambiguation algorithm and evaluating the NER and normalization performance for mouse. All four of these variations perform worse than our full system. Variant generation made the smallest difference, giving an F-measure improvement of 2.0%. Ambiguity resolution improves the F-measure 2.8%. The 300 most common word stop list contributed an improvement of 6.8%. Lastly, separation into case-sensitive and case-insensitive dictionaries made the largest improvement of 15.6%. Removing all of the pre-processing features at once and using the combined Entrez/MGI database as a "raw" term list performs very badly, with good recall but a precision of only 30.1%.

Table 6 compares the results of our system to the participants of BioCreative Task 1B for the mouse and yeast corpora. On both mouse and yeast, our system performs above the median F-measure. On mouse the difference in F-measure between our system and the top scoring system is less than 5%. On the yeast corpus, our approach has among the highest precision, with recall slightly below the median, and F-measure about 3% below the highest scoring system.

While ambiguity resolution resulted in a modest improvement, we wanted to get an idea of the magnitude of the ambiguity within our automatically created dictionaries. Table 7 shows the number and percentage of ambiguous terms and genes with at least one ambiguous term in the dictionaries that we created using Entrez in combination with the MGI database, as well as MGI alone.

The system runs very rapidly. On a 1.7GHz Pentium 4m laptop with 512M RAM, the 18,000 sentences in the GENIA corpus were processed in about 30 seconds. The 250 abstracts in the BioCreative corpora were processed in less than 5 seconds.

5 Discussion

The Entrez Gene database was identified as the best general-purpose source of gene and protein terms for use in a dictionary-based NER and normalization. Including data from other databases did not improve NER performance. It appears that the producers of Entrez Gene are doing an excellent job in finding and curating this information from the available sources. One of the most common difficulties cited in recognizing gene and protein names is that the vocabulary of terms is continuously expanding (Hirschman *et al.*, 2002). Online databases, such as Entrez Gene provide a curated central repository for these terms, making the task of keeping gene/protein NER and normalization systems up to date on new genes and proteins somewhat easier.

All three of our dictionary pre-processing enhancements improved performance, as did the ambiguity resolution algorithm. Surprisingly, variant generation made the smallest difference in F-measure. This may be due to the tendency for genes to be mentioned multiple times within an abstract, or that authors are keeping to the forms collected in the genomics databases, or that the database curators are doing a good job in keeping up with the terms used by authors. The BioCreative test collection scores normalization at the level of an entire abstract. It is possible that variant generation might have made a larger difference if the test collection was scored at a sentence level. On the other hand, it may be that the

Entrez database itself contains sufficient variants. In either case, the small improvement gained from variant generation suggests that computationally expensive approximate string matching techniques may not be worth the effort.

The next largest improvement was made by ambiguity resolution. Precision increased almost 8%, while recall dropped only about 2%. While an F-measure improvement of 2.8% is small, this figure is highly dependent upon the make up of the test corpus.

Certainly, as seen in Table 7, there are a large proportion of mouse genes with ambiguous terms in our dictionary. How often these ambiguous terms actually appear in the literature is an open question. Additional and larger test collections may be necessary to accurately measure the overall importance of ambiguity resolution.

Table 5. NER and normalization performance results when removing dictionary pre-processing features and ambiguity resolution for mouse.

System	Precision	Recall	F-measure	Difference
full system	0.775	0.726	0.750	-
- case	0.493	0.746	0.594	-15.6%
- stop	0.643	0.726	0.682	-6.8%
- variant	0.771	0.693	0.730	-2.0%
- ambiguity	0.697	0.748	0.722	-2.8%
- all	0.301	0.713	0.423	-32.7%

Table 6. Comparison with results from BioCreative on mouse and yeast corpora.

Organism	System	Precision	Recall	F-measure
Mouse	biocreative-highest	0.765	0.819	0.791
	cohen	0.775	0.726	0.750
	biocreative-median	0.765	0.730	0.738
	biocreative-lowest	0.418	0.898	0.571
Yeast	biocreative-highest	0.950	0.894	0.921
	cohen	0.950	0.837	0.890
	biocreative-median	0.940	0.848	0.858
	biocreative-lowest	0.661	0.902	0.763

Table 7. Term ambiguity measurements for mouse genes.

	Entrez/MGI	MGI
# All Distinct Genes	57185	57180
# All Distinct Terms	336353	250435
# Ambiguous Terms	6585 (1.96%)	2104 (0.84%)
# Genes w/ Ambiguous Terms	8036 (14.05%)	2619 (4.58%)

The stop-list made the next largest improvement in F-measure, 6.8%. Use of the stop list improved precision greatly and did not change recall. Case-sensitivity using the common word file made the largest improvement of 15.6%. While making a large, almost 30% difference in precision, case sensitivity decreased recall by only 2%.

Overall, all three of the dictionary pre-processing methods we applied worked well, as did ambiguity resolution. Each method resulted in improvement in either precision or recall, and did not greatly degrade the other measure. Together the three techniques gave an F-measure improvement of over 30% as compared to using a plain unprocessed dictionary.

Chen et al. investigated the ambiguity of official gene names within and across organisms and found the level of shared official names within an organism to be low (~0.02%) but the level of ambiguity when considering all terms associated with a gene to be higher, about 5% (Chen *et al.*, 2005). Our results are similar, with about 5% of genes in the MGI database having terms also associated with other genes. This rises to 14% when combined with the information in the Entrez database. As previously noted, inter-organism ambiguity is much higher. Further work is needed to determine the extent of the problem present within in the actual literature.

We did not apply our method to fly, the other organism in the BioCreative Task 1B test collection. We were unable to find direct mappings between identifiers in the fly database and Entrez Gene. Moreover, the fly corpus would present special problems for our method. Unlike for mouse and yeast, the fly genome contains many genes that have the same names as common English words, and the use of these words as gene names are not commonly delineated using capitalization as they are with mouse. For fly at least, methods such as ours are at a disadvantage compared to trained systems.

However, the literature of one of the most important and interesting genomes (at least to us), human, does appear to follow the practice of differentiating common English words from gene and protein names by uppercase or initial capitalization similar to the mouse literature (Chen *et al.*, 2005). Therefore we expect that our unsupervised approach will be useful for human genomics literature as well.

Unfortunately at the present time we are unable to test this hypothesis. We are unaware of any human gene NER and normalization test collection. While there are several test collections widely available for NER alone (Franzen *et al.*, 2002, Kim *et al.*, 2003, Hu *et al.*, 2004), the same cannot be said for the essential normalization step. More and larger collections, covering additional organisms such as human and rat, are necessary to measure and motivate progress in gene and protein NER and normalization.

6 Conclusions and Future Work

These results demonstrate that an unsupervised dictionary-based approach to gene and protein NER and normalization can be effective. The dictionaries can be created automatically without human intervention or review. Dictionary-based systems such as ours can be set up to automatically update themselves by downloading the database files on the Internet and pre-processing the files into updated dictionaries. This could be done on a nightly basis if necessary, since the entire dictionary creation process only takes a few minutes. One general database, combined with an organism-specific database for each species, is sufficient.

Our work is distinguished from other dictionary-based work such as Tsurukoka and Tsujii, and Hanisch et al. in several ways. Unlike both of these prior investigators, we use on-line curated information as our primary source of terms, instead of deriving them from a training set, and have shown both which databases to use and how to process them into effective sources of terms for NER. Our textual variants are generated by simple rules determined by domain knowledge instead of machine learning on training data. Lastly, the disambiguation algorithm presented here is unique and has been shown to have a positive impact on performance.

The system is as accurate as other more complex approaches. It does not require training, and so may be less sensitive to specific characteristics of a given text corpus. It may also be applied to organisms for which there do not exist sufficient training and test collections. In addition, the system is very fast. This may enable some text mining tasks to be done for users in real time, rather than the batch processing mode that is currently most common in biomedical text mining research.

Dictionary-based approaches are likely to remain an essential part of gene and protein normalization, even if the NER step is handled by other methods. Further work is necessary to determine the best manner to combine automatically created dictionaries with trained NER systems. It may be the case that different approaches work best for different organisms, depending upon the specific naming conventions of scientists working on that species.

References

Brill, E. (1992) A simple rule-based part of spech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*.

Carroll, J. B., Davies, P. and Richman, B. (1971) The American heritage word frequency book. Houghton Mifflin, Boston,.

Chang, J. T., Schutze, H. and Altman, R. B. (2004) GAPSCORE: finding gene and protein names one word at a time. *Bioinformatics*, **20**, 216-25.

Chen, L., Liu, H. and Friedman, C. (2005) Gene name ambiguity of eukaryotic nomenclatures. *Bioinformatics*, **21**, 248-56.

Cherry, J. M. (1995) Genetic nomenclature guide. Saccharomyces cerevisiae. *Trends Genet*, 11-2.

Cohen, A. M. and Hersh, W. (2005) A Survey of Current Work in Biomedical Text Mining. *Briefings in Bioinformatics*, **6**, 57-71.

Cohen, A. M., Hersh, W. R., Dubay, C. and Spackman, K. (2005) Using co-occurrence network structure to extract synonymous gene and protein names from MEDLINE abstracts. *BMC Bioinformatics*, **6**,

Collier, N. and Takeuchi, K. (2004) Comparison of character-level and part of speech features for name recognition in biomedical texts. *J Biomed Inform*, **37**, 423-35.

Franzen, K., Eriksson, G., Olsson, F., Asker, L., Liden, P. and Coster, J. (2002) Protein names and how to find them. *Int J Med Inf*, **67**, 49-61.

Hanisch, D., Fundel, K., Mevissen, H. T., Zimmer, R. and Fluck, J. (2004) ProMiner: Organism-specific protein name detection using approximate string matching. In *BioCreative: Critical Assessment for Information Extraction in Biology*.

Hirschman, L., Morgan, A. A. and Yeh, A. S. (2002) Rutabaga by any other name: extracting biological names. *J Biomed Inform*, **35**, 247-59.

Hirschman, L., Colosimo, M., Morgan, A., Columbe, J. and Yeh, A. (2004) Task 1B: Gene List Task BioCreAtIve Workshop. In *BioCreative: Critical Assessment for Information Extraction in Biology*.

Hu, Z. Z., Mani, I., Hermoso, V., Liu, H. and Wu, C. H. (2004) iProLINK: an integrated protein resource for literature mining. *Comput Biol Chem*, **28**, 409-16.

Kim, J. D., Ohta, T., Tateisi, Y. and Tsujii, J. (2003) GENIA corpus - a semantically annotated corpus for bio-textmining. *Bioinformatics*, **19**, i180-i182.

Krallinger, M. (2004) BioCreAtIvE - Critical Assessment of Information Extraction systems in Biology. http://www.pdg.cnb.uam.es/BioLINK/BioCreative.eval.html

Lee, K. J., Hwang, Y. S., Kim, S. and Rim, H. C. (2004) Biomedical named entity recognition using two-phase model based on SVMs. *J Biomed Inform*, **37**, 436-47.

Tanabe, L. and Wilbur, W. J. (2002) Tagging gene and protein names in biomedical text. *Bioinformatics*, **18**, 1124-32.

Tsuruoka, Y. and Tsujii, J. (2004) Improving the performance of dictionary-based approaches in protein name recognition. *J Biomed Inform*, **37**, 461-70.

Tuason, O., Chen, L., Liu, H., Blake, J. A. and Friedman, C. (2004) Biological nomenclatures: a source of lexical knowledge and ambiguity. *Pac Symp Biocomput*, 238-49.

Ward, G. (2000) Grady Ward's Moby. http://www.dcs.shef.ac.uk/research/ilash/Moby/mwords.html

Yu, H. and Agichtein, E. (2003) Extracting synonymous gene and protein terms from biological literature. *Bioinformatics*, **19**, i340-i349.

Zhou, G., Zhang, J., Su, J., Shen, D. and Tan, C. (2004) Recognizing names in biomedical texts: a machine learning approach. *Bioinformatics*, **20**, 1178-90.