HLT-NAACL 2004 Workshop: Biolink 2004,
Linking Biological Literature, Ontologies and Databases, pp. 25-32.
Association for Computational Linguistics.

A Study of Text Categorization for Model Organism Databases

Hongfang Liu
University of Maryland at Baltimore County
hfliu@umbc.edu

Cathy Wu
Georgetown University Medical Center
wuc@georgetown.edu

Abstract

One of the routine tasks for model organism database curators is to identify and associate research articles to database entries. Such task can be considered as text categorization which has been studied in the general English domain. The task can be decomposed into two text categorization subtasks: i) finding relevant articles associating with specific model organisms, and ii) routing the articles to specific entries or specific areas. In this paper, we investigated the first subtask and designed a study using existing reference information available at four well-known model organism databases and investigated the problem of identifying relevant articles for these organisms. We used features obtained from abstract text and titles. Additionally, we studied the determination power of other MEDLINE citation fields (e.g., Authors, MeshHeadings, Journals). Furthermore, we compared three supervised machine learning techniques on predicting to which organism the article belongs.

1 Introduction

With the accelerated accumulation of genetic information associated with popularly used genetic model organisms for the human genome project such as laboratory mouse, C. elegans, fruit fly, and Saccharomyces, model organism databases that contain curated genetic information specifically to the associated organism have been initialized and evolved to provide a central place for researchers to seek condense genetic information (Flybase, 2003; Blake, 2003; Misra, 2003). At the same time, a rich amount of genetic and biomedical information associated with these model organisms are published through scientific literature. One of the routine curation tasks for the database curators is to associate research articles to specific genetic entries in the databases and identify key information mentioned in the articles. For example, a regular practice of mouse genetic database curators is to scan the current scientific literature, extract and enter the relevant information into databases (Blake, 2003). In Saccharomyces Genome Database (SGD, *Saccharomyces cerevisiae*: http://www.yeastgenome.org), database curators are currently in the process of revising the information associated with the Description field of an entry to ensure that the Description (which usually is a concise summary of the function and biological context of the associated entry) contains the most up-to-date information and is written in a consistent style. One of the current objectives of WormBase (http://www.wormbase.org) is to systematically curate the C. elegans literature. However, manually scanning scientific articles is a labor intensive task. Meanwhile, the outcome may be incomplete, i.e., curators may miss some critical papers. Additionally, more than 2000 completed references are added daily to MEDLINE alone. It seems impossible to be always up-to-date.

The task of associating research articles with specific entries can be decomposed into two subtasks: i) categorizing articles into several categories where articles with the same category are about the same model organism, and ii) associating the articles to specific entries or specific areas. Finding relevant articles specific to a particular model organism is a case of information retrieval. A simple way to retrieve relevant articles about a model organism is to retrieve articles containing terms that represent that organism. For example, if the term “C. elegans” appears in a paper, most

likely, the paper is relevant to C. elegans. Another way is to apply supervised machine learning techniques on a list of category-labeled documents. In this paper, we designed a study on retrieving relevant articles using MEDLINE reference information obtained from four model organism databases aiming to answer the following questions:

- Can we just use keywords to retrieve relevant MEDLINE references instead of using complicated machine learning techniques?
- How accurate is the retrieval when we use various MEDLINE fields such as Authors, MeshHeadings etc to retrieve articles? Which kind of feature representations has the best performance?
- Which kind of machine learning algorithm is suitable for categorizing the articles to the appropriate categories?
- How good is the MEDLINE citation information when we used category-labeled documents obtained in the past to predict the category of new documents?

In the following, we first provide background information about applying supervised machine learning techniques on text categorization. We then describe materials and methods. Finally, we present our results and discussions.

2 Background and Related Work

Using keywords to retrieve relevant articles is to use a list of keywords to retrieve articles containing these keywords. The main component here is to derive a list of keywords for each category. Using supervised machine learning techniques to retrieve relevant articles requires a collection of category-labeled documents. The objective is to learn classifiers from these category-labeled documents. The construction process of a classifier given a list of category-labeled documents contains two components. The first component transfers each document into a feature representation. The second component uses a supervised learning algorithm to learn classification knowledge that forms a classifier.

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

Several supervised learning algorithms have been adapted for text categorization: Naïve Bayes learning (Yang and Liu, 1999), neural networks (Wiener, 1995), instance-based learning (Iwayama and Takunaga, 1995), and Support vector machine (Joachims, 1998). Yang and Liu (1999) provided an overview and a comparative study about different learning algorithms. In previous studies of applying supervised machine learning on the problem of word sense disambiguation, we investigated and implemented several supervised learning algorithms including Naïve Bayes learning, Decision List learning and Support Vector Machine for word sense disambiguation. There is not much difference between word sense disambiguation task and text categorization task. We can formulate a word sense disambiguation task as a text categorization task by considering senses of a word as categories (Sebastiani, 2002). We can also formulate a text categorization task by considering there is a hidden word (e.g., TC) in the text with multiple senses (i.e., categories). Note that in word sense disambiguation task, one occurrence of a word usually holds a unique sense. While for text categorization task, sometimes one document can be in multiple categories. After verifying that there were less than 1% of documents holding multiple categories (shown in detail in the following section), for simplicity, we applied the implementations of supervised machine learning algorithm (used for word sense disambiguation) directly for text categorization by considering the disambiguation of a hidden word (TC) in the context. The following summarizes the algorithms used in the study. For detail implementations of these algorithms, readers can refer to (Liu, 2004).

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

prior probabilities for different categories as well as the probabilities of each category for each feature.

The **Decision List method** (DLL) (Yarowsky, 1994) is equivalent to simple case statements in most programming languages. In a DLL classifier, a sequence of tests is applied to each feature vector. If a test succeeds, then the sense associated with that test is returned. If the test fails, then the next test in the sequence is applied. This continues until the end of the list, where a default test simply returns the majority sense. Learning a decision list classifier consists of generating and ordering individual tests based on the characteristics of the training data.

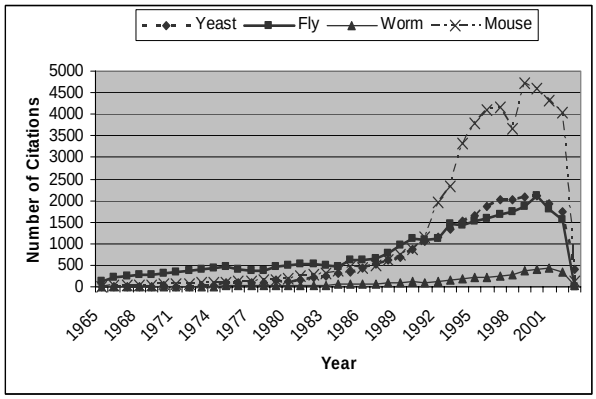
Support vector machine (SVM) (Vapnik, 1998) is a supervised learning algorithm proposed by Vladimir Vapnik and his co-workers. For a binary classification task with classes {+1, –1}, given a training set with n class-labeled instances, (x1, y1), (x2, y2), ..., (xi, yi), …, (xn, yn), where xi is a feature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

Using a list of keywords to retrieve relevant articles has been used frequently for NLP systems in the biological domain. For example, Iliopoulos et al. (2001) used keywords pertinent to a biological process or a single species to select a set of abstracts for their system. Supervised machine learning has been used by Donaldson et al. (2003) to recognize abstracts describing bio-molecular interactions. The training articles in their study were collected and judged by domain experts. In our study, we compared keywords retrieving with supervised machine learning algorithms. The category-labeled training documents used in our study were automatically obtained from model organism databases and MEDLINE.

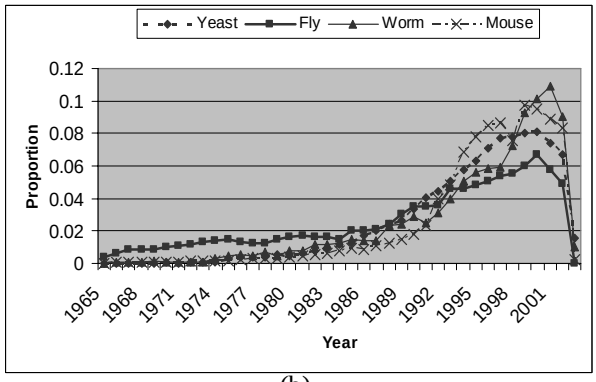
3 Material and Methods

3.1 Model Organism Databases

The research done here is based on MEDLINE references associated with four model organisms



(a)



(b)

Figure 1. References for four organism databases from 1966 to 2002. X-axis represents years from 1965 to 2003 in ascending order. The Y axis in figure (a) represents the number of citations. The Y-axis in figure (b) represents the proportion of each year comparing to the total number of citations for a specific organism.

(i.e., mouse, fly, worm and yeast) obtained from Mouse Genome Informatics (MGD, *Mus musculus*: http:// www.informatics.jax.org, FlyBase (*Drosophila melanogaster*. http://www.flybase.org), WormBase (*Caenorhabditis elegans*: http://www.wormbase.org), and Saccharomyces Genome Database (SGD, *Saccharomyces cerevisiae*: http://www.yeastgenome.org). We downloaded literature reference information from each database on March 2003. All databases provide PMID (unique identifier for MEDLINE citations) information except WormBase where some references use MEDLINEID (another unique identifier for MEDLINE citations) as reference identifiers, some references use PMID as reference identifiers. Meanwhile, about two thirds of the references in WormBase do not have reference identifiers to MEDLINE, which we eliminated in our study since we were not able to get the MEDLINE

citation information. We then used e-Fetch tools provided by Entrez (http://entrez.nlm.nih.gov) and fetched complete MEDLINE citations in XML format from MEDLINE.

Finally, we obtained 31,414 MEDLINE citations from Flybase, 26,046 from SGD, 3,926 from WormBase, and 48,458 from MGD. Figure 1 lists the statistical information according to the publication date for each organism, where X-axis represents year, Y-axis in Fig. 1(a) represents the number of citations and Y-axis in Fig. 1(b) represents the percentage of the number of citations to the total number of citations for each organism. Note that there were 1,005 citations holding multiple categories (15 of them were referred by mouse, fly and yeast, 1 referred by fly, worm, and mouse, 338 referred by mouse and yeast, 282 referred by fly and yeast, 310 referred by fly and mouse, 9 referred by worm and yeast, 36 referred by fly and worm, 5 referred by mouse and worm). However, comparing to the total of 109,844 citations, there were less than 1% of citations with multiple categories. For simplicity, we defined our categorization task as a single category text categorization task.

3.2 Methods

We studied Taxonomy from NCBI (http://www.ncbi.nlm.nih.gov) and UMLS knowledge sources (http://umlsks.nlm.nih.gov) and derived a list of keywords for each organism and used them to retrieve relevant articles. If the title, the abstract or Mesh Headings of a MEDLINE citation contains these keywords, we considered it as a relevant article. Table 1 shows the list of keywords we obtained for each model organism.

MEDLINE citations also contain other information such as Authors, Mesh Headings, and Journals etc besides abstracts and titles. Based on the intuitions that biologists tend to use the same organism for their research and a specific journal tend to publish papers in a limited number of areas, we also evaluated Authors and Journals as features. Additionally, Mesh Headings which were assigned manually by librarians to index papers represent key information of the papers, we also evaluated the categorization power of Mesh Headings in determining which organism the paper belongs to. We then combined some or all features together and evaluated the prediction power.

Year	Tra	Te	AbT	ArT	Aut	Jou	MH
1990	21563	2963	251	0	0	11	3
1991	24526	3394	250	0	2	21	0
1992	27920	4355	300	0	1	8	2
1993	32275	5267	391	0	1	46	0
1994	37542	6474	450	0	3	34	1
1995	44016	7178	490	0	4	12	1
1996	51194	7782	613	0	2	3	4
1997	58976	8115	593	0	4	8	10
1998	67091	7726	519	0	2	8	12
1999	74817	9057	599	0	10	15	23
2000	83874	9234	587	0	4	16	28
2001	93108	8479	362	0	4	16	309
2002	101587	7688	237	0	6	7	1366
2003	109275	569	6	0	0	0	146
Total	NA	88281	5647	0	43	206	1905

Table 1. The number of citations for training (Tra), testing (Te) for each year. Note that some fields in certain MEDLINE citations may be empty (e.g., not all references have abstracts), the number of these non-applicable citations for feature representations abstracts (AbT), titles (ArT), authors (Aut), Journals (Jou), and Mesh Headings (MH) for each year.

MOUSE	Mouse, mice, mus muscaris, mus musculus, mus sp
YEAST	Saccharomyces, yeast, yeasts, candida robusta, oviformis, italicus, capensis, uvarum, erevisiae
FLY	drosophila, fly, flies
WORM	Elegans, worm, worms

Table 2. Keywords used to retrieve relevant articles for four model organisms mouse, yeast, fly and worm.

3.3 Experiments

For each year from 1990 to 2003, we trained a classifier using citations published in all previous years and tested using citations in the current year. Table 2 lists the detail about the training set and the test set for each year. We experimented the following feature representations: stemmed words from AbstractText, stemmed words from Title,

Author, MeshHeading, and Journals. Since some of the MEDLINE fields may be empty (such as some citations do not contain abstracts), Table 2 also provides the number of non-applicable references each year for a given feature representation method. From Table 2, we found that every citation has a title. However, there are about 6.4% of citations (5,647 out of 88,281) that do not have abstracts. For each feature representation, we applied three supervised learning algorithms (i.e., Naïve Bayes learning, Decision List learning, Support Vector Machine).

For each combination of machine learning algorithm and feature representation, we computed the performance using the F-measure, which is defined as 2*P*R/(P+R), where P is the precision (the number of citations predicted correctly to the total number of citations being predicted) and R is the recall (the number of citations predicted correctly to the total number of citations).

We then sorted the feature representations according to their F-measures and gradually combined them into several complex feature representations. The feature vector of a complex feature representation is formed by simply combining the feature vector of its members. For example, suppose the feature vector of feature representation using stemmed words from the title contains an element **A** and the feature vector of feature representation using stemmed words from the abstract contains an element **B**, then the feature vector of the complex representation obtained by combining stemmed words from title and stemmed words from abstracts will contain the two elements: **Title: A** and **Abstract: B**. These feature representations were then combined with the machine learning algorithm that has the best overall performance to build text categorization classifiers. Similarly, we evaluated these complex feature representations using citations published in all previous years as training citations and tested using citations published in the current year.

4 Results and Discussion

Table 3 shows the detail F-measure obtained for each combination of machine learning algorithm, year, and feature representation. Among them, Support Vector machine along with stemmed words in abstracts achieved the best F-measure (i.e., 90.5%). Decision list learning along with stemmed words in titles achieved the second best F-measure (i.e., 90.1%). Feature representation using Mesh Headings along with Decision list learning or Support Vector machine has the third best F-measure (i.e., 88.7%). Feature representation using Author combined with Support Vector Machine has an F-measure of 71.8%. Feature representation using Journals has the lowest F-measure (i.e., 62.1%). From Table 3, we can see that Support Vector Machine has the best performance for almost each feature representation.

Note that the results for feature representation Authors were significantly worse for year 2002. After reviewing some citations, we found that the format of the author field has changed since year 2002 in MEDLINE citations. The current format results in less ambiguity among authors. However, we could not use the author fields of citations from previous years to predicate the category of documents for year 2002. Also, since a lot of citations in years 2002 and 2003 are in-process citations (i.e., people are still working on indexing these citations using Mesh Headings), feature representation using Mesh Headings had worse performance in these two years comparing to other years.

According to the reported performance, we explored the following feature representations: i). stemmed words from titles and stemmed words from abstracts, ii) Mesh Headings, stemmed words from titles, and stemmed words from abstracts, iii) Authors, Mesh Headings, stemmed words from titles, and stemmed words from abstracts, and iv) Journals, Authors, Mesh Headings, stemmed words from titles and stemmed words from abstracts. Figure 2 shows the performance of these feature representations when using support vector machine as machine learning algorithm. Note that the F-measures for complex feature representations that contain Abstract, Title, and Mesh Headings are indistinguishable. The inclusion of addition features such as Authors or Journals does not improve F-measure visibly. Figure 2 also includes the measure for keyword retrieving, which is different from the measure for each complex feature representation. The performance of keyword retrieving is measured using the ratio of the number of citations in each organism that contain keywords from the list of keywords obtained for that organism to the total number of citations for the organism. The measure for each complex feature representation is the F-measure obtained using support vector ma.

Year	MeshHeading			Journal			Author			AbstractText			ArticleTitle		
	DLL	NBL	SVM	DLL	NBL	SVM	DLL	NBL	SVM	DLL	NBL	SVM	DLL	NBL	SVM
1990	94.3	88.8	93.7	56.4	56.1	58.5	82.6	80.9	84.7	89.1	88.7	91.8	94.5	90.4	94.4
1991	92.7	88.3	93.5	55.4	56.6	57.3	81.3	77.9	83.4	88.7	88.4	91.7	92.2	89.5	92.3
1992	92.7	88.4	92.7	55.0	59.8	57.2	76.8	71.2	78.6	88.1	88.6	91.8	91.9	89.9	91.7
1993	93.0	88.6	92.9	60.0	60.3	58.6	76.8	70.6	76.2	87.5	87.1	91.0	91.5	89.1	91.3
1994	93.0	89.7	92.9	61.1	61.9	60.6	74.0	66.8	74.2	88.9	88.9	91.0	92.3	89.7	91.5
1995	93.3	90.7	92.5	63.2	63.4	63.1	76.5	67.2	73.3	88.0	88.6	91.2	92.0	89.2	89.4
1996	92.0	88.9	90.8	64.1	64.2	63.8	75.7	65.4	74.1	85.8	87.2	90.0	90.8	88.5	87.8
1997	90.6	87.5	90.5	63.9	64.0	64.4	75.8	65.2	72.8	85.1	86.0	89.7	89.8	86.7	87.8
1998	90.6	87.6	91.2	61.1	62.1	61.8	76.0	65.4	73.7	84.1	86.4	89.8	89.7	85.6	87.9
1999	89.7	87.1	88.9	61.8	63.3	62.4	73.3	64.1	69.5	84.3	86.2	89.6	88.5	85.3	85.8
2000	88.4	84.2	88.0	60.8	61.0	62.4	74.0	63.0	71.0	83.5	85.7	88.9	87.7	84.2	86.5
2001	87.0	84.9	87.7	62.4	62.7	62.4	74.4	62.5	68.3	85.0	86.8	91.0	89.1	85.3	87.2
2002	63.8	19.9	67.3	62.7	64.5	63.6	3.8	8.8	53.9	86.2	88.2	91.7	88.7	84.8	86.4
2003	77.6	76.0	76.0	48.0	48.0	54.5	62.2	53.1	63.1	85.3	89.0	92.8	83.3	87.0	83.7
Overall	88.7	82.1	88.8	61.3	62.1	61.8	69.3	61.6	71.8	86.0	87.2	90.5	90.1	87.0	88.5

Table 3. The F-measure of supervised text categorization study on different combination of supervised machine learning algorithms and feature representations. The classifiers trained using citations published previous years and tested using citations published in the current year.

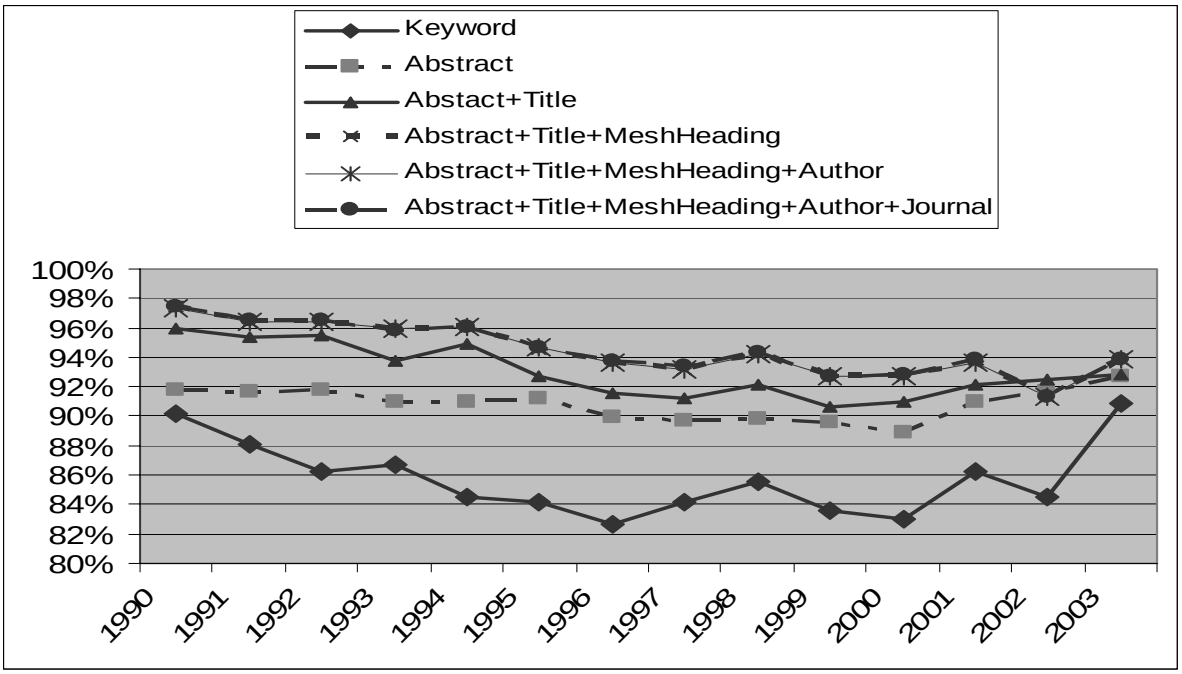


Figure 2. The F-measure of classifiers using complex feature representations learned using Support Vector Machine and the percentage of the number of citations containing keywords associated with the corresponding organism comparing to the total number of citations associated with that organism. Note that F-measures for complex feature representations Abstract+Title+MeshHeading, Abstract+Title+MeshHeading+Author, and Abstract+Title+MeshHeading+Author+Journal are overlapped with each other.

chine which was trained using citations from all previous years and tested using citations in current year.

From the study, we answered at least partially the questions. We cannot just simply use keywords to retrieve MEDLINE citations for model organism databases. From Figure 2, we can see that using keywords to retrieve citations may miss 20% of the citations. However, when combining all feature representation together, using citations from previous years could correctly predict to which organism the current year citations belong with an overall F-Measure of 94.1%.

For the supervised learning on text categorization task, different MEDLINE citation fields have different power on predicting to which model organism the paper belongs. Feature representation using stemmed words from abstracts has the most stable and highest predicting power with an overall F-measure of 90.5%. Authors alone can predict the category with an overall F-measure of 71.8%.

Among three supervised machine learning algorithms, support vector machine achieves the best performance. For feature representations where there are only a few features in a feature vector with non-zero values, decision list learning achieved comparable performance with (sometimes superior than) support vector machine. For example, decision list learning achieved an F-measure of 90.1% when using stemmed words from titles as feature representation method, which is superior than support vector machine (with an F-measure of 88.5%). Consistent with our findings in (Liu, 2004), the performance of Naïve Bayes learning is very unstable. For example, when using stemmed words from abstracts, the performance of Naïve Bayes learning is comparable to the other two machine learning algorithms. However, when using Mesh Headings as feature representation methods, the performance of Naïve Bayes learning (with an F-measure of 82.1%) is much worse than decision list learning and support vector machine (with F-measures of over 88.0%).

One limitation of the study is that we used only abstracts that are about one of the four model organisms. The evaluation would be more meaningful if we could include abstracts that are outside of these four model organisms. However, such evaluation would involve human experts since we can not grantee that abstracts that are not included in these four model organism databases are not about one of the four model organisms. That is also the reason we cannot provide F-measures when we evaluated the performance of keyword retrieving since we cannot grantee that abstracts associated with one organism are not related to another organism since the list of references in each organism database is not complete.

We could use previous published articles together with their categories to predict categories of the current articles where the list of categories is not limited to model organisms. It could be other categories such as the main themes for each paragraph in each paper. We will conduct a serial of studies on text categorization in the biomedical literature under the condition of the availability of category-labeled examples. One future project would be to apply text categorization on citation information for the protein family classification and annotation in Protein Information Resources (Wu, 2003).

As we know, homologous genes are usually represented in text using the same terms. Knowing to which organism the paper belongs can reduce the ambiguity of biological entity terms. For example, if we know the paper is related to mouse, we can use entities that are specific to mouse for biological entity tagging. Future work will be combining text categorization with the task of biological entity tagging to reduce the ambiguity of biological entity names.

5 Conclusion

In this paper, we designed a study using existing reference information available at four well-known model organism databases and investigated the problem of identifying relevant articles for these organisms using MEDLINE. We compared the results obtained using keyword searching with supervised machine learning techniques. We found out that keyword searching retrieved about 80% of the citations. When using supervised machine learning techniques, the overall F-measure of the best classifier is around 94.1%. Future work would be applying the supervised machine learning technique to the whole MEDLINE citation to retrieve relevant articles. Also we plan to apply text clustering techniques or text categorization techniques for the routing problem inside a specific model organism database (such as routing to curators in a specific area).

Acknowledgement

We thank anonymous referees for their valuable comments and insights. This work was supported in part by grant EIA-031 from the National Science Foundation.

References

The FlyBase Consortium. *The Flybase database of the Drosophila genome projects and community literature*, Nucleic Acid Research 2003, Vol 31 (1) 172-175

Blake J, Richardson J.E., Bult C.J., Kadin J.A., Eppig J.T. *MGD: the Mouse Genome Database*, Nucleic Acid Research, 2003, Vol 31 (1) 193-195

Donaldson I, Martin J, de Bruijn B, et al. *PreBIND and Textomy--mining the biomedical literature for protein-protein interactions using a support vector machine* BMC Bioinformatics. 2003 Mar 27; 4(1): 11.

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

Harris,T.W., Lee,R., Schwarz,E., et al. *WormBase: a cross-species database for comparative genomics.* Nucleic Acids Research 2003, Vol 31, 133–137.

Iliopoulos I, Enright AJ, Ouzounis CA. *Textquest: document clustering of Medline abstracts for concept discovery in molecular biology*, Pac Symp Biocomput. 2001; 384-95.

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

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

Liu H, V. Teller, C. Friedman. *A Multi-Variable Comparison Study of Supervised Word Sense Disambiguation*. Submitted to JAMIA

Misra S, Madeline A.C., Mungall C.J., et al. *Annotation of the Drosophila melanogaster euchromatic genome: a systematic review*, Genome Biology 2002, 3 (12)

Sebastiani F. *Machine learning in automated text categorization*. ACM Computing Surveys, 2002, Vol 34 (1) 1-47

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

Wiener ED, Pedersen JO, Weigend AS. *A neural network approach to topic spotting*, SDAIR 1995, 317-332

Wu CH, L Yeh, H Huang, L Arminski, et al. *The Protein Information Recource* Nucleic Acids Research, 31: 345-347

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

Yarowsky D. *Decision lists for lexical ambiguity resolution: Application to accent restoration in Spanish and French*. 1994; ACL 32: 88-95