CLASSIFYING SCIENTIFIC TEXTS IN BIOLOGY FOR FOCUS SPECIES

Qi Wei

DOCTOR OF PHILOSOPHY

DEPARTMENT OF INFORMATICS
SCHOOL OF MULTIDISCIPLINARY SCIENCES
THE GRADUATE UNIVERSITY FOR ADVANCED STUDIES (SOKENDAI)

2011 (SCHOOL YEAR)

A dissertation submitted to the
Department of Informatics,
School of Multidisciplinary Sciences,
The Graduate University for Advanced Studies (SOKENDAI)
in partial fulfillment of the requirements for
the degree of DOCTOR OF PHILOSOPHY.

Research Supervisor:

Nigel Collier, Assoc. Prof.	NII, SOKENDAI

Advisory Committee:

Nobuhiro Furuyama, Assoc. Prof.	NII, SOKENDAI
Asanobu Kitamoto, Assoc. Prof.	NII, SOKENDAI
Fabio Rinaldi, Dr.	University of Zurich
Hideaki Takeda, Prof.	NII, SOKENDAI
Seiji Yamada, Prof.	NII, SOKENDAI

Table of Contents

List of Tables

List of Figures

Abstract

In recent years high throughput methods have led to a massive expansion in the free text literature on molecular biology. Automated text mining has developed as an application technology to organize this wealth of published results into structured database entries. Presently, there are more than 10,000 species and taking the marbled lungfish (*Protopterus aethiopicus*) as an example, there are 132.8 billion base pairs in this fish genome. In a typical systems biology abstract, there are 4-5 genes mentioned on average. Thus, recording and encoding them manually would take prohibitive amounts of time and human resources. Building intelligent tools to help authors and database curators integrate published results into databases has therefore become a major goal of research in biomedical natural language processing. However, the multiplicity of interpretations of meanings makes the specification of the authorś intended meaning extremely challenging for automated natural language processing.

In this dissertation, the contribution is presented through a series of three experiments for identifying the focus species in biological papers as an aid to classifying and summarizing the experimental result. The focus species presents the authorś major claim in reporting their own results. I report a new method to identify the focus species with novel features providing optimized performance on full text papers and abstracts. I also report a new knowledge model based on a typed citation function and show its application to focus species identification.

In the experiments, 3 model organisms are classified in full papers selected based on the Biocreative 1b dataset and 4 model organisms are classified in abstracts selected from the DECA corpus. With three experiments, I showed a best F-score of

90.7% for classifying the full papers by using internal features. I also showed that when only using internal features, full papers perform much better than abstracts. By using external features from related publications, I demonstrated a best F-score of 91.14% for classifying abstracts. Finally I developed a new typed citation scheme and showed that among the four citation classes of background, method, results and data, the strongest relation for aiding the focus species classification was the one relating author results to the target paper.

Acknowledgements

I would like to thank my people for their support, encouragement and guidance during my graduate school years in NII.

First of all, I would like to thank Nigel Collier, my supervisor, for his many suggestions and constant support during this research. Without his encourage, advices and suggestion, this dissertation would not have happened.

I also want to thank my other committee members, Nobuhiro Furuyama, Asanobu Kitamoto, Fabio Rinaldi, Hideaki Takeda and Seiji Yamada, I'd like to thank them for their kindly discussion and comments regarding my research and this dissertation.

Moreover, I'd like to thank all researchers in my research group for their support in solving research problems.

All friends in NII and in China always cheer me up at all times when I need them.

Most of all, my husband has always supportive, understanding and encouraging at all times through my graduate school years, and he also provided help when I had problems in hardware and problems in writing scripts.

Of course, I am grateful thank to my three daughters for their smiles and love everyday especially when I feel tired.

Finally, I'd like thank to my parents in China for their patience, support and love. Without them this work would never have come into existence.

January 1st, 2012

Chapter 1

Introduction

1.1 Motivations & Thesis question

1.1.1 Motivations

Over the last two decades databases have become central stores of organized knowledge for life scientists. However, too much knowledge is still locked away in free texts and is therefore inaccessible by computational techniques that require structured data. As high throughput experiments drive a massive expansion in the literature, life scientists are finding it more challenging to keep up to date with the wealth of newly published information, slowing the pace of progress and risking duplication of work. Building intelligent tools to help authors and curators integrate their published results into databases has been a major goal of research in biomedical natural language processing.

Identifying the focus species for an experiment is a specialized subtask of topic classification that requires precise identification of semantic features such as the gene/protein which is the topic of discourse. In earlier work on focus species identification, word-level features were explored by F. Rinaldi et al.[4] for target organisms including *Homo sapiens*, *Mus musculus*, *Saccharomyces cerevisiae* and *C. elegans*.

Rinaldi et al. devised a system that created a ranked list of species for each MEDLINE abstract and demonstrated its effectiveness for disambiguating gene references. They showed that the number of possible gene references was reduced to 45012 (p = 0.0308, r = 0.5763) from 283556 in the initial annotation step (p = 0.0072, r = 0.7469). Wang and Matthews [5] created a system for gene entity recognition and identification that used a combination of species name and gene name features co-occurring in the same sentence. They showed significant improvements of 11.6% on accuracy for the gene mention task. In order to study term level species identification, Wang et al. [1] manually created an annotated MEDLINE dataset, providing a species ID for each gene mention. Kappeler et al. [6] devised a system to detect the focus organisms in biomedical papers. Their approach used the NCBI taxonomy to make a protein-organism list and this was used to detect the focus organism in full-text articles showing a top F-score of 74.0%.

In order to identify the focus species, several features inside the given papers were used. In addition to these, resources outside the given papers can also be used. An important clue is the citation network. Previous work on citation analysis in the biomedical literature includes work on bibliometrics and enhanced ranking of search. I. Tbahriti et al. 2006 [7] for example, looked for related articles using argumentative categories in Medline abstracts and measured this with co-citation lists. P. Nakov et al., 2004 [8], used text surrounding citations for grammatical paraphrase extraction and S. Teufel et al., 2006 [2] explored automatic classification of citations, i.e. the reason why a work was cited. The application of citation analysis to text classification was also attempted within the computer science domain where B. Zhang et al., 2005 [9] reported a 7% improvement using citation information. T. Delbecque and P.

Zweigenbaum [10] showed the successful use of the cited articles and cited authors in indexing MEDLINE full papers.

1.1.2 Thesis question

The thesis explores the general question " ***What features are most effective for resolving conflicting evidence about focus organism in biomedical abstract and full text?*** " Since the question is potentially open-ended, I break this down into three specific sub-questions.

1. What level of classification performance is achievable using state-of-the-art lexical semantic features for focus species in full papers and abstracts?

2. Of the abstracts which are cited or archived in the PubMed database, do bibliographic features provide enhanced classification accuracy?

3. Of the abstracts which are cited does a typed citation function provide enhanced classification accuracy? Also what citation types prove the most useful?

1.1.3 Contribution of this dissertation

In the dissertation, I present a new method to identify focus species with novel features in full-text papers and abstracts. I present a new knowledge model for species citations in biomedical papers. With this scheme, I developed a tool to provide authors and curators with a high-throughput method capable of determining the focus species in experimental papers. Unlike previous studies my approach does not consider target documents in isolation but makes use of a network of citation relationships, amplifying information which is implicit in the target document. The various features explored in the thesis questions are evaluated on gold standard data sets that have been constructed by external groups for community evaluation exercises.

1.1.4 Organization of this dissertation

This Ph.D. dissertation presents a method for identifying the focus species of full-text papers and abstracts and a new citation scheme for biomedical papers.

This dissertation consists of seven chapters. Chapter 1 gives the introduction, and Chapter 2 presents the related work. Chapter 3 describes the first experiment on focus species classification for full-text papers. Chapters 4 describes the second experiment on focus species classification for abstracts. Chapter 5 discusses the new citation scheme for biomedical papers and its application to focus species classification. And chapter 6 discusses the difficult cases for the task and online tools. Chapter 7 concludes this dissertation and discusses future work.

There is one set of experiments for each thesis question. Hypothesis 1 is explored in a series of experiments in chapter 3. Based on the findings of this experiment which showed the relative merits of various in document lexical semantic features, I conducted Hypothesis 2 experiments which are reported in chapter 4. Based on the findings of experiments in chapter 4 that showed the effectiveness of bibliographic features, I conducted Hypothesis 3 experiments which are reported in chapter 5.

Chapter 2

Background

High throughput methods have led to a massive increase in the literature on molecular biology. Formalization of the published results and automated text processing are required in order to register all this new information in a database. S. Yeh et al. [11] pointed out two purposes for biological databases: (1) Databases are places for experts to consolidate data on a single organism or a single class of organisms which often including DNA sequence information; (2) Databases have made information searchable by using a variety of automatical techniques. Biological experts can formalize their results by registering them in databases, such as MSD (Mouse Genome Database) [12],FlyBase [13], DictyDB [14] and Wormpep [15]. However, there is still quite a lot of knowledge locked away in unstructured format which is hard to share,organize and acquire. Figure 2.1 shows the rapid growth rate of PubMed records, reflecting the size of the biomedical scientific literature. In order to keep up to date with new findings and avoid duplication of work,there is a need for formalizing new experiment results into structured data. In the biological domain, biological experiments are yielding more and more results that can be formalized by registering them in databases such as MGD (Mouse Genome Database) [12], FlyBase [13], DictyDb [14],

and Wormpep [15]. Curation of literature in databases ensures that the data stored in them reflects scientific facts. In particular, database curation in life sciences helps to ensure data quality to enable quick access to the latest experimental results. The problem is that curation is a time-consuming task requiring high level understanding of the domain and expert skills. For example, MGD curators have to ensure that the stored publication data that can be used to validate expressions of genes under certain conditions.The importance of database curation is growing, and a number of communities have become established to support development of gold-standard shared tasks. The Knowledge Discovery and Data Mining (KDD) Challenge CUP task in 2002 [11] focused on automating the work of curating Flybase, by identifying what papers need to be curated for Drosophila gene expression information. The Goal of the BioCreative [16] challenge was to pose tasks that would result in scalable systems for use by biology researchers and end users such as annotation database curators. BioCreative tried to address the database curation task by challenging participants to identify papers according to the evidence they contained for assigning GO codes to human proteins.

However, curating results manually will take lots of time and human resources. Consequently, building intelligent tools to help authors and DB curators integrate their published results into databases has been a major goal of research in biomedical natural language processing. Winnenburg et al. (2008) [17] explored the effective use of text mining techniques in helping the curation task. Figure 2.2 shows a typical curation process.

(1) gene products was identified in the publication papers. This step can be treated as Named Entity Recognition (NER), i.e. a task seeks to locate and classify

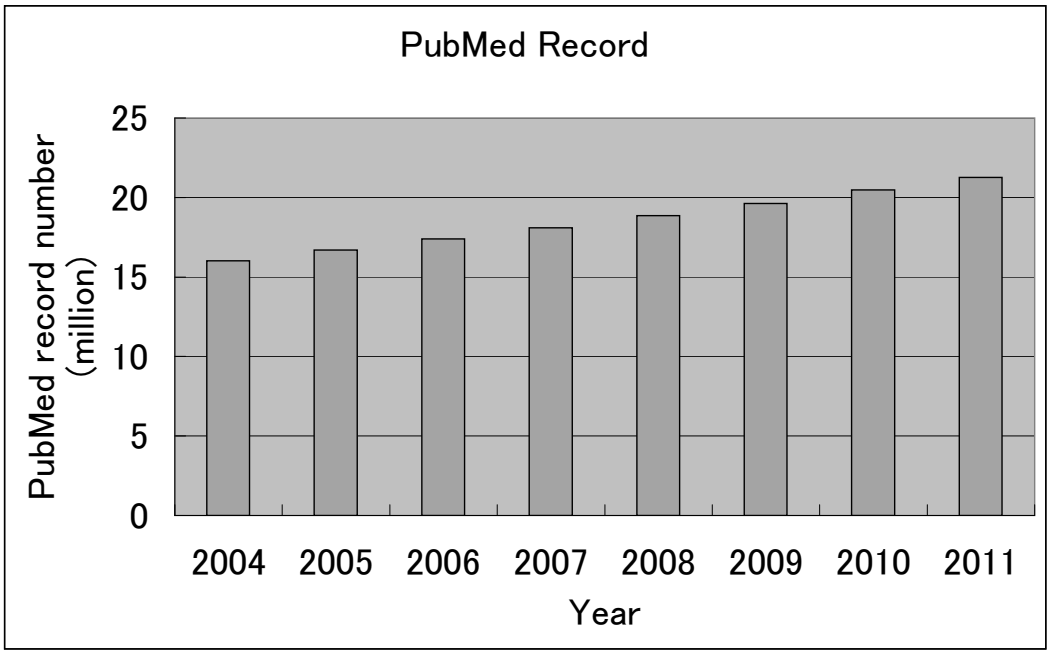


Figure 2.1: Growth rate of PubMed records

atomic elements in text into predefined categories such as the names of gene names, protein products, which can be solved using dictionary based method [18], rule based method [19]or machine learning method [20].

(2) the gene products were link to the unique database ID according to the ontologies. This step can be called gene normalization which also can be solved using dictionary look up method , machine learning method etc.

(3) With the result of step (2) and some other information such as external resources, the focus organisms were identified. This is called focus species recognition. This step also can be solved using machine learning methods and also other methods.

In that case, we can see that the machine learning methods can help in the curation task and reduced the amount of time and human effort.

In recent years, biomedical text mining community have paid huge efforts to meet these three goals, such as JNLPBA [21], BioNLP shard task, the Knowledge Discovery and Data Mining (KDD) Cup Chanllege [11], BioCreAtIvE [16] and the TREC Genomics track [22]. There still remain many challenges in such tasks, such as the ambiguous vocabularies of genes as well as identifying target relations within complex sentence and discourse structures. This thesis aims to contribute to our understanding of the third task, focus species identification, by a thorough investigation of models and features as well as proposing a novel citation schema for use in leveraging external bibliographic features. The remainder of this chapter is organized as followed. I start by introducing the component tasks that lead up to focus species identification including Named Entity Recognition and gene normalization.Then I introduce the task of focus species identification. I end by introducing the feature selection and citation analysis in the task.

2.1 BioNER and gene normalization

The aim of gene normalization task is to link gene mentions to indexes in standard biomedical databases such as Entre Gene or NCBI identifier. By doing this, it will improve the document indexing and support more sophisticated knowledge discovery tasks. Gene normalization can be treated as the first step of identifying the focus species of the whole article. Gene normalization task can be treated in two steps: Named Entity Recognition and database identifiers linkage. I now briefly survey some of the more important works that have taken place in NER and gene normalization

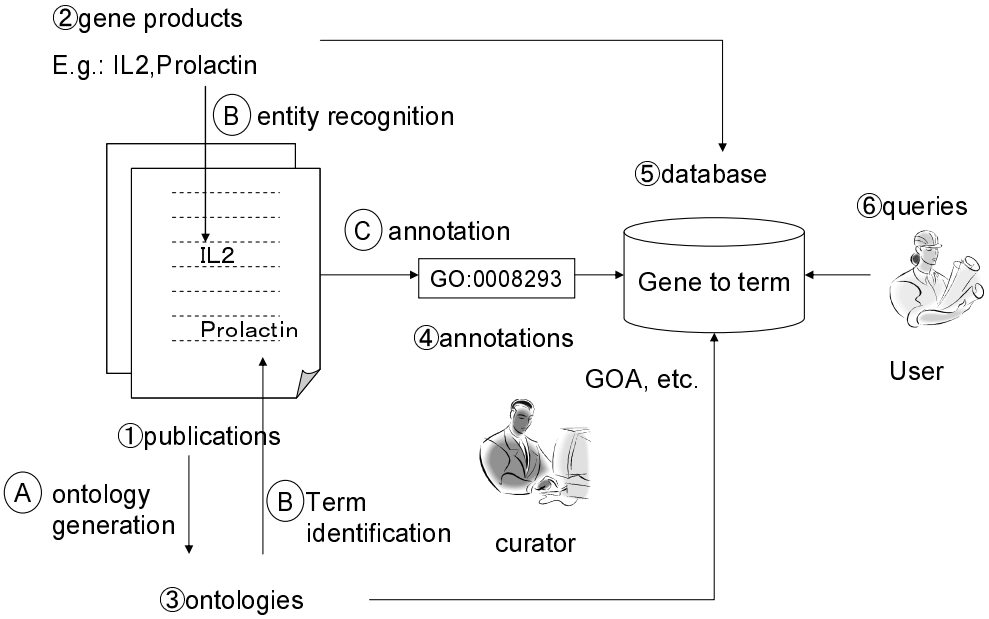


Figure 2.2: Integration of text mining and ontology development
Integration of text mining and ontology development to curation process: the curator reads papers (1) and identifies gene products (2) and terms from ontologies (3), which have been proposed by text mining methods (A-C). Annotations (4) are formulated and added to a database (5), which can be queried by the end user (6).

within the biomedical text mining community.

2.1.1 Biomedical Named Entity Recognition

Named Entity Recognition (NER) originated from the Message Understanding Conferences (MUC) [23] in 1990s. The task in MUC is to identify terms such as person name, organization name, location name and etc, in the Newswire domain. NER is the foundation stage of information extraction, Question Answering, Ma chine Translation as well as many other applications requiring semantic understanding of text. The evaluation of NER system is targeted in a serial of conferences, such as MUC-6,MUC-7, COLING2002,COLING2003 and etc. With the rapid growth of biomedical knowledge, NER was introduced into the biomedical domain and was called BioNER. The purpose of BioNER system is to identify terms such as gene, protein and etc. Figure 2.3 shows an example of BioNER. Till now, BioNER task was still involved as a part of many share tasks in biomedical domain. The main share tasks in BioNER are:

(1) JNLPBA [21]: JNLPBA is a share task for bio-entity recognition. The aim of this task is to assign the technical terms in the domain of molecular biology such as protein, gene and etc. The data set was a subset of the GENIA version 3.02 corpus. Among the 7 participating systems the best F-score (harmonic mean of recall and precision) of 72.6% was achieved by combining Support Vector Machines (SVMs) and Markov Models [24]. In addition to traditional lexical features such as surface word, morphological patterns and part of speech, semantic triggers, name aliases and cascading entity features were also considered as well as external resources such as SwissProt and LocusLink lists. Systems also made use of in domain part of speech

tagging using the GENIA part of speech corpus for training.

(2) BioCreAtIvE [16]: The BioCreAtIvE (Critical Assessment of Information Extraction systems in Biology) challenge consisted 2 tasks. In task 1a, it mainly focused on the BioNER task. The dataset was extracted from MEDLINE corpus. The participants using techniques such as Hidden Markov Models (HMM) or Support Vector Machines (SVMs) were showed a F-score over 80%. Task 1b is discussed later on in this chapter.

(3) CALBC challenge [25]. CALBC (Collaborative Annotation of a Large Biomedical Corpus) was a new challenge hold in 2010.The objective of the CALBC challenge was to produce a very large scale corpus using semi-automated techniques. The task 1 in CALBC is to identify the biomedical terms (gene/protein ,organism, chemical and disease). The significant specific of this challenge is that the challenge use a very large corpus with 50000 abstracts for training and 100000 abstracts for annotation. With such a large corpus, the evaluation was compared to the harmonized corpus. The best system showed a F-score of 86% on average compared to the harmonic corpus using dictionary look up method.

From a methodological perspective, the NER system can be divided into dictionary based, rule based and machine learning based. These broad categories of techniques were also used in BioNER systems. For instance, a dictionary-based system was developed by Y. Tsuruoka (2003) [26] and K.Zhou (2005) [27]; a rule based system was developed by D. Hanisch et al.(2005) [19]; and machine learning based systems were developed by GD Zhou (2004) [28], S. Zhao (2004) [29], Tsai et al.(2005) [30], Dingare et al.(2005) [31], Y.P. Li et al. (2009) [20] and etc.

(1) The dictionary based approach

<article>
<articleinfo>
<bibliomisc>MEDLINE:95343554</bibliomisc>
</articleinfo> <title>
<sentence>
<cons sem=other_name><conssem=DNA_domain_or_region>E1A gene</cons> expression</cons> inducessusceptibility to killing by <cons sem=cell_type>NK cells</cons> followingimmortalization but not <cons sem=other_name><conssem=virus>adenovirus</cons> infection</cons> of <cons sem=cell_type>humancells</cons>.
</sentence>
</title>
<abstract>
<sentence>
<cons sem=other_name><cons sem=virus>Adenovirus</cons> (Ad)infection</cons> and <cons sem=other_name> <conssem=protein_molecule> E1A</cons> transfection</cons> were used to model changesin susceptibility to <cons sem=other_name>NK cell killing</cons> caused by transientvs stable <cons sem=other_name> <cons sem=protein_molecule>E1A</cons>expression</cons> in <cons sem="cell_type">human cells</cons>.
</sentence>
...
</abstract>
</article>

Figure 2.3: An example of BioNER task data taken from the GENIA corpus of annotated PubMed abstracts. The record shows inline text annotation for named entity classes DNA, cell line, virus, cell type, protein, protein family group and other.

By using a dictionary-based method, an entity is classified by searching a dictionary or database and matching it with similar words found there. The advantage of the dictionary based method is that the dictionary look-up procedure is easy to implement, thus, many systems attempting more sophisticated tasks use a dictionary based methods as the first step of their NER system.

The dictionary based methods usually have low recall. Hirschman et al.[32] reported a dictionary based system with a low precision about 2%. The disadvantage of dictionary based systems is as follows: (1) The dictionary quickly gets out of date. Take protein name as an example, a new protein is discovered every month, hence, it is hard to update a dictionary on time. (2) Homonymy and polysemous cases. While the Newswire domain does not suffer much from homonymy and polysemous names, problems are widespread in the biomedical domain. For example, some protein names are the same as common English words, such as 'by' and 'can' and many proteins are named after the genes that encoded it.

Hence, a dictionary based method often combined with other methods, such as edit distance, machine learning method is often used. The accuracy of an NER system is especially improved when a dictionary based method is combined with a machine learning method. For example, Z. Kou et al.[18] created a system combining a dictionary based method and Hidden Markov models (HMM) together. They report their system had a higher recall than dictionary lookup algorithm and achieved a slight improvement in F-score. Tsuruoka and Tsujii [33] showed an 10.8% improvement in F-score by introducing a naive Bayes classifier filter and another 1.6% improvement of F-score by expanding the dictionary with a probabilistic variant generator.

(2) Rule-based approach

The rule based approach is to craft a set of rules by hand using a rule based language and domain knowledge. The main approach of Rule based is to develop rules that describe common naming structures for certain entity classes by using either orthographic or lexical clues, or more complex morphoea syntactic features. In most of cases, the rules are created manually. For example, there is a rule in Newswire domain:

⟨proper-noun⟩+ ⟨corporate designator⟩ -> ⟨corporation⟩

In the rule showed above, items between < and > represent a dictionary list. The left hand side represents the rule to match the surface text and the right hand side represents the annotation to be inserted into the text around the matching string. Using a rule like this, an organization name such as APOLLO CO. can be easily recognized where APOLLO is a proper noun and CO is the corporate designator. There are several rule languages, such as Simple rule language (SRL) [34].

Several systems use the pattern-based approach, e.g. the system developed by Fukuda and colleagues [35] were among the earliest to use it for BioNER. The advantage of the rule based approach is that the rules can be adapted, added and extended as needed. However, the analysis of the text in the target domain and the task of creating rules manually are time consuming especially in some specific domain which need expert to create the rules. Moreover, the rule-based approaches are hard to adapt to other domains since the individual rules are difficult to adapt. Also, it is impossible for the rule writer to keep track of thousands of rules and hence inconsistencies and gaps will begin to appear without effective tool support.

(3)Machine learning based approach

Within text mining, machine learning (ML) has been commonly used for tasks

1. Create guidelines
2. Annotate a representative sample of texts
3. Select training model and features
4. Train and test model
5. Analyze the output
6. Select the optimal model and features for unlabeled data

Figure 2.4: Procedure of machine learning method in tagging task

such as clustering, classification, sequence labelling, trend and anomaly detection.

Some clustering (flat and hierarchical) and topic models are: Latent Dirichlet Allocation(LDA) Pachinko allocation, k-means, agglomerative hierarchical clustering.

Classification (flat and hierarchical) models include: Naive Bayes (NB), kernel methods, Maximum entropy Models (MEM), Decision Trees (DT)

Some sequence tagging models are : Hidden Markov Models (HMM), Maximum Entropy Markov Models (MEMM), Conditional Random Fields (CRFs), Markov Logic Networks (MLN), stochastic CFG, tree-based kernel methods, ensemble methods, transformation based error driven learning.

Compared with dictionary based and rule based methods, machine learning based methods are more flexible and adaptable. They are easy to apply it to a specific domain if one provides a proper model and training data set. However, the training data set should be manually created,which is time consuming and may require expert involvement.

A typical process for model selection in supervised machine learning for BioNER is shown in Figure 2.4

(a) Create Guidelines: this is always the first step of a supervised machine learning system. NER guidelines are created by collaboration with expert in the task domain. In fact, creation of guidelines is not exactly a science, it contains many human factors, i.e. the idiosyncratic choices of human. A well prepared set of guideline can help annotators converge their opinions on the task and thereby improve the accuracy of the NER system. On the other hand, a set of guidelines with poor schema can significantly reduce the performance of the system by instantiating inconsistencies.

(b) Annotate a representative sample of texts: the annotators label the raw text according to the guideline in the first step. There are also many human factors in this step. Although the annotators have guidelines, each one annotates the text according to his/her own understanding of the guidelines. In addition to the cost of creating the training data care also needs to be taken that the training and testing corpus are representative of the task domain and that the annotation guidelines are consistent and easy to follow. It's also necessary to make sure that if multiple annotators are used in annotation,they are well trained and are performing according to the guidelines, so that the annotate results can be consistent.

(c)Select training model and features. Here, first, the experiment type was decided. The experiment type influences the split of the data. There are mainly two ways, one is using cross-fold validation and one is to split the annotated data set into a training data set and test data set. 10 cross-fold validation are widely used. The database is split into 10 folders and 9 folders are using for training and 1 folder is used for test, 10 different combinations are tested and result are based on the 10 different combinations. Training model and features are also decided.There are several machine learning methods, e.g. biomedical text classification (rule-based classifiers

such as decision trees [36], logical rules [37]. Linear classifiers such as logistic regression [38], Naive Bayes methods [39, 40], boosted linear classifiers [41]. Non-linear classifiers also successful applied in text classification task, such as Support Vector Machine (SVM) [42, 43, 44], k-nearest neighbor methods (kNN) [45], Boosting [41], [46]) , biomedical entity recognition (HMM [28, 29], SVMs [28], CRFs [30]) etc. Also, feature selection is an important part in this step; lexical features such as word, POS and morphological features such as brief word shape are widely used.

Feature selection is an important part in machine learning method, either in Named Entity Recognition and in text classification.

(d) Training model and test: In this step, the machine learning system uses training data to learn the features useful for entity recognition. The machine learning system is tested after it has been trained.

(e) Analyze result: Evaluate the test results. Adjust the model parameters according to the test results.

F-measure is a commonly used evaluation of the text classification system. Consider a binary label case where one entity has two statuses: positive(+) or negative(-). As shown in figure2.5, label stands for its gold label and assignment stands for the label given by the model. Here, TP stands for true positive, TN stands for true negative, FP stands for false positive and FN stands for false negative. By using these, precision(P) and recall(R) are defined as:

$P = \frac{TP}{TP+FP}$

$R = \frac{TP}{TP+FN}$

Then, F-score is defined as:

$F = \frac{2PR}{P+R}$

		Assignment	
		+	-
label	+	TP	FN
	-	FP	TN

Figure 2.5: F score example

(f) After the final model is created, this model can be used to annotate new unlabeled data which is assumed to come from a broadly similar source to the training data.

With the improvement of genomics databases, e.g, Entrez Gene or UniProt, these kinds of databases were used as external resources to extract dictionaries and enhance the accuracy of BioNER systems. These external recourses have been widely used in recent years [25].

Although BioNER was well studied these years, BioNER is still a challenging task because:

(1) Ambiguities exist in the biomedical domain. Since different species have various naming rules, one entity may have different meanings. The ambiguities in biomedical domain have various causes. (a) One entity may refer to several different genetic entities, either from the same species or from other organisms. (b) One entity may refer to another type of biological entity, such as protein or phenotype. (c) Some entities are the same as common English words, for example, a drosophila gene called 'can'.

Ex1. The string 'CAT' represents different genes in cow, chicken, fly, human, mouse, pig, deer and sheep;

Ex2. The mouse gene 'hair loss' is a common phenotype.

Ex3. The mouse gene 'diabetes' is also used in other domain, clinical domain.

Ex4. Drosophila genes called 'can', 'lie' are common English words.

(2) In the biomedical domain, new protein names and gene names are continually being created. There is no complete dictionary that includes entries on every biomedical entity. Thus, the simple dictionary-based method cannot work well.

(3) The orthographic combinations in the biomedical domain are complex. Terminology is encoded using specific combinations of capitals, puncuation and digits as well as Greek letters.

(4) There is widespread use of abbreviations in the biomedical domain. For example, ATL stands for adult T-cell leukemia, beta-EP stands for beta-endorphin. Chang et al.[47] showed that in MEDLINE abstracts, 42.8% of abstracts have at least one abbreviation and 23.7% of abstracts have two or more. It is also shows that there is one new abbreviation in every 5-10 abstracts on average and the growth rate of new abbreviations is increasing.

(5) Many descriptive phrases exist in biomedical entities. I counted the distribution of entity lengths in GENIA corpus, over 25% of the entities include more than 4 words,for example, 'primary human bone marrow cultures', 'normal thymic ephithelial cells'.

(6) Some entities use conjunction and disjunctions. Two or more entity names in the biomedical domain may share parts by using conjunction and disjunctions. (a) Sharing using an ellipse. In word sequence 'protein kinas C-alpha, -epsilon, and -zeta', three entities shared one part, the full forms are 'protein kinas C-alpha', 'protein kinas C-epsilon' and 'protein kinas C-zeta'. (b) Sharing using the word 'and'. For example,

'LMP1 and 2' shares the first part 'LMP', the full form is 'LMP1' and 'LMP2'.

2.1.2 Gene normalization task

Gene normalization is a key step in an accurate search in biomedical literature. With the result of BioNER system, gene normalization task is to link the gene mentions with the unique database identifiers. The importance of gene normalization task was recognized these years, this growing trends is shown by wide scale participation in shared tasks such as BioCreAtIvE I and II [16, 48], CALBC challenge.

(1) In BioCreAtIvE I task [16], the task was to link the gene with Entrez identifers. Figure 2.6 shows an example of gene normalization task in BioCreAtIvE. After identified the gene "esteraze 6", the synonym list extracted from Flybase was checked and unique identifier "FBgn0000592" was linked to this gene mention.

Three organisms were chosen in BioCreAtIvE 1b task: fly, mouse and yeast. The abstracts were collected from MEDLINE articles. Eight groups participated in this task and a highest F-score of 92% in yeast 82% in fly and 79% in mouse was reported. The analysis showed that the differences of the accuracy in three organisms were caused by several factors such as the ambiguity in names, the complex of gene names.

(2)BioCreAtIvE II [48]: The aim of BioCreAtIvE II gene normalization task is to link the EntrezGene (Locus Link) identifiers to human genes and direct gene products. The dataset is collected from MEDLINE abstract. Polysemy in gene and protein names created additional complexity both within and between organisms. The best F-score achieved 79% in this task. In that case, compared to BioCreAtIvE I task 1b, the F-score is much lower than the result in mouse, yeast and fly.

(3) CALBC challenge is a new challenge in gene normalization task.

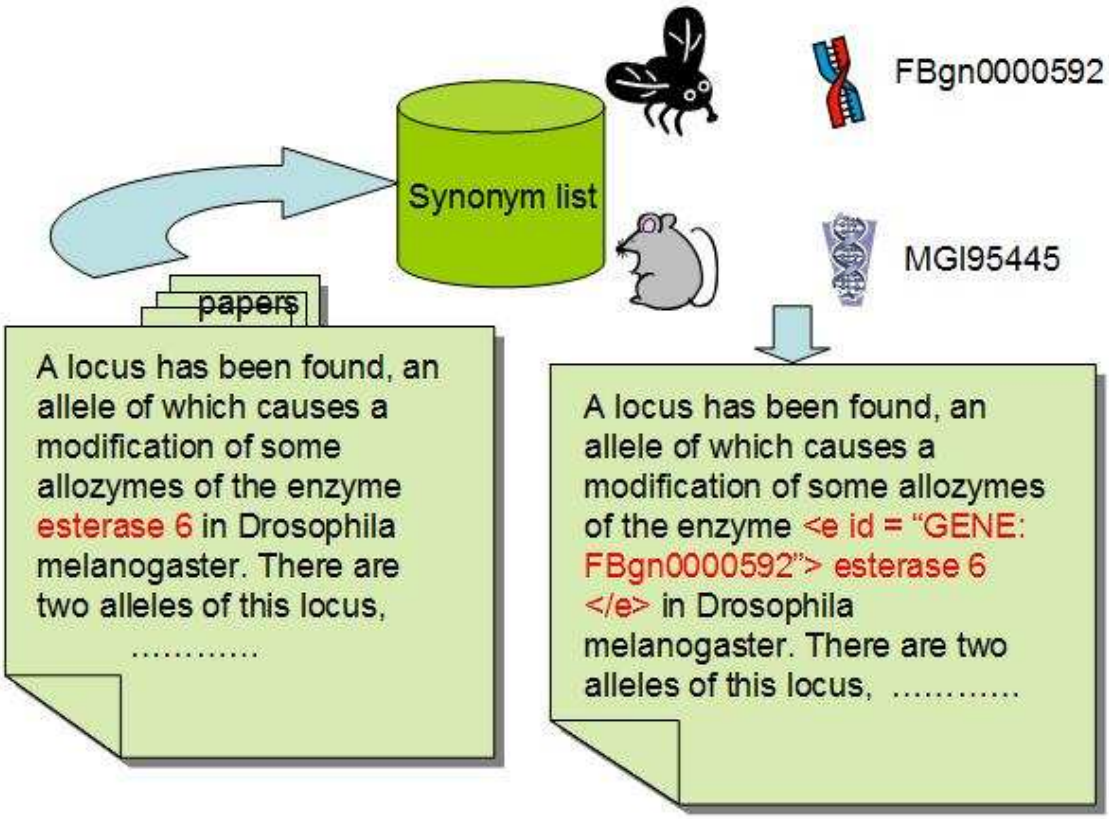


Figure 2.6: An example of Gene normalization

Many approaches were reported in the gene normalization task. H. Fang et al. (2006) [49] reported a rule based system in gene normalization task in BioCreAtIvE I. They first extracted an dictionary automatically and then built several rules to prune the uninformative synonyms, a dictionary match step was performed afterwards. They reported a best 70.1% F-score in such task. D. Hanisch et al.2005 [19] showed a system called ProMiner in CALBC challenge, they first generate a dictionary and then used rule-based classification method to identify the related ID to gene mentions. They showed a best F-score of 81.6% for fly, 79% for mouse and 89.9% for yeast. J. Crim et al. 2005 [50] showed a maximum entropy classification systems in gene normalization task. They compared classification-based system to patten match system and found that classification system outperforms the pattern matching system for fly-related documents. They reported a F-score of 74.2% for fly, 75.8% for mouse and 91.7% for yeast.

Several difficulties in gene normalization task were mentioned by A. Morgan et al. (2007) [48].

(1) Gene mentions are elided under various forms of conjunction, which cause more problem in normalization task. For example, "protein kinase C-alpha , - epsilon , and - zeta" which stands for three forms of PKC gene,PKC alpha, PKC epsilon and PKC zeta. It is difficult to identify the boundaries for the names of the different forms of PKC in such case. Furthermore, more difficult cases exist. For example, "AKR1C1-AKR1C4" includes four gene mentions: "AKR1C1", "AKR1C2", "AKR1C3" and "AKR1C4".

(2) Large gray area in gene and protein nomenclature between a description and a name. For example, the text "Among the various proteins which are induced when

human cells are treated with interferon, a predominant protein of unknown function, with molecular mass 56 kDa, has been observed" mentioned the protein which is known as "interferon-induced protein 56", as the text only describes this protein without listing the name of the protein. The question of what should be tagged was raised.

(3) Ambiguities in gene and protein names both within and between organisms. For example, the text showed in Figure 2.6, "esterase 6" can be a gene in Drosophila with an id FBgn0000592 and a gene in Mouse with an id MGI95445. However, considering the whole text, "esterase 6" was mentioned as a Drosophila gene. Current years, many researchers add the disambiguation step in their gene normalization system. F. Rinaldi et al. [4] showed a system called OntoGene solving the gene normalization task in BioCreAtIvE II. They showed that in the corpus they used, the main organisms mentioned in abstracts were humans (56.3%), mice (9.3%), yeast (6.5%) and C. elegans (6%). They devised a system with two steps, first with a high-recall annotation followed by a disambiguation steps. In disambiguate step, they created a ranked list of species for each article and showed that such a list was good for disambiguation; the number of possible gene references was reduced to 45012 ($p = 0.0308$, $r = 0.5763$) from the initial annotation step 283556 ($p = 0.0072$, $r = 0.7469$). X. Wang and M. Matthews [5] created a system that used surround words of gene mentions. They used a combination of species name and gene name in the same sentence in disambiguation and showed an improvement significantly by up to 11.6%.

(4) As the gene normalization was mainly based on the gene identification result, the accuracy of BioNER also effect the accuracy of gene normalization task.

In study of the gene normalization task, there are three corpora for this task:

(1) BioCreAtIvE I task 1b corpus. This corpus contains manual selections from three model organism databases: Fly [13] (Drosophila melanogaster), Mouse[12] (Mus musculus), Yeast[51] (Saccharomyces cerevisiae). PubMed IDs were selected from the databases and MEDLINE abstract was selected out according to these PubMed Ids to make up the BioCreative I task 1B corpus.

(2) BioCreAtIvE II corpus. The corpus contained 20000 sentences and approximately 44500 GENE and ALTGENE annotations (A boundary alternated GENE annotation made by human annotators). The token specifications of all previous annotations were changed to character specifications. It became possible to annotate a gene that is hyphenated to another word, the combination of which is not a gene mention.

(3) DECA corpus [1]. Abstracts for the DECA corpus were selected from the BioCreAtIvE I & II dataset. In total 644 MEDLINE abstracts have been manually annotated by assigning NCBI species IDs for each gene mention. Mentions of gene and gene products are annotated and a species ID has been assigned to every entity mention. The species tags are identifiers from the NCBI Taxonomy of model organisms (http://www.ncbi.nlm.nih.gov/sites/entrez?db=taxonomy). Table 2.1 illustrates the distribution of species IDs given in DECA corpus. Abstracts focusing on *Drosophila melanogaster*, *Mus musculus* and *Saccharomyces cerevisiae* were selected from BioCreAtIvE I task 1b development test corpus and abstracts focusing on *Homo sapiens* were selected from the BiocreAtIvE II dataset.

Table 2.1: Distribution of NCBI IDs in the DECA corpus indicating the degree of ambiguity

Species name	NCBI Species ID	Freq	Percentage
Homo sapiens	ncbitaxon: 9606	3201	50.01%
Mus musculus	ncbitaxon:10090	1504	23.50%
Drosophila melanogaster	ncbitaxon:7227	636	9.94%
Saccharomyces cerevisiae	ncbitaxon:4932	508	7.94%
Other	ncbitaxon:-1	366	5.72%
Other2	ncbitaxon:0	66	1.03%
Rattus norvegicus	ncbitaxon:10116	70	1.09%
Escherichia coli K-12	ncbitaxon:83333	18	0.28%
Xenopus tropicalis	ncbitaxon:8364	19	0.30%
Caenorhabditis elegans	ncbitaxon:6239	7	0.11%
Bos taurus	ncbitaxon:9913	3	0.05%
Arabidopsis thaliana	ncbitaxon:3702	2	0.03%
Martes zibellina	ncbitaxon:36722	1	0.02%

Table 2.2: Data sources in the DECA corpus

Main Species	Source	Abstracts
Fly	BC1 Devtest	108
Mouse	BC1 Devtest	250
Yeast	BC1 Devtest	110
Human	BC2 Test	262

2.2 Focus species recognition

Identifying the main model organism of an article and linking the genes in the article to their unique identifiers in a database is one part of curation task. This "link" step can be considered as a text classification task which was well studied in the general English domain [52]. If we treat the task of NER as a 'linkage task' at the word level, then we can treat the task of focus species recognition as a 'linkage task' in text level. The two tasks can be considered similar as the text is constructed by a sequence of words.

Consider the text classification task in the biomedical domain, the main purpose of task is to identify the organism of special interest within the given paper. There are several ways to achieve this purpose. (1) cluster the document into different classes and the papers in the same class will be about the same organism. (2) Using keyword to link the paper to special organisms. For example, if "Drosophila" appeared in the paper, the paper is much likely to talk about drosophila. (3) Using machine learning method to combine the several evidences together and decided the focus species of the paper.

H. Liu and C. Wu (2004) [53] studied the text classification for four organism (fly, mouse, yeast and worm) of 31414 MEDLINE papers. This dataset was low ambiguity (1%), ie. lower than 1% of papers had genes from more than one species mentioned in the abstract. They created a keyword list from NCBI (http://www.ncbi.nlm.nih.gov) and UMLS knowledge sources (http://umlsks.nlm.nih.gov). They assume that if the title,abstract or Mesh Headings of the MEDLINE papers contains the words in the list, the paper was a relevant article. The list they used was shown in table 2.3. The feature they used were stemmed words from Abstract, stemmed words from title,

Table 2.3: List of species synonyms

ORGANISM	KEYWORDS
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

Author of the paper ,Mesh Headings and Journals. The papers in previous years was used for training of the SVM model. They reported a best F-score around 94.1%.

Kappeler et al., (2009)[6] devised a system to detect the focus organisms in biomedical papers. Their approach used the NCBI taxonomy to make a protein-organism list and this was used to detect the focus species in full text articles. They counted the number of different organism and use a statistic method to create a ranked list of the focus organisms of the paper. In their experiments, they assumed that the organisms occurred in abstract is much important than the ones occurred in full text. They also refer to the frequency of the organism in IntAct. Results showed a top F-score of 74.0% (Precision: 74.2%, Recall: 73.8%).

From the previous result we can see, in a low ambiguous dataset, text classification achieved a high F-score even in abstract. However, in a high ambiguous dataset, disambiguation is an important part in focus species identification. Kappeler et al. (2009) [6] chooses full text to include more gene information. For full papers, J. Lin 2009 [54] compared full text and abstract in IE task. He showed the value of full text collections for text retrieval.

Identifying the focus species for an experiment is a specialized subtask of topic classification that requires precise identification of semantic features such as the

gene/protein which is the topic of discourse. In earlier work on focus species identification, word level features were explored by F. Rinaldi et al., 2008 [4] for target organisms including *Homo sapiens*, *Mus musculus*, *Saccharomyces cerevisiae* and *C. elegans*. Rinaldi et al. devised a system that created a ranked list of species for each MEDLINE abstract and demonstrated its effectiveness for disambiguating gene references. Results showed that the number of possible gene references was reduced to 45012 (p = 0.0308, r = 0.5763) from 283556 in the initial annotation step (p = 0.0072, r = 0.7469). Wang and Matthews, 2008 [5] created a system for gene entity recognition and identification that used a combination of species name and gene name features co-occurring in the same sentence. They showed significant improvements of 11.6% on accuracy for the gene mention task. In order to study term level species identification, Wang et al. [1] manually created an annotated MEDLINE dataset, providing a species id for each gene mention.

Sometimes, the focus species was not easily classified by only considering the surface clue of the given paper. For example, Figure 2.7 shows part of an abstract from Bignon et al. (PMID: 8370518). Here the authors discuss homolog experiments on mouse which has the potential to associate with the human RB protein. One of the messages of the abstract is that the human RB gene can functionally complement the mouse homolog. It is clear that the mouse is the experimental model but the results have important implications for Human gene function. However, without care the strong mentioning of the human RB protein would bias a naive model towards classifying the focus species in this article as Homo sapiens.

Title: Expression of a retinoblastoma transgene results in dwarf mice.
Abstract:
Introduction of the normal retinoblastoma gene (RB) into different tumor cells possessing inactivated RB genes suppresses their tumorigenicity in nude mice. These results suggest that RB replacement is a potential strategy for developing future clinical treatments of cancer. In a transgenic mouse model, we found that the quantity of RB protein in a given cell may play an important role in dictating its effect. Four founder mice containing 1-7 copies of a human RB cDNA transgene under the transcriptional control of the human RB promoter were generated. Most of the transgenic mice were smaller than nontransgenic littermates. This effect was found as early as embryonic day 15. The degree of dwarfism correlated roughly with the copy number of the transgene and the corresponding level of RB protein. The expression pattern of the transgene products was similar to that of the endogenous mouse RB gene with regard to tissue and temporal distribution. Transferring the transgene to RB deficient mice, which are nonviable, resulted in the development of normal, healthy mice, indicating that the human RB gene can functionally complement the mouse homolog. These studies demonstrate that the effect of RB on overall mouse development is closely dependent upon its dosage.

Figure 2.7: Abstract from Bignon et al. (PMID: 8370518)

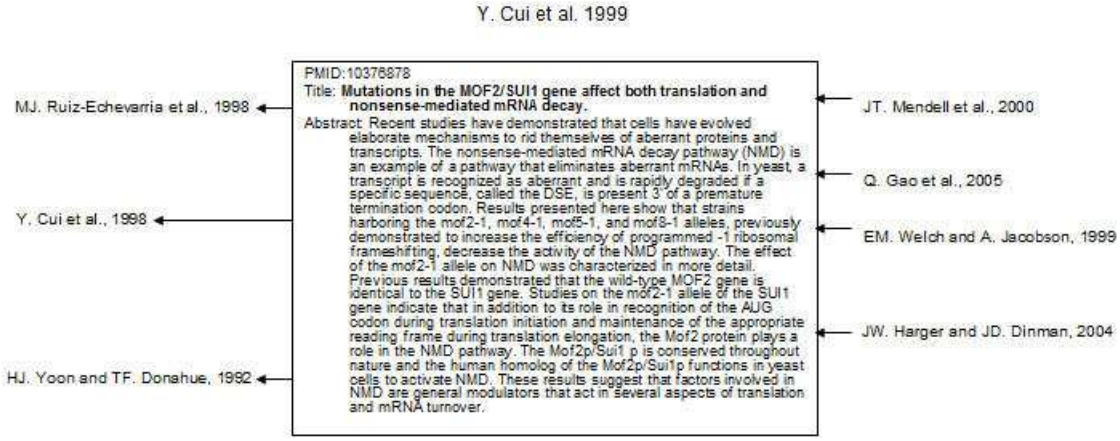


Figure 2.8: An example of citation structure

2.3 Citation analysis

Citation plays a central role in the progress of writing a paper. Many researchers are interested in the purpose of the citations. For example, citations are used to introduce the starting point where the authors' research work started [55]. Moravcsik and Murugesan [56] divided the citations into four purposes:

(1) Conceptual or Operational use. The purpose of the citation is use the theory of use the technical method in the citation paper.

(2) Evolutionary or juxtapositional use. The purpose of the citation is to show that the author's work is based on the cited work or the author's work is an alternative to the citation one.

(3) Organic or perfunctory. The author's work needs the understanding of the citation article or the citation is a general acknowledgement.

(4) Affirmative vs. negation. To confirm or correct the findings in the citations.

And they showed that 40% of the citations are in case 2.

Figure 2.8 shows an example of the structure of citations. The citation includes two ways, *cited* and *citing*. *Cited* is the paper was cited by other papers and *citing* is the paper citing other papers. It is well know that biomedical papers had longer reference lists on the average than papers in mathematics and engineering [57].

According to Moravcsik and Murugesan [56] definition, the citations with the purpose evolutionary or juxtapositional use can be treated as a related work of the original paper, in other words, such kinds of citations can be assumed to have the same focus species as the original paper.

As citations can be freely obtained from website such as Google scholar or CiteSeer [58], there are many algorithms for uncovering the strength of hidden relations inside citation networks. One of the citation relations can be called citation function. Citation function was defined as the reason of author to cite a given paper. Why one paper was cited was an interesting question for scientific researchers for many years. J. Swales [59] showed that the scientific writer cited the papers with a special structure. S.B. Shum [60] showed that researchers are often interested with the relation of the citations. Case and G M. Higgins [61] argued that authors tended to cite "concept markers" representing a genre of work. To research the relationship of the citation, many researchers gave their own citation functions. For example, the scheme of M. Weinstock [62], the scheme of J. Swales [63] , the scheme of C. Oppenheim and S P. Renn [64] and etc. One of the well-known schemes was explored by S. Teufel et al., 2006 [2]. He explored automatic classification of the citation function, i.e. the reason why a work was cited. With the development of natural language processing, researchers began to move the focus to automatically classify the citation functions. For example, M Garzone and R E. Mercer [65] developed a classifier to

automatically classify the citation function in scientific papers. In recent years, citation analysis was extended from scientific literatures to biomedical domain. Previous work on citation analysis in the biomedical literature includes work on bibliometrics and enhanced ranking of search. I Tbahriti et al. 2006 [7] for example, looked for related articles using argumentative categories in MEDLINE abstracts and measured this with co-citation lists; P. Nakov et al., 2004 [8], used text surrounding citations for grammatical paraphrase extraction; and S. Teufel et al., 2006 [2] explored automatic classification of the citation function, i.e. the reason why a work was cited. The application of citation analysis to text classification was also attempted within the computer science domain where B. Zhang et al., 2005 [9] reported a 7% improvement using citation information.

2.4 Discussion

According to the survey in this chapter we have seen that focus topic identification of biomedical papers has the potential to help authors and curators integrate their published results into databases much more easily. To achieve this automatically would reduce the ambiguity in gene name identification. However using only document internal information it seems there is still a significant performance gap. Since citation information is freely accessible I will explore the use of citation analysis through a dedicated citation scheme which I will propose for biomedical papers. In contrast to previous studies, my approach will not consider target documents in isolation but will try to leverage the use of a network of citation relationships, amplifying information which is implicit in the target document.

Chapter 3

Experiment 1: Focus topic identification for full papers

As noted by J. Lin (2009) [54], full papers seem to be valuable for various information extraction tasks compared to abstracts since full papers contain much more information than abstract. Before exploring document external information I wanted to see if using full papers would allow us to identify focus species more easily. In this chapter I present a series of experiments designed to compare the two text types.

Hypothesis one: ***What level of classification performance is achievable using state of the art lexical semantic features for focus specifies in full papers and abstracts?***

3.1 Experiment setup

3.1.1 Data set

The dataset I employ was based on the BioCreAtIvE I task 1B corpus which was manually selected from three model organism databases: Fly [13] (*Drosophila melanogaster*), Mouse [12] (*Saccharomyces cerevisiae*), Yeast [51] (*Mus musculus*). PubMed IDs were

selected from the databases and MEDLINE abstracts were selected according to these PubMed identifiers to make up the BioCreAtIvE I task 1B corpus. There are 4 gene mentions in each abstract on average. I manually collected the corresponding full papers for the abstracts from PubMed and Google search. The final corpus contained 3761, 3572, 3725 papers for fly, yeast and mouse respectively.

3.1.2 Work flow

The workflow for the experiment is shown in Figure 3.1. (1) Documents were cleaned and saved in a standard format; (2) Documents were then classified using a rule-based classification model. The purpose of this step was to choose the easiest cases in the dataset and classify them first. The heuristic rule was simple: if a title contained only one organism mention then the text was tagged according to that organism. In this way 5% of documents were classified, and the remaining documents were resolved in the following steps; (3) AbGene [66] was used to annotate the gene names in each document and which part of the document should be used was determined by using a content selection model. One-hundred articles with similar structures (abstract, introduction, result, experiment, discussion, and conclusion) were selected manually and a gene-section distribution for these 100 articles was created which I called as gold-standard gene distribution. Based on this analysis the abstract, introduction, result and conclusion sections were selected, and other sections were excluded. If an article contained sub-title of each section, then the four sections were selected out automatically. If an article contained no significant sub-title to show these four sections, the genes were annotated by AbGene and gene number of each section were calculated. Compared to the gold-standard gene distribution, the four sections were

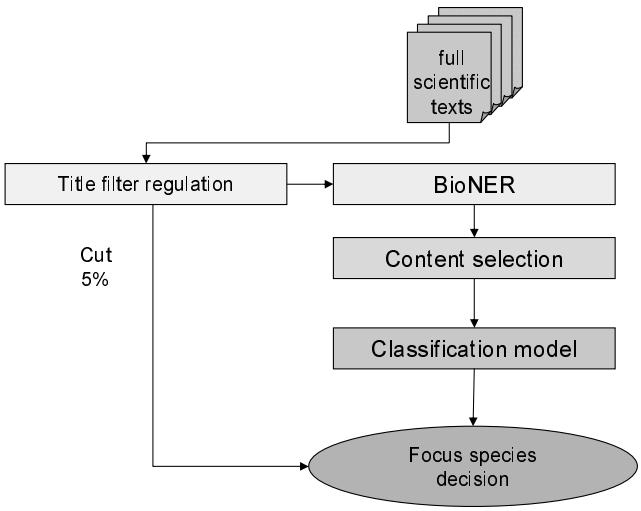


Figure 3.1: Structure of FFS model

decided. (4) Additional features such as title and journal name were then added; (5) Eight supervised models were used to classify the documents. In this step, the data remaining undecided from step (2) were used. I then analyzed the model's performance using ablation experiments on various combinations of features.

3.1.3 Models

In my experiments, I compared eight supervised classification methods: Naive Bayes, Conditional Random Fields, support vector machines (SVMs), Decision table, Decision trees, Logistics Regression as well as Adaboost and Bagging on the best performing models.

1. Naive Bayes: The Naive Bayes model is a simple probabilistic classifier based on Bayes's theorem with strong independence assumptions that is widely used in text classification. The Naive Bayes implementation I used was included in the Weka toolkit [67], default parameters were used for training.

2. The conditional random fields algorithm (CRF): Conditional random fields (CRF) [68] is a discriminative probabilistic framework that is used for labeling and segmenting sequential data. A CRF is an undirected graphical model that defines a single log-linear distribution over labeled sequences given a particular observation sequence. Recently Hirohata et al. [69] showed success in applying CRF for a document classification task. I applied the same broad methodology as Hirohata et al. in my implementation. I formulated the document classification task as a sequence labeling task by firstly labeling each document section with its focus species and then labeling the focus species for the whole document based on the sequence of section labels. The CRF++ toolkit [70] was used. The hyper-parameter to set the trade-off between over-fitting and under-fitting was set at 10. Default values were used for the other parameters.

3. Support Vector Machine (SVM): SVMs were introduced by Vapnik [71] in 1995 as a learning system that uses a hypothesis space of linear functions in a high dimensional feature space, trained with a learning algorithm from optimization theory that implements a learning bias derived from statistical learning theory.

4. Boosting and Bagging: Boosting [41] and bagging [72] are generic methods aimed at aggregating classifiers for improved prediction performance using sample weighting and re-sampling respectively on the original training data. Both

techniques can be applied to a variety of base learners and have been shown to give substantial gains in accuracy for classification tasks. In my experiments Naive Bayes was chosen as the base learner for its high level of performance in the stand alone task.

5. Decision tables : Decision tables [73] contain two major components, a list of attributes and a set of labeled instances on those attributes. Labeling is done by default on majority class matching and then by arbitrary tie breaking and attribute elimination. They have a close relation to rule-based knowledge bases.

6. Decision trees : Decision trees [36] are potentially powerful predictors and explicitly represent the structure of a rule set in tree form with leaf nodes functioning as classification decisions and transitions along branches taking place according to attribute values.

7. Logistic regression [38]: Logistic regress [38] is a popular discriminative classifier for modeling binary data.

In my experiment, AdaBoost, Bagging, Decision tables, Decision trees and Logistic regress were implemented from the Weka toolkit.

3.1.4 External resources

For the named entity recognizer, AbGene was used to annotate the gene names in the document. AbGene [74] was pre-trained on annotated MEDLINE abstracts with a reported F-score of 98%. Tanabe [66] showed that it is possible to use AbGene on full text articles from PubMed Central (PMC) with a reduced level of performance at 72.6% precision and 66.7% recall. Since my abstracts were selected from MEDLINE

and the full texts were selected from PMC and Google search, we can expect broadly similar levels of performance with this earlier experiment.

3.1.5 Features

The experiment tested several linguistic features which I describe in detail below:

1. GN:Gene name terms

   Following gene name annotation with AbGene, genes were listed according to their frequency in the document and the top n genes were selected as features to train the model. Here, n is a fixed number decided before the experiment. I varied n from 1 to 100 in preliminary experiments, with the results indicating that the larger n was, the better the results were. As n > 100 was difficult to handle using my CRF software due to machine memory limitations, n=100 was used in the experiment.

2. OF:Organism frequency

   Organism name mentions were used as a reference for classifying the text into different model organisms. The organism names included not only mice, fly and yeast but also synonym words such as mouse, drosophila, and saccharomyces. This list was compiled by hand according to the NCBI taxonomy.

3. MH:MeSH headings

   Mesh heading has been proven effective across many tasks in the bioNLP application domain.

   Medical Subject Headings (MeSH)[75] is a comprehensive controlled vocabulary. The purpose of MeSH is to index journal articles and books in the life sciences. It

was used in MEDLINE/PubMed databases. Most of the MeSH terms are short descriptions or definitions, linked to related descriptors and synonyms or similar terms. Every journal articles in MEDLINE was indexed with some 10-15 MeSH terms. In that that cases, MeSH can be treated as an index clue of the focus organisms of the article. Bloehdorn and Hotho [41] report that MeSH headings improved the accuracy of classification by 3% to 5%. I therefore selected the three frequently mentioned MeSH headings for each based on frequency in the training data.

4. DT: Document title terms

   Some of the document titles contained organism name mentions and gene name mentions which were then used as features in the rule classification model and NLP classification model.

5. TS: Term-species

   If one sentence contained a gene name and a species name, the weight of the species name was counted by using the distance between the species name and gene name. The total weight was tallied for each article, and the weight of the species name was used as a feature.

6. JN: Journal Name

   This was the name of the journal in which the abstract or article was published.

7. NT: Number of gene terms

   First, gene list was extracted from the training corpus and sorted by the frequency of the gene. Then the number of genes in the top-100 frequent gene list

was counted.

8. AGN: Additional gene name terms

   When there was a gene-species pair in one sentence, the gene name and species name was used to find an additional gene name in UniProt. For example, if there was a gene named "IL2", by looking it up in UniProt, the additional gene name "Interleukin" could be found.

3.2 Result of experiment one

3.2.1 Experiment 1.1: Comparison on different learner models

In the first sub-experiment, eight different models were selected: Naive Bayes, AdaBoost, Bagging, Decision table, Decision tree, Logistics Regression, CRF and SVMs. Table 3.1 compares the 10-fold cross evaluation of the different models. NB had the highest F-score (84.8% for fly, 73.9% for mouse and 73.8% for yeast), and CRF had the second highest (80.2% for fly, 73.0% for mouse and 72.3% for yeast). AdaBoost and Bagging both used Naive Bayes as the base learner, but we did not observe a significant improvement when using the basic feature set. Logistics Regression performed well on fly (79.6%) but not so well on the other two species. SVMs gave high precision but low recall in fly and yeast; high recall but low precision in mouse.

The model comparison used only the basic feature set (MeSH headings, journal name, gene name, and article title). I also did feature analysis on MeSH headings and journal name in this experiment. The analysis showed that by using MeSH headings

Table 3.1: Classification performance across 8 machine learning models

		F1			F1-JN			F1-MH		
		P	R	F	P	R	F	P	R	F
NB	fly	0.780	0.929	0.848	0.685	0.890	0.777	0.530	0.808	0.640
	mouse	0.810	0.680	0.739	0.697	0.620	0.656	0.683	0.410	0.776
	yeast	0.750	0.727	0.738	0.646	0.515	0.573	0.747	0.657	0.828
AdaBoost	fly	1.000	0.182	0.308	0.530	0.808	0.640	0.370	0.717	0.488
	mouse	0.359	0.900	0.513	0.683	0.410	0.513	0.333	0.090	0.142
	yeast	0.310	0.091	0.141	0.747	0.657	0.699	0.443	0.354	0.393
Bagging	fly	0.371	0.232	0.286	0.371	0.232	0.286	0.371	0.232	0.286
	mouse	0.342	0.400	0.369	0.342	0.400	0.369	0.342	0.400	0.369
	yeast	0.345	0.414	0.376	0.345	0.414	0.376	0.345	0.414	0.376
Decision Table	fly	0.532	0.667	0.592	0.556	0.354	0.432	0.532	0.667	0.592
	mouse	0.515	0.520	0.517	0.388	0.870	0.537	0.515	0.520	0.517
	yeast	0.740	0.545	0.628	0.727	0.081	0.145	0.740	0.545	0.628
Decision tree	fly	0.637	0.798	0.709	0.500	0.687	0.579	0.341	0.606	0.436
	mouse	0.557	0.640	0.595	0.500	0.550	0.524	0.339	0.400	0.367
	yeast	0.729	0.434	0.544	0.596	0.313	0.411	1.000	0.040	0.078
Logistics Regression	fly	0.878	0.727	0.796	0.932	0.697	0.798	0.663	0.576	0.603
	mouse	0.586	0.750	0.658	0.721	0.490	0.583	0.541	0.730	0.621
	yeast	0.705	0.626	0.663	0.513	0.808	0.627	0.740	0.545	0.628
CRF	fly	0.762	0.868	0.802	0.688	0.879	0.764	0.503	0.830	0.621
	mouse	0.789	0.700	0.730	0.725	0.640	0.669	0.766	0.390	0.511
	yeast	0.734	0.725	0.723	0.652	0.566	0.598	0.744	0.650	0.685
SVM	fly	0.925	0.641	0.757	0.619	0.684	0.650	0.200	0.240	0.217
	mouse	0.403	1.000	0.574	0.309	0.406	0.351	0.250	0.147	0.185
	yeast	0.636	0.194	0.297	0.400	0.207	0.273	0.356	1.000	0.525

as a feature, a 2% improvement in F-score was achieved by using Naive Bayes and CRFs. The journal name feature improved the F-score by 1% by using Naive Bayes and CRFs.

3.2.2 Experiment 1.2: Comparison of different feature sets

NB and CRF were selected as the two best performing models from Experiment one. This time I used an extended set of features that included TS (term-species) and OF (organism frequency) in 10-fold cross evaluation experiments. The best performing combination achieved an average F-score of 90.7%. As shown in Table 3.2, classification for fly achieved the best among the three kinds of organisms (97.1%) followed by mouse (88.6%) and yeast (85.5%). I considered that the reason for this is that for fly focussed experimental papers, the gene-species pairing gave a clear signal, whereas in mouse the organism was often considered as the experiment model for human so the gene-species pair and organism frequency became highly ambiguous. In yeast the species name of yeast was rarely mentioned in the paper. The most significant result was that by using TS, OF and AGN features; an improvement of 10% was achieved.

3.2.3 Experiment 1.3: Comparison on full texts and abstracts

Large-scale collections of abstracts are often used in life science classification experiments, whereas full text articles are rarely used due to difficulties in sourcing them from publishers and converting them into plain text format. This trend is now changing due to the availability of open source publications. However, the highly detailed experimental information contained in full text papers reveals new challenges

Table 3.2: Classification performance across different feature set

		F1			F1+TS		
		P	R	F	P	R	F
NB	fly	0.78	0.929	0.848	0.97	0.97	0.97
	mouse	0.81	0.68	0.739	0.826	0.95	0.884
	yeast	0.75	0.727	0.738	0.929	0.788	0.852
CRF	fly	0.762	0.868	0.802	0.965	0.952	0.958
	mouse	0.789	0.7	0.73	0.814	0.878	0.845
	yeast	0.734	0.725	0.723	0.902	0.786	0.84
		F1+OF			F1+NT		
		P	R	F	P	R	F
NB	fly	0.792	0.931	0.856	0.775	0.825	0.799
	mouse	0.821	0.685	0.747	0.823	0.621	0.708
	yeast	0.762	0.731	0.746	0.752	0.723	0.737
CRF	fly	0.771	0.87	0.818	0.753	0.877	0.81
	mouse	0.791	0.71	0.748	0.865	0.698	0.773
	yeast	0.739	0.73	0.734	0.729	0.727	0.728
		F1+ADN			F1+TS+OF+NT+ADN		
		P	R	F	P	R	F
NB	fly	0.812	0.931	0.867	0.971	0.972	0.971
	mouse	0.823	0.712	0.763	0.827	0.953	0.886
	yeast	0.786	0.987	0.875	0.931	0.791	0.855
CRF	fly	0.773	0.887	0.826	0.966	0.954	0.96
	mouse	0.792	0.714	0.751	0.817	0.878	0.846
	yeast	0.762	0.751	0.756	0.901	0.788	0.841

for biomedical document classification. For example, Tanabe [66] showed that entities like restriction enzyme sites, laboratory protocol kits, primers, vectors, molecular biology supply companies, and chemical reagents are rarely mentioned in abstracts, but plentiful in the methods section of the full article. Their appearance adds to the previously mentioned morphological, syntactic and semantic ambiguities. To mitigate this issue, content selection was applied to filter data in the full articles according to sections. Secondly, the full text, especially the Method and Introduction sections, contain larger numbers of associated gene/protein mentions in comparison with the abstracts. Again, this can be partially mitigated by content selection.

On the other hand, there are also some advantages to using full texts over abstracts. Potential redundancy of information allows models with lower levels of recall to have several chances to discover reported facts such as the species-gene/protein features that we observed to be highly valuable when making decisions about focus species.

To confirm the value of using full texts I compared classification performance of the full texts from my corpus of abstracts to the original abstracts. The comparison is shown in Table 3.3. I performed a two tailed paired sample t-test to show that there is an improvement of 11 points in F-score. In these experiments 10x10 cross validation was used in conjunction with two-tailed corrected resample t-test ($p < 0.001$) as presented by Bouckaert and Frank 2004 [76].

3.3 Discussion

In this experiment, I presented a system to identify focus species using combinations of lexical semantic features and comparing across biomedical full text papers and

Table 3.3: Classification performance across full text papers and abstract

		full text (F1+TS+RN+NT+AND)			abstract (F1+TS+RN+NT+AND)		
		P	R	F	P	R	F
NB	Fly	0.971	0.972	0.971	0.812	0.892	0.850
	Mouse	0.827	0.953	0.886	0.755	0.763	0.759
	Yeast	0.931	0.791	0.855	0.791	0.748	0.769
CRF	fly	0.966	0.954	0.960	0.820	0.898	0.857
	mouse	0.817	0.878	0.846	0.732	0.741	0.736
	yeast	0.901	0.788	0.841	0.757	0.750	0.753

abstracts. By comparing different novel models in full papers and compare two best models with full papers and abstract, I conclude that by using state of the art lexical semantic features, an F-score of over 90% was achieved for full text papers and the F-score is less than 80% for abstracts. From this results, we can see that although lexical semantic features performed well on full text papers, for abstracts, the lexical semantic features is not enough. As the copyright problem of collecting the full papers, the size of my database is small and it is hard to extend it, one reasonable consideration is to introduce external features to improve the performance in abstracts.

3.3.1 Content selection

As discussed above, one difficulty for focus species classification on full text articles is that of content selection. Deciding which part of the document is the most valuable and developing a strategy to select it is quite a difficult issue given that documents in my collection come from different journals which have different section structures. As a proxy for explicit section headings I decided to use the gene mention distribution as a clue for partitioning the full text papers. However, this approach proved weak in cases where the test document contained more sections than the standard one (

i.e. the four sections mentioned in the methods). During analysis I found that using such section selections showed no improvement in F-score.

3.3.2 Feature selection

Another challenge was feature selection. Rinaldi et al. [4] used the species name appearing in a document as a clue to find the correct topic organism. My experiment built on Rinaldi's findings in that not only did it use the species word itself as a feature, it also used species-gene pairs appearing together in one sentence and weighted the species according to the distance between the gene and species. Doing so improved the average F-score by 12% compared to that for the basic feature set. Compared with Rinaldi's work, my approach showed an average 3% improvement in the F-score.

In the feature set, I used one feature called additional gene name terms. The additional gene name term is only existed when there is a gene-species pairs and the additional gene name is searched based on the species information. However, the additional gene name terms still provide ambiguous in some cases. How to reduce these ambiguous still remained as a problem.

3.3.3 Discussion: multi-species mentioned in one paper

Although many researchers have focused on text classification in biology, their experiments have mainly been targeted at extracting information about single organisms. Considering the task in the real world; texts are often not clean data on specific organisms.

The most difficult cases I encountered were when the text contained multiple species names. As the abstract below (PMID: 11018518) illustrates, four kinds of

species were mentioned: fly (*Drosophila melanogaster*), mouse, zebrafish and silkworm (*Bombyx mori*).

Coatomer is a major component of COPI vesicles and consists of seven subunits.The gamma-COP subunit of the coatomer is believed to mediate the binding to the cytoplasmic dilysine motifs of membrane proteins. We characterized cDNAs for Copg genes encoding gamma-COP from mouse, zebrafish, Drosophila melanogaster and Bombyx mori. Two copies of Copg genes are present in vertebrates and in B. mori. Phylogenetic analysis revealed that two paralogous genes had been derived from a single ancestral gene by duplication independently in vertebrates and in B. mori. Mouse Copg1 showed ubiquitous expression with the highest level in testis. Zebrafish copg2 was biallelically expressed in hybrid larvae in contrast to its mammalian ortholog expressed in a parent-of-origin-specific manner. A phylogenetic analysis with partial plant cDNA sequences suggested that copg gene was also duplicated in the grass family (Poaceae).

This is a special case, but approximately 5% of articles in my collection reported multiple species. In the future I will need to consider how to handle these special cases more efficiently.

Although in these experiments I achieved a high level of accuracy in focus species identification there are still some disadvantages that can be seen. Firstly, the automated sourcing of full papers is difficult to achieve given copyright restrictions. Secondly only document internal features were so far considered, ignoring the potential for external features to contribute to the classifier.

3.4 Conclusion

In the beginning of this chapter, we raised a sub thesis question on what level of classification performance is achievable using state-of-the-art lexical semantic features for focus species in full papers and abstracts? In this experiment, totally 11058 full papers were contained in the corpus, 10-fold cross method was used in the evaluation. Eight types of features were used in training, and three species were identified. In this experiment, by using state-of-the-art lexical semantic features for identifying the focus species, an average F-score of 90.4% is achieved for full text papers and compared to the full text papers, a lower level of F-score (79.3% on average) is achieved.

Chapter 4

Experiment 2: A system to identify focus topic in abstract

As shown in the previous chapter, I compared the accuracy of species identification in MEDLINE abstracts and full text papers with a best F-score of 97.1% for *Drosophila melanogaster*, 88.6% for *Mus musculus* and 85.5% for *Saccharomyces cerevisiae*. My findings indicate that the classification performance for the focus species in abstracts was much lower than in full text papers when using only document internal features. However in practice full text papers are not always available, e.g. due to copyright reasons, so I considered to study the focus species identification in abstracts. As shown in the previous chapter, internal features did not performed well in abstract which F-score less than 80%, new features are needed to identify the focus species in abstracts. In this new set of experiments, I explore similar document internal features to my previous study as a baseline. The contribution of this chapter to my thesis is to expand the investigation to see if the use of features harvested from external resources such as citing papers and associated papers can contribute to classification performance. By associated papers I mean those that are the results of the PubMed search engine. I also expanded the investigation from 3 to 5 focus species including

Homo Sapiens and *Rattus norvegicus*. Finally I develop a practical species tagger called FS tagger which is based on the best features discovered so far for automatically identifying the focus species in an abstract.

Hypothesis two: ***Of the abstracts which are cited or archived in the PubMed database, do bibliographic features provide enhanced classification accuracy?***

4.1 Experimental set up

4.1.1 Data collection

As a gold standard, I leveraged the newly released DECA corpus [1], which contains a wider range species than my previous study. In first stage experiments I look at recognizing and identifying gene/gene product mentions for their species taxon identifiers and show significantly improved performance compared to Wang et al.'s maximum entropy model. 15 different kinds of Taxon identifier was identified. In the second stage, I identified the 4 different focus species in abstracts using the combination of internal lexical features and bibliographic features. The study contributes to work on biological text classification and database curation. A novel characteristic of my approach is the analysis of various linguistic features in combination with bibliographic features and PubMed related citations to achieve state of the art performance.

Abstracts for the DECA corpus were selected from the BioCreAtIvE I & II collection. BioCreAtIvE I dataset contained three species, mouse, fly and yeast and BioCreAtIvE II dataset contained one species, human. In total 644 MEDLINE abstracts were manually annotated by assigning NCBI species IDs for each gene mention

by Wang [1] . Mentions of gene and gene products are annotated and a species ID assigned to every entity mention. The species tags are identifiers from the NCBI Taxonomy of model organisms.

Table 2.1 illustrates the distribution of species IDs for gene and gene products in the DECA corpus given by Wang et al., 2010 [1]. In my experiment, the species IDs was used as the gold standard in the first step. Abstracts focusing on *Drosophila melanogaster* (96 papers), *Mus musculus* (204 papers) and *Saccharomyces cerevisiae* (92 papers) were selected from the BioCreAtIvE I task 1b development test corpus [77] and abstracts focusing on *Homo sapiens* (252 papers) were selected from the BioCreAtIvE II dataset [78]. I annotated the focus species using the classes in the BioCreAtIvE sources. For example, the data selected from the BioCreAtIvE II dataset is considered to focus on *Homo sapien*s.

The reason I chose DECA corpus was:

- The abstract in DECA corpus was selected from 4 organisms and the gene and protein mentions in the abstract contains 11 kinds of organisms. That is to say, the number of organism is fit to use in this task.

- The NCBI species id is widely used in the gene normalization task. Although the corpus was small, each of the NCBI id was annotated manually. The accuracy of the annotation was believable, I showed an F-score over 85% for identifying the Taxon ID for four top sequence species.

- The abstracts in the DECA corpus were selected from MEDLINE abstracts. It is easy to find related information such as Mesh, Title and etc. from PubMed database.

4.1.2 Workflow

An outline of the workflow for the experiment is shown in Figure 4.1 with a more detailed description following. FS tagger includes three stages: the GT model (Gene mention Taxon ID annotation model), the EIE model (External resource information extractor) and the FSD model (Focus species decision model). In the GT model, the GENIA named entity tagger [79] was used to tag the gene and gene product mentions. Using the GENIA tagger's output, a Conditional Random Fields (CRFs) model was used to annotate the NCBI Taxon ID for each gene mention. I choose CRF model because the tagging is treated as a sequence labelling task and CRF showed a good performance in such tasks [30]. Next, in the EIE model, related citation papers were automatically downloaded from PubMed and citing papers were downloaded from PubMed and Google search and species information was extracted from the related citations and citing papers. With the result of the GT model and EIE model, some basic features such as the gene mention and species mentions in title and journal name were also added. Another CRF model was then used to identify the focus species of the abstract. This is described in more detail below.

4.2 GT model

In the GT model, the goal is to assign the Taxon ID to each gene and gene product mention that has already been identified by GENIA tagger. In the example below, RB is a gene tagged with Taxon ID 10090.

Ex 1. The expression pattern of the transgene products was similar to that of the endogenous mouse $< TaxonId = 10090 >$ RB $< /TaxonID >$ gene with regard to

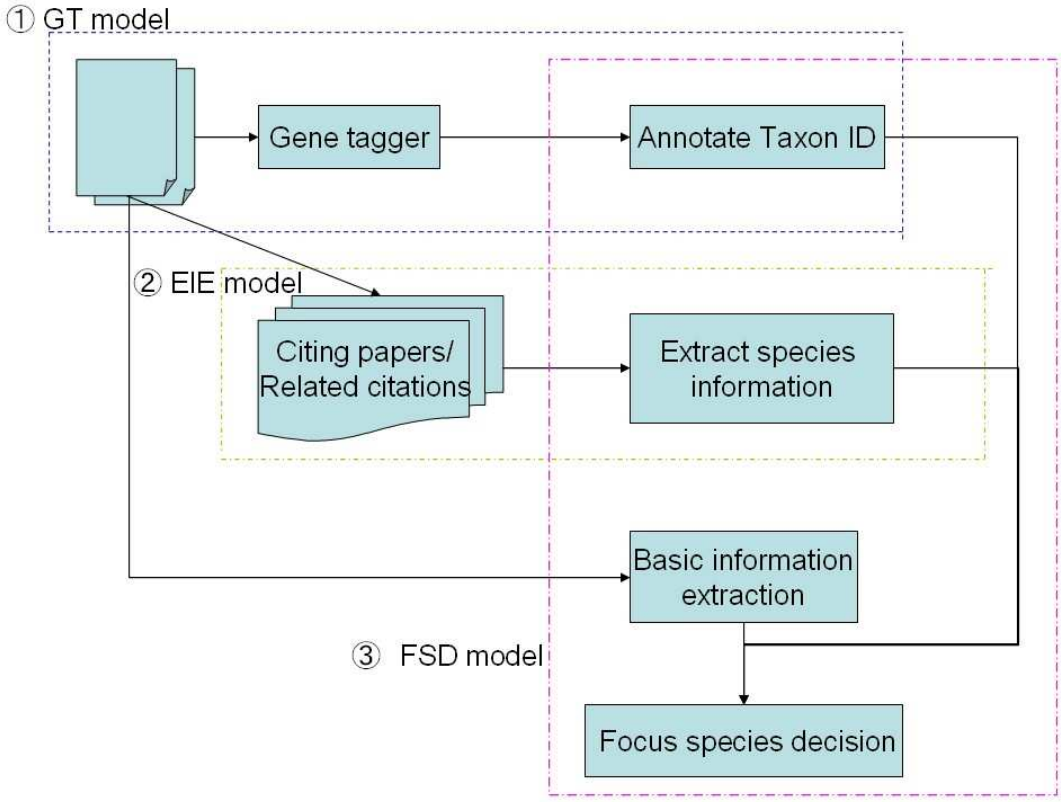


Figure 4.1: Structure of FS tagger

Table 4.1: Features used in GT model

Features	Descriptions
Word tokens	the surface form of word itself
Part of speech (POS)	POS was given by GENIA named entity tagger
Suffix//prefix	the suffix of the word, for example, 'ex', 'im' and etc A suffix list was made manually.
Orthography	a normal used feature in natural language processing
Brief word shape	String are changed to 'a' and number are changed to 'o'. For example, IL2 will changed to a0
GENIA named entity tagger	The result of GENIA named entity tagger. There are 5 types of entity types: Protein, DNA, RNA, Cell line and Cell type. The entities with the entity type Protein was used as one feature in the GT model.

tissue and temporal distribution.

No explicit label difference between gene and gene products was made either in the DECA corpus or the GT model. In the experiment, evaluations were carried out using 5-fold cross-validation.

The model I used is based on a machine learning method called conditional random fields (CRF) [68] with a feature set of word tokens, part of speech, suffix/prefix, orthography, brief word shape and GENIA named entity tagger. The feature set is shown in Table 4.1.

CRF itself is a discriminative model for labelling structured data. It has been widely used in NLP tasks and has been proven to offer state-of-the-art performance in sequence labelling tasks such as part of speech tagging and named entity resolution as well as a number of real-world bioinformatics tasks such as protein structure prediction and RNA structure alignment. This is due to its relaxed independence assumptions over hidden Markov Models and Maximum Entropy Markov Models. The CRF++ toolkit [70] was used with default values for the parameters. In GT model, one

sentence is treated as a sequence labelling data and each mention in the sentence is labeled by CRF model. With the result of CRF model, the mentions which were tagged as Protein by GENIA named entity tagger were picked out for evaluation.

The result of the GT model, which is the taxon ID for each of the gene names, is used as one of the features in the FSD model described later.

4.2.1 EIE model

The purpose of the EIE model is to identify the closest matching related paper to the target abstract using bibliographic data or search engine associations.

In the EIE model, two kinds of external resources were used: related citations and citing papers.

(1) Related citations provided by PubMed [80]. PubMed uses a word-weighted algorithm to compare words from the title and abstract for each target abstract, as well as the MeSH Main headings assigned. The best matches by PubMed [80] for each abstract in the collection were pre-calculated and designated as "related citations". In the EIE model, all related citations were downloaded from PubMed automatically. As shown in Figure 4.2, related citations provided by PubMed were divided into two sets according to the time of publication relative to the target article. Abstracts published earlier than the target article are marked as II and those published later than the target article are marked as I.

(2) Citing papers. The set of articles citing the target article. Citing papers were downloaded using PMC's "cited papers in PMC". As shown in Figure 4.2, the citing papers are designated as IV.

Set III in Figure 4.2 shows the set of papers which are found to be associated with

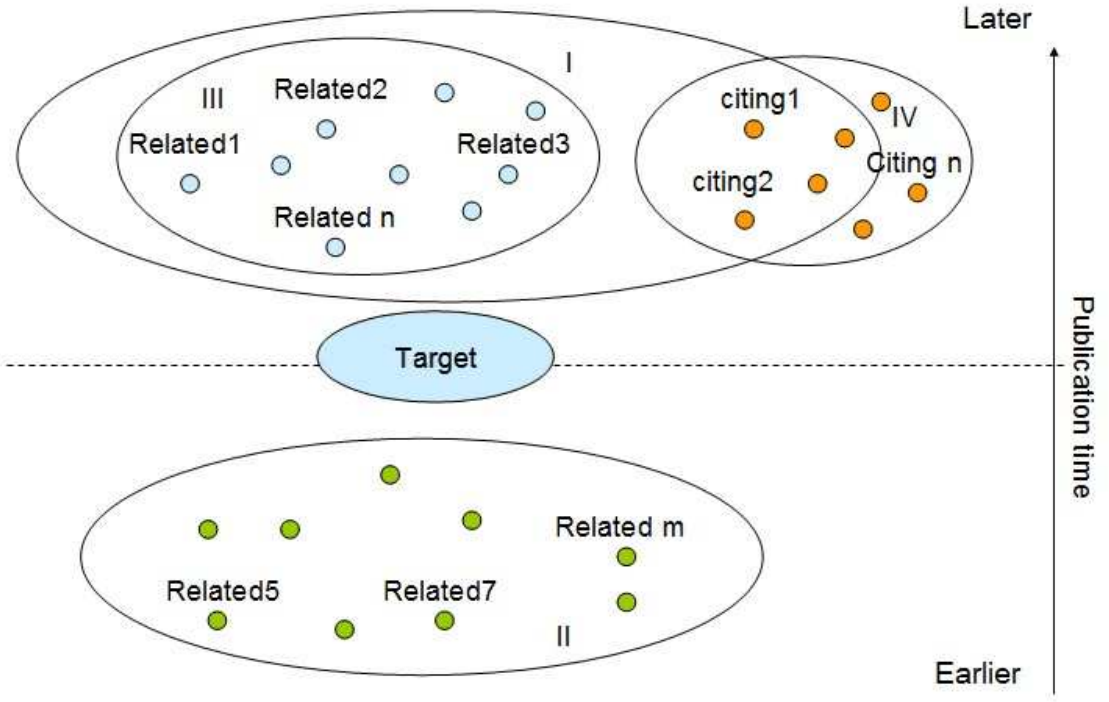


Figure 4.2: External resources

the target abstract by PubMed but which have no explicit citation relating them to the target paper.

I assumed that the species mentioned in related citations and citing paper was similar to the species mentioned in the target and that this would become clear by aggregating information across papers. Such features should provide higher accuracy for species identification by helping to (a) reinforce internal clues about the focus species in the target abstract and/or (b) making explicit any hidden understanding on behalf of the reader. In these experiments, I evaluated a distance metric called Hierarchical Bayesian Clustering (HBC) for clustering external resources.

HBC [81] is a widely used method in document clustering that I evaluated for selection of citation papers. Given a collection of documents D, a binary tree is

Input:
$D = \{d_1, d_2, ..., d_N\}$: a collection of N data;
Initialize:
$C_0 = \{c_1, c_2, ..., c_N\}$: a set of clusters;
$c_i = \{d_i\}$ for $1 \leq i \leq N$
calculate $SC(c_i)$ for $1 \leq i \leq N$
calculate $SC(c_i \cup c_j)$ for $1 \leq i \leq j \leq N$
for $s = 1$ to $N - 1$ do

Figure 4.3: Hierarchical Bayesian Clustering algorithm (M. Iwayama and T. Tokunaga, 1995)

constructed. In the first step, each document is treated as one cluster Ci (also called a tree). Then the Maximum Likelihood P(C—D) is calculated for each pair of trees and the two clusters (trees) with the largest likelihood are merged into one. The procedure is repeated until one cluster remains. The algorithm is shown in Figure 4.3.

An example (PMID: 10376878) is shown in Figure 4.4. In my experiments, I assumed that the two nearest documents belonged to the same species. In this case, the nearest document to the target abstract (PMID: 10376878) is the brother leaf (PMID: 9488467) of the tree. Then the tree is cut in the first merge step and the first round result (PMID: 10376878) of merging step was used.

As a further part of my investigation I compared the performance of selecting the closest matching paper with HBC against using species features from all citation papers.

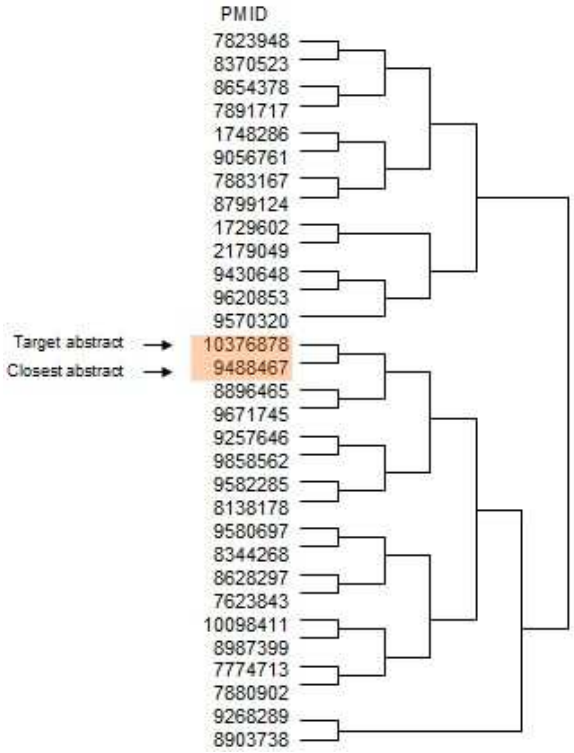


Figure 4.4: Example showing clustered related citations for abstract PMID 10376878.

Table 4.2: Internal Features used in FSD model

Features	Descriptions
Gene name terms (GN)	The GENIA named entity tagger was used to annotate genes and gene products in the target abstract. Using the result of the GT model, each tagged gene name was assigned a Taxon ID which was also used as a feature.
Species words (SW)	The list of species words was extracted manually from the NCBI Taxonomy of model organisms.
Document title terms (DT)	Species words and gene names were extracted from the abstract title.
Journal Name (JN)	a An index value was assigned indicating the name of the journal in which the abstract was published.
MeSH Main headings (MH)	S. Bloehdorn and A. Hotho, 2004 [41] report that MeSH Main headings improves the accuracy of classification by 3%-5%. I used the top three most frequent MeSH Main headings for each paper. Frequency was calculated from the DECA corpus abstracts.

4.2.2 FSD model

In the FSD model, the objective was to tag the focus species for the target abstract. In this model, a CRF with both document internal and external features was used. Internal features are shown in Table 4.2. By using CRF, I treated the abstract as different sentences and the task is to give the label to different sentences and the label of the whole abstract is given by combination of the label of different sentences.

External features consisted of species words from the EIE abstracts. I didn't using taxon id in EIE abstracts as external features because that I'm afraid that the performance of identifying the taxon id in EIE abstracts is not good enough. Of course, if performance of identifying taxon id in EIE performance can be ensured, it is better to include such value as external features. EIE abstracts were selected as

Table 4.3: Performance comparison of classification on 5 species against Wang et al. [1]

Species Name (TaxonID)	X. Wang et al. (ME)	X. Wang et al. (HYBRID)	My system (ME)	My system (CRF)
Homo sapiens (9606)	85.6	86.48	92.48	93.53
Mus musculus(10090)	79.38	80.41	80.21	89.61
Drosophila melanogaster(7227)	87.07	87.37	70.11	85.38
Saccharomyces cerevisiae(4932)	82.66	84.64	42.56	86.86
Other	0	25	0	22.9
Rattus norvegicus(10116)	48.42	59.41	44.61	22.2
Average	82.69	83.8	82.1	90.12

explained in previously and I conducted ablation experiment to discover the best set of features. Experimental results are reported below.

In the corpus, some papers contained mentions of more than one target species. However, the focus species was decided using the classes in the BioCreAtIvE sources as described earlier.

4.3 Result of Experiment Two

4.3.1 GT model

In the GT model, the main task is to assign the Taxon ID to each gene and gene product mention. In the experiment, evaluations were carried out using 5-fold cross-validations. Each combination of features was compared to Wang's system [1] using micro-average F1 score. Table 4.3 shows a comparison of the results.

In Wang's ML method, a maximum entropy (ME) model was used. The features employed were lexical feature which are the same as those used in my GT model, but also included neighboring species ID and all species IDs occurring in the document. However, neighboring species ID in my GT model were not employed, because the species ID was used in a later step and I wanted to avoid the duplicate usage of the same feature. The results in Table 4.3 show that my system performed better for higher frequency species. For lower frequency species, performance was not so high because neighboring species IDs were not used in my feature set. This was because neighboring species IDs provided some species information to which species was the gene belongs to. I conclude that overall differences in performance between the two approaches can largely be an attributed to the learner model, i.e. CRF rather than ME.

4.3.2 FSD model

The contribution of external features is shown in Figure 4.2. I first selected the abstracts from each set of external resources I-IV shown in Figure 4.2 and applied the HBC method and then extracted species information which was described in the Method section.

Ablation experiments were conducted to compare different sets of features as shown in Table 4.4. Species is indicated as H (*Homo sapiens*, NCBI taxon 9606), M (*Mus musculus*, NCBI taxon 10090), F (*Drosophila melanogaster*, NCBI taxon 7227) and Y (*Saccaromyces cerevisiae*, NCBI taxon 4932).

After introducing external document features, the performance increased greatly, especially for *Saccharomyces cerevisiae*. This is caused by that the title of selected

Table 4.4: Micro-averaged 10-fold cross validation comparison of features for focus species classification

Spec.	GN			SW		
	P	R	F	P	R	F
Average	9.78	24.99	14.01	49.47	29.27	35.43
H	38.14	99.96	55.65	40.62	98.23	56.85
F	0.00	0.00	0.00	96.08	9.91	17.50
M	0.00	0.00	0.00	31.42	5.50	8.65
Y	0.00	0.00	0.00	29.75	3.43	5.94
	GN+JN			GN+DT		
Average	31.22	27.22	30.09	12.09	25.97	16.24
H	39.56	95.58	55.34	40.34	97.01	56.22
F	58.80	5.93	10.34	8.03	6.87	6.65
M	26.51	7.35	10.85	0.00	0.00	0.00
Y	0.00	0.00	0.00	0.00	0.00	0.00
	GN+JN+DT			GN+JN+DT+MH		
Average	39.79	33.30	37.18	43.14	38.78	42.83
H	45.15	90.66	59.51	50.42	91.04	64.21
F	52.56	28.64	33.13	56.73	43.12	47.03
M	51.55	12.45	19.06	55.50	19.52	28.12
Y	9.90	1.43	2.47	9.90	1.43	2.47
	GN+JN+DT+MH+SW			GN+JN+DT+MH+SW (related citation)		
Average	46.34	39.52	43.40	**76.96**	**74.50**	**77.93**
H	52.66	87.55	64.90	77.21	79.63	87.59
F	53.40	50.85	49.95	79.92	98.45	76.85
M	68.98	19.13	28.57	87.84	86.98	86.66
Y	9.90	0.53	0.99	62.89	32.95	41.63

associated/citing papers for the papers of *Saccharomyces cerevisiae* is much likely contain the species word. The best performing combination of features achieved an overall F-score of 91.14%, which is 47.74% above the best model without features extracted in the EIE model. It can be argued that the size of the DECA corpus leads to low frequency feature counts particularly for the rare species such as *Rattus norvegicus* and that there is an inbuilt bias towards a method that uses an external knowledge source like MEDLINE [54]. However I note that the internal features included the results of the GT model with over 90% F-score for gene name disambiguation. The result indicates that this potentially rich source of information, such as the related citation provided by PubMed, contained a high degree of ambiguity that could not easily be resolved using internal clues alone.

I also noticed that although the journal name and title by themselves do not appear to improve the F-score, combining them with other features improves performance. This was especially noticeable in *Drosophila melanogaster* where the F-score increased by about 9%.

4.3.3 Comparison of performance for external resources in the EIE model

Different combinations of external resources were tested with the results shown in Table 4.5. The best performance was achieved using a combination of related citations (F-score: 91.14%).

Drill down analysis revealed that one of the external resources, associate papers provided by Pubmed, achieved the best performance. Those designated as IV in Figure 4.2 performed worst among combinations of external resources. One interesting

Table 4.5: Micro-averaged 10-fold cross validation comparison for different combination of external resources

Spec.	I+II all related citations			II related citations (earlier than target)		
	P	R	F	P	R	F
Average	**92.10**	**90.24**	**91.14**	85.64	84.63	85.10
H	92.68	95.24	93.28	85.63	79.96	81.91
M	87.26	92.22	88.86	77.27	87.10	81.00
F	93.32	87.02	88.78	88.46	82.88	84.52
Y	95.15	86.50	89.80	91.21	88.58	88.91
	I related citations (later than target)			III related citations (later than target) excluded citing		
Average	77.32	69.41	73.03	80.65	68.99	74.33
H	70.72	91.02	78.75	66.88	87.59	75.07
M	84.04	84.79	83.63	78.67	74.65	75.40
F	89.81	83.79	85.90	88.25	64.94	73.81
Y	64.70	18.06	27.04	88.80	48.77	61.33
	IV citing paper(HBC)			IV citing paper(all)		
Average	60.37	39.04	46.93	62.66	48.40	54.40
H	47.75	94.91	62.78	54.12	85.94	65.56
M	78.45	33.73	44.24	62.44	52.35	55.46
F	68.86	17.53	27.21	82.45	40.07	52.19
Y	46.41	9.98	15.82	51.61	15.24	22.68
	II+IV related citations (earlier than target) +citing paper(HBC)			II+IV related citations (earlier than target) +citing paper(all) (all citation)		
Average	87.54	84.88	86.15	85.97	84.69	85.30
H	85.57	85.55	84.82	85.97	79.59	81.76
F	81.71	87.56	83.72	78.22	89.49	82.40
F	91.73	80.72	85.10	92.14	85.04	87.50
Y	91.12	85.68	87.11	87.55	84.63	85.16
	II+III+IV related citations +citing paper(HBC)			II+III+IV related citations +citing paper(all)		
Average	88.36	86.86	87.56	87.93	86.58	87.19
H	86.25	85.57	85.26	86.95	83.31	84.23
F	81.84	88.69	84.45	80.51	88.88	83.60
M	93.40	87.51	89.15	93.47	87.85	89.32
Y	91.94	85.68	87.50	90.78	86.27	87.28

Table 4.6: Micro-averaged 10-fold cross validation comparison for different combination of external resources(Cont.)

	III+IV related citations (later than target) +citing paper(HBC)			III+IV related citations (later than target) +citing paper(all)		
Average	81.87	68.01	74.24	83.68	74.67	78.80
H	65.11	90.80	75.00	71.03	87.12	77.28
F	82.89	72.72	76.10	80.78	77.24	77.73
M	90.50	64.04	73.53	90.84	77.46	81.47
Y	88.97	44.49	57.15	92.08	56.87	68.10
	II+III related citations excluded citing paper					
Average	87.44	86.23	86.78			
H	87.24	82.26	84.06			
F	77.95	87.97	82.04			
M	93.40	87.51	89.15			
Y	91.15	87.17	87.83			

result is seen when I compare the F-score using set III and set IV. When excluding the citing papers, performance improved greatly especially for *Saccharomyces cerevisiae*. This is because the associated paper for *Saccharomyces cerevisiae* more likely to contain the species words and citing papers contained less information of such species words.

The average date of the selected papers compared to the target paper for set III is 2.1 years; however, for set IV, it was 6.3 years. This raises the interesting possibility that selection should be adjusted for recency.

4.4 Discussion

In this experiment, I am the first people to introduce bibliographic features into focus species identification task. I conclude that external features, such as bibliographic features, are effective to focus species identification in biomedical abstract.

There are several causes of low performance for rare species entity classification such as *Rattus norvegicus* (F-score: 22.6). Compared to Wang et al.'s work, HYBRID generally performed better for low frequency species than my system. Secondly, in the FSD step, as the number of abstracts belonging to *Rattus norvegicus* is quite small (less than 1%), tests showed a high degree of confusion between the *Rattus norvegicus*/*Homo sapiens* and *Rattus norvegicus*/*Mus musculus* pairs. As the output of the GT model was used in other models, I decided not to test the low frequency species in the latter steps of FS tagger.

From the result, I found that using abstracts designated as III in Figure 4.2, the performance improved about 20%, especially for *Saccharomyces cerevisiae*, where the performance improved by 39%, compared to using abstracts designated as IV in Figure 4.2. The drill down analysis result was shown in the previous section. I concluded that the selected paper in set III is much closer to the paper select in citing papers.

I found that bibliographic features improved classification of some low frequency species. An example is shown in Figure 4.5, where the target abstract focused on *Saccharomyces cerevisiae*. The classification is complicated because *Homo sapiens* appears as a species word and the abstract also contains a protein with a *Homo sapiens* Taxon ID. By only considering internal features, the model cannot obtain the correct result. Using the closest associated papers strongly suggested the abstract is

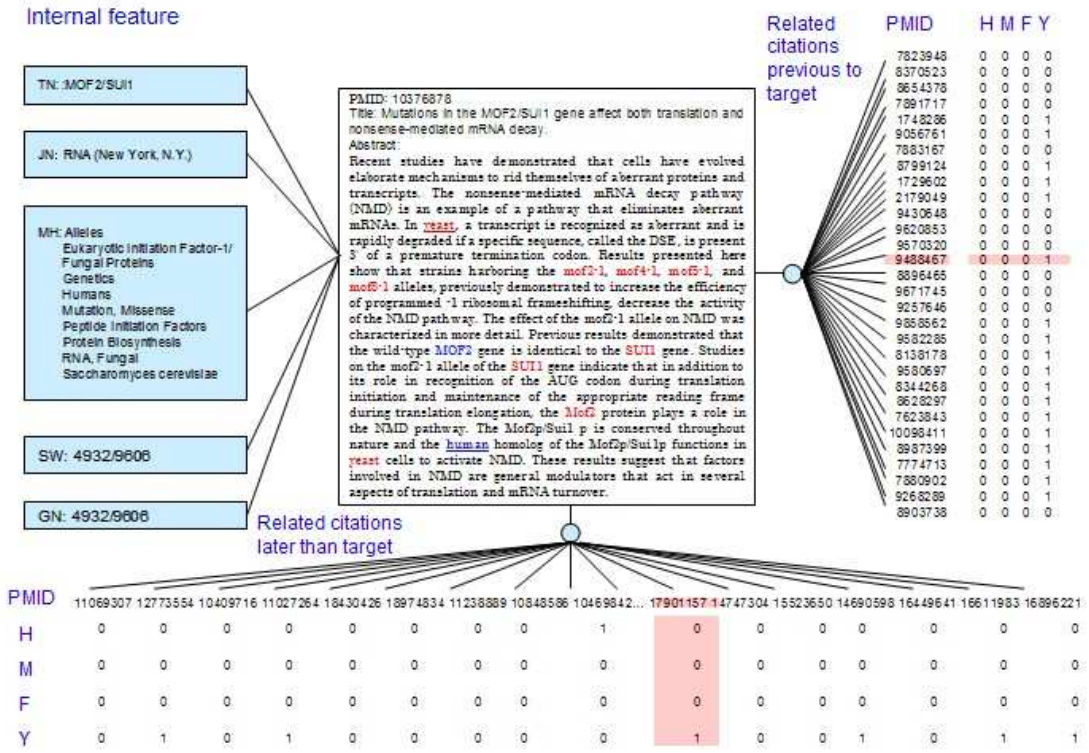


Figure 4.5: An example of using associated citation features for identifying focus species for an abstract (PMID: 10376878)

focused on *Saccharomyces cerevisiae*.

It is encouraging to use the external resources, however, there are still remain challenge to study further for the citation relations. A promising line of work is to use typed relations between citing articles and it is interested to study whether the semantics of these relations can contribute to the task of focus species classification.

After introducing the bibliographic features, the performance increased greatly, especially for *Saccharomyces cerevisiae*. The best performing combination of features achieved an overall F-score of 91.14%, which is 47.74% above the best model without EIE features. It can be argued that the size of the DECA corpus leads to a fragmented probability distribution; particularly for low frequency species and that

there is an inbuilt bias towards a method that uses an external knowledge source like MEDLINE. However I note that the internal features included the results of the GT model with over 90 F-score for gene name disambiguation. The result indicates that this potentially rich source of information contained a high degree of ambiguity that could not easily be resolved using internal clues alone.

I also see that although the journal name and title by them-selves do not appear to improve the F-score, whereas combining them with other features improves performance. This was especially noticeable in *Drosophila melanogaster* where the F-score increased by about 9%.

4.5 Conclusion

The sub thesis question in this chapter is that of the abstracts which are cited or archived in the PubMed database, do bibliographic features provide enhanced classification accuracy? In this experiment, by using of the citing papers and associated papers provided by PubMed, I extracted the species information to help the identification task for focus species, such kinds of bibliographic features were showed to enhance the F-score of the identification task in abstracts.

Chapter 5

Citation Scheme Development

As mentioned in Chapter 2, there has been growing interest in citation analysis within the biomedical text mining community and several researchers have developed their own citation schemes. From my results in the previous experiments I have found that the use of bibliographic information offers some important clues about the focus species class, helping to focus the classifier on the correct species. I am interested to see whether selective differentiation of the citations would yield improved performance on focal species classification. For this purpose, I developed a new citation scheme based on the citation function in biomedical papers.

One of the most widely considered citation scheme is S. Teufel et al. 's [2] applied to computer science papers. As shown in Figure 5.1, we can see that the citation function is more concentrated on method part which is because the structure of computer science papers are much more concentrated in method description when citing other papers. However, the purpose and structure of biomedical papers are quite different, i.e. the biomedical papers are much more concentrated on result descriptions when citing other papers.

As Teufel's scheme is only partly suitable for this task, I searched Y. Mizuta and

Category	Description
Weak	Weakness of cited approach
CoCoGM	Contrast/Comparison in Goals or Methods (neutral)
CoCo-	Authors work is stated to be superior to cited work
CoCoR0	Contrast/Comparison in Results (neutral)
CoCoXY	Contrast between 2 cited methods
PBas	Author uses cited work as basis or starting point
PUse	Author uses tools/algorithms/data/definitions
PModi	Author adapts or modifies tools/algorithms/data
PMot	This citation is positive about approach used or problem addressed (used to motivate work in current paper)
PSim	Authors work and cited work are similar
PSup	Authors work and cited work are compatible/provide support for each other
Neut	Neutral description of cited work, or not enough textual evidence for above categories, or unlisted citation function

Figure 5.1: S. Teufel et al. schema of Citation functions [2]

N. Collier's scheme shown in Figure 5.2. Their scheme is based on biomedical papers which is in same domain. However, their scheme is used for zone analysis, the scheme is in the view of section analysis. Zone analysis investigates the global rhetorical status of each sentence in terms of argumentation and intellectual attribution. Citation analysis investigates the purpose of a citation. Although their purpose is different to my research, their scheme is well studied in biomedical domain, I took their scheme as a starting point for my own citation scheme development.

5.1 Scheme development

Citations are important in scientific papers. According to Teufel et al, there are different reasons for each citations, including "mentioned similar method", "mentioned

- BACKGROUND (BKG): an assumption referring to previous work or a generally accepted fact.
- PROBLEM SETTING (PBM): a problem to be solved and/or the goal of the present work/paper.
- OUTLINE (OTL): a characterization or a summary of the content of the paper
- TEXTUAL (TXT): the organization of the paper.
- OWN: the authors own work. Sub classes:
    1. METHOD(MTH): methodology and materials;
    2. RESULT (RES): the results of the experiment performed;
    3. INSIGHT (INS): the insights/findings obtained (e.g. the authors interpretation of the result);
    4. IMPLICATION (IMP): the implications of the experimental result, including conjectures, assessment, applications, and future work;
    5. ELSE (ELS): anything else about the authors work;
- DIFFERENCE (DFF): a contrast or inconsistency between data and/or findings.
- CONNECTION (CNN): a relation or consistency between data and/or findings.

Figure 5.2: Scheme of Y. Mizuta and N. Collier [3]

the data used", "neutral description" and etc. Investigating why papers cite another paper can help us to find the papers with similar topics. Mizuta [3] widely cited scheme was developed based on the rhetorical function each clause has in reporting a scientific result in a biomedical paper. To meet our purpose of identification focus species in biomedical abstract, I avoid using complex scheme. My redefinition of the categories aims at reliable annotation; at the same time, the categories should be informative enough for document selection in citing paper pools.

I modified the scheme by specializing it to the citation function. My categories are as follows: One of the categories is the BKG function which stands for the background function. In the background function, the citation function includes neutral description of gene mentions, introduction of other's works and such kinds of general introduction citations. This function is mainly in background and introduction section, but it can also be found as neural descriptions in other sections.

The next category is DAT (data) function which mainly appears in data and method section and describe the data used in the current paper. This function always mentioned that the citation papers use the same or similar experimental model, experiment medicine and etc as the original papers.

The third category is EXP (Experiment) function. EXP (Experiment) function mainly appeared in experiment section which described an experiment did in citation paper.

The final category is RSL (Result) function which mainly appeared in result of discussion section which described a result mentioned in citation paper. There are different kinds of result: (1) the citing paper has similar results of the original paper; (2) the citing paper's result support the result of original paper; (3) the citing paper

has positive results to the original paper. I didn't category the result function in detailed because to meet our purpose, different kinds of result doesn't make any different.

In my scheme, BKG is same as BKG function in Mizuta's scheme, the PBM, OTL, TXT function in Mizuta's scheme are not contained in our scheme because in such sections, there are no citation mentioned. I divided Mizuta's MTH function into two functions: DAT and EXP, because the materials and methodology in the experiment are different in citation view. The RSL function is similar to Mizuta's RES function. The INS, IMP function in Mizuta's are not considered in my scheme because these sections contained less information of citations. I didn't use DFF and CNN function in Mizuta's scheme because that in our purpose, the citation function is same whether it is different or similar in the data/findings.

Though my scheme is quite simple compared with the citation function of Teufel's (2006) [2] and Y. Mizuta's [3], its aim is to help to find the similar topic for one citation papers. The simple one can achieve this purpose much easier.

Citation function is hard to annotate because the citation represents the author's intention when citing another work. My principle of citation function is to avoid ambiguities in citation functions.

5.1.1 Set of Citation Functions

In Figure 5.3, there is a target paper T1 which we want to classify for its focal species and a paper citing this target paper which is called C1. My scheme considers why C1 cites T1.

The set of citation functions are shown in Table 5.1.

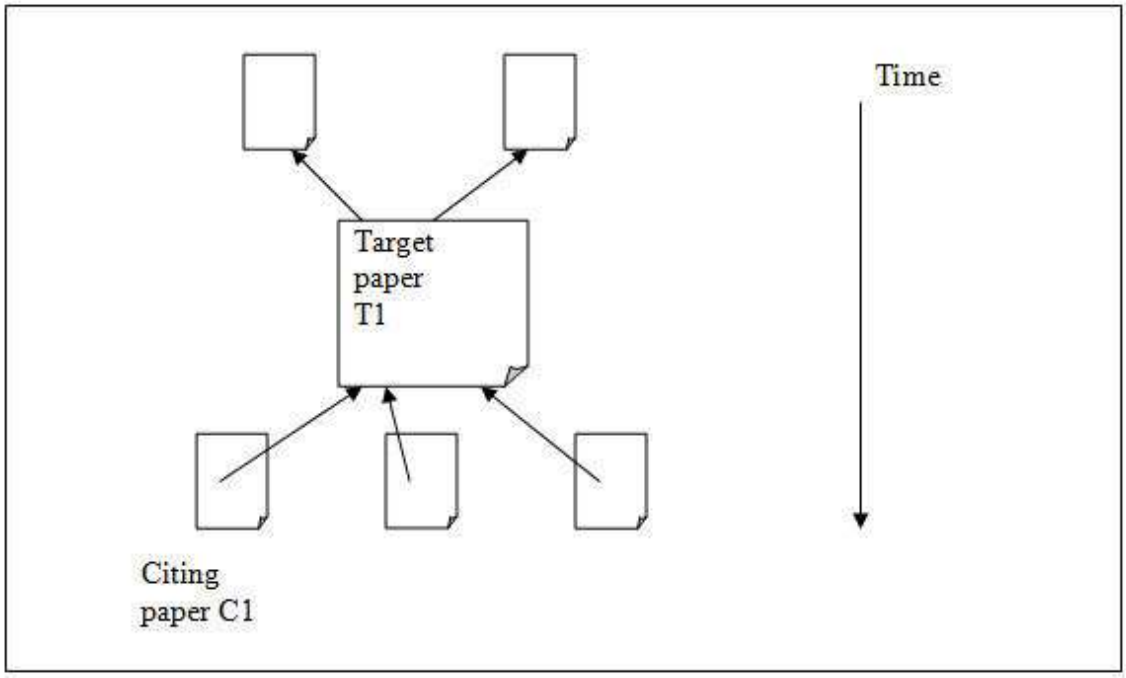


Figure 5.3: An example of citation structure
T1: the target paper.
C1: the citing paper cited T1.

Table 5.1: Citation function

Citation function	Description
BKG (Background)	The citation was caused by general introduction of the work T1 or a neutral description.
DAT (Data and data structure)	C1 cites T1 in the data section to describe the data used in C1
EXP (Experiment)	C1 cites T1 in the experimental section. The reason of citation was C1 mentioned an experiment in T1.
RSL (Result)	C1 cites T1 in the result section to describe an experimental result in T1.

The basic citation function is a citation sentence with the section information which the citation sentences appeared in citing paper. To analyze a citation sentence, I explored two level annotations.

5.1.2 Citation Principle

To annotate a citation sentence, first the section information was considered; And then rule is applied in annotation. To show an example here:

(1)[Experiment section] Similar results were obtained following parenteral immunization of B cell-deficient mice with Schistosoma mansoni eggs, in that immune deviation from a Th2 to a Th1 response was observed ($< CITATION > 17 < /CITATION >$).

(2)[Discussion section] Two Rb-related proteins, p107 and p130 (14, 20, 36, $< CITATION > 39 < /CITATION >$), did not interact with hsHec1p.

(3)[Result section] In fact, the subpopulations changed in a manner similar to normal mice and, as previously reported, <CITATION> 47-49 </CITATION> with the exception of the CD8 α α and CD8 α β IEL subpopulations.

The section information was first examined and the citation function was first generally classified in accordance with the section functions: for example, in the first round annotation, example (1) is an EXP function, example (2) is an RSL function and example (3) is an RSL function.

The second step is rule based annotation. The sentence in example (1), (2) and (3) is analyzed by parser. In accordance with the results of the parser, keywords such as time and verb are identified. With the rule explored, the citation function are determined. Two rules were developed for the annotation. (a) if there are keywords in

the set of keywords DAT, EXP, RSL, the citation function remained. (b) if the citation mark directly followed by gene names, the citation function is BKG. According to the parser, the (1) included the keywords obtain, (3) included the keyword previously and report. In (2), the citation was directly followed the gene names and the citation function was changed to BKG functions.

5.2 Experiment and Result

Hypothesis three: ***Of the abstracts which are cited does a typed citation function provide enhanced classification accuracy? Also what citation types provided the most useful?***

To test the citation function I developed, the FS tagger software presented in the last chapter was used. As mentioned in the last chapter, FS tagger included three models: the GT model (Gene mention Taxon ID annotation model), the EIE model (External resource information extractor) and the FSD model (Focus species decision model). In this experiment, I modified the EIE model, so that only citing papers are selected. I selected the citing papers according to the citation functions. The citing papers of same function are all selected out and the species words in abstracts and titles are extracted to use as the external features in FSD model.

5.2.1 Citation function selection

Citation functions are classified by a Machine learning classifier. First, citation sentences are picked up in citing papers and the section information is also picked up in this step. Second, citation sentences were parsed by Miyao and Tsujii's Enju parser

[82]. Third, a rule-based classifier was developed for classifying citation functions. There are three rules in classifier. (A) The section location of the citation sentence. There are six kinds of section location were included: introduction, background, data, method, result and discussion. (B) Keywords search, the key words with specific verb and time are considered. There are 56 keywords included. These consisted of two kinds of keywords. The first group is the verb to describe the relationship between the two papers, such as describe, report and etc. The second group is the word to show the time such as previous, recently and etc. (C) Citation location, the location of the mark of the citation. For example, whether the citation mark directly follows the gene/ gene products.

For example,

<Result section> In fact, the subpopulations changed in a manner similar to normal mice and, as previously reported, <CITATION> 47 - 49 </CITATION> with the exception of the CD8 α α and CD8 α β IEL subpopulations.

In the example, the citation sentence was first picked out with the section information and then the sentence was analyzed by the parser. Then the sentence was examined by rules. By first rule, the sentence was belonged to RSL function. By second rule, the keywords were previously and report, then the citation function was remained as RSL function. By third rule, the citation function was remained as RSL function. Then the citation function of the sentence was RSL function.

As mentioned upon, I have a list of keywords to search. There are two kinds of keywords existed. One is related to time, such as "in previous research", "to our knowledge". I treated such kinds of keywords as a clue to the comparison of experiment and result. The other kind of keyword is verb, such as "report", "show",

Table 5.2: Accuracy of Citation function classification

	P	R	F
BKG	0.94	0.98	0.96
DAT	0.93	0.95	0.94
MTH	0.95	0.93	0.94
RSL	0.96	0.96	0.96

"present", these are the verbs often used in scientific goal of a paper is defined.

5.2.2 Results

As the number of citation sentences was quite large, to test the accuracy of the rule-based classifier, 200 citation sentences for each of citation function were randomly selected out and the citation function were annotated manually. The test result was shown in Table 5.2. From the result, the F-score of four types of citation functions were all upon 94%.

In my previous work, I used the HBC method to select citing papers. In this experiment, citing papers were first classified in accordance with their citation functions. In addition, citing papers with same types of citation functions were selected out and the species information was extracted from these papers and used to classify the focus species of the whole document. The results are shown in Table 5.3. The results show that BKG function was not effective for focus species classification. However, RSL function was most useful in classification, followed by the MTH function.

5.3 Discussion

In this chapter, I present a citation function for biomedical papers. The performance of focus species identification changed much by selecting citing papers with different

Table 5.3: Micro-averaged 10-fold cross validation comparison for citation functions

	Citing paper(HBC)			Citing paper(all)		
	P	R	F	P	R	F
H	47.75	94.91	62.78	54.12	85.94	65.56
M	78.45	33.73	44.24	62.44	52.35	55.46
F	68.86	17.53	27.21	82.45	40.07	52.19
Y	46.61	9.98	15.82	51.61	15.27	22.68
	Citing paper(BKG)			Citing paper(DAT)		
	P	R	F	P	R	F
H	41.23	91.23	56.79	46.22	93.21	61.80
M	74.23	36.44	48.88	79.34	45.22	57.61
F	65.43	19.31	29.82	68.92	20.34	31.41
Y	47.21	8.77	14.79	51.33	22.18	30.98
	Citing paper(MTH)			Citing paper(RSL)		
	P	R	F	P	R	F
H	56.21	93.22	70.13	79.23	92.34	85.28
M	79.33	56.71	66.14	72.16	65.43	68.63
F	78.21	34.27	47.66	72.87	44.35	55.14
Y	52.14	47.28	49.59	57.12	57.32	57.22

citation functions. The citing papers cited the target papers in result sections performed best among four different citation functions. In this experiment, I showed that different citation functions provides different distance of the citing paper to the target papers.

In this experiment I analyzed the effectiveness of the newly proposed citation scheme. There is a assumption that the citations with the purpose evolutionary or juxtapositional use have the same focus species as the original paper. In fact, this is not always true, for example, the citation talking the experimental model in the original paper and the influenced organism in the citation paper. To avoid such kinds of case, there is a need to specific the citation functions in more detailed way.

5.4 Conclusion

The sub-question in this chapter is that of the abstracts which are cited does a typed citation function provide enhanced classification accuracy? Also what citation types prove the most useful? In this experiment, a citation scheme contained 4 different citation functions was developed. By this citation scheme, typed citation function provide an enhance of 29.05% of F-score for classification of focus species in abstracts. The citation type called RSL provided the most useful among the four kinds of citation functions. Although we showed the improvement by using such kinds of citation scheme, the performance of the focus species classification in abstracts was still lower than using external features such as PubMed related papers. As the current citation scheme was simple and relied on the section of the citing papers, whether a more detailed citation scheme can achieve a relevant performance as using external features extracted from PubMed related papers remained as a challenge work.

Chapter 6

Discussion

The series of experiments described in this dissertation, showed a F-score of 90.4% for classifying focus species in full-text papers and 79.3% for classifying focus species in abstracts by using lexical semantic features. An F-score of 91.14% was achieved for classifying focus species in abstracts by introducing external resources. I also demonstrated an improvement of 17.60% by selecting citing papers according to the citation function.

6.1 FS tagger

FS tagger is an online tool I have provided based on the experiments shown in this dissertation. FS tagger is available as a demonstration service at http://www-collier.nii.ac.jp/fstagger/index.php. The software is freely available on request from the author of this dissertation. A screen shot is shown in Figure 6.1. FS tagger is in English. The user can input a PubMed ID and then the detailed results will be shown by FS tagger. The default results include PMID, the title of the requested paper, the abstract of the paper, and the focus species of the requested paper. A more detailed result is provided by clicking the drop-down menu in the results page. The journal

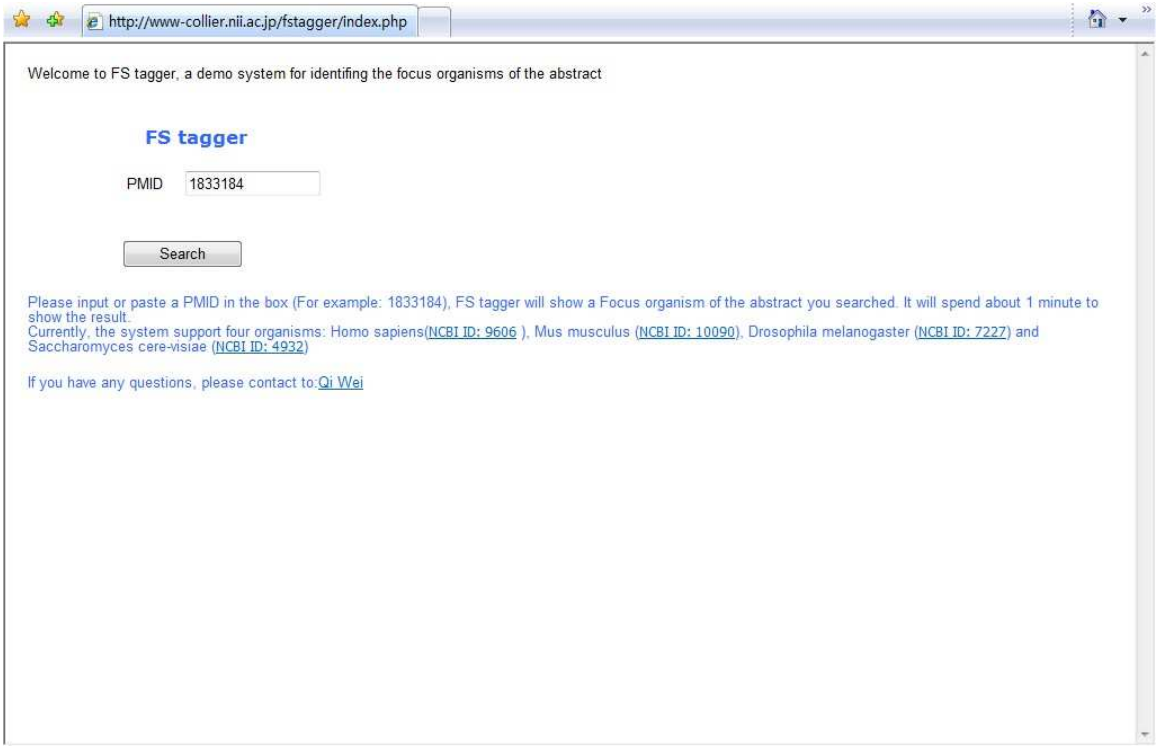
information, MeSH Headings, and species words in the given paper and the PMID of related bibliographic papers are also included.

The model is trained on the entire DECA corpus and the best feature set with GENE name, species words, document title, journal name, MeSH Main headings and species words extracted from associated papers shown in Chapter 2.

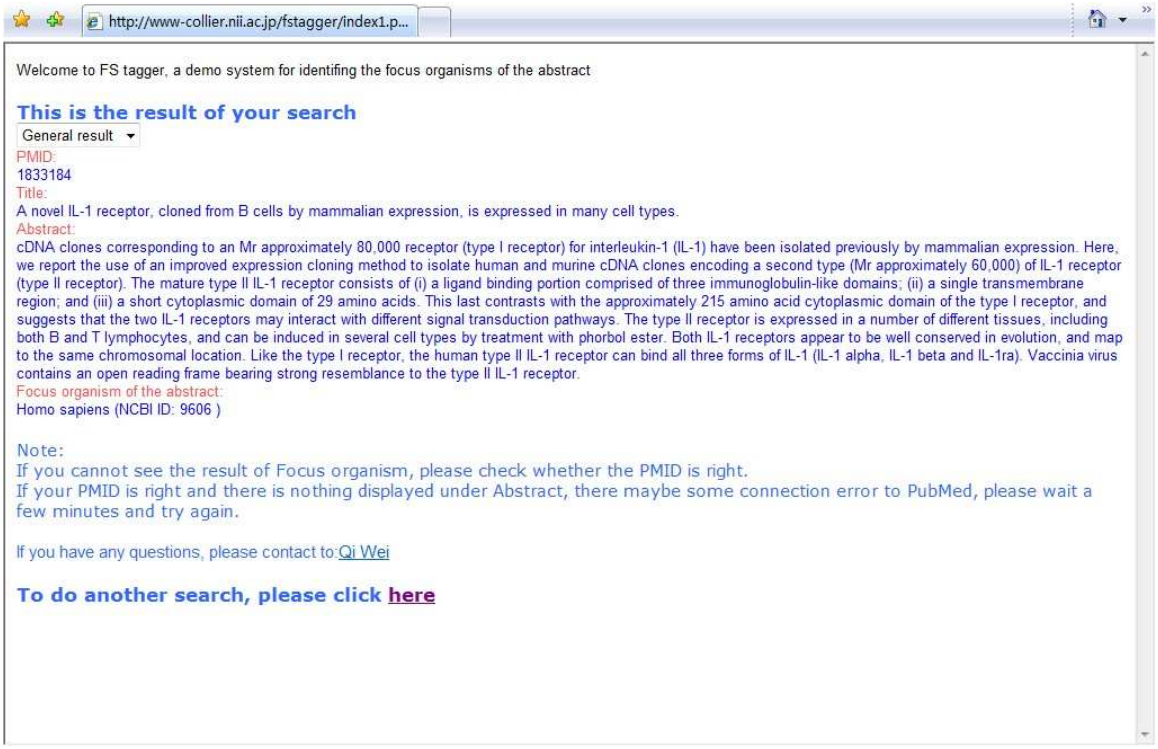
6.2 Citation selection

As I mentioned in Chapter 3, abstracts contained less information than the full-text papers regarding the focus species. That being the case, for abstracts, external resources can provide valuable additional clues. In my drill-down analysis I found that bibliographic features improved the F-score of some low-frequency species. An example is shown in Figure 4.5, where the target abstract focused on *Saccharomyces cerevisiae*. The classification is complicated because *Homo sapiens* appears as a species word and the abstract also contains a protein with a *Homo sapiens* TaxonID. By only considering internal features, the model cannot obtain the correct result. As internal features are extracted from the paper itself, if there are several species mentioned in the paper, features such as species words would contain several values. As the machine-learning classifier makes decisions based on features, the learning model may give the wrong focus species as a result. Using citation features however, the closest cited and citing papers both strongly suggested that the abstract was focused on *Saccharomyces cerevisiae*.

Including associated papers to identify the focus species was successful when applied to the abstract, however, a deeper analysis is needed. As shown in the previous chapter, only using citation papers with abstracts does not perform well compared



(a) Search page of FS tagger



(b) Result of search PMID 1833184

Figure 6.1: Screenshot of FStagger showing the request of PMID 1833184

to using PubMed related papers. This raised the question of whether the low performance was caused by the selection method or the citation papers themselves. Experiment three explored this question by differentiating citations using a citation function.

6.3 Simple citation scheme

In the development of the citation scheme, first, I tried to adapt my scheme from S. Teufel's [2], a scheme that is based on computer science papers which were much more focused on method. However, the structure of the biomedical papers seems quite different. Then I moved to Y. Mizuta's [3] scheme which is based on biomedical papers. However, the scheme was designed for zone analyze and my work was more directed towards the citation function. Based on these differences, I adapted Mizuta's scheme and developed a new citation scheme, because my purpose in developing the citation scheme was to help select the citing papers to use in the task of focus species identification task. I argue that the advantage of using a simple scheme is that: (1) A simple scheme can make the selection of the citing papers better focused. (2) A simple scheme can avoid ambiguities in citation function annotation.

After testing the citation scheme on FS tagger, I found that different citation functions provided different levels of performance in the focus species identification task. The RSL function, in which the citing paper cites the original papers because that the citing paper described the experiment result in the original paper, were most effective in identifying the focus topic. The reason the RSL function performs best is because it looks for cases in which the citation is mentioned in a comparison of the result of two experiments, and such comparisons are quite likely to be between

the same species, thus, the two papers most likely deal with the same species. The BKG function performed worst because the citations in the BKG function are general description of a gene or gene products, there are rare species relations between two papers.

When applying the citation scheme on FS tagger, selecting the citation paper based on the citation function was much more effective than selecting it based on the superficial words. A new question was raised: the citation function I created was strongly reliant on the section information of the citation papers, is this rely positive or negative to the selection of the citation papers.

6.4 Papers with multiple focus species

In the experiments reported in this dissertation, I assumed that one paper only deals with one focus species, but some papers involve several focus species. An example is shown in Figure 6.2, where the abstract mentions four species: *Drosophila*, *Mus musculus*, *Danio rerio* and *Bombyx mori*. Approximately approx 5% of the articles involved multiple species in my full-text paper corpus and 2% of articles mentoned multiple species in the DECA corpus I used. Error analysis showed that some errors across *Mus musculus* and *Homo sapiens* were caused by this kind of multiple focus species. To deal with such errors, I plan in future work to extend the framework and allow the assignment of multiple focus species, possibly also taking into account the subsumption relations within a taxonomy of focus species.

In future work, it may be necessary to develop a classifier to output a focus species list, perhaps as a list of major and minor species or a ranked species list with a probability assigned to each species. To achieve this, first, I need to adjust my training

> Coatomer is a major component of COPI vesicles and consists of seven subunits. The gamma-COP subunit of the coatomer is believed to mediate the binding to the cytoplasmic dilysine motifs of membrane proteins. We characterized cDNAs for Copg genes encoding gamma-COP from mouse, zebrafish, Drosophila melanogaster and Bombyx mori. Two copies of Copg genes are present in vertebrates and in B. mori. Phylogenetic analysis revealed that two paralogous genes had been derived from a single ancestral gene by duplication independently in vertebrates and in B. mori. Mouse Copg1 showed ubiquitous expression with the highest level in testis. Zebrafish copg2 was biallelically expressed in hybrid larvae in contrast to its mammalian ortholog expressed in a parent-of-origin-specific manner. A phylogenetic analysis with partial plant cDNA sequences suggested that copg gene was also duplicated in the grass family (Poaceae).

Figure 6.2: An example of abstract with multiple focus species

data, in the current experiment,for each paper from DECA corpus,I annotated the focus species using the classes in the BioCreAtIvE sources, by providing the species list, I need to provided the species list of each paper. Second, in the testing model, I need to adjust the models to give the all possibility of the focus species.

6.5 Identification of species mentions

In the current experiments, species words are identified by a species dictionary I created manually, which include different forms for 4 species. However, to manually extend the species dictionary to include more organisms would be a difficult task. A species mention identifier is needed. One recently released species mention tagger is Organism Tagger, developed by Naderi et al. [83]. Their system achieved an accuracy of 97.5% on the OT corpus. To extend my system to more organisms, an identifier

like this should be added.

Chapter 7

Conclusion

In this dissertation, I have described a series of experiments on focus topic identification: (1) First I explored a series of experiments, then based on the results of these I showed the relative merits of various in document lexical semantic features in full-text papers and abstracts; (2) Then, considering the limitations of full-text papers, new features were developed for abstracts. External resources were explored for the abstracts for the focus species identification task. Using the best feature set, an online tool called FStagger was developed. In this experiment, different feature sets and different external resources were compared, and I showed the value of using bibliographic resources. (3) Finally,the question was raised whether the citation function was effective in the selection of external resources and a citation function scheme was developed and tested on the same task. Experiment showed that different citation functions produced performance different in the focus species identification task.

I demonstrated a system that automatically categorizes full-text papers into 3 organism categories and another system (FStagger) that automatically categorizes abstracts into 4 organism categories: *Homo sapiens*, *Mus musculus*, *Drosophila melanogaster*

and *Saccharomyces cerevisiae*. Different combinations of bibliographic features were tested in the experiment and a best F-score of 91.14% was achieved. My analysis has shown that (a) species words from PubMed related citations published after the target abstracts were better predictors of the target species than those from papers citing the target abstracts; (b) bibliographic features can improve the performance on low-frequency species. As the bibliographic information was useful. I chose to explore citation functions as a way of differentiating between them. Further analysis showed that the citation papers that cited the original paper because they described the results of the original paper were most effective in the categorization task.

There are still several limitations of my experiment which needs to be improved in the future. (1) My experiment assumed that each paper only has one focus species. In the future, I will consider ways to introduce a species list to deal with multiple species in one paper. (2) The current simple scheme of the citation function was not experimentally compared to alternative schemes such as the more complex approaches outlined by Teufel or Mizuta for zoning tasks. (3) Also, in my experiment, I only used basic learning models such as CRFs. In the future, I will consider how to introduce more new learning models such as collective classification using networked data. (4) For the external resources selection, the HBC selection method was used. In the future, I will consider whether other selection methods can be used and what the best selection method for this task is. Another limitation is that I only used species words extracted from bibliographic resources as external features, and if taxon identifiers were available for all citing papers and associated papers, the performance could likely be improved by 3%.

In the present experiments, I identify 3 species in full-text papers and 4 species in

abstracts. To answer the thesis question, I should consider working on more species. With the features I used in the present experiments, extending the work to more species requires considering the following points: (1) Gene names and species mentions are the features that are useful for the classification task. For the gene names, the identification tools used in the present experiment are useful. For species mentions, as in the present experiment, the species mentions are identified using a dictionary-based method, and to extend the work to more species, automatic tools should be provided. (2) Document titles contain information on species, however, in my experiment, I showed that by using only document titles gave no performance improvement. However, using the combination of document titles and journal names gave an improved F-score. I used journal names because I assumed that some of the journal names might contain the species information. To extend the work to more species, using the combination of document titles and journal names may not yield much obvious improvement. (3) MeSH headings show the value in my experiment. As the MeSH headings are used as an index clue for focus species, they are considered to be useful when extending the work to more species. (4) Taxon ID shows the species information for each gene mentions, so if I can extend the use of Taxon ID not only in target abstracts but also in the external abstracts, the accuracy of the classification task may be improved. However, the method of identifying the Taxon ID should be changed, and should not only rely on the DECA corpus because that corpus is quite small. (5) Bibliographic features were shown to be useful in the classification task. As such kinds of external resources can easily be obtained through PubMed, I believe that they will be useful when extending the task to more species. In the present experiment, only citing papers are tested, so considering that there are no citing papers

for a newly published paper, in the future, the contract part, cited papers should also be used. In a conclusion, if I can improve the points I mentioned earlier such as Taxon ID and species word identification, I believe that the current method using external features and citation scheme can give the similar level of performance if I extend the work into more species.

Bibliography

[1] Xinglong Wang, Jun'ichi Tsujii, and Sophia Ananiadou. Disambiguating the species of biomedical named entities using natural language parsers. *Bioinformatics*, 26(5):661–667, 2010.

[2] Simone Teufel, Advaith Siddharthan, and Dan Tidhar. Automatic classification of citation function. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 103–110, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[3] Yoko Mizuta and Nigel Collier. An annotation scheme for a rhetorical analysis of biology articles. In *Proceedings of the Forth Intl. Conference on Language Resources and Evaluation (LREC2004)*, 2004.

[4] Fabio Rinaldi, Thomas Kappeler, Kaarel Kaljurand, Gerold Schneider, Manfred Klenner, Simon Clematide, Michael Hess, Jean-Marc von Allmen, Pierre Parisot, Martin Romacker, and Therese Vachon. Ontogene in biocreative ii. *Genome Biology*, 9(Suppl 2):S13, 2008.

[5] Xinglong Wang and Michael Matthews. Distinguishing the species of biomedical named entities for term identification. *BMC Bioinformatics*, 9(Suppl 11):S6, 2008.

[6] T Kappeler, K Kaljurand, and F Rinaldi. Tx task: Automatic detection of focus organisms in biomedical publications. In *Workshop on BioNLP*, pages 80–88, June 2009.

[7] Imad Tbahriti, Christine Chichester2, Frdrique Lisacek, and Patrick Ruch1. Using argumentation to retrieve articles with similar citations: an inquiry into improved related articles search in the medline digital library. *Int. J. Med. Inf.*, 75(6):488–495, 2006.

[8] Preslav I. Nakov, Ariel S. Schwartz, and Marti A. Hearst. Citances: Citation sentences for semantic analysis of bioscience text. In *Proceedings of the SIGIR04 workshop on Search and Discovery in Bioinformatics*, 2004.

[9] Baoping Zhang, Marcos Andre Goncalves, Weiguo Fan, Yuxin Chen, Edward A Fox, Pavel Calado, and Marco Cristo. *Intelligent Fusion of Structural and Citation-Based Evidence for Text Classification*, pages 667–668. Number TR-04-16. ACM Press, 2005.

[10] Thierry Delbecque and Pierre Zweigenbaum. Using co-authoring and cross-referencing information for medline indexing. *AMIA Annu Symp Proc*, 2010:147–151, 2010.

[11] Alexander S. Yeh, Lynette Hirschman, and Alexander A. Morgan. The evaluation of text data mining for database curation: lessons learned from the kdd challenge cup. *Bioinformatics*, 19:331–339, 2003.

[12] Judith A. Blake, Joel E. Richardson, Carol J. Bult, Jim A. Kadin, Janan T. Eppig, and The Mouse Genome Database Group. Mgd: the mouse genome database. *Nucleic Acids Research*, 31(1):193–195, 2003.

[13] Susan Tweedie, Michael Ashburner, Kathleen Falls, Paul Leyland, Peter McQuilton, Steven Marygold, Gillian Millburn, David Osumi-Sutherland, Andrew

Schroeder, Ruth Seal, Haiyan Zhang, and The FlyBase Consortium. Flybase: enhancing drosophila gene ontology annotations. *Nucleic Acids Research*, 37(suppl 1):D555–D559, 2009.

[14] *DictyDB (An ACeDB Database for Dictyostelium) BMC Ltd, BM Central - 2004 - en.scientificcommons.org.*

[15] *Wormpep (C. Elegans Protein Database) BMC Ltd, BM Central - 2003 - en.scientificcommons.org.*

[16] Lynette Hirschman, Alexander Yeh, Christian Blaschke, and Alfonso Valencia. Overview of biocreative: critical assessment of information extraction for biology. *BMC Bioinformatics*, 6(Suppl 1):S1, 2005.

[17] Rainer Winnenburg, Thomas Wchter, Conrad Plake, Andreas Doms, and Michael Schroeder. Facts from text: can text mining help to scale-up high-quality manual curation of gene products with ontologies? *Briefings in Bioinformatics*, 9(6):466–478, 2008.

[18] Zhenzhen Kou, William W. Cohen, and Robert F. Murphy. High-recall protein entity recognition using a dictionary. *Bioinformatics*, 21(1):266–273, January 2005.

[19] Daniel Hanisch, Katrin Fundel, Heinz-Theodor Mevissen, Ralf Zimmer, and Juliane Fluck. Prominer: rule-based protein and gene entity recognition. *BMC Bioinformatics*, 6(Suppl 1):S14, 2005.

[20] Yanpeng Li, Hongfei Lin, and Zhihao Yang. Incorporating rich background knowledge for gene named entity classification and recognition. *BMC Bioinformatics*, 10(1):223, 2009.

[21] Jin-Dong Kim, Tomoko Ohta, Yoshimasa Tsuruoka, Yuka Tateisi, and Nigel Collier. Introduction to the bio-entity recognition task at jnlpba. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, JNLPBA '04, pages 70–75, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

[22] W. Hersh, A. Cohen, P. Roberts, and H. Rekapalli. TREC 2006 genomics track overview. In *TREC Notebook*, 2006.

[23] Ralph Grishman and Beth Sundheim. Message understanding conference-6: a brief history. In *Proceedings of the 16th conference on Computational linguistics - Volume 1*, COLING '96, pages 466–471, Stroudsburg, PA, USA, 1996. Association for Computational Linguistics.

[24] Zhou GuoDong and Su Jian. Exploring deep knowledge resources in biomedical name recognition. In Nigel Collier, Patrick Ruch, and Adeline Nazarenko, editors, *COLING 2004 International Joint workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP) 2004*, pages 99–102, Geneva, Switzerland, August 28th and 29th 2004.

[25] Dietrich Rebholz-Schuhmann, Antonio Yepes, Chen Li, Senay Kafkas, Ian Lewin, Ning Kang, Peter Corbett, David Milward, Ekaterina Buyko, Elena Beisswanger, Kerstin Hornbostel, Alexandre Kouznetsov, Rene Witte, Jonas Laurila, Christopher Baker, Cheng-Ju Kuo, Simone Clematide, Fabio Rinaldi, Richard Farkas, Gyorgy Mora, Kazuo Hara, Laura I Furlong, Michael Rautschka, Mariana Neves, Alberto Pascual-Montano, Qi Wei, Nigel Collier, Md Chowdhury, Alberto Lavelli, Rafael Berlanga, Roser Morante, Vincent Van Asch, Walter Daelemans, Jose Marina, Erik van Mulligen, Jan Kors, and Udo Hahn. Assessment of NER solutions against the first and second CALBC Silver Standard Corpus. *Journal of Biomedical Semantics*, 2(Suppl 5):S11, 2011.

[26] Yoshimasa Tsuruoka and Jun'ichi Tsujii. Probabilistic term variant generator for biomedical terms. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, SIGIR '03, pages 167–173, New York, NY, USA, 2003. ACM.

[27] GuoDong Zhou, Jie Zhang, Jian Su, Dan Shen, and ChewLim Tan. Recognizing names in biomedical texts: a machine learning approach. *Bioinformatics*, 20(7):1178–1190, 2004.

[28] Guodong Zhou. Recognizing Names in Biomedical Texts using Hidden Markov Model and SVM plus Sigmoid. In Nigel Collier, Patrick Ruch, and Adeline Nazarenko, editors, *COLING 2004 International Joint workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP) 2004*, pages 1–7, Geneva, Switzerland, August 2004.

[29] Shaojun Zhao. Named Entity Recognition in Biomedical Texts using an HMM Model. In Nigel Collier, Patrick Ruch, and Adeline Nazarenko, editors, *COLING 2004 International Joint workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP) 2004*, pages 87–90, Geneva, Switzerland, August 2004.

[30] Tzong han Tsai, Shih hung Wu, and Wen lian Hsu. Exploitation of linguistic features using a crf-based biomedical named entity recognizer. In *ACL Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics (BioLINK-05)*, 2005.

[31] Shipra Dingare, Malvina Nissim, Jenny Finkel, Christopher Manning, and Claire Grover. A system for identifying named entities in biomedical text: how results from two evaluations reflect on both the system and the evaluations: Conference papers. *Comp. Funct. Genomics*, 6(1-2):77–85, February 2005.

[32] Lynette Hirschman, Alexander A. Morgan, and Alexander S. Yeh. Rutabaga by any other name: extracting biological names. *Journal of Biomedical Informatics*, 35(4):247 – 259, 2002.

[33] Yoshimasa Tsuruoka and Junichi Tsujii. Improving the performance of dictionary-based approaches in protein name recognition. *Journal of Biomedical Informatics*, 37(6):461 – 470, 2004.

[34] Ronen Feldman, Yonatan Aumann, Yair Liberzon, Kfir Ankori, Jonathan Schler, and Benjamin Rosenfeld. A domain independent environment for creating information extraction modules. In *Proceedings of the tenth international conference on Information and knowledge management*, CIKM '01, pages 586–588, New York, NY, USA, 2001. ACM.

[35] K. Fukuda, T. Tsunoda, A. Tamura, and T. Takagi. Toward information extraction: Identifying protein names from biological papers. In *Proceedings of the Pacific Symposium on Biocomputing*, pages 707–718, 1998.

[36] Haijian Shi. Best-first decision tree learning. Master's thesis, University of Waikato, Hamilton, NZ, 2007.

[37] Wlodzislaw Duch, Rafal Adamczak, Krzysztof Grabczewski, and Grzegorz Zal. Hybrid neural-global minimization method of logical rule extraction. *JACIII*, pages 348–356, 1999.

[38] le S. Cessie and van J. Houwelingen. Ridge estimators in logistic regression. *Applied Statistics*, 41(1):191–201, 1992.

[39] Pat Langley, Wayne Iba, and Kevin Thompson. An analysis of bayesian classifiers. In *Proceedings of the Tenth National Conference on Artificial Intelligence*, pages 223–228. MIT Press, 1992.

[40] Sang-Bum Kim, Kyoung-Soo Han, Hae-Chang Rim, and Sung-Hyon Myaeng. Some effective techniques for naive bayes text classification. *IEEE Trans. Knowl. Data Eng.*, pages 1457–1466, 2006.

[41] Stephan Bloehdorn and Andreas Hotho. Boosting for text classification with semantic features. In *Proceedings of the 6th international conference on Knowledge Discovery on the Web: advances in Web Mining and Web Usage Analysis*, WebKDD'04, pages 149–166, Berlin, Heidelberg, 2006. Springer-Verlag.

[42] M.A. Hearst, S.T. Dumais, E. Osman, J. Platt, and B. Scholkopf. Support vector machines. *Intelligent Systems and their Applications, IEEE*, 13(4):18 – 28, Jul/Aug 1998.

[43] Hagit Shatkay, Nawei Chen, and Dorothea Blostein. Integrating image data into biomedical text categorization. *Bioinformatics*, 22(14):e446–e453, 2006.

[44] Aaron M. Cohen. An effective general purpose approach for automated biomedical document classification. *Proceedings of the American Medical Informatics Association (AMIA) 2006 Annual Symposium*, pages 161–165, 2006.

[45] Cui Yu, Beng C. Ooi, Kian-Lee Tan, and H. V. Jagadish. Indexing the Distance: An Efficient Method to KNN Processing. In *VLDB '01: Proceedings of the 27th International Conference on Very Large Data Bases*, pages 421–430, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[46] S Haykin. *Neural Networks: A Comprehensive Foundation*, volume 13. Prentice Hall, 1999.

[47] Jeffrey T. Chang, Hinrich Schtze Ph. D, Novation Biosciences, Russ B. Altman, and Ph. D. Creating an online dictionary of abbreviations from medline. *Journal of the American Medical Informatics Association*, 9:612–620, 2002.

[48] A. A. Morgan, B. Wellner, J. B. Colombe, R. Arens, M. E. Colosimo, and L. Hirschman. Evaluating the automatic mapping of human gene and protein mentions to unique identifiers. In *Proceedings of Pacific Symposium on Biocomputing*, 2007.

[49] Haw-ren Fang, Kevin Murphy, Yang Jin, Jessica S. Kim, and Peter S. White. Human gene name normalization using text matching with automatically extracted synonym dictionaries. In *Proceedings of the Workshop on Linking Natural Language Processing and Biology: Towards Deeper Biological Literature Analysis*, BioNLP '06, pages 41–48, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[50] Jeremiah Crim, Ryan McDonald, and Fernando Pereira. Automatically annotating documents with normalized gene lists. *BMC Bioinformatics*, 6(Suppl 1):S13, 2005.

[51] J. Michael Cherry, Caroline Adler, Catherine Ball, Stephen A. Chervitz, Selina S. Dwight, Erich T. Hester, Yankai Jia, Gail Juvik, TaiYun Roe, Mark Schroeder, Shuai Weng, and David Botstein. Sgd: Saccharomyces genome database. *Nucleic Acids Research*, 26(1):73–79, 1998.

[52] Jeffrey T. Chang, Hinrich Schtze Ph. D, Novation Biosciences, Russ B. Altman, and Ph. D. Creating an online dictionary of abbreviations from medline. *Journal of the American Medical Informatics Association*, 9:612–620, 2002.

[53] Hongfang Liu and Cathy Wu. A study of text categorization for model organism databases. In Lynette Hirschman and James Pustejovsky, editors, *HLT-NAACL 2004 Workshop: BioLINK 2004, Linking Biological Literature, Ontologies and Databases*, pages 25–32, Boston, Massachusetts, USA, May 6 2004. Association for Computational Linguistics.

[54] Jimmy Lin. Is searching full text more effective than searching abstracts? *BMC Bioinformatics*, 10(1):46, 2009.

[55] Susan Bonzi. Characteristics of a literature as predictors of relatedness between cited and citing works. *Journal of the American Society for Information Science*, 33(4):208–216, 1982.

[56] M. J. Moravcsik and P. Murugesan. Some results on the function and quality of citations. *Social Studies of Science*, 5:86–92, 1975.

[57] Francis Narin. *Evaluative bibliometrics: The use of publication and citation analysis in the evaluation of scientific activity.* Computer Horizons, 1976.

[58] C. Lee Giles, Kurt D. Bollacker, and Steve Lawrence. Citeseer: an automatic citation indexing system. In *Proceedings of the third ACM conference on Digital libraries*, DL '98, pages 89–98, New York, NY, USA, 1998. ACM.

[59] John Swales. Citation analysis and discourse analysis. *Applied Linguistics*, 7(1):39–56, 1986.

[60] Simon Buckingham Shum and Simon Buckingham Shum. Evolving the web for scientific knowledge: "first steps towards an hci knowledge web". In *Interfaces, British HCI Group Magazine*, pages 16–21, 1998.

[61] Donald O. Case and Georgeann M. Higgins. How can we investigate citation behavior? a study of reasons for citing literature in communication. *Journal of the American Society for Information Science*, 51(7):635–645, 2000.

[62] Melvin Weinstock. Citation indexes. *In Encyclopedia of Library and Information Science*, 5:16–40, 1971.

[63] John Swales. *Genre analysis: English in academic and research settings*, volume 11, page 272. Cambridge University Press, 1990.

[64] Charles Oppenheim and Susan P. Renn. Highly cited old papers and the reasons why they continue to be cited. *Journal of the American Society for Information Science*, 29(5):225–231, 1978.

[65] Mark Garzone and Robert Mercer. Towards an automated citation classifier. In Howard Hamilton, editor, *Advances in Artificial Intelligence*, volume 1822 of *Lecture Notes in Computer Science*, pages 337–346. Springer Berlin / Heidelberg, 2000.

[66] Lorraine Tanabe and W. John Wilbur. Tagging gene and protein names in full text articles. In *Proceedings of the ACL-02 workshop on Natural language processing in the biomedical domain - Volume 3*, BioMed '02, pages 9–13, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[67] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November 2009.

[68] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[69] Kenji Hirohata, Naoaki Okazaki, Sophia Ananiadou, and Mitsuru Ishizuka. Identifying sections in scientific abstracts using conditional random fields. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP 2008)*, pages 381–388, 2008.

[70] Taku Kudo. Crf++: yet another crf toolkit. http://crfpp.sourceforge.net/.

[71] Vladimir N. Vapnik. *The nature of statistical learning theory.* Springer-Verlag New York, Inc., New York, NY, USA, 1995.

[72] Leo Breiman and Leo Breiman. Bagging predictors. In *Machine Learning*, pages 123–140, 1996.

[73] Ron Kohavi. The power of decision tables. In *Proceedings of the European Conference on Machine Learning*, pages 174–189. Springer Verlag, 1995.

[74] Lorraine Tanabe and W. John Wilbur. Tagging gene and protein names in biomedical text. *Bioinformatics*, 18(8):1124–1132, 2002.

[75] Stuart J. Nelson., Michael Schopen, Jacque-Lynne Schulman, and Natalie Arluk. An interlingual database of mesh translations. In *8th International Conference on Medical Librarianship*, London, UK., 2000 Jul 4.

[76] Remco R. Bouckaert and Eibe Frank. Evaluating the replicability of significance tests for comparing learning algorithms. In *PAKDD'04*, pages 3–12, 2004.

[77] Marc Colosimo, Alexander Morgan, Alexander Yeh, Jeffrey Colombe, and Lynette Hirschman. Data preparation and interannotator agreement: Biocreative task 1b. *BMC Bioinformatics*, 6(Suppl 1):S12, 2005.

[78] Alexander Morgan, Zhiyong Lu, Xinglong Wang, Aaron Cohen, Juliane Fluck, Patrick Ruch, Anna Divoli, Katrin Fundel, Robert Leaman, Jorg Hakenberg, Chengjie Sun, Heng-hui Liu, Rafael Torres, Michael Krauthammer, William Lau, Hongfang Liu, Chun-Nan Hsu, Martijn Schuemie, K Bretonnel Cohen, and Lynette Hirschman. Overview of biocreative ii gene normalization. *Genome Biology*, 9(Suppl 2):S3, 2008.

[79] Yoshimasa Tsuruoka, Yuka Tateishi, Jin-Dong Kim, Tomoko Ohta, John McNaught, Sophia Ananiadou, and Junichi Tsujii. Developing a robust part-of-speech tagger for biomedical text. In *Advances in Informatics - 10th Panhellenic Conference on Informatics*, volume 3746 of *LNCS*, pages 382–392. Springer-Verlag, Volos, Greece, November 2005.

[80] Jimmy Lin and W John Wilbur. Pubmed related articles: a probabilistic topic-based model for content similarity. *BMC Bioinformatics*, 8(1):423, 2007.

[81] Makoto Iwayama. Hierarchical bayesian clustering for automatic text classification. In *IJCAI*, pages 1322–1327. Morgan Kaufmann Publishers, 1995.

[82] Yusuke Miyao and Jun'ichi Tsujii. Deep linguistic analysis for the accurate identification of predicate-argument relations. In *Proceedings of the 20th international conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

[83] Nona Naderi, Thomas Kappler, Christopher J. O. Baker, and Ren Witte. Organismtagger: detection, normalization and grounding of organism entities in biomedical documents. *Bioinformatics*, 27(19):2721–2729, 2011.