EXPLORING SEMANTIC ROLES FOR NAMED ENTITY RECOGNITION IN THE MOLECULAR BIOLOGY DOMAIN

Tuangthong Wattarujeekrit

DOCTOR OF
PHILOSOPHY

Department of Informatics,
School of Multidisciplinary Sciences
The Graduate University for Advanced Studies

2005 (School Year)

September 2005

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

Supervisor:

Nigel Collier, *Assoc. Prof.*	National Institute of Informatics, SOKENDAI

Advisory Committees:

Fujiyama Asao, *Prof.*	National Institute of Informatics, SOKENDAI
Noriko Kando, *Prof.*	National Institute of Informatics, SOKENDAI
Asanee Kawtrakul, *Assoc. Prof.*	Kasetsart University (Thailand)
Ken Satoh, *Prof.*	National Institute of Informatics, SOKENDAI
Hideaki Takeda, *Prof.*	National Institute of Informatics, SOKENDAI

Abstract

Named entity recognition (NER) in the molecular biology domain, the task of identifying and categorizing molecular entities appearing in text, is one of the most important tasks in a biological text mining engine. In general, this task is taken as the first step towards the more ambitious task of molecular event extraction (relation extraction) and, eventually, pathway discovery. However, NER in this scientific domain, which seems to be the easiest task among others in text mining, still achieves quite low performance. As can be seen from the most recent shared-task evaluations of NER in this domain (JNLPBA-2004), the best performance in terms of F1-score is only 72.6. This result is far below what is achieved by NER system in newswire domain (F1-score of about 96%) which is near the human level of performance. At present, most NER systems employ term internal features (e.g., lexical and morphology) and co-occurrence information as term external features. Due to the lack of molecular naming convention, which leads to the difficulty of terminological variations as well as the difficulty of polysemy (i.e. the sharing of names between different entities), such features are insufficient to handle the difficulties for NER in the molecular biology domain. To obtain a complete set of rules for lexical patterns of molecular names seem impossible, thus to use term external features other than co-occurrence information is of interest.

In this thesis, the semantic relationships between a predicate and its arguments in terms of semantic roles are proposed to enhance NER system in the molecular biology domain. The semantic role information is derived from a predicate-argument structure (PAS) which is a higher sentence representation level than syntactic relation and surface form levels. Thus, the use of semantic roles is more consistent than co-occurrence information derived from a surface level. To employ the semantic role for NER system, it is realized in various sets of syntactic features which were used by a machine learning model to explore the most efficient way in allowing this knowledge to provide the highest positive effect on the NER.

As a result, the best feature set is composed of the 6 lexical features (i.e., *surface word*, *lemma form*, *orthographic feature*, *part-of-speech*, *phrase-chunk* and *head word of NP-chunk*) and 4 PAS-related features for representing an argument's semantic role (i.e., *predicate's surface form*, *predicate's lemma*, *voice* and the united feature of *subject-object head's lemma* and *transitive-intransitive sense*). Moreover, the use of semantic roles can show the positive effects for only the predicates conforming to the criteria as follows. A predicate must have its

arguments as both *agent* and *theme* with a higher probability of belonging to a named entity class than non-named entity class; otherwise, a predicate must have its arguments as both *agent* and *theme* with a lower probability of belonging to a named entity class than non-named entity class and the number of training examples for this predicate should be large enough (by observing from empirical evidences, at least 270 sentences). The improvement in performance obtained from the NER system using PAS-related features, compared to not using these features, affirms that the using of semantic roles can enhance NER system.

Acknowledgements

I would like to thank many people for their support, encouragement and guidance during my years as a graduate student of Sokendai, but working at NII building.

First and foremost, this dissertation represents a great deal of time and effort not only on my part, but on the part of my supervisor, Nigel Collier. He has helped me shape my research from day one, pushed me to get through the inevitable research setbacks, and encouraged me to achieve to the best of my ability. Without him, this dissertation would not have happened.

I also thank my other committee members, Ken Satoh, Hideaki Takeda, Hiroko Satoh, as well as Noriko Kando and Fujiyama Asao for their valuable time to discuss and make comments regarding my research.

Moreover, I owe thanks to all researchers in my research group, Yoko Mizuta, Ai Kawazoe, Tony Mullen, and Anna Kohonen for their support in solving research problems and in enriching my fatigue during study period.

My special thanks go to Parantu K Shah for his general support in grounding me molecular biology knowledge.

My appreciation also goes to development teams of GENIA corpus to make the shared resources publicly available.

Thank you also many other people at NII, especially Reiko Okano, Futabako Urakawa, Nahoko Iwanaga and Aiko Kawamura who helped me to carry on my research smoothly without irritating on any office works.

All friends in NII, Tokyo University, even in Thailand always cheer me up at all times I need them.

And most of all, my parents, my aunt and my elder brother have always been supportive, understanding, and encouraging at all times through the entire graduate period.

Contents

List of Figures

List of Tables

Chapter 1
Introduction

Recently, the field of molecular biology has brought about the rapid growth in the volume of scientific literature published online in order to report experimental results. To make use of these free-text articles which are readable only by humans for further analysis (i.e., to find connected information among research or to discover information implicit conveyed in the text), the articles require to be transformed into machine-readable formats such as data base or ontology. The need for a structured representation can be seen from the human efforts to construct databases such as BIND (Bader et al., 2001), KEGG (Kanehisa et al., 2002), DIP (Xenario et al., 2002), MINT (Zanzoni et al., 2002). As the production rate of literature is very high, it is hard for human curators to maintain up-to-date database resources. Information extraction (IE) systems that aim to identify and extract required facts mainly from documents, as well as relate and integrate these facts from multiple sources are considered to be an important remedy for biology researchers.

At present, most researches of biomedical IE systems pay their attentions to two main targets: recognition of molecular named entities and recognition of relationships among named entities. These two targets are on the way to reach the goal of discovering biological pathways which is a network of interactions and events between biological molecules (e.g., proteins, drugs). Although IE systems in the molecular biology domain benefit from the techniques of traditional IE which are used efficiently in the news domain, the overall performance of the molecular biology IE systems for both named entity recognition (NER) and relation extraction are still far from the levels where they can be used to replace the human curator.

This thesis focuses to enhance traditional NER system in the molecular biology domain by using the deeper knowledge than the knowledge derived from syntactic and surface form levels. The semantic relationships between a predicate an its arguments in terms of semantic roles are proposed to enhance NER system.

1.1 Motivation

A proposition conveyed in a sentence can be represented in a semantic representation level such as a predicate-argument structure (PAS). The relationships in terms of semantic roles between a predicate dictating the event and its arguments as containing entities participating in the event are represented in a PAS. For example, the PAS representing a sentence "*John loaded the truck with hay*" says that "*John*" plays the semantic role as "*an agent or loader*" of the event driven by the predicate *load*, "*the truck*" plays the semantic role as "*a beast of burden*", and "*hay*" plays the semantic role as "*a cargo*".[1] The semantic representation at the level of PAS has its important property that the same PAS will be used to represent different surface forms if these surface forms convey the same proposition. Thus, a sentence "*John loaded hay on the truck*" is represented in the same PAS frame as the sentence "*John loaded the truck with hay*". The capability of PAS to unambiguously represent the semantics of an event motivates us to enhance IE systems using PAS.

With regard to the recognition of relationships among named entities, to extract proteins participating in protein-protein interaction event from a sentence "*These findings suggest that Msp1p is a component of the secretary vesicle docking complex whose function is closely associated with that of Dec1p*" by using a surface syntactic form of "*A associate with B*", an incorrect pair of proteins "*Msp1p*" and "*Dec1p*" is likely to be extracted. The knowledge from PAS of predicate *associate* that "*Dec1p*" is not an argument in this event would help to avoid the extraction of "*Msp1p*" and "*Dec1p*" from the above sentence.

In the case of NER which is the target application in this thesis, the hypothesis is that an argument's semantic role should impose type restrictions on the entities within the argument. This is founded on the basis observation that a biological event can be realized as a predicate and its participating named entities (NEs) as the predicate's arguments. So far, various methods to solve NER problem have been proposed (Fukuda et al., 1998; Krauthammer et al., 2000; Kazama et al., 2002; Takeuchi and Collier, 2002; Settles, 2004; Finkel et al., 2004; Zhang et al., 2004). Most methods rely on two types of evidences: the internal evidence and external evidences. The internal evidence is related

[1] These semantic role labels are taken from PAS of the predicate *load* proposed in the PropBank project (Kingsbury and Palmer 2002).

to lexical information of a term (i.e., orthographic and morphological information). The external evidence is the information of co-occurrence of terms appearing in the local context of a target term. The overall performances of these systems are still quite low. As can be seen from the most recent shared-task evaluations of NER in this domain (JNLPBA-2004), the best performance in terms of F1-score is only 72.6. This result is far below what is achieved by NER system in newswire domain (F1-score of about 96%) which is near the human level of performance. The reason is that the term external and internal evidences using so far are sufficient to handle the difficulty of terminological variations as well as the difficulty of polysemy (i.e. the sharing of names between different entities) in the newswire domain, but not in the molecular biology domain which contains higher difficulty. The semantic role which is proposed in this thesis is counted as term external evidence as same as co-occurrence information. However, it is derived from a PAS which is a higher level than a surface form level where the co-occurrence is derived. According to this, the semantic role is rigid to the variation at the surface and syntactic levels. The use of this semantic knowledge should be able to enhance existing NER systems in the molecular biology domain.

1.2 Objectives and Approach

1.2.1 Thesis Question

The principal question addressed in this thesis is:

Can the semantic information describing the relationships in terms of semantic roles between a predicate and its arguments enhance NER?

1.2.2 Approach

The general approach to answering the thesis question has been to apply semantic relationships represented in PAS to the molecular biology NER system based on a state-of-the-art machine learning approach. More specifically, two main subsidiary works have been done:

- **Construction of PASBio resource**: In order to completely take advantage of a set of semantic roles for a predicate, a PAS frame is used as a reference. In each PAS frame, a typical set of semantic roles for a predicate used in a

particular sense is represented. PAS frames for general English are already being constructed in several projects. However, several predicates found in the biological text, which is also written in English, have been used in different either meaning or behavior from what they have been used in general domain. Therefore, it is important to construct PAS frames for this scientific domain.

- **Employment of semantic roles in machine-learning based NER system**: The objective of this task is to prove the hypothesis that the semantic role is useful to improve the performance of traditional NER system. In this, how to apply the semantic information of semantic roles for NER system and how much the improvement in performance can be gained from employing each predicate's semantic roles are investigated.

1.3 Contributions

This thesis makes two distinct contributions to the fields of bioinformatics and computational linguistics:

- **PASBio resource**: This resource contains frames of predicate-argument structure analyzed from the literatures in molecular biology domain and a set of annotated sentences corresponding to the frames. Available to download at: http://research.nii.ac.jp/~collier/projects/PASBio/.
- **Enhancing NER system by employing deep knowledge represented in PAS**: Semantic relationships between a predicate and its arguments have been applied to improve performance of the state-of-the-art NER system. Theoretically, this thesis has shown the relation between semantic roles and named entity type. Empirically, this thesis has shown how to efficiently represent semantic roles in terms of machine learning features as well as the criteria for obtaining positive effect from semantic roles.

1.4 Reader's Guide to the Thesis

The remaining chapters of this thesis are as follows.

- **Chapter 2** provides background knowledge of the IE in molecular biology domain, especially the methodology to recognize molecular biology named

entities. Also, the chapter includes background knowledge about predicate-argument structure. The related works are discussed in this chapter as well.

- **Chapter 3** describes in detail the work in this thesis to construct predicate-argument frames for molecular biology domain
- **Chapter 4** describes in detail about another work in this thesis which is to enhance NER by applying the knowledge of semantic relationships represented in PAS. Additionally, how to transform PAS knowledge into a set of features for machine learning is illustrated in this chapter.
- **Chapter 5** presents conclusions and future works.

Chapter 2

Background and Related work

In the last decade, the field of molecular biology, aiming to understand about the origin, function and structure of living systems has experienced a massive growth in the peer-reviewed publications reporting experimental results. Most publications are stored in computer-based resources retrievable through the Internet such as the PubMed[2] organized by the National Library of Medicine. To make optimal use of these free-text articles readable only by humans, the articles must be transformed into a structured format such as a database or ontology. The need of data in the computer-readable format has been shown in the construction of several databases such as BIND (Bader et al., 2001), KEGG (Kanehisa et al., 2002), DIP (Xenario et al., 2002) and MINT (Zanzoni et al., 2002). A lot of human efforts have been taken to build these databases. Because the production rate of literature is very high, for example, it is noted in a survey of Cohen and Hunter that 1500 abstracts per day are added to Medline (Cohen and Hunter, 2004), it is hard for human curators to keep database systems up-to-date. As an IE system has its global aim to extract the information mainly from documents and relate the pieces of information by filling a structured template or a database record, it is considered to be an important remedy for biology researches concerning above need .

This chapter is organized as follows. Firstly, how IE from text, particularly the named entity extraction task, has been applied in the molecular biology domain will be discussed. Secondly, one of the most important levels of linguistic knowledge that is the PAS level will be discussed. Finally, how the PAS seems to be helpful for the task of NER in the molecular biology domain will be discussed.

2.1 Information extraction in the molecular biology domain

The goal of information extraction (IE) is to provide instances of structured knowledge representations from unstructured free-form text. IE systems in general are capable of identifying and extracting useful information, as well as relating and integrating

[2] http://www.ncbi.nlm.nih.gov/PubMed/

information from multiple sources (Cowie and Lehnert, 1996). Traditionally, IE systems with a high influence from the series of evaluation exercises called the Message Understanding Conferences (DARPA, 1995; DARPA, 1998) have been successfully used to extract information from text in newswire domain. The success of IE systems depends on the performance of several subtasks of which the pipeline shown in Figure 2-1(a). A traditional IE system generally consists of 5 subtasks: (1) named entity extraction, also called named entity recognition, to identify and categorize proper names appearing in text such as person names, email addresses, location names; (2) template element task to extract instances of features related to each named entity such as age values of persons; (3) co-reference task to find and link together all references to the same entity in a given text such as identifying the antecedent of each pronoun; (4) template relation task to extract instances of relations among entities such as extracting employment relation between entity *company* and entity *person*; (5) scenario template task to extract instances of events or facts of predefined frames and slots (i.e., to integrate and relate extracted facts from all the tasks explained before). IE systems that have actually been deployed in the general domain are; for example, ATRANS (Lytinen and Gershman, 1993), JASPER (Andersen et al., 1986) and SCISOR (Rau, 1991).

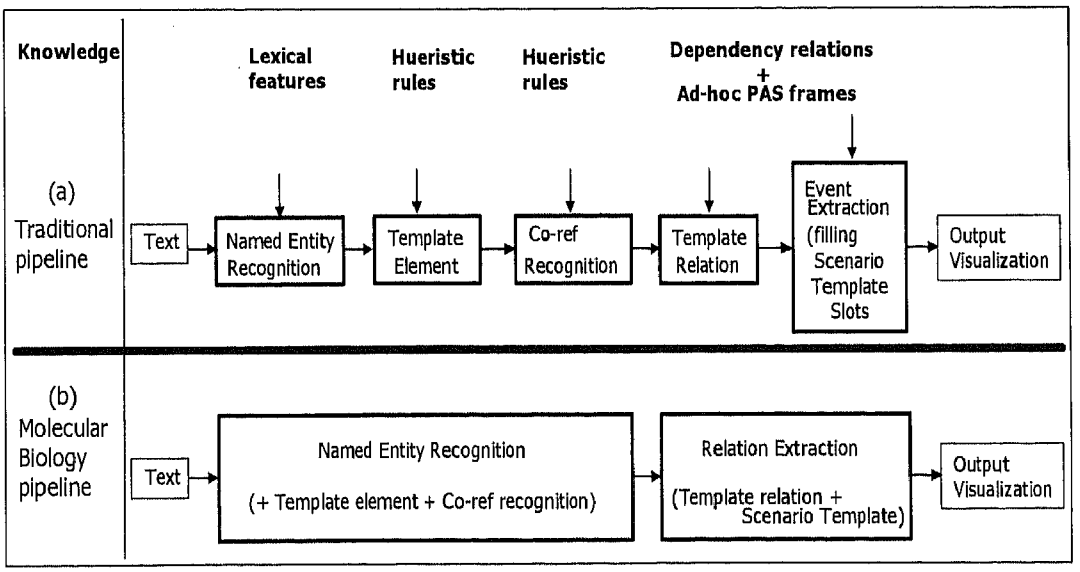


Figure 2-1: Subtask pipelines of (a) traditional and (b) molecular biology IE system

In the molecular biology domain, IE systems are developed in the similar way of the traditional IE to reach the goal of discovering biological pathways which is a network of interactions and events between proteins, drugs, and other molecules. However, the biological IE system is normally composed of only 2 main subtasks as shown in Figure 2-1(b). The molecular NER systems work in the wider scope than the traditional NER systems. Because each molecular entity has its unique identity which has been identified in the domain ontology such as Gene Ontology GO)[3], the extraction of molecular entities in some works (Couto et al., 2005; Ehrler et al., 2005) involve also the link between the extracted entities from text to the GO concept. This is considered as doing the co-reference task. In addition, in order to classify the entity to the correct identity concept, the information related to the entity (such as which organism this entity is found, what the product of this entity is, etc.) described in text must be extracted also. This accounts for doing the template element task (i.e. the instances of features related to the entity is extracted). Another main task of the molecular IE system is the relation extraction task of which the task scope already cover the template relation and scenario template tasks. For a traditional IE system which is applied to extract information from the literature in general domain, the template relation task captures a relation between two entities; for example, the relation "employee of" between person "John" and company "ABC"), next a number of these two binaries relation and also properties of the entities are used to fill the slots of the scenario template that is predefined to explain the occurrence of a particular event. In the general domain, the binary relation between entities itself is often not an event. On the contrary, most relations between molecular entities are identified as event (e.g., protein-protein, protein-gene and protein-drug). Thus, template relation task and scenario template task are merged together to extract relations describing the events in the molecular biology domain. The examples of the IE systems in the molecular biology domain are summarized below.

The Highlight system (Thomas et al., 2000) works based on the techniques from SRI Menlo Park's Fastus (Hobbs et al., 1996), a leading performer in the Message Understanding Conferences (MUCs)'s evaluation. To capture protein interactions, hence only the interactions associated with the verb phrase *interact with*, *associate with*, and *bind to*, it uses part-of-speech tagging and partial parsing. Also, discourse analysis is

[3] http://www.geneontology.org/

employed to identify co-referring noun phrases. Finally, predefined domain-specific patterns are used to map relevant information in the literature. EMPathIE and PASTA (Humphreys et al., 2000) are systems that aim to capture enzyme reactions and the system to capture information concerning the role of amino acids in protein molecules respectively. Similar to the Highlight system, these systems have also been developed through five separate component subtasks as defined by MUCs. Pustejuuovsky and colleagues (Pustejovsky et al., 2002) propose the system to extract *inhibition*-relations by using UMLS Thesaurus (Humphreys et al., 1998) as a reference knowledge source for named entity extraction and coreference resolution tasks. To extract relations, syntactic grammars are defined from intensive analysis over the corpus. Although only *inhibition*-relation extraction is examined, the authors claim that the system is applicable for any binary relations. Prior systems cover the extraction of relation found between sentences, but the system such as SUISEKI (Blaschke, 1999) can only extract a binary relation found at a sentence level. Also, the GENIES system (Friedman et al., 2001) and the GeneWay system (Rzhetsky et al., 2004) which is an extended version of GENIES do not cover cross-sentence relations. However, these systems can extract complicated relations, such as relations of relations. For example, the extraction of relations from the sentence "*phosphorylated* Cbl *coprecipitated with* CrkL, *which was constitutively associated with the* C3G" will result in a form like "[action, attach, [protein, Cbl, [state, phosphorylated]], [protein, CrkL, [action, attach, [protein, CrkL], [protein, C3G]]]]" meaning that the final relation between "phosphorylated Cbl" and "CrkL" occur after the prior relation between "CrkL" and "C3G".

Due to the restriction in the access to full length articles imposed by copyright and the availability to access publicly MEDLINE abstracts[4], most of the systems mentioned above have been developed solely on abstracts. It has been reported that abstracts contain higher information density (information content divided by document length) than full texts, however a lot of critical information is contained in the body of the text (Shah et al., 2003; Schuemie et al., 2004). Thus, biological IE systems should aim for further development to extract information from full text (Friedman et al., 2001; Corney et al., 2004; Shi et al., 2005). As the difficulties will increase in full text, the IE system would

[4] http://www.ncbi.nlm.nhih.gov/PubMed/

require more sophisticated linguistic processing than it uses to extract information from abstracts.

So far, many efforts have been taken to solve the NER problem as it is the first step in biological IE and it seems to be the simplest task compared to others. Various approaches have been applied, for instance, dictionary-based approaches (Krauthammer et al., 2000; Hirschman et al., 2002; Tsuruoka et al., 2003; Tuason et al., 2004), rule-based approaches (Fukuda et al., 1998; Proux et al., 1998; Gaizauskas et al., 2000; Cohen et al., 2002; Franzen et al., 2002; Tanabe and Wilbur, 2002; Yu et al., 2002; Narayanaswamy et al., 2003), and machine learning approaches (Andrade and Valencia, 1998; Collier et al., 2000; Hatzivassiloglou et al., 2001; Lui et al., 2001; Kazama et al., 2002; Takeuchi and Collier, 2002; Lee et al., 2003, Morgan et al., 2003; Shen et al., 2003; Yamamoto et al., 2003).

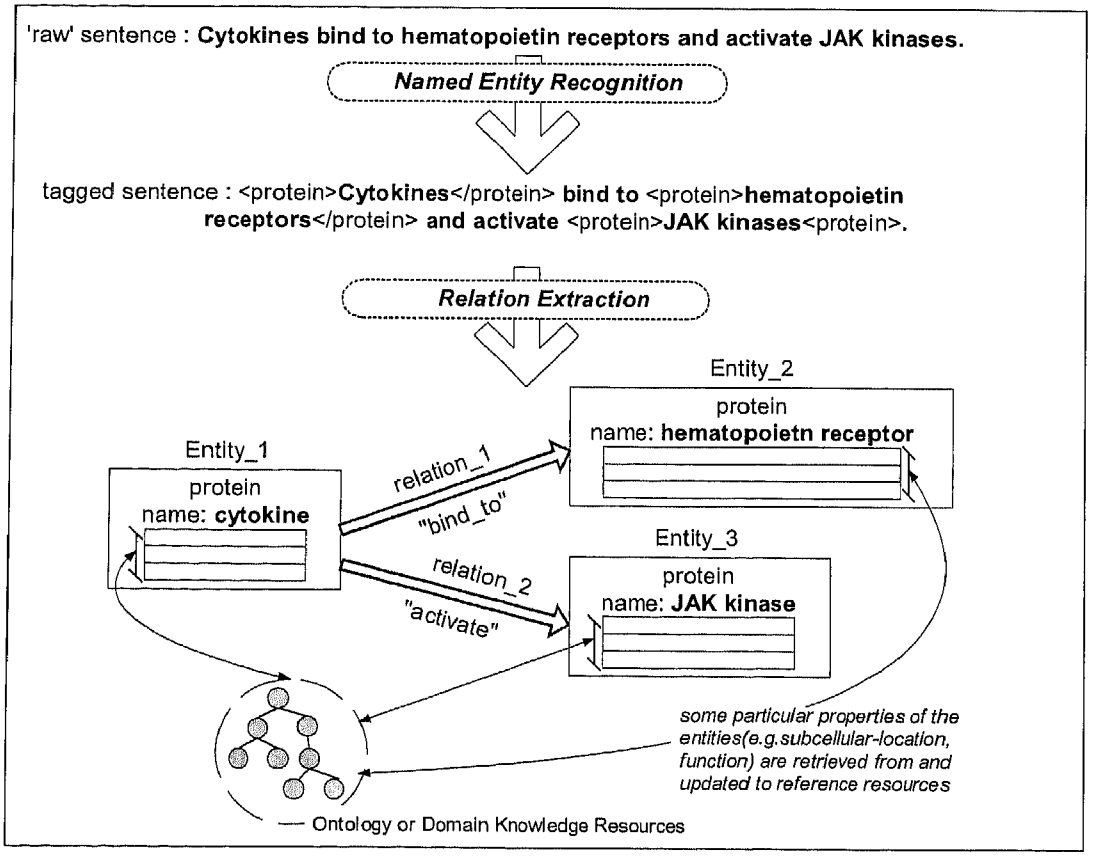


Figure 2-2: An example of outcomes after named entity recognition and relation extraction

With respect to the overall target of extraction, most of the biomedical IE systems aim at present on two main targets: (1) recognition of molecular named entities and (2) recognition of relationship among named entities. A simplified example of the input and output of these two processes is shown in Figure 2-2. These two targets are *en route* to help in molecular pathway understanding. A detailed discussion of named entity recognition (NER) is presented in Section 2.1.1 and biological relation extraction in Section 2.1.2. The shared-tasks for biomedical IE systems are summarily discussed in Section 2.1.3.

2.1.1 Recognition of molecular named entities

In the molecular biology domain, named entity recognition (NER) is used to identify within the text which text constituents refer to molecular named entities, and then to classify the entity into relevant biology concept classes. Molecular named entities include genes, proteins, small molecules, chemical molecules, tissues, etc. From the sentence shown in Figure 2-2, the result from this NER process is the fact that "*Cytokines*", "*hematopoietin receptors*" and "*JAK kinases*" are protein names. Not only is NER an important component of biological relation recognition systems, the task can benefit for other applications of biological text mining. For instance, document retrieving where a pertinent subset of documents are obtained (Stapley and Benoit, 2000) and document clustering where similar documents are grouped together (Willett, 1988).

Although NER in the molecular biology domain has been receiving attention by many researchers for a decade, the task remains very challenging. Its challenges are caused mainly by the complex structure of molecular names and the lack of naming convention. Organisms that have nomenclatures which are highly controlled by groups of researchers will tend to have smaller variations than those without control, making them easier to identify. This is affirmed by results reported in the task 1B of BioCreAtIvE (Hirchman et al., 2005). Among three model organism databases (i.e., mouse, fly and yeast), genes or proteins in the fly database contain the highest ambiguity; followed by mouse and yeast. Therefore, the NER systems got the lowest performances when they are applied to fly database compared to other two databases. Genes or proteins in the fly database contain the highest ambiguity; follow by mouse and yeast (Hirschman et al., 2004). More details about the ambiguity or difficulty in recognizing molecular named entity are explained in the following section.

2.1.1.1 Difficulty of named entity recognition in the molecular biology domain

Several factors that have made the task of biological NER difficult are shown as follows.

- **Lack of naming conventions in biology**: Some efforts have been made to standardize the naming of biological entities (e.g., Guidelines for Human Gene Nomenclature[5], Drosophila Gene Nomenclature[6] and Standardized Genetic Nomenclature of Mice[7]); however, many biologists often do not follow these recommended nomenclatures. This factor is the fundamental cause of other difficulties, which will be described in the following.
- **Various patterns of terminology**: Some names may be named with standard English words, for instance "light", "map", "complement" and "Sonic hedgehog" are used to name human genes. Some names may be named by using alphanumeric, such as "9-cis retinoic acid". Some names may be named like symbols or codes, such as "M(2)201". Some names, especially protein names, may be named by using an amino-acids sequence, such as "amino acids [aa] 1 to 25".
- **Term nesting**: Names may be formed by nesting of terms such as "[leukaemic[T [cell line]] Kit225]". The term nesting brings into the question that at what level of fine-grained distinctions should be processed. Krauthammer and Nenadic mentioned that to also recognize and highlight the sub-terms (i.e., "cell line" and "T cell line") when recognizing the term "leukaemic T cell line Kit225" would be valuable in the subsequent term identification process (Krauthammer and Nenadic, 2004). The semantic categories in the ontology must play an important role for this granularity problem.
- **Term coordination**: Sometimes two entities are coordinated by their arguments, such as "B and T cells" refer to two entities: 1) B cells and 2) T cells. Sometimes two entities are coordinated by their heads, such as "adrenal glands and gonads" refer to two entities: 1) adrenal glands 2) adrenal gonads.

[5] http://www.gene.ucl.ac.uk/nomenclature/guidelines.html
[6] http://tinman.vetmed.helsinki.fi/eng/drosophila.html
[7] http://www.informatics.jax.org/mgihome/nomen/

Also, sometimes there is no coordinating conjunction neither "and" nor "or" which makes more ambiguity. For instance, it is difficult to know that "Toll-6, -7, -8" refer to "Toll-6", "Toll-7", and "Toll-8", but not "Toll-6", "-7", and "-8". In addition to the ambiguity to distinct entities within coordinated term, to differentiate term coordination from term conjunction is also highly ambiguous since term coordination and term conjunction share the same structures. As noted by Nenadic and colleagues (Nenadic et al., 2004), "adrenal glands and gonads" may be recognized by way of the term conjunction between "adrenal glands" and "gonads".

- **Homonymy**: Homonymy, one kind of polysemy[8], is the ambiguity that occurs when two or more entities have the same name but refer to unrelated meanings or objects (Buitelaar, 1998). The homonymy ambiguity is mostly caused by the overlapping of acronyms of different entities. For example, the acronym "THA" is used for 98 meanings including "total hip arthroplasty", "tetrahydroaminoacridine", tetrahydro-9-aminoacridine, and so forth; or "GR" can refers to both glucocorticoid receptor and glutathione reductase.

- **Systematic polysemy**: This kind of polysemy is the ambiguity regarding the sameness of name for referring to the objects which systematically relate to each others, especially through the complementary of senses (Buitelarr, 1998). A gene and its produced protein often have the same name; for instance, Hatzivassiloglou and colleagues mentioned in their work (Hatzivassiloglou et al., 2001) that they found from the same article the use of term "SBP2" is a protein in the sentence "*By UV cross-linking and immunoprecipitation, we show that SBP2 specifically binds selenoprotein mRNAs both in vitro and in vivo*" and is a gene in the sentence "*The SBP2 clone used in this study generates a 3173 nt transcript (2541 nt of coding sequence plus a 632 nt 3' UTR truncated at the polyadenylation site)*". Furthermore, the names of some genes are from the related diseases such as "FHM" comes from family hemiplegic migraine disease (Erhardt et al., 2005).

[8] Polysemy is the ambiguity of an individual word or phrase that can be used (in different contexts) to express two or more different meanings.

- **Many alias names for each entity (synonym)**: The reason for containing many alias names of a molecular entity is mainly because most biologists will shorten the name of the prior mentioned entity when mention it again in the literature, such as "-150 CD28 response element (CD28RE)/AP-1 site" is the same DNA as "-150 CD28RE/AP-1 site". Interestingly, two entities with similar names like one name is the shorter name of another one, but they can refer to two distinct entities. For instance, "epidermal growth factor" and "epidermal growth factor receptor" are two distinct proteins. It is mentioned in the survey paper of Dickman that there are 50-100 alterations every week to the nomenclature section of mouse genome database Web page (Dickman, 2003). This huge increasing volume of alterations for just one organism implies the difficulty for named entity extraction among various organisms in the molecular biology domain.

Due to the factors mentioned above, NER task in the molecular biology domain seems to be more complex than in the newswire or general domain. More naming convention can be found in general domain, for instance, a word initialized with a capital letter must be a proper noun, a word containing "Co." must be an organization name, and a word containing "Mr." or "Ms." must be a person name. Also, many closed groups of words for particular types of named entities are present, such as a group of words referred to months are "January", "February", "March", etc.

2.1.1.2 Existing methods to extract biological named entity

To cope with the above-mentioned ambiguous and variable nature of names in the molecular biology domain, a number of techniques have been proposed. These techniques can be roughly divided into three categories: (1) dictionary looking up approach, (2) heuristic rule-based approach and (3) machine learning approach.

2.1.1.2.1 Dictionary looking up approach

By comparing a word or a string in text with each entry in the existing, manually created dictionaries or knowledge resources (e.g., ontology or databases) that contain lists of molecular entities is a straightforward approach to recognize entity names. However, if only a direct map between a term in text and a term in a reference knowledge source is

used, such an exact match is likely to fail due to the term variant problem. As mentioned before in section 2.1.1.1, most of the concepts in the molecular biology domain have more than one term (e.g., four possible variations of protein name spelling "TNFalpha1R", "TNFalpha-1 R", "TNF alpha1R"). Thus, the term used to represent a concept in terminology resource is perhaps not the one used in its text mention. At least, the text mention should be normalized for being compatible to the resource term regarding to case, inflection and hyphen variation, as well as word order variation (Bodenreider et al., 2002). Another method is generalizing a dictionary entry that is replacing dictionary terms with generic placeholders (Bunescu et al., 2004). Two dictionaries used in Bunescu and colleagues' work include the Human Proteome Initiative (HPI) [9] and the Gene Ontology Database[10]. For each term in both dictionaries, Bunescu and colleagues isolate and replace numbers with <*n*>, Roman letters with <*r*> and Greek letters with <*g*>. So, the term "interleukin-1 beta" would be transformed to "interleukin <*n*> <*g*>". So far, the generic dictionary is used instead of the original dictionary. A canonical dictionary is also created when more coverage is required. From previous example, the final term of "interleukin-1 beta" would be "interleukin". The method gains higher coverage while precision is compensated.

Another one is the use of the DNA sequence-like strings to represent both input text and a dictionary entry (Krauthammer et al., 2000). Then, the Basic Local Alignment Search Tool (BLAST)-based identification algorithm is used to compare the DNA sequence-like strings of the input text to of a dictionary entry. The recall of 79% and the precision of 71% have been achieved.

The EDGAR system (Rindflesch, 2000) which aims to extract drugs, genes and relations is a kind of hybrid technique. The EDGAR is based mainly on direct mapping to UMLS with support from the ancillary gene and cell lists. Rindflesch states that UMLS, Metathesaurus has wide coverage for drug names, but not genes and cells. This corresponds to what is reported from Bodenreider and colleagues (Bodenreider et al., 2002). Moreover, the ancillary lists are also incomplete, particularly for cell lines. Therefore, EDGAR makes use of head noun information to be its clue to identify gene

---

[9] Available to be downloaded at http://us.expasy.org/sprot/hpi/hpi_ftp.html.
[10] Available to be downloaded at http://www.godatabase.org/dev/database/archive.

and cell names. For example, if in a string contain the word "cell" as its head then a substring before this word would be a cell name. Similar way is applied for gene.

The success of the dictionary-based approach depends on how efficiently the method can do the matching terms between text and dictionary as the method mentioned above have focused. Furthermore, its success also depends on the availability and the coverage of the dictionaries, as well as how up-to-date dictionaries are. If a dictionary has low coverage, even the best matching method has been applied, it would cause high false negatives [11]. So, the system should not rely on just a dictionary as shown with the EDGAR system.

2.1.1.2.2 Heuristic rule-based approach

A second approach relies on heuristic rules aiming to tackle the problem of false negatives when named entity terms are missed from the reference dictionary. One of the first methods based on hand-built rules was the PROPEP algorithm, Protein Proper-noun phrase Extracting Rules (Fukuda et al., 1998). This method used surface clue on character strings (e.g., the patterns of capitalization, numbering, and the use of hyphens or special characters in terms) to identify protein names from MEDLINE abstracts. This system achieved the precision of 94.70% and the recall of 98.84% from the experiments on 30 abstracts related to SH3 domain[12].

Proux and colleagues (Proux et al., 1998) analyzed the list of gene names for Drosophila and showed that Fukuda et al.'s method cannot hold for these groups of genes. They make use of linguistic knowledge (i.e., word morpheme and part-of-speech) derived from their finite state-based tagger, together with a set of local dictionaries. After this first stage of the algorithm, typical words such as species names, units, or common protein names are recovered. Then, the contextual clues such as the unknown word, for example the "Antp" or "esp" will be validated as gene name if it is located near the word "gene". For their small-scope test set extracted from FlyBase, they achieved good results (91.4% precision and 94.4% recall).

---

[11] "False negatives" refers to the terms supposed to be relevant NEs, but the NER system cannot recognize from the text.

[12] Hence, "domain" means a discrete portion of a protein with its own function, but does not mean a particular field of thought, activity, or interest.

The use of contextual clues and dictionaries in addition to lexical patterns is aimed at resolving false negatives and false positives [13] problems, respectively. Similarly, Ng and Wong's automatic pathway discovery system (Ng and Wong, 1999) has been augmented the lexical rules, which are mainly adapted from the method of Fukuda and colleagues, with dictionaries and context clues.

The ABGENE system (Tanabe and Wilbur, 2002) employs a number of dictionaries not only to keep out false positives, but also to recover false negatives. In this system, Tanabe and Wilbur use Brill's part-of-speech tagger (Brill, 1992) to learn transformation rules for a single-word gene and protein name recognition. These rules are based on the occurrences and part of speech of word and its neighboring words. Then, false positives (i.e., wrong genes or protein names) from these results are filtered out if a word matches a term in a precompiled dictionary of general biological terms (acids, antigen, etc), amino acid names, restriction enzymes, cell lines, and organism names from the NCBI Taxonomy Database (Wheeler et al., 2000). Interestingly, this system recovered false negative names, which were failed to be recognized by the lexical and contextual rules of the Brill tagger, by looking up from a dictionary as Locuslink (Pruitt and Maglott, 2001) and an ontology as the Gene Ontology (The Gene Ontology Consortium, 2000). It seems to conflict with a prior mentioned that the false negatives are generated from the recognition of names based on dictionary. From my viewpoint, the reason could due to an imbalance in the corpus used to train the Brill tagger. More precisely, only high frequency rules or patterns to capture genes or proteins would be generated, thus even simple names could be omitted because of their low frequency. Therefore, dictionaries can cover these missing names. On the contrary, the manual built rules (Fukuda et al., 1998; Proux et al., 1998; Ng and Wong, 1999) are capable of correct capturing simple names, but need dictionaries to clean out the wrong complex names.

Although, the use of rules with a contribution of contextual clues and reference resources are very helpful to disambiguate ambiguous bio-molecular names, yet a rule-based approach inevitably generates a large number of false positives. In biology domain, numerous terms are associated with multiple meanings since in many cases a protein shares a same name with its associated gene as well as a gene always shares its name with

[13] "False positives" refers to the wrong answers given by the NER system (i.e., terms are not required NEs, but are suggested by the system as NEs).

its transcripts, such as mRNA, rRNA and tRNA (Hatzivassiloglou et al., 2001). Supplementally, some terms could be other types of concepts in related domains, like the mouse gene "*diabetes*" has its homonym in the clinical field. It seems to conflict with the high performances shown by prior mentioned methods. These outcomes resulted from the small coverage of their test corpus, e.g., the FlyBase corpus used by Proux and colleagues contain only a small percentage of multi-word gene names (Proux et al. 1998) or only 30 MEDLINE abstracts on SH3 domain are used in the experiments of Fukuda and colleagues (Fukuda et al., 1998). Also, only genes and proteins are the targets for recognition. It is widely recognized that the rules would be very complicated or even cannot be constructed when the recognition focus widens to extract different kinds of molecular entities concurrently in order to understand their relations.

2.1.1.2.3 Machine learning approach

The prominent weak points of the dictionary-based approach are the difficulty to maintain up-to-date dictionaries or knowledge resources and keep them well organized in order to avoid ambiguities among terms. In case of the rule-based approach, besides the weak point that handcrafted rules are often incomplete, other weak points are the difficulty to modify the rules when new rules corresponding to new names are needed as well as when rules are required to be used with a new biology sub-domain which is different from the ones in which the rules were constructed. Moreover, the maintenance of rule sets becomes increasingly difficult over time. The interaction between tens of rules may be understandable, but this becomes impossible for hundreds of rules. With respect to these expandability problems, the machine learning approach is often seen as the best alternative.

It should be noted that the machine learning approach itself needs expensive and time-consuming annotation of corpora by domain experts, which are used as training examples. However, to construct training examples by domain experts seem to be done with less effort than to analysis complex rules or to complete knowledge resources in a wide and dynamic domain such as the molecular biology domain. In addition, the machine learning approach has clearer direction regarding to the performance improvement of the system. Even though human cannot have a clear picture in what happens inside the model, but it is at least apparent where is the point to be improved (e.g., the corpus or the mathematic algorithm or the selection of features). Therefore, in this work, the NER system is

developed following the machine learning direction. The proposed system focuses on feature engineering (i.e., feature design and feature selection), but not on making the choice of machine learning algorithms. As such, the following discussion focuses mainly on feature sets employed by disparate machine learning methods.

Among various machine learning methods implemented on various choices of algorithms, for example, Hidden Markov Models (Collier et al., 2000; Zhang et al., 2004; Zhao, 2004; Zhou et al, 2004), Support Vector Machines (Kazama et al., 2002; Takeuchi and Collier, 2002; Lee et al., 2003, Park et al., 2004), Decision Trees (Nobata et al., 1999.), Conditional Random Fields (Settles, 2004), Maximum Entropy (Finkel et al., 2004; Lin et al., 2004), and the hybrid model (Zhou and Su, 2004; Song et al., 2004; Rossler, 2004), the significant features that have been used are listed below.

- **Surface word**: As proven by Nobata and collegues (Nobata et al., 2000), this feature is very powerful clue about the class to which it belongs. However, using only this feature, a classification model will not achieve high accuracy because of the potentially very large size of the vocabulary, the large number of new terms being created and the limited coverage of the annotated corpus. In other word, to use this feature separately from other features causes the so-called data-sparseness problem. The *surface word* feature has to be combined with other features which provide more generalizable information such as part-of-speech.
- **Part-of-speech**: This feature is more generalizable than lemma and surface word features. A part-of-speech or syntactic word class is defined as the role that a word plays in a sentence. In traditional English grammar, there are eight parts of speech (e.g., noun, verb, adjective, adverb, pronoun, preposition, conjunction and interjection), but this list is extended for using in text applications such as 45 types of part-of-speech categories that were used in Nobata and colleagues' system (Nobata et al., 1999). Although not only ambiguities from sentence structure as in general domain but also ambiguities from overlaps between names of molecular entities and words in general language[14], various part-of-speech taggers or shallow parsers have been used to parse biological language and have been shown to be flexible enough for this distinctive sublanguage. For instance,

[14] Proux and colleagues reported that from their analysis of the list of gene names for Drosophila, many names belong to the (English) natural language, e.g. vamp, ogre, zip, zen, etc. (Proux et al, 1998)

the Brill's part-of-speech tagger (Brill, 1992), the Conexor FDG parser (Tapanainen and Jarvinen, 1997) in systems, HMM-based part-of-speech tagger (Zhou et al., 2004) etc. Some studies reported better accuracy when these statistical parsers have been trained on biological corpus (Zhou et al., 2004). On the contrary, some found lower accuracy (Collier and Takeuchi, 2004). Some systems claim very significant improvements in accuracy using this feature (Lee et al., 2003; Zhou et al., 2004).

- **Orthography**: Orthographical information can be considered to be the set of rules of how a word is spelled. It has been widely used in NER, such as (Collier et al., 2000; Nobata et al., 2000; Takeuchi and Collier, 2002; Lee et al., 2003; Lin et al., 2004; Zhang et al., 2004). This feature has been claimed to be very significant for NER in the molecular biology domain (Collier and Takeuchi, 2004). Each system has defined its own sets of orthographical information. For examples, features *AllCaps*, *CapMixAlpha*, *LowMixAlpha*, and *SingleCap* are reported by Lin and colleagues as more useful orthographical features than others in their set. But, these first three features are not use in some other systems. However, they are actually all minor variations. Basically, they are not very different from each other. I think that to be able to efficiently define a set of orthographical features, the nomenclature rules (e.g., Guidelines for Human Gene Nomenclature, Drosophila Gene Nomenclature, and Standardized Genetic Nomenclature of Mice) should be consulted.

- **Morphology**: This information about minimal meaning units in words is obviously important for NER to some extent. Some examples are shown in Table 2-1. In example no.1, a word which has its suffix "~cyte" is most likely referring to a particular cell_line or cell_type. The example no. 2 shows that the suffix "~nase" indicates that the word containing it is a name of a protein. Besides named entities in the molecular biology domain, special suffixes used within an individual sub-domain can be found as an example shown in example no. 3. The suffix "~emia" is always used for naming any sicknesses which are related to a condition of the blood. The using of morphology feature results in a generalization at the word level. This feature has been assessed by pervious works (Yamamoto et al., 2004) to be useful for NER.

Table 2-1: Examples of morphological patterns observed from GENIA corpus

No.	Suffix	Named Entity Class	Example
1	~cyte	cell_line or cell_type	monocyte, lymphocyte
2	~nase	enzyme (protein)	biocytinase, chitinase, dextranase, glucokinase, kinase
3	~emia or ~emic	sickness related to a condition of the blood	leukemia, anemia

- **Head noun**: In the text, named entities in some classes are always mentioned by using their proper names plus head noun showing their classes. From my observation in the GENIA corpus version 3.02 [15], the corpus containing MEDLINE abstracts already annotated named entity classes manually, most of RNA names are ending by the head word "RNA" or "mRNA". Moreover, most named entities belonging to class cell_line are ending by the head word "cell_line" or "cells" and most named entities of cell_type are ending with the word "cells" or "cell". On the contrary, named entities of class protein or class DNA are rarely ended by its class name (i.e., 9% for class protein and 1% for DNA). However, not only does class name play a role as head word of named entities in its class, but also can other nouns as shown by some examples in Table 2-2. So, this feature seems to be salient for recognizing named entity. The experiments done by Zhang and colleagues has proved the head noun is very useful by improving the F-measure at least 7.3 in their work (Zhang, 2004). Also, it is worth to note that the head noun was shown to be significant firstly in rule-based NER systems, where it was called a function term or f-term instead (Fukuda et al., 1998; Franzen et al., 2002; Torii et al., 2003).
- **Context words**: This feature includes words occurring in the same sentence as the named entity being classified. Context words can be nouns, verbs, or words of any type of part-of-speech. The basic idea is probably based on the concept of sublanguage [16]. Harris said in one of his works that "There is a particular

[15] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/

[16] Harris introduced the term 'sublanguage' for a portion of natural language differing from other portions of the same language syntactically and/ or lexically (Harris, 1968).

structure to science information in general, and to the information of each subscience in particular because for each subscience there are particular subsets of nouns that occur with particular subsets of verbs of other words" (Harris, 2002). The underlined strings seem to convince the relevance of contextual information to NER. Some systems concern a set of words before and after a named entity in current focus, as well as other features of these context words to be used as features (Takeuchi and Collier, 2002), or just a set of words before and after a focusing named entity (Lee et al., 2003).

Table 2-2: Examples of head noun for some named entities observed from GENIA corpus

No.	Head Noun	Named Entity Class	Example
1	activator	protein	heterodimeric activator, Ets activator, urokinase_like plasminogen activator, transcriptional activator
2	inhibitor	protein	cytoplasmic inhibitor, cytosolic inhibitor
3	promoter	DNA	engenous E2A promoter, lg promoter, V kappa gene promoter, IL-2 gene promoter, stem cell promoter
4	transcript	RNA	AML1/ETO transcript, PML/RAR alpha transcript

Moreover, some particular verbs [17] which occur adjacent to a focusing named entity have been used as contextual information for a named entity being classified as well (Zhou et al., 2004). In Zhou and colleagues' work, a binary feature is used to represent whether a verb occurs in the context around the term in focus or not. By combining information related to verbs into NER system, the performance is expected to be improved. It is straightforward to think that particular verbs may provide the evidence on the boundaries and the classes of biomedical entity names, for example, the verb "bind", "interact", and "inhibit"

[17] Eight verbs including activate, express, bind, induce, inhibit, interact, regulate, and stimulate are used in Zhou and colleagues' work.

are often used to indicate the protein-protein interaction. This knowledge has been widely used to extract instances of relations between biological entities in the text (Ono et al., 2001; Blaschke et al., 1999; Putejovsky et al, 2002; Thomus et al., 2000 ; Sekimizu et al., 1998). Furthermore, there exists a prior work (Spasic et al., 2003) which tried to classify terms into their potential classes by using the context given by verb. This method employs the verb complementation patterns which have been learnt automatically by combining information found in a corpus and ontology. The verb complementation patterns are the co-occurrence between each verb and the arguments which their concept classes are known. Once the verb patterns are obtained, an unknown class term will be classify to the potential class based on the similarity measure between this new term's verb complementation patterns to the pre-analyzed for known class term. However, against expectation, both efforts get not preferable results for using contextual verbs to classify named entities. The overall F-measure has decreased by 1.8 in Zhou and colleagues. F-measures of 40.68%, 26.28%, 21.85%, and 19.69% resulted from using bind, inhibit, interact, and mediate respectively in the work of Spasic and collegues.

In my opinion, I think it needs an efficient adaptation for using this verbal context. The key point to be discovered is how to represent the semantic relations between a verb and terms being its arguments. To employ verb features able to represent just the knowledge that if verbs exist in the context of the term in question or not would be insufficient.

As mentioned above, these features derived from name-internal sources of information (e.g., surface word, part-of-speech, orthography, morphology, and head noun) and from name external sources of information (e.g., context words) are important to recognize molecular named entities. However, it should be noted that these features have to be used with care because some features may interfere with other features resulting in performance degradation.

Furthermore, there is no existing system which has explored these features that can achieve the performance, F-measure, over 75%. The highest performance reported in the most recent shared-task, JNLPBA [18], is obtained by Zhou and Su with the F1-score of

[18] http://www.genisis.ch/~natlang/JNLPBA04/

72.6 (Zhou and Su, 2004). This is far behind the state of the art in news-based NER where an F-score of over 95% on recognition of people and place name etc. is the standard. I believe there should be some other features else that can enhance NER in molecular biology domain. New valuable features are what I am are exploring in this thesis.

2.1.2 Extraction of relations between molecular named entities

Following on from NER, the higher level task is to recognize relations between named entities, such as protein-protein relation (Ono et al., 2001; Thomas et al., 2000; Friedman et al., 2001), protein-drug relation (Rindflesch et al., 2000), or protein-subcellular location relation (Craven and Kumlein, 1999). From

Figure **2-2**, the facts that *Cytokines* has a binding relationship [19] to *hematopoietin receptors* and *Cytokines* has an activation relationship [20] with *JAK kinases* are extracted at this recognition stage.

Relation extraction includes not only relation of named entities in the basic event (e.g., transcription event, translation event and post translational modification event), but also the relation between the relationships between the basic events. It can be the extraction within a single sentence or even more ambitious to extract relations that span over multiple sentences. In this latter case, the techniques for coreference resolution are essential (Pustejovsky et al., 2002). At present, most interest of relation extraction is to extract protein-protein interaction relation. Various techniques for extracting protein-protein interaction have been proposed; perhaps these approaches belong to 3 groups as discussed below.

The first approach is based on the statistical estimation of the occurrence of surface words in literature. This approach is practical for the system which aims merely to discriminate documents that are likely to contain interaction information from the others (Stapley and Benoit, 2000; Jenssen T. K. et al., 2001; Marcotte et al., 2001;Donaldson et al. 2003). It is obvious that the performance of the system using this approach will be decreased for the case that the interaction occurs in a sentence with multiple names. This

[19] Keyword *bind* is classified into *attach* category of interaction keyword. (Friedman et al., 2001; Temkin and Gilder, 2003)

[20] Keyword *activate* is classified into *activate* category of interaction keyword.

should be because the system cannot recognize the real number of interaction events occur in such sentences.

Secondly, a more practical approach is the use of regular expressions predefined based on the ordering in a sentence of interaction elements (i.e., protein names) and the verb indicating a specific type of interaction (e.g., bind, inhibit and interact). No deep linguistic analysis is used here (Blaschke et al., 1999). Only simple patterns such as "protein A – action – protein B" were used in this system. Variants of such pattern (i.e., "action – protein A – protein B" and "protein A – protein B – action") cannot be extracted from the system. Moreover, this approach seems unable to cope with a sentence in which a subject or object is distanced from a verb by parenthetical commas, relative clauses, etc.

Third, some rule-based approaches employ deeper syntactic aspects such as syntactic categories of a particular sentence's constituent (part-of-speech) and syntactic roles (e.g., subject and object) to construct a set of rules (Sekimizu et al., 1998; Rindflesch et al., 2000; Ono et al., 2001; Pustejovsky et al., 2002). The coverage and precision of this group outperforms others.

Furthermore, machine-learning-based approach to extract instances of relations between named entities also exists (Craven and Kumlien, 1999). Craven and Kumlien's interest are not protein-protein interaction, but other 5 types of relations: the relation between protein and subcellular-structure, the relation between protein and cell-type, the relation between protein and tissue, the relation between protein and disease, as well as the relation between protein and pharmacologic-agent. The system benefits from a relational learning algorithm (Quinlan, 1990).

2.1.3 Evaluation events

With regard to system performance, overall performances of systems for both NER and relation extraction are still far from the levels where they can be used to replace the human curator. Although the performances reported from some systems are very high, it is not significant enough to conclude that the methods used in the systems practical. First reason would be related to the coverage scope of the corpus. The set of corpus corresponding to a particular sub-domain would contain fewer ambiguities in names than wider one. For instance, the system of Proux and colleagues (Proux et al, 1998) obtained F1-score of 93% which is as high as human-level performance from using the Flybase

corpus, but the precision can be decreased by 20% from using the larger scope corpus such as the Medline. Moreover, the ambiguities in names of different types of organisms are not equal as discussed before in section 2.1.1. This is also one key point to the system performance. In addition to the aspect of corpus scope, the size of corpus impacts the system performance as well. This is affirmed by the experiments reported in Collier and Takeuchi's work (Collier and Takeuchi, 2004). In that, F1-score of 60.7% and 40% obtaining from applying the system with base features to 100% and 20% of Bio1 training corpus[21]. The machine learning system always gets the better result from the larger set of training data.

Furthermore, the difference of criterion or what an error is in an evaluation of each research also needs to be considered. The different assumptions in evaluation have already been concerned by some systems; for instance, Franzen and colleagues (Franzen et al., 2002) have shown the efficient way to evaluate their system compared to other systems with several notions of correct matching. As the need of gold standard evaluations like in general domain (i.e., the evaluation series of MUCs), the shared-tasks offering standard evaluations for many types of IE in the molecular biology domain had been established. For each shared-task, all participants must use the same data proposed by the task organization with their own methods to reach the same extraction goal of the task. Therefore, performances among different participants can be compared fairly. The short descriptions of some main evaluation events in the molecular biology domain are shown in Table 2-3 and the summarization of results from each shared-task is given below.

- **JNLPBA-2004** [22]: This International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA) only focused on named entity recognition. Only 5 classes (e.g., protein, DNA, RNA, cell line and cell type) annotated in GENIA V.3.02 corpus were used. In this shared-task, the best system resulted in the F-score of 72.6% by using the well designed classification model which is the integration of a Support Vector Machine (SVM) into a Hidden Markov Model (HMM). In addition, the knowledge deeper than lexical-level knowledge were explored in this best system, such as the name alias

[21] http://research.nii.ac.jp/~collier/resources/bio1.1.xml
[22] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/ERtask/report.html

phenomenon, the cascaded entity name phenomenon, the use of closed dictionary which is constructed by extracting all entity names from the training data, the use of open dictionary from the database term list Swiss-Prot[23] and the alias list LocusLink[24], the abbreviation resolution as well as part-of-speech tagging which is trained on the GENIA corpus V3.02p.

- **TREC Genomics Track-2004**[25]: In ad hoc retrieval task, 50 information-need statements and a database of Medline 1993-2004 (~4.6 million abstracts) are given. Then, the IE system must answer all documents satisfying the information-need. The examples of 50 information-need statements are such as "What are the stem cell markers in different tissues?", "Information on neil1." and "Similarities of metabolic pathways for yeast and ecoli". The best result for this task is 0.4075 for mean average precision (MAP), 6.04 for mean number of relevant document @10 and 41.96 for mean number of relevant document @100.

  Another task is to retrieve from the collection of about 20,000 full-text articles, all relevant documents to a particular gene. Then, these documents must be categorized corresponding to the GO top category, i.e. function, process and locus. The best F-score for this task is Triage: 0.2681 and GO categorization: 05611.

- **BioCreAtIvE-2004**[26]: This Critical Assesment of Information Extraction in Biology (BioCreAtIvE) focused on two main tasks: task 1-extraction of entities and task 2-funtional annotation. Task 1 was also divided into task 1A and task 1B of which details are as follows.

---

[23] http://ca.expasy.org/sprot/
[24] http://www.ncbi.nih.gov/entrez/query.fcgi?db=gene
[25] http://ir.ohsu.edu/genomics/2004workshop.html
[26] http://www.mitre.org/public/biocreative/

Table 2-3: List of shared-task evaluation events for IE systems in the biology domain

Evaluation Event	Data	Shared Task Description
JNLPBA-2004 *Place/Date: Geneva, August 28-29, 2004*	the extended version of the GENIA corpus version 3.02 with reduced numbers of classes	Allow participants to use whatever methodology and knowledge sources they liked to identify and classify technical terms in the molecular biology domain
TREC Genomics Track-2004 *Place/Date:NIST, November 18, 2004*	- MEDLINE bibliography -full text of articles about mouse genomics biology	2 main tasks: 1) retrieve documents from a large subset of the MEDLINE bibliographic database using topics obtained from scientists 2) retrieve documents relevant to a particular gene and categorize these documents to GO top category
BioCreAtIvE-2004 *Place/Date: Spain, March 28-31, 2004*	- text annotated with gene names from three model organism: Fly, Yeast and Mouse - text annotated with GO term	2 main tasks: 1) extract of gene-related name and gene name 2) link descriptive mentions in text with a GO concept
TREC Genomics Track-2003 *Place/Date: NIST, November 19, 2003*	text annotated with the GO classes	For gene X, find all MEDLINE references that focus on the basic biology of the gene or its protein products from the designated organism. Basic biology includes isolation, structure, genetics and function of genes/proteins in normal and disease states.
KDD Cup-2002 *Place/Date: Canada, July 23-26, 2002*	a set of papers on genetics or molecular biology, and for each paper, a list of the genes mentioned in that paper	2 main tasks: 1) determine if a paper contain any curatable gene 2) determine if that paper has experimental results for transcripts of that gene or proteins of that gene

BioCreAtIvE Task 1A focused on extraction of gene-related names in sentences; for example, genes, binding sites, motifs and proteins. A term (mutant-related term "*p53 mutant*" or any of words *codon*, *antibody*, etc.) with a

gene name were extracted. But, the generic terms such as "*zinc finger*" or mutation terms such as "*p53 mutations*" were not required. The evaluation in this task showed that the best methods achieved F-scores of 80%, with both the best precision and recall values of around 80%. It was concluded that this performance is inherently limited by the accuracy of the human annotator. Also, the main concerning point for the improvement of overall performance is about tokenization. Tokenization or word segmentation for biological terms is non-trivial; for instance, from this portion of text "...a protein kinase A-mediated pathway...", the needed token is "protein kinase A" but not "A-mediated". It is important to have a parser with the special rules for token text in the molecular biology domain.

BioCreAtIvE Task 1B focused on extracted normalized gene names (i.e. to link all the genes mentioned in text to a list of unique identifiers) for 3 model organisms (mouse, fly and yeast). The system produced lists were evaluated by comparison of the lists produced by human efforts. The result has shown that there is a high correlation between degree of achievement and ambiguity nature of the organisms. Among these 3 databases, the yeast database consists of smallest vocabulary, shortest names and least ambiguity. The mouse database consists of largest vocabulary, longest names, but less ambiguity than fly. Contrarily, the fly database has most ambiguity with medium-length names and large vocabulary. Therefore, the comparison of the approaches' performances from the organism aspect indicated that the highest F-scores of most systems are when identifying yeast's gene names, followed by mouse and fly respectively. Only the system extracting yeast's gene names can obtain the best F-score of about 90%.

BioCreAtIvE Task 2 addressed issues of automatic functional annotation using GO classes. This task required systems to link specific text passages with GO concepts when full text articles were given. As this task is very difficult, top systems returned 250-300 correct passages with recall and precision only about 30%. It was expected that the creation of lexical resources for GO terms and paraphrases will help to improve these results.

- **TREC Genomics Track-2003**[27]: In TREC-Genomics track, the best results to extract all PubMed abstracts discussing a given gene's function from 525,938 abstracts (dating 4/1/2002-4/1/2003) were 0.4165 for mean average precision (MAP), 3.16 for mean number of relevant documents @ 10, and 4.84 for mean number of relevant document @ 20. These best results are achieved by a research group from the National Library of Medicine. The key methodologies to achieve these good results were: (1) identifying species through use of MeSH terms and other simple rules, (2) recognizing terms or their synonyms or lexical variants in non-text fileds, in particular MeSH and substance name and (3) using additional general key words, such as genetics, sequence, etc.
- **KDD Cup-2002** [28] : This Knowledge Discover and Data Mining (KDD) competition aimed to construct automatically extraction system to assist genome annotators (task 1) and to construct models that can characterize the behavior of individual genes described in text. The best performances (resulted from the team using manually generated rules and patterns to perform the tasks) were the F-scores of 78% and 67% for the first and second tasks, respectively.

2.2 Predicate-argument structure (PAS): a frame describing semantic roles

An event is described in each sentence by a composition of a verb and its arguments. A verb, which indicates a type of an event expressed in a sentence, can exist in its verbal form, its participial modifier format or its nominal form. For example, the normal form of a verb used to describe the event "making something active" would be *activate*, its participial modifier format would be *activating* or *activated*, and its nominal format would be *activation*. Beyond a verb, sentence constituents holding semantic roles to complete the meaning of an event indicated by the verb are called arguments. The semantic roles played by the set of arguments with respect to the particular verb are represented in the PAS frame of that verb.

---

[27] http://ir.ohsu.edu/genomics/2003meeting.html
[28] http://www.biostat.wisc.edu/~craven/kddcup/index.html

PropBank	VerbNet	FrameNet
PAS for Verb: SELL **Arguments:** 0: seller 1: thing sold 2: buyer 3: price paid 4: beneficiary **Sentence 1:** [All Brownstein]$_0$ **sold** [it]$_1$ for [$60 a bottle]$_3$. **PAS for Verb:** RENT **Arguments:** 0: landlord 1: thing rented 2: renter 3: rent 4: term **Sentence 2:** [Mary]$_0$ **rented** [a room]$_1$ to [John]$_2$ for [a week]$_4$ then evicted him.	**PAS for Verb Group:** GIVE **Verb Members:** give, sell, rent, render, refund, peddle, pass, loan, lend, lease **Arguments :** 0: agent 1: theme 2: recipient **Sentence 1:** [All Brownstein]$_0$ **sold** [it]$_1$ for *$60 a bottle*. **Sentence 2:** [Mary]$_0$ **rented** [a room]$_1$ to [John]$_2$ for *a week* then evicted him.	**PAS for Event:** Commerce_sell **Event Definition:** Basic commercial transactions from the perspective of the seller **Verb Members:** sell, rent, charge, lease, retail, vend **Arguments :** 0: seller 1: goods **Sentence 1:** [All Brownstein]$_0$ **sold** [it]$_1$ for *$60 a bottle*. **Sentence 2:** [Mary]$_0$ **rented** [a room]$_1$ to *John* for *a week* then evicted him.

Figure 2-3: Predicate-argument structures of PropBank, VerbNet and FrameNet

Recently several major projects have been proposed in providing resources of an English predicate-argument lexicon. These projects include VerbNet (Kipper et al., 2000), FrameNet (Baker et al., 1998), and PropBank (Kingsbury and Parmer, 2002; Kingsbury et al., 2002). There are significant differences in approach among these 3 projects. For example, PAS of verbs *sell* and *rent* are proposed as two distinct structures in the case of

PropBank and only a single structure for both verbs in the case of VerbNet and FrameNet (Figure 2-3).

VerbNet defines general PAS for a group of verbs that share similar syntactic behavior, underlying Levin's alternations theory (Levin, 1993). VerbNet's PAS for *give* contains *sell* and *rent* as members. Argument roles for all of the *give* verb members are assigned for *agent*, *theme*, and *recipient* illustrated by example sentences 1 and 2. In the case of FrameNet, PAS is defined based on the underlying principal of what users or applications expect to see for a specific event definition. FrameNet's PAS for event *Commerce_sell* shown in Figure 2-3 expects only argument *seller* and *goods* from the event driven by any verb in a set of verb members. Considering the annotation on sentence 1 in these 3 projects, "All Brownstein" is annotated as *seller*, *agent*, and *seller* in PropBank, VerbNet, and FrameNet respectively. Similarly, there is also an argument to support the annotation of "it" in all projects. But, only the PropBank scheme has an argument labeled *price paid* to support element "$60 a bottle" of sentence 1 which is likely to be an important participant of the event describing a selling activity. Moreover, a constituent "a week" in sentence 2 is considered to be an argument labeled as *term* only by the PropBank scheme. I consider that arguments like *price paid* for the events involving the verb *sell,* and an argument *term* for events involving the verb *rent,* are important for user applications.

In contrast to VerbNet and FrameNet, PropBank defines individual verb-specific PAS frames which are likely to contain more detailed specifications of arguments than are possible for verb groupings. Moreover, PAS construction in a more verb-specific manner than either VerbNet or FrameNet would assist explicitly in discovering rules for mapping from surface syntactic structures to underlying semantic propositions. In this thesis, I decide to use PropBank's scheme as a basic starting point and examined sentences containing interesting verbs from a variety of molecular biology journal articles. Thus, more detail only about PropBank is illustrated.

In PropBank a verb may get more than one PAS frame if the verb can be used in several senses. There will be one set of arguments labeled with semantic roles for each verb sense. For example, PropBank defines three distinctive PAS frames (Figure 2-4) for the verb *run* on account of sense variation. A semantic role of an argument represents a semantic relationship between the argument and its related verb. In a sentence, it is possible that not all arguments defined in a PAS frame of a particular verb sense are

mentioned. The example sentence in Figure 2-4(a) illustrates this point (i.e., only *Arg0* and *Arg1* occur in this sentence without the occurrence of *Arg2*, *Arg3* and *Arg4*). In each PAS, arguments are labeled ranging from *Arg0* up to *Arg5* with a mnemonic label indicating its predicate-dependent role.

(a)	(b)	(c)
PAS for Verb: RUN **Sense:** operate, proceed **Arguments:** Arg0: operator Arg1: machine, operation, procedure Arg2: employer Arg3: coworker Arg4:instrumental **Example:** Mr. Stromach wants to resume a more influential role in running the company. Arg0: Mr.Stromach REL: running Arg1: the company	**PAS for Verb:** RUN **Sense:** walk quickly **Arguments:** Arg0: runner Arg1: course, race, distance **Example:** John ran the Boston Marathon. Arg0: John REL: ran Arg1: the Boston Marathon	**PAS for Verb:** RUN **Sense:** encounter **Arguments:** Arg0: encounterer Arg1: thing encountered **Example:** John ran into problems with his dissertation. Again. And again. Arg0: John REL: ran Arg1: problems with his dissertation

Figure 2-4: PropBank's three distinct predicate-argument structures of *run*

Besides these core arguments defined in PAS are adjuncts which are traditionally not defined in PAS because they can potentially take multiple values and not required to minimally define the event. PropBank does consider adjuncts when annotating sentences, and provides labels such as ArgM plus tags such as TMP for temporal information, LOC for locative information, PRP for a reason or motivation, etc. After manually defining

PAS, PropBank has annotated the Penn TreeBank II Wall Street Journal corpus, which contains constituency and dependency information from the TreeBank project (Marcus, 1994).

2.3 Predicate-argument structure and the molecular named entity recognition

This section aims to illustrate why this thesis proposes that the semantic relation between a predicate and its arguments represented in PAS has a power to enhance NER systems. The PAS is the deeper level than lexical and syntactic representation levels of the proposition conveyed in a sentence. In each PAS frame, a set of semantic relationships in terms of the specified roles of the arguments of the predicate indicating the event is formed.

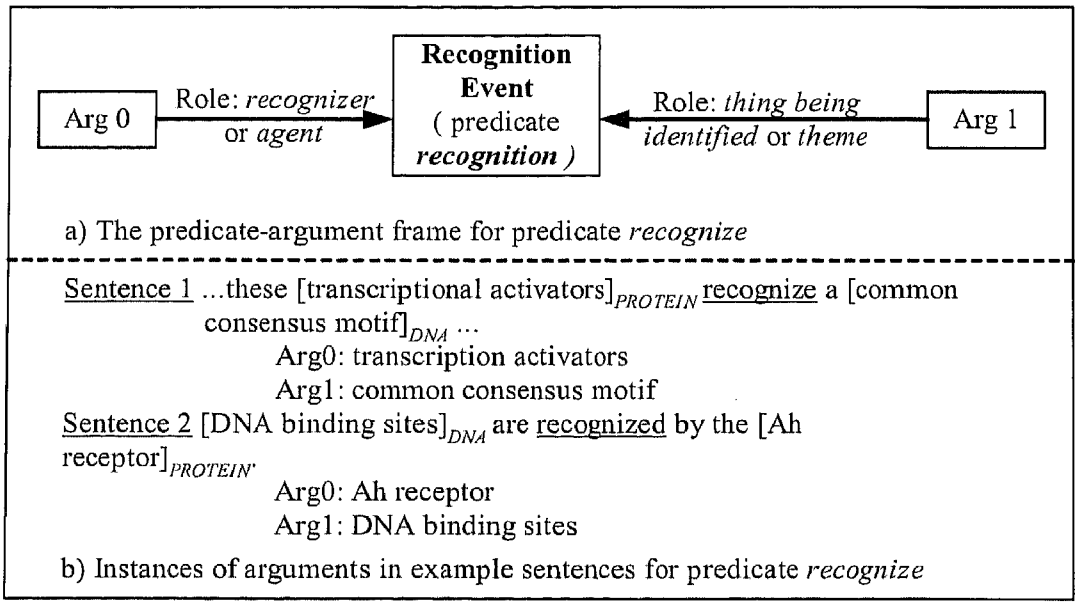


Figure 2-5: Semantic relationships between a predicate *recognize* and its arguments

Figure 2-5(a) shows the predicate-argument frame of the predicate *recognize* which is used to express the recognition event[29] in the molecular biology. From consulting to this

[29] It is the recognition for an antigen or a substrate.

PAS frame, the semantic knowledge at the PAS level of sentences 1 and 2 can thus be obtained as shown in Figure 2-5(b). That is the occurrence of a recognition event has two participants (i.e., Arg0 and Arg1). The first argument (Arg0) has a relationship to the predicate *recognize* as a *recognizer* or *agent* of the event, and the second argument (Arg1) plays the role of the *thing being identified* or *theme* in the event. Sentence 1 shows the usage of the predicate *recognize* in active voice. The sentence's surface subject, which is "*transcriptional activators*", plays the role of *agent* and its surface object "*common consensus motif*" plays the role of *theme*. On the contrary, a surface subject of sentence 2, which is "*DNA binding sites*", plays the role of *theme* and a surface object "*Ah receptor*" plays the role of *agent* as the predicate *recognize* is used in *passive voice*.

The constituent "*transcriptional activators*" in sentence 1 and the constituent "*Ah receptor*" in sentence 2 have the same semantic role of *agent* and these two entities belong to the same semantic class of *PROTEIN*, but their syntactic functions are different (i.e., the former functions as the subject whereas the latter functions as the complement of the preposition *"by*"). Similarly, the constituent "*common consensus motif*" in sentence 1 and the constituent "*DNA binding sites*" in sentence 2 have the same semantic role of *theme* and the same semantic class of *DNA*, but not the same syntactic function (i.e., the former functions as the direct object while the latter functions as the subject). Thus, it can be concluded that the semantic role of an argument would impose to a particular type of named entity[30]. This is a key idea to employ semantic relations in PAS for enhancing NER system. However, semantic roles in the PAS of a predicate will contribute to improve NER system if they relate to a particular type of named entities for which the NER system is involved.

The evidences used in the existing NER systems can be categorized into 2 groups: the internal and external evidences. The internal evidences are provided by the words in the named entity term itself such as orthographic and morphological information. The external evidences are provided by the context in which a term appears. The co-occurrence of terms appearing in the local context of a target entity term is so far the main external evidences used in NER systems. In this thesis, the evidence proposed to be used is the semantic roles represented in a PAS. This evidence is considered as a term external

---

[30] The empirical evidence observed on GENIA V3.02 corpus (http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/topics/Corpus) shows that the frequency of occurrence for *PROTEIN* to be *agent* in a recognition event is about 53% and for *DNA* to be *theme* is about 26%.

evidence as same as the co-occurrence. The external evidences are helpful for NER system especially when the internal evidences are weak in determining in which class the named entity is. Sometimes the internal evidences are totally useless. For instance, by using only the term internal evidence (i.e. the lexical information), the NER systems will not be able to identify that the term "*China*" in a sentence "*China announced a new policy regarding North Korea*" does not refer to the country. Instead, it should be classified as a government's spokesman. In this case, the co-occurrence information which is a kind of the external evidence plays an important role to identify that the term "*China*" refers to a person (not a country) because the co-occurrence between the term in the class person and the word "*announce*" (a logical form of "*announced*") would have higher frequency than between the term in the class country and the word "*announce*". However, most of the sentences used in the molecular biology literature are compound[31] or complex[32] sentences. The external evidence in the form of co-occurrence information can easily mislead the NER systems.

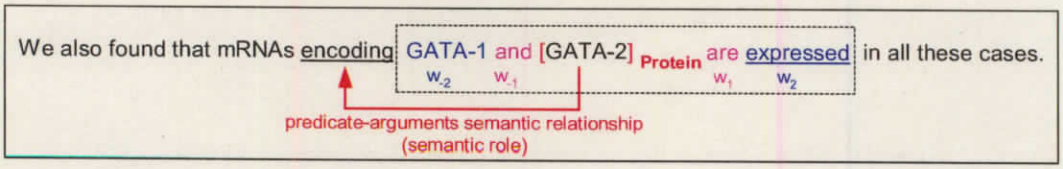


Figure 2-6: The external evidences (co-occurrence and semantic roles) related to the target term "*GATA-2*"

The sentence in Figure 2-6 is an example of a complex sentence found in the molecular biology domain. The target term to be classified to a particular named entity type is "*GATA-2*". By using the co-occurrence information which is usually related to 2 terms before the target term (i.e., "*GATA-1*" and "*and*") and after (i.e., "*are*" and "*expressed*"), the term "*GATA-2*" tends to be classified to a wrong class that is a gene name due to the high frequency of co-occurrence between a gene name and the word "*express*" (a logical form of "*expressed*"). On the contrary, by considering the semantic roles of the term "*GATA-2*", this term tends to be correctly classified as a protein name. The term "*GATA-*

[31] A compound sentence is a sentence composed of independent clauses connected by a co-coordinating conjunction, semi-colon or an independent marker such as *however*, *therefore* and *moreover*.

[32] A complex sentence is a sentence which has one independent clause and one or more dependent clauses that rely on the some component of the independent clause for their completeness.

2" has no semantic relationship to the predicate "*expressed*", but it is an argument of the predicate "*encoding*" and has the semantic role as "*the product of encoding event*". In the molecular biology domain, proteins but not genes are counted as the encoding product. From this example, it can be concluded that to know what a semantic role (or a semantic relation to a predicate) of a target term is would lead to the correct result for NER system.

As describe above, this thesis proposes to take into account the semantic relationship between a predicate and its argument in terms of semantic role for enhancing NER in the molecular biology domain. The investigation of this proposal has been evaluated and reported in chapter 4.

Chapter 3
PASBio: an analysis of PAS frames from literatures in the molecular biology domain

In the molecular biology domain, a gene and its products are the center of the study. Therefore, the assertion of genes, their products and functions at the cellular level, as well as the combined effects at the organism level can be seen in the literature of this domain. As explained in section 2.2, PAS is a representation of a set of semantic roles played by the arguments participating in the event indicated by the predicate. Genes or gene products, the participants of the molecular event, possibly are the arguments playing either the role of *agent* or *theme*. In addition, different molecular level or phenotypic effects are described as the other arguments of such events.

To conceptualize a surface sentence describing an event into PAS is a nontrivial task. There is often no simple mapping rule between syntactic knowledge (also called grammatical information) and semantic knowledge in the form of PAS. Domain knowledge is very important for the identification of an argument's boundary from the syntactic components in a sentence. Also, it helps to correctly interpret the semantic role of the argument. A sentence's constituent functioning as a subject would have its semantic role as *agent* or *causer* and an object would have its semantic role as *theme* or *patient* in most of the cases. Not only are machines confronted with the difficulties in transforming a syntactic component of a sentence into an argument with its semantic role, but also human. Domain expertise is essential but it does not guarantee the absence of misinterpretation. Therefore, the PAS frames as reference knowledge for annotating semantic roles in text are needed. The reference resource of PAS frames describing semantic roles of a predicate's arguments should be constructed for human annotators before machines are trained. Furthermore, the PAS frames can also be used as a guideline on defining extraction templates for text mining application.

To construct a lexicon of PAS frames is not new, but all available resources such as PropBank (Kingsbury and Palmer, 2002; Kingsbury et al., 2002), VerbNet (Kipper et al., 2000) and FrameNet (Baker et al., 1998) which are described in the previous chapter are

grounded on the general domain. The scientific sublanguage[33] used in the molecular biology literature has its special characteristic that makes it distinct from general language. Some nouns (entities) belong to some semantic classes and verbs are used in some particular senses can be found in only the molecular biology sublanguage. Due to the constraint relations between participants of a molecular event, a word can have subtly different meanings between being used in the molecular biology domain and in general domain. The differences in the meaning and the usage of some nouns and verbs between the molecular sublanguage and general English language motivate the PASBio project that aims to construct PAS frames specifically for molecular biology domain.

This chapter is organized as follows. Firstly, the biological events being the focus of the PASBio project are explained. Secondly, the reason supporting the requirement of an analysis of PAS frames specific for the molecular biology domain is shown. That is how it is very difficult even for human in order to derive the understanding on a sentence in the form of semantic relationships represented in PAS will be illustrated. Then, the scheme and method used in the analysis of PAS frames are described, following with the resulted frames. Finally, the general idea to utilize PASBio frames as reference resources for researchers in the field of bioinformatics is introduced.

3.1 Events in the molecular biology domain

According to the Gene Ontology, the term *biological process* refers to a broad category of biological tasks accomplished via one or more ordered assemblies of molecular entities (genes or gene products). The biological process often involves transformation, in the sense that something goes into a process and something different comes out of it. The biological processes are cell growth and maintenance, signal transduction, metabolism and biosynthesis, etc.

A biological process can be composed of several molecular events. Each molecular event is carried out by one molecular entity or well-defined assemblies of several entities. For example, *phosphorylation* of a protein molecule by a protein kinase is a molecular event, which is a part of the cellular signaling process or *transcription* of a gene by a

[33] A sublanguage is the particular language used in a body of texts dealing with a circumscribed subject area (often reports or articles on a technical speciality of science subfield, in which the authors of the documents share a common vocabulary and common habits of word usage (Hirchman and Sager, 1982).

polymerase is a part of the gene expression process. Consequence of a molecular event or a disruption of it will be a local effect to the event or an overall effect to the entire process. For instance, a *mutation* in the coding region of a gene that introduces a stop codon into the open reading frame would lead to a pre-mature termination of transcription considered as the local effect and may be responsible for a disease state of an organism due to deficiency of that protein as the phenotypic effect. Different events are described by different predicates with the associated sets of arguments as illustrated in Figure 3-1.

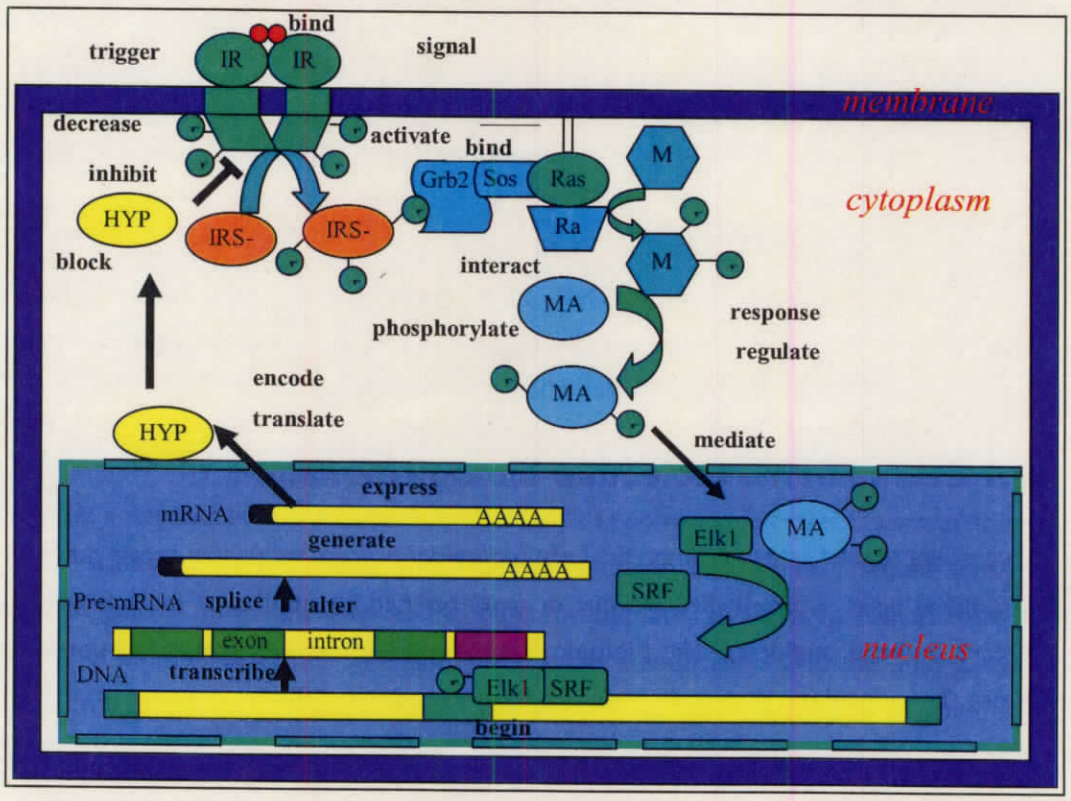


Figure 3-1: Molecular events and predicates (*bold letter*) used to describe the events[34]

The figure shows a hypothetical signal transduction pathway of an idealized cell and its corresponding predicate for each event. The signal is triggered at the outer membrane ligand-binding to receptor dimers. This signal is mediated (by various proteins) to the

[34] This figure is drawn by Parantu K. Sha, a researcher at EMBL (European Molecular Biology Laboratory), Heidelberg, Germany.

nucleus of the cell using various events (protein-protein interactions, phosphorylation, etc) and initiates transcription of a gene. The protein product (after splicing, translation and synthesis) of the gene inhibits receptor signaling. Thus, it regulates its own expression levels via a negative feedback loop. The direction of information flow is shown with arrows. The predicates listed in this figure cover events in gene expression, regulation and signaling processes which are of the current interest.

3.2 Predicate-argument relationships conveyed by the statements in molecular biology literatures

As PAS is a representation of semantic relationships between a predicate relating to a particular event narrated in a sentence and its arguments. These semantic relationships are described in terms of the arguments' semantic roles. With regard to the biological events described in section 3.1, the molecular entities are mostly related to the arguments of a predicate indicating the event and the functions of the molecular entities in the event are related to the semantic roles of the arguments.

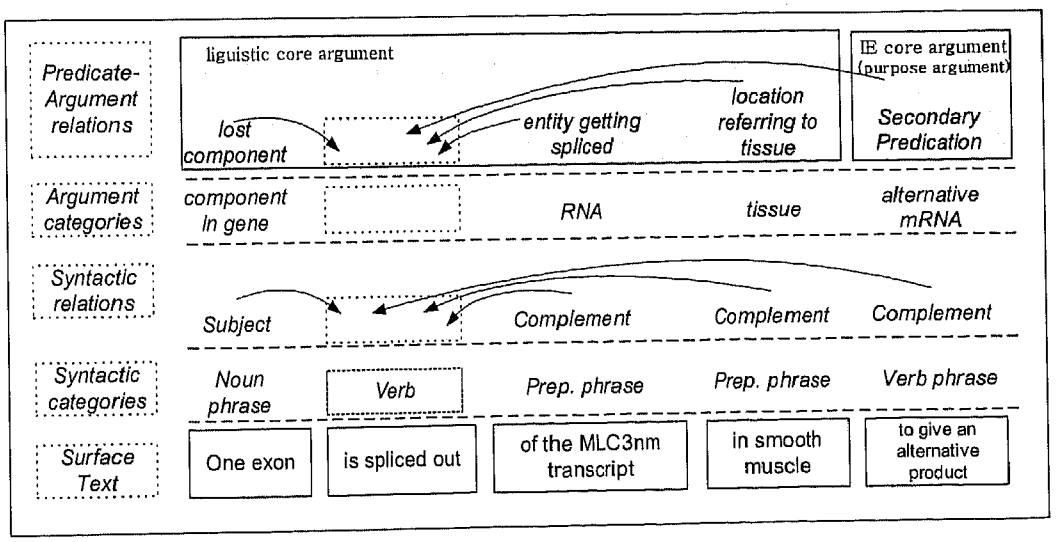


Figure 3-2: Syntactic and semantic level representation of the surface text "*One exon is spliced out of the MLC3nm transcript in smooth muscle to give an alternative product*"

The semantic roles are represented at the PAS level which is the level higher than surface and syntactic representation of a sentence. The semantic knowledge in the PAS level is considered to be deeper than the semantic and syntactic knowledge in other lower levels. Fundamentally, underneath the understanding on a sentence, a hierarchy of sentence representation containing various kinds of conceptual knowledge needs to be derived as an example in Figure 3-2.

From above figure, the understanding made on a sentence "*One exon is spliced out of the MLC3nm transcript in smooth muscle to give an alternative product*" is shown in different levels (i.e., from the surface text level up to the PAS level). The first higher level than the surface level is the level describing syntactic categories, also called linguistically as parts of speech. The sentence's constituents "*One exon*", "*is spliced out*", "*of the MLC3nm transcript*", "*in smooth muscle*", and "*to give alternative product*" have their syntactic categories as *noun phrase*, *verb*, *prepositional phrase*, *prepositional phrase*, and *verb phrase* respectively. Next higher level is the syntactic relations level in which the syntactical function of each constituent in the sentence is described. From the figure, "*One exon*" functions as the *surface subject* of the passive form verb "*is spliced out*" and other constituents function as the *subject's complement*.

The levels higher than the two syntactic levels mentioned above are semantic levels including *argument categories* level and *predicate-argument relations* level. In contrast to the syntactic levels, these semantic levels are highly related to domain knowledge. Both the categories and the semantic roles, used in the level of *argument categories* and *predicate-argument relations* respectively, would be the concept classes of objects and the set of relationships between objects. In this example, at the argument categories level, "*One exon*", "*the MLC3nm transcript*", "*smooth muscle*" and "*alternative product*" constituents pertain to the domain concept classes of *a gene product (RNA)*, *tissue* and *alternative mRNA* respectively. At the highest level of the scheme proposed, the PAS representation contains the most abstract information motivated from conceptual in the real world. Semantic roles played by the constituents aside from the predicate in focus are represented at this level. Thus, the process of *removal of an exon from mRNA* (alternative splicing) is indicated by the predicate *splice out* which is in verbal form for this example. Here, the verb's arguments play the semantic roles of *lost component* ("One exon"), *entity getting spliced* ("the MLC3nm transcript"), *location referring to tissue* ("smooth muscle"), and *secondary predication - showing purpose or reason in this example* ("to

give an alternative product"). The semantic role of *secondary predication* is assigned to the argument "*to give an alternative product*" because this by itself is capable of instantiating a PAS frame and at the same time it is considered in this scheme to possibly be a core argument due to being a key component of the occurrence of the event.

In accordance with IE systems, it is substantial to be able to understand the sentence at the level higher than syntactic levels in order to efficiently extract required facts. At least, the knowledge from the level describing argument categories which relates to the facts explaining types of domain objects participating in the event should be obtained. At the PAS level, the knowledge of semantic roles increases the degree of completeness of the knowledge in order to understand what occurs in the event. The intervention between participants and the outcomes of the event are represented in the PAS level. Some recent efforts to apply this PAS knowledge for their applications affirm the importance of the PAS knowledge. Text processing applications in general domain such as machine translation tools make use of PAS knowledge as a key knowledge in the sentence representation shared between source language and target language (Han et al., 2000; Hajic et al., 2004). In the molecular biology domain, the PAS knowledge has been used for extracting interaction or relation between gene and gene products (Sekimizu et al., 1998; Rindflesch et al., 2000).

In order to take advantages of PAS knowledge, a constituent at the surface level must be mapped to an argument defined in the PAS frame. It is naturally for molecular biology experts to conceive the PAS semantics of the sentence from using pre-constructed syntactic information (grammatical knowledge), but it is not true for people without domain knowledge. In the following, selected sentences from MEDLINE abstracts and EMBO[35] Journal articles are used to illustrate what makes the conceptualizing on a surface sentence into semantic relationships between a predicate and its argument difficult.

In Figure 3-3, the sentences (1)-(3) are the examples of surface sentences describing the event indicated by the predicate *eliminate*. Here, we consider 3 different arguments: A, B and C. Semantic roles assigned to each argument include A – causal agent of the event, B – the entity being removed, and C – location at molecular (sequence) or cellular level where the entity is being removed. Sentence (1) shows simple indicative form of which the surface-based pattern to map surface form to PAS level would be "A eliminates B in

[35] The European Molecular Biology Organization [http://www.nature.com/emboj/]

C" (where A=*One mutation*, B=*the BamHI site* and C=*exon7*); sentence (2) shows the passive form, without the mention of A and C, of which the mapping pattern would be "B are eliminated" (where B=*all three sites*); sentence (3) shows a form, using a different preposition compared to sentence (1) in order to mention C, of which the mapping pattern would be "A would eliminate B within C" (where A=*a 3-bp in-frame deletion*, B=*an*

(1) [One mutation]$_A$ eliminates [the BamHI site]$_B$ in [exon 7]$_C$ and ...

(2) The same high level of activation of B-Raf occurs only when [all three sites]$_B$ are eliminated.

(3) One of the remaining three families carried [a 3-bp in-frame deletion]$_A$ that would eliminate [an asparagines residue]$_B$ within [a kinase domain of the product]$_C$; the other two...

(4) Northern blot analysis with mRNA from eight different human tissues demonstrated that [the enzyme]$_A$ was expressed exclusively in [brain]$_C$, with [two mRNA isoforms of 2.4 and 4.0 kb]$_B$.

(5) [Two equally abundant mRNAs for il8ra]$_A$, [2.0 and 2.4 kilobases in length]$_B$, are expressed in [neutrophils]$_C$ and arise from usage of two alternative polyadenylation signals.

(6) This "functional allelic exclusion" is apparently due to control of the TCR assembly process because these [T cells]$_C$ express [RNA and protein for all four transgenic TCR proteins]$_A$.

Figure 3-3: Examples of the surface forms describing events corresponding to the predicates *eliminate* and *express*. Semantic roles of the predicates' arguments are marked as $[\ldots]_A$ or $[\ldots]_B$ or $[\ldots]_C$

asparagines residue and C=*a kinase domain of the product*). As can be seen, each sentence requires a specific pattern based on surface form to map into the same PAS. It would not be possible to construct surface-based mapping rules to cover all possible alternation. However, for these 3 sentences, using of the knowledge from the syntactic relations level is practical to build up semantic relations from these 3 surface texts. The corresponding mapping rule based on semantic relations would be (A=a subject of

eliminate, B=a direct object of *eliminate*, C=a complement of the preposition functioning as adverbial of *eliminate*) for active voice[36].

The examples of sentences describing the event *express*[37] are shown as sentences (4)-(6). These sentences illustrate that only the syntactic knowledge is unable to directly link the surface level to PAS level. All participants of the event and their semantic roles when PAS-based representation of the sentences is applied are shown. The arguments participated in the occurrence of the event *express* consist of A – entity expressed, B – physical property of the expressed entity, and C – location referring to organelle, cell or tissue. In sentence (4), (where A= *the enzyme*, B=*two mRNA isoforms of 2.4 and 4.0kb*, C=*brain*) the information needed to describe the event with respect to argument B is marked by using a prepositional phrase, but in sentence (5), (where A=*two equally abundant mRNAs for il8ra*, B=*2.0 and 2.4 kilobases in length*, C=*neutrophils*), argument B is described in an appositive which is seemingly not playing an important role in the description of the event. For these 2 sentences, a common mapping rule based on syntactic relations cannot be obtained. The constituents of argument B in (4) and (5) have different syntactic functions (i.e., adverbial and apposition respectively).

Sentence (6), (where A=*RNA and protein for all four transgenic TCR proteins* and C=*T cells*, without mentioning B) shows a different kind of problem. When reading through sentence (6), without deep domain knowledge, one may think that "T cells" has semantic role as an agent to activate the expression of "RNA and protein for all four transgenic TCR proteins" because it is a subject of the sentence. However, biologists would not think of "T cells" as an agent in this context, perceiving it as information about location. With the knowledge about molecular biology, "T cells" cannot be perceived of as an agent but the appropriate condition happening in "T cells" would be a reason.

To discuss linguistically, the usage in the molecular biology literature of the predicate *express* is analogous to the usage of the predicate *develop* in PropBank as shown in Figure 3-4. The constituent "John's neck" which locates in sentence (7) as a complement of preposition "on" is moved to be a subject in sentence (8) without changing its semantic

---

[36] The direct object in active voice sentence is promoted to function as subject in passive voice sentence, and the subject is demoted to a complement of the preposition "by" (that may be left out).

[37] Expression or gene expression is the process by which a gene's coded information is converted into the structures presenting and operating in the cell.

roles. This transition is similar to the movement of the argument C, a location argument, of sentence (4) or (5) to a subject position in sentence (6).

As discussed above, the semantic information in PAS can be directly obtained from the grammatical knowledge (i.e., syntactic categories and syntactic relations) just in some cases. To build up the complex rules for mapping surface level into PAS level from various sources of knowledge (e.g. syntactic information and domain knowledge) would need the machine learning as the rules can be too complex to complete by manual. Therefore, the PAS frame describing a set of significant semantic roles for a predicate is required to be a reference frame for constructing a semantic role-annotated training data. The availability of PAS frame resources for using in general domain (e.g., PropBank, VerbNet and FrameNet) supports the claim.

(7) [A big spot]$_{Arg2}$ developed on [John's neck]$_{Arg1}$.

(8) [John's neck]$_{Arg1}$ developed [a big spot]$_{Arg2}$.

Figure 3-4: Examples of sentences annotated by Propbank project. PAS frame of the predicate *develop* consists of 2 arguments (i.e., Arg1: non-intentional theme and Arg2: thing developed)

The molecular biology sublanguage has some specific properties compared to general domain as the scientific conceptual would have strong influence to the word-meanings. In PASBio project, the analysis results show that some verbs have been used in only molecular biology domain or some have been used with the sense slightly different (or obviously different from non-scientific domain. The particular types of participants (entities) of the molecular events described in molecular biology literature bring into a set of PAS frames specialized for the molecular biology domain as discussed in the section 3.4.

3.3 Defining PAS frames for the molecular biology domain

This section contributes to explanation of how predicates are choose, how example sentences are obtained, using scheme and methods, as well as constructed PAS frames.

Only some PAS frames are selected to discuss in this section. All of theme are available to be download at http://research.nii.ac.jp/~collier/projects/PASBio/.

3.3.1 Data collection

3.3.1.1 Selection of verbs

The English language used in research articles of biological and biomedical sciences is a sublanguage of written natural language. While most of its vocabulary is similar to that of general English, some verbs are domain-specific in nature. The main focus here is the verbs that are used for describing molecular events in biology. Various researchers have different areas of interest and new concepts are added in the literature continuously. However, the areas of cellular signaling, gene expression, regulation and disruption of expression events are very important for the larger community of investigators involved in basic biomedical research and those involved in high throughput analysis. They are discussed throughout different parts of papers as possible cause of normal and disease states of different organisms. Hence, ignoring the normal distribution (frequency) of different verbs in the literature the verbs are chosen from those involved in the above-mentioned processes (events). Most of the verbs are shown in Figure 3-1.

3.3.1.2 Selection of example sentences

Most of IE applications are still largely carried out using PubMed abstracts. Using abstracts is advantageous because they contain the highest density of keywords compared to other sections of research articles. However, the bio-text mining should scale-up to cover full journal articles where most of the detailed results are contained along with the supporting evidences, the comparisons to other works, the background information, etc. Recent investigations have shown that Introduction and Discussion sections apart from paper abstracts may be viewed as interesting sources of important biological information (Shah et al., 2003). Thus, the PAS-frames are defined by analysis on sentences from MEDLINE abstracts and from all other sections except the Method section of full text journals EMBO, PNAS[38], NAR[39] and JV[40]. Sentences from the Method section are not

[38] Proceedings of the National Academy of Sciences of the United States of America [http://www.pnas.org/]
[39] Nucleic Acids Research Articles [http://nar.oupjournals.org/]
[40] Journal of Virology [http://jvi.asm.org/]

used in this analysis because they are limited in terms of biomedical information, as well as they have generic written styles and verb sense usage tending to overlap with general language.

Sentences were carefully chosen to cover a broad usage of each verb under study from the MEDLINE and full text journal corpora as described before. The equal numbers of sentences containing a particular verb in its verbal format and its participial modifier format are in control. Before starting an analysis on each sentence, a sentence was parsed using Connexor Parser (Tapanainen and Jarvinen, 1997) that uses Functional dependency Grammar (FDG), to give parse tree, word, lemma, syntactic function and dependency links between words in order to help in determining the boundary of each argument exists in a sentence. This parse tree served as a useful guide in hand analysis, but was not considered by any means as a gold standard. At least 10 sentences were selected to determine PAS of the verb under study. The use of the parser considerably reduces the manual labors involved in defining arguments.

3.3.2 Guidelines to define PAS frame in the PASBio project

The PropBank's scheme (with necessary adaptations) is used to define PAS for the molecular biology domain. To define PAS for any verb, a survey about the usages of the verb from a set of sample sentences in a representative corpus is made. Examining the usage of an individual verb will indicate if it needs to be divided into several senses. In PASBio, these senses are divided with the aim of obtaining fine-grained semantic senses using the WordNet (Miller, 1990) lexical database. Each of PASBio's PAS contains a set of core arguments. A core argument is an argument shown by its usage to be important to complete the meaning of the event. Nevertheless, if an argument is considered important but there is no evidence to show that the argument exists together with the predicate in at least 20% of the selected sentences, this predicate may not be assigned as a core argument. There are two different types of core argument: the first type plays a role during the main event denoted by the predicate while the second type plays a role after the main event and aims to express results or consequences of the main event. Further details are given in the next section (Figure 3-5) illustrated with the PAS for *mutate*. *Arg X* (with *X*, a cardinal number, starting from *0* and then incremented for each additional argument) is used for labeling the first type of core argument and *ArgR* is used for the second type. A mnemonic label is added after *Arg X* and *ArgR* in order to give a short

description of the semantic role played by the argument. Biological function and usage of the argument are used to describe the semantic role in PAS. No attempt is made to ensure the consistency of mapping between argument labels (argument name) and the roles (the mnemonic labels) played by the arguments across verb frames, except *Arg0*. *Arg0* is reserved for only the argument playing the semantic role of *agent*. In some cases, this agent argument is not found in the usage of some verbs. Thus, PAS frames of such verbs will contain the core argument from *Arg1*. See PAS frames for *mutate* (Figure 3-5), *express* (Figure 3-11) and *transform.02* (Figure 3-13) as examples.

In addition to annotating a sentence's constituents corresponding to core-arguments with the tag *Arg X* or *ArgR*, the sentence's constituents which do not play the role of core arguments but fall into three types (i.e., adverbial, negation and modality) are annotated with the tag *ADV* or *MAN* in the case of an adverbial, *NEG* in the case of negation, and *MOD* in the case of modality. At the current stage of this project, only adverbials in terms of adverbs are considered to be annotated as *MAN* (for a manner adverb) or *ADV* (for other types of adverbs). If any adverbials in terms of phrases or clauses are mandatory for expressing events indicated by particular predicates, these adverbials will be defined as core arguments within PAS frames. For example, an adverbial phrase playing the role of locative modifier is included in the set of core arguments in the frame for predicate *initiate*. (Refer to example sentence "Apparently HeLa cells either initiate transcription *at multiple sites within RPS14 exon 1*."). Moreover, a manner adverb deserves special distinction from other adverb types because it shows how a certain action is performed which is very important to understand facts in a biological sentence. For example, "*normally*" in the sentence "Mice have previously been shown to develop *normally*" is necessary for IE in order to understand that there is no problem in the development of the mice. Other types of adverbs for example play the roles of aspectual modifiers that give information about whether some event or state of affairs is completed or is still going on, and so forth (e.g., "*still*" in the sentence "Wanda *still* would like to talk about the music festival."), adverbs playing roles as frequency modifiers that indicate the frequency of a certain type of event (e.g., "always" in the sentence "One *always* hears rumors."), adverbs playing roles as focusing modifiers that consist of the four words *even, only, also,* and *too* (e.g., "The transcription is initiated *only* in female blastoderm embryos."), and so on will be all tagged as *ADV*. In case of negation and modality, *NEG* and *MOD* are given directly to a negator word (i.e., not or n't) and a modal verb (i.e., will, may, can, shall, must,

might, should, could and would) respectively. Though negations (operating at the sentence level) and modality (operating at various levels) are not defined as core arguments (mandatory arguments) within any PASBio's PAS frames because linguistically both of them cannot even be considered as any types of predicate's arguments, they are all worth annotating from an IE perspective if they exist in the same clause where a focused predicate exists. Similarly, adverbials which are not mandatory enough to be core arguments are also considered worthy of being annotated when found in the text. They should not be ignored as they can significantly alter or even reverse the meaning of the sentence.

3.4 Examples of PAS frames with the explanation

In this subsection, some examples of PASBio's PAS frames and how each frame is defined by examples of sentences relevant to it are illustrated. There are three important cases that are examine in comparison to PropBank: (1) verbs that are rarely used in general language (e.g., *splice*) or have a unique biological interpretation (e.g., *express* and *translate*), (2) verbs that have a similar meaning used in the newswire domain and biology domain but show different patterns of usage (e.g., *alter* and *initiate*), and (3) verbs that are used with the same meaning and usage style in both domains (e.g., *abolish* and *delete*).

Table 3-1: Examples of predicates in each group

Group A : same sense, more arguments
alter, begin, develop, disrupt, inhibit, initiate, mutate, proliferate, skip
Group B : same sense, less arguments
generate, block, decrease, lose, modify
Group C : same sense, same structure
abolish, confer, eliminate, lead to, result, delete
Group D : different sense or not occur
splice, express, truncate, translate, encode, transform, catalyze, transcribe, recognize

The usage of different verbs in biology influence PAS for biological domain falls into four groups: A – same sense, more arguments; B – same sense, fewer arguments; C – same sense, same structure; D – different sense or does not occur. Table 3-1 shows the examples of verbs for each group. The frames of PAS of two verbs are given as examples for each group.

3.4.1 Group A

Verbs in this group have been used in biology documents with the same semantic sense as in PropBank, but they required more core arguments in their structures.

Consider the event of mutation, one of the most important biological events and a general cause behind genetic diseases. The verb *mutate* is used to describe the changes in an entity (gene or gene product) and mutations can be natural or engineered. PropBank defines two arguments for this verb which are *Arg0: agent* and *Arg1: entity undergoing mutation*, but from analysis four arguments are proposed in this work for the PAS frame of the verb *mutate*. As mentioned in the section 3.3.2, *Arg0* is reserved only for the argument playing the semantic role of agent. From all observed the examples, passive forms are used to describe *mutate* events which mean that the agent does exist in the event but it is unnecessary to be explicitly stated because it is commonly known by the domain experts. This results in PASBio's core arguments for *mutate* starting from *Arg1* and a position for agent is left because the agent possibly could be mentioned in other biological sub-domains. The PASBio's *Arg2* describing event participating entities (referred to as 'Named Entities') is analogous to PropBank's *Arg1*. Thus PASBio's *Arg1*, *Arg3*, and *ArgR* are extra arguments compared to PropBank. The arguments *Arg1*and *Arg3* are captured conforming to linguistic criterion (Mayers et al., 1996) which considers that a sentence element which plays a particular role to a predicate will be considered to be a core argument in a PAS frame even though it does not exist in every sentence in which the predicate appears. The existence of the omitted element is implied by the existence of the predicate. For example, in the sentence "John is eating", the existence of a core argument of *eat* which denotes a type of food will be assumed. Similarly, Figure 3-5 shows that *Arg1* and *Arg3* do not exist in all sentences 1.1 to 1.3, but are assigned as core arguments by their intuitive presence in the domain models of biologists. Noticeably, consequences of the event driven by verb *mutate* are often seen in examples. Apart from "changes at molecular level" assigned as *Arg3*, the consequence, "changes at phenotype

Frame 1: Predicate MUTATE

Argument Structure for Biology	PropBank Argument Structure
Arg1: physical location where mutation happen *//exon, intron//* Arg2: mutated entity *// gene //* Arg3: changes at molecular level ArgR: changes at phenotype level	Sense = to undergo and cause to undergo mutation Arg0: agent Arg1: entity undergoing mutation

Match to MUTATE senses in WordNet: sense 1 - undergo mutation

Sentence 1.1 The exon 5 **mutated** allele with the premature translation termination resulted in severe deficiency of Hex A.

Pred: mutate
Arg1: exon 5
Arg2: allele
Arg3: [with] the premature translation termination
ArgR: resulted in severe deficiency of Hex A

Sentence 1.2 The gene **mutated** in variant late-infantile neuronal ceroid lipofuscinosis (CLN6) and in nclf mutant mice encodes a novel predicted transmembrane protein.

Pred: mutate
Arg1: -
Arg2: gene
Arg3: [in] variant late-infantile neuronal ceroid lipofuscinosis (CLN6) and in nclf mutant mice
ArgR: encodes a novel predicted transmembrane protein

Sentence 1.3 Transient expression of the exon 8 **mutated** alpha-chain cDNA in COS-1 cells resulted in deficiency of enzymatic activity.

Pred: mutate
Arg1: exon 8
Arg2: alpha-chain cDNA in COS-1 cells
Arg3: -
ArgR: resulted in deficiency of enzymatic activity

Figure 3-5: Predicate-argument frame for *mutate*, belonging to group A

level" is suggested as *ArgR* (explained below). Sentence 1.1, 1.2, and 1.3 support this explanation.

The argument *ArgR:results/consequences* is an argument giving information about consequences after the event denoted by the predicate occurs. For *mutate*, most of the example sentences describing this event contain an *ArgR* argument, revealing the necessity of it. The requirement of this argument from an observation perspective coincides with biologist's viewpoint, thus this is considered as a core argument (more precisely an IE core argument) and named as *ArgR* instead of *Arg X* (a core argument from a purely linguistic perspective). The argument named as *Arg X* has to play a role during the event but not after the event. This condition is depicted by a formula like "mutation event = (*Arg X* + mutation + *Arg X*) + *ArgR*". Empirically, it is found that this result argument (*ArgR*) is used with verbs relating to an abnormal biological phenomenon. Examples of other verbs that need this argument are *skip*, *delete*, etc.

Frame 2: Predicate INITIATE

Argument Structure for Biology	**PropBank Argument Structure**
Arg0: agent *//gene//* Arg1: entity created *//transcription or translation//* Arg2: specific location on gene *//exon or intron//* Arg3: location as tissue or cell Arg4: method	Sense = begin Arg0: agent Arg2: theme (-creation) Arg3: instrument
Match to INITIATE senses in WordNet: sense 1 - bring into being	
Sentence 2.1 Apparently HeLa cells either **initiate** transcription at multiple sites within RPS14 exon 1, or capped 5'oligonucleotides are removed from most S14 mRNAs posttranscription. **Pred: initiate** **Arg0: -** **Arg1: transcription** **Arg2: [at] multiple sites within RPS14 exon 1**	

Arg3: HeLa cells
Arg4: -

Sentence 2.2 I kappa B-epsilon translation **initiates** from an internal ATG codon to give rise to a protein of 45 kDa, which exists as multiple phosphorylated isoforms in resting cells.

Pred: initiate
Arg0: -
Arg1: I kappa B-epsilon translation
Arg2: [from] an internal ATG codon
Arg3: -
Arg4: -

Sentence 2.3 Since RTKs **initiate** signaling by recruiting downstream components to the activated receptor, proteins that are immediately downstreamof an activated RTK can be identified by first identifying sequences in the RTK that are necessary to activate downstream signaling (Schlessinger and Ullrich, 1992; Pawson, 1995).

Pred: initiate
Arg0: RTKs
Arg1: signaling
Arg2: -
Arg3: -
Arg4: [by] recruiting downstream components to the activated receptor

Figure 3-6: Predicate-argument frame for *initiate*, belonging to group A

Verb *initiate* also takes additional arguments as core arguments. As shown in Figure 3-6, *Arg2* (sentences 2.1 and 2.2) describes the point of transcription initiation and *Arg3* provides information about the tissue/cell where the gene (or product) is expressed. In PropBank, the sentence's segments defined by the parser with functional tag as LOC (location) will be considered as non-required elements. However, the extraction of spatial information is very important from the perspective of biological description. Furthermore, another interesting point that can be seen from the examples in Figure 3-6 is that authors in biology not only put the agent but also various other kinds of semantic roles in the subject position. In Sentence 2.1 "*HeLa cells*" is syntactically the subject that seems to be the agent of an *initiate* event, but domain knowledge suggests that the agent can be only a protein (usually polymerases bound to the gene being transcribed) in this case. "*HeLa cells*" is annotated as *Arg3:location as tissue or cell* instead of *Arg0:agent*. In sentence 2.2, "*I kappa B-epsilon translation*" is also a subject as in the previous example, but it is

Frame 3: Predicate BLOCK

Argument Structure for Biology	PropBank Argument Structure
Arg0: agent, causer Arg1: theme *//entity or process being stopped//*	Sense = oppose, halt, stop Arg0: agent Arg1: theme (action or object being stopped) Arg2: secondary predication Arg3: instrument

Match to BLOCK senses in WordNet: sense 3 - stop from happening or developing

Sentence 3.1 Tagetin is more specific for distinguishing between different RNA polymerases because it **blocks** RNA polymerase during elongation.

Pred: block
Arg0: it
Arg1: RNA polymerase during elongation

Sentence 3.2 Membranes were **blocked** in TBST (Tris-buffered saline, 0.05% Tween-20) containing 5% bovine serum albumin (for anti-phosphoryrosine blots) or skimmed milk and probed with antibodies.

Pred: block
Arg0: -
Arg1: Membranes

Sentence 3.3 Mutations at the 3′ splice site that specifically **block** step II do not affect the association of hPrps 16 and 17 with the spliceosome, indicating that these factors may function at a stage of step II prior to recognition of the 3′ splice site.

Pred: block
Arg0: Mutation at the 3′ splice site
Arg1: step II
MAN: specifically

Figure 3-7: Predicate-argument frame for *block*, belonging to group B

"entity created" assigned as *Arg1*. Only in Sentence 2.3 (describing initiation of signaling event), the subject of the sentence fills the semantic role "agent", so a subject "*RTKs*" can be annotated as *Arg0*. Additionally, the point to note is "the entity created" in sentence

2.3 is different from sentence 2.1 and 2.2 as it is a signaling event that is initiated, but not a transcription or translation event.

Frame 4: Predicate GENERATE

Argument Structure for Biology	PropBank Argument Structure
Arg0: agent, causer *//gene, protein//* Arg1: thing created	Sense = create Arg0: creator Arg1: thing created Arg2: source Arg3: benefactive Arg4: attribute, secondary predication

Match to GENERATE senses in WordNet: sense 1 - bring into existence

Sentence 4.1 Prnd **generates** major transcripts of 1.7 and 2.7 kb as well as some unusual chimeric transcripts generated by intergenic splicing with Prnp.

Pred: generate
Arg0: Prnd
Arg1: major transcripts of 1.7 and 2.7 kb

Sentence 4.2 The bidentate RNase III Dicer cleaves microRNA precursors to **generate** the 21-23 nt long mature RNAs.

Pred: generate
Arg0: The bidentate RNase III Dicer
Arg1: the 21-23 nt long mature RNAs

Sentence 4.3 Human leukocyte antigen (HLA)-G molecules are **generated** by an alternative splicing of the primary transcript of the gene and display specialized function in regulating the immune response.

Pred: generate
Arg0: an alternative splicing of the primary transcript of the gene
Arg1: Human leukocyte antigen (HLA)-G molecules

Figure 3-8: Predicate-argument frame for *generate*, belonging to group B

3.4.2 Group B

Verbs in this group have been used in biological texts with the same semantic sense as in PropBank, but they required fewer arguments in their structures in the PAS frame proposed here.

Verb *block* both in biomedical texts and in business news texts has very similar semantics. However, an event described by verb *block* in the biomedical domain may not mention information about secondary predication and instrument most of the time. The semantic role *secondary predication* is assigned to the argument that is in itself capable of instantiating another PAS frame. The sentence "[*John*$_{Arg0}$] *blocked* [*Mary* $_{Arg1}$] *from* [*completing her dissertation* $_{Arg2}$] *with* [*his constant pestering* $_{Arg3}$]." is annotated by PropBank's PAS frame. An argument Arg2-secondary predication is annotated for "completing her dissertation" because this contains in itself the PAS of the verb *complete.*

From this PropBank example, the meaning of the event denoted by *block* cannot be completely understood if the sentence just states as "[*John*$_{Arg0}$] *blocked* [*Mary* $_{Arg1}$]." as it is necessary to mention the action being stopped. In contrast in the biology domain, by mentioning only the entity being stopped (Sentence 3.1-3.3), the expert reader can understand that the event which applies to that entity is being stopped without providing an explanation of the action being stopped at the position of secondary predication. Similarly, an instrument used to block is encoded in the nature of an agent or causer. The structure of *block* and its examples are given in Figure 3-7. Only core arguments as defined in the structure exist in Sentence 3.1 and 3.2 (the agent is not mentioned). In sentence 3.3, *MAN* is used to label "specifically" as this adverb plays the role of a manner modifier.

In Figure 3-8 the PAS frame of *generate* is similar to that of *block.* Only *Arg0-agent* and *Arg1-entity created* are expressed in all observed sentences from the biology corpus.

3.4.3 Group C

Verbs in this group have been used in biological documents with the same semantic sense as in PropBank. Moreover, their usage in both the biological corpus and PropBank indicates that their PAS frames are identical. Specialization of domain does not seem to affect verbs in this group.

Frame 5: Predicate CONFER	
Argument Structure for Biology	**PropBank Argument Structure**
Arg0: agent *//mechanism, process, entity//* Arg1: given biological property Arg2: entity receiving biological property *//gene product, cell//*	Sense = grant, give Arg0: agent Arg1: gift Arg2: given to
Match to CONFER senses in WordNet: sense 2 - present	
Sentence 5.1 Besides these side chain interactions with the O6-alkyl group, structure-based analysis of mutational data suggests that substitutions at Gly156 and Lys165 **confer** resistance to O6-BG through backbone distortions. **Pred: confer** **Arg0: substitutions at Gly156 and Lys165** **Arg1: resistance** **Arg2: [to] O6-BG** **Sentence 5.2** The portion of the STATs **conferring** specificity for either a MAPK or a MAPK substrate kinase (MAPKAP) has not been determined. **Pred: confer** **Arg0: The portion of the STATs** **Arg1: specificity** **Arg2: [for] either a MAPK or a MAPK substrate kinase (MAPKAP)**	

Figure 3-9: Predicate-argument frame for *confer*, belonging to group C

In Figure 3-9 and Figure 3-10 show PAS for *confer* and *lead*. In both biology and newswire corpora, *confer* is used with semantic "to give (as a property or characteristic) to someone or something" and *lead to* is used in the sense of "to tend toward or have a result".

Frame 6: Predicate LEAD

Argument Structure for Biology	PropBank Argument Structure
Arg1: factor/cause Arg2: result	Sense = resulted Arg1: factors/cause Arg2: result

Match to LEAD TO senses in WordNet: sense 3 - tend to or result in

Sentence 6.1 In this homologous part of the genes, GPB lacks one exon due to a point mutation at the 5' splicing site of the third intron, which inactivates the 5' cleavage event of splicing and **leads to** ligation of the second to the fourth exon.

Pred: lead
Arg1: a point mutation at the 5' splicing site of the third intron
Arg2: [to] ligation of the second to the fourth exon

Sentence 6.2 Genetic deficiency of GM2 activator **leads to** a neurological disorder, an atypical form of Tay-Sachs disease (GM2 gangliosidosis variant AB).

Pred: lead
Arg1: Genetic deficiency of GM2 activator
Arg2: [to] a neurological disorder

Figure 3-10: Predicate-argument frame for *lead*, belonging to group C

3.4.4 Group D

Verbs in this group have been used in biology documents with a different semantic sense compared to PropBank, or PAS frames for them are not found in PropBank. More than one semantic sense is found in the corpus for some verbs. PAS frames for express and transform are presented in Figure 3-11, Figure 3-12, and Figure 3-13, respectively to illustrate predicate-argument structures for this group.

Frame 7: Predicate EXPRESS

Argument Structure for Biology	PropBank Argument Structure
Arg1: named entity *//gene or gene products//* Arg2: property of the existing named entity Arg3: location refering to organelle, cell or tissue	Sense = say (express.01) Arg0: speak Arg1: utterance Arg2: hearer Sense = send very quickly (express.02) Arg0: sender Arg1: thing sent Arg2: sent to

Match to TRANSFORM senses in WordNet: sense 5 - manifest the effects of a gene or genetic trait

Sentence 7.1 Northern blot analysis with mRNA from eight different human tissues demonstrated that the enzyme was **expressed** exclusively in brain, with two mRNA isoforms of 2.4 and 4.0 kb.

Pred: express
Arg1: the enzyme
Arg2: [with] two mRNA isoforms of 2.4 and 4.0 kb
Arg3: [in] brain
ADV: exclusively

Sentence 7.2 Two equally abundant mRNAs for il8ra, 2.0 and 2.4 kilobases in length, are **expressed** in neutrophils and arise from usage of two alternative polyadenylation signals.

Pred: express
Arg1: mRNAs for il8ra
Arg2: 2.0 and 2.4 kilobases in length
Arg3: [in] neutrophils

Sentence 7.3 T cells from double TCR transgenic mice **express** only one or the other of the two available TCRs at the cell surface.

Pred: express
Arg1: one or the other of the two available TCRs
Arg2: -
Arg3: T cells from double TCR transgenic mice
ADV: only

Figure 3-11: Predicate-argument frame for *express*, belonging to group D

Frame 8: Predicate TRANSFORM.01

Argument Structure for Biology	PropBank Argument Structure
Sense = to cause (a cell) to undergo genetic transformation Arg0: agent/causer of transformation Arg1: entity undergoing transformation Arg2: effect of transformation/end state	Sense = change Arg0: causer of transformation Arg1: thing changing Arg2: end state Arg3: start state

Match to TRANSFORM senses in WordNet: sense 2 - change or alter in form, appearance, or nature

Sentence 8.1 We and others have found that FGF8b can **transform** the midbrain into a cerebellum fate, whereas FGF8a can promote midbrain development.

Pred: transform
Arg0: FGF8b
Arg1: the midbrain
Arg2: [into] a cerebellum fate
MOD: can

Sentence 8.2 Phospholiipase D (PLD) is known to stimulate cell cycle progression and to **transform** murine fibroblast cells into tumorigenic forms, although the precise mechanisms are not elucidated.

Pred: transform
Arg0: Phospholipase D (PLD)
Arg1: murine fibroblast cells
Arg2: [into] tumorigenic forms

Sentence 8.3 Overexpression of the retroviral oncoprotein v-Rel can rapidly **transform** and immortalize a variety of avian cells in culture.

Pred: transform
Arg0: Overexpression of the retroviral oncoprotein v-Rel
Arg1: a variety of avian cells in culture
Arg2: -
MOD: can
ADV: rapidly

Figure 3-12: Predicate-argument frame for *transform (sense 1)*, belonging to group D

Frame 9: Predicate TRANSFORM.02 (TRANSFORM INTO)

Argument Structure for Biology	PropBank Argument Structure
Sense = to transfer gene from source organism into target organism Arg1: entity being inserted Arg2: organism or cell undergoing transformation	Sense = change Arg0: causer of transformation Arg1: thing changing Arg2: end state Arg3: start state

Match to TRANSFORM senses in WordNet: sense 6 - change (a bacterial cell) into a genetically distinct cell by the introduction of DNA from another cell of the same or closely related species)

Sentence 9.1 This construct was **transformed into** the yeast strain HF7c (Clontech).

Pred: transform
Arg1: This construct
Arg2: [into] the yeast strain HF7c (Clontech)

Sentence 9.2 For expression of the recombinant protein, pET28a-5 was **transformed into** Escherichia coli strain BL21(DE3).

Pred: transform
Arg1: pET28a-5
Arg2: [into] Escherichia coli strain BL21(DE3)

Sentence 9.3 To generate GST fusion proteins, the relevant DNA fragments were cloned into pGex2T (Pharmacia) and **transformed into** the bacterial strains BL21 or TOPP (Stratagene).

Pred: transform
Arg1: the relevant DNA fragments
Arg2: [into] the bacterial strains BL21 or TOPP (Stratagene)

Figure 3-13: Predicate-argument frame for *transform (sense 2)*, belonging to group D

The verb *express* is used in the biology domain with the meaning "to manifest the existence of a gene or gene product" (or detection of the same by the experimenter) unlike its normal usage with the meaning of "give an opinion or send quickly". The PAS of *express* is given as Figure 3-11.

In the case of *transform*, two senses are used in biology papers: "to cause (a cell) to undergo genetic (or neoplasmic) transformation" as shown in Figure 3-12 and "to transfer a gene from source organism into target organism" as shown in Figure 3-13. Even though the first meaning of *transform* found in biological corpus is similar to the sense of "change" found by PropBank, there is still a huge gap between them. In the biological literature, illustrated by examples in sentences 8.1-8.3, this genetic transformation mentions only the agent or causer, what entity is getting transformed, and what will be the effect after transformation. It will not mention the start state of the entity undergoing transformation because it is known from the expert reader's domain 'common sense' knowledge that the start state refers to a normal condition of that entity. *Transform* in the second sense always occurs in a sentence connected by preposition *into*, and in the passive voice form in which no mention is made with regard to the agent.

3.5 Utilization of PASBio's frames

The semantic representation at the PAS level does not depend on the variation of the sentences' surface forms. If two surface forms carry the same proposition, they will be represented into the same PAS. Whatever syntactic function each constituent has in one surface form, its semantic role (i.e. the semantic relation to the predicate) does not change in another surface form. For instance, the constituent "*LMP-1*" has a syntactic function as the subject of the predicate "*activates*" in the sentence "*LMP-1 activates the pathways*" and has a syntactic function as the object in the sentence "*The pathways is activated by LMP-1*". Because the proposition transmitted by the former sentence is as same as by the latter, the constituent "*LMP-1*" has the semantic role as *agent or causer* of the activation event in both sentences although its syntactic function in the former sentence differs from in the latter.

This property of the PAS interests to various information processing applications: machine translation, text summarization, relation extractions between molecular entities, etc. Machine translation requires encoding a surface sentence of a source language into a language independent logical form of a clause meaning, and then generating from this logical representation a surface sentence in a target language. PAS has been used as such a logical representation in machine translation (Han et al., 2000; Hajic et al., 2004). For

text summarization applications, PAS could be employed as the basic unit of a discourse representation, before being summarized (Marcu, 2000).

In case of the extraction of events or relations between molecular entities, PAS were simply used in the work of Sekimizu and colleagues (Sekimizu et al., 1998). In this work, the template for relation extraction had two slots for two arguments of the predicate corresponding to the target relation. This PAS template supported only binary relations between two participants of the event (i.e. only the arguments functioning as *subject* and *object*). More complete notion of PAS is employed in MedScan system (Novichkova et al., 2003). This system can transform the syntactic tree of an entire sentence into the set of logical relationships between the words in a sentence. This transformation would result in a sentence's normalized semantic tree counted as the sentence representation at the PAS level.

The potentiality of PAS that can be represented by a single form for various surface sentences containing the same information have been taken advantage directly by above text processing applications. However, the NER system which is a focus application in this thesis has shown the indirect usage of PAS. Not the meaning of the whole sentence is the focus of the NER system, but some portions of a sentence containing the named entities of interest. In this thesis, the NER system uses the semantic role of a sentence's constituent as one of the evidences to classify the named entity into the correct semantic class. This is practicable because the semantic role that each constituent (an argument of a predicate) plays in the event partly imposes type restrictions on the entities within the constituent[41].

As mentioned above, the representation of a sentence in the form of PAS benefits for several kinds of the text processing applications. Either the manual methods or the automatic methods for transforming a sentence from its surface level to the PAS level, a set of PAS frames is necessary. A lexical resource of PAS frames available for public allows the common agreement among experts. The manual methods will construct the mapping rules based on the information of a sentence in lower levels (i.e. lexical and syntactic information) to obtain the target templates defined in a shared PAS resource. The automatic methods (the machine learning-based methods) construct the training examples by using the shared PAS frames as the reference frames to annotate texts. Then,

[41] Please see section 2.3 for more details.

these annotated texts are learnt by the machine learning to build the automatic semantic role identification model.

Chapter 4

Applying semantic roles in PAS to enhance named entity recognition (NER)

In the molecular biology domain, NER is the task of identifying and classifying text constituents referring to the bio-molecular entities into their concept classes. The terms used for referring to the entities in the molecular biology domain are significantly different from the terms in the newswire domain in which the original definition of named entities is introduced. In 1995, the named entities to be extracted from the newswire literature were defined by MUC-6 as the proper names of person (e.g., *Anan*), organization (e.g., *Toyota*), location (e.g. *Southern Thailand*), etc. However, the terms counted as the bio-molecular named entities can be either the proper names or the descriptive names. For instance, the proper name "*I kappa B alpha*" and the descriptive name "*growth factor*" are both protein names. The descriptive names of bio-molecular entities are mainly derived from the biological functions of the entities. The *growth factor* is a protein named from its function of cell division stimulation when it binds to its cell-surface receptor. Another example is *elongation factor*, it is a name of protein that ribosome needs to make polypeptide chains longer. As these molecular entities are named by the terms containing general words (not proper nouns), they are ambiguous to be identified as named entities. Many terms containing similar words are not molecular named entities; for example, *bioaccumulation factor* means the concentration of a chemical in living tissue divided by its concentration in the animal's diet, or *risk factor* means anything that raises the chance that a person will get a disease. To name the molecular entities with the descriptive names multiply the difficulties from the factors described in section 2.1.1.1 (e.g., lack of naming conventions in biology, various patterns of terminology, term nesting, term coordination, homonymy, systematic polysemy and synonym).

Although, it is unclear how to compare the difficulties between NER in newswire domain and in the molecular biology domain, by the above characteristics of terms for molecular named entities as well as the ambiguity factors described in section 2.1.1.1 the

bio-molecular NER seems to be more difficult than the newswire NER. Furthermore, the experiments conducted by Gaizauskas and colleagues (Gaizauskas et al., 2003) resulted in the F-score of 93% and 83% when the same NER system is applied to newswire domain and molecular biology domain, respectively. This lower F-score of about 10% in case of the molecular biology domain compared to the newswire domain supports the previous mentioned idea that the bio-molecular NER seems to be more difficult. Furthermore, from the final results reported in the most recently shared-task of NER for the molecular-biology domain (JNLPBA-2004), the best performance in terms of F-score is only 72.6 (Kim et al., 2004). In contrast, according to MUC-6 the accuracy in general news-based NER is about 96%.

The gap between the performances of the NER systems in the molecular biology domain and in the newswire domain brings into the question that why the traditional NER systems are successful for newswire domain, but not the molecular biology domain. This would be explainable that compared to the general named entities, the molecular named entities have much more lexical ambiguities especially from the use of the descriptive names, homonymy and systematic polysemy[42]. The term internal evidence, the lexical information, is too weak to handle. The external evidence will play an important role for disambiguating molecular named entities. As discussed in section 2.3, the use of co-occurrence, one kind of the term external evidences, is practical when the target term exists near the informative terms[43] in a sentence. On the other hand, the use of semantic roles is more consistent because the semantic role of a term to its related predicate is independent of the distance between a target term's position and its related predicate's position in a sentence.

In this thesis, underlying the idea that an argument's semantic role should impose type restrictions on an entity within the argument, the use of semantic roles as term external evidences for NER systems is proposed. The semantic role knowledge is formed into features of machine learning based NER systems. What is the effective way to form semantic roles into features and how much each predicate's semantic role set can contribute to improve tradition NER systems are investigated in this work.

---

[42] Please see section 2.1.1.1 for details and examples.
[43] The informative terms refer to the terms carrying clues to the named entity class of the target term.

This chapter is organized as follows. Firstly, each of the processes and knowledge components composing the system is explained in detail. Secondly, the experimental results are shown and the contribution of PAS-related features to NER is discussed on the basis of the used features and the predicate groups. Then, the sources of errors and the factors that impede the NER system to get high performance improvement is discussed. Finally, the effectiveness of an argument's semantic role is illustrated

4.1 Method

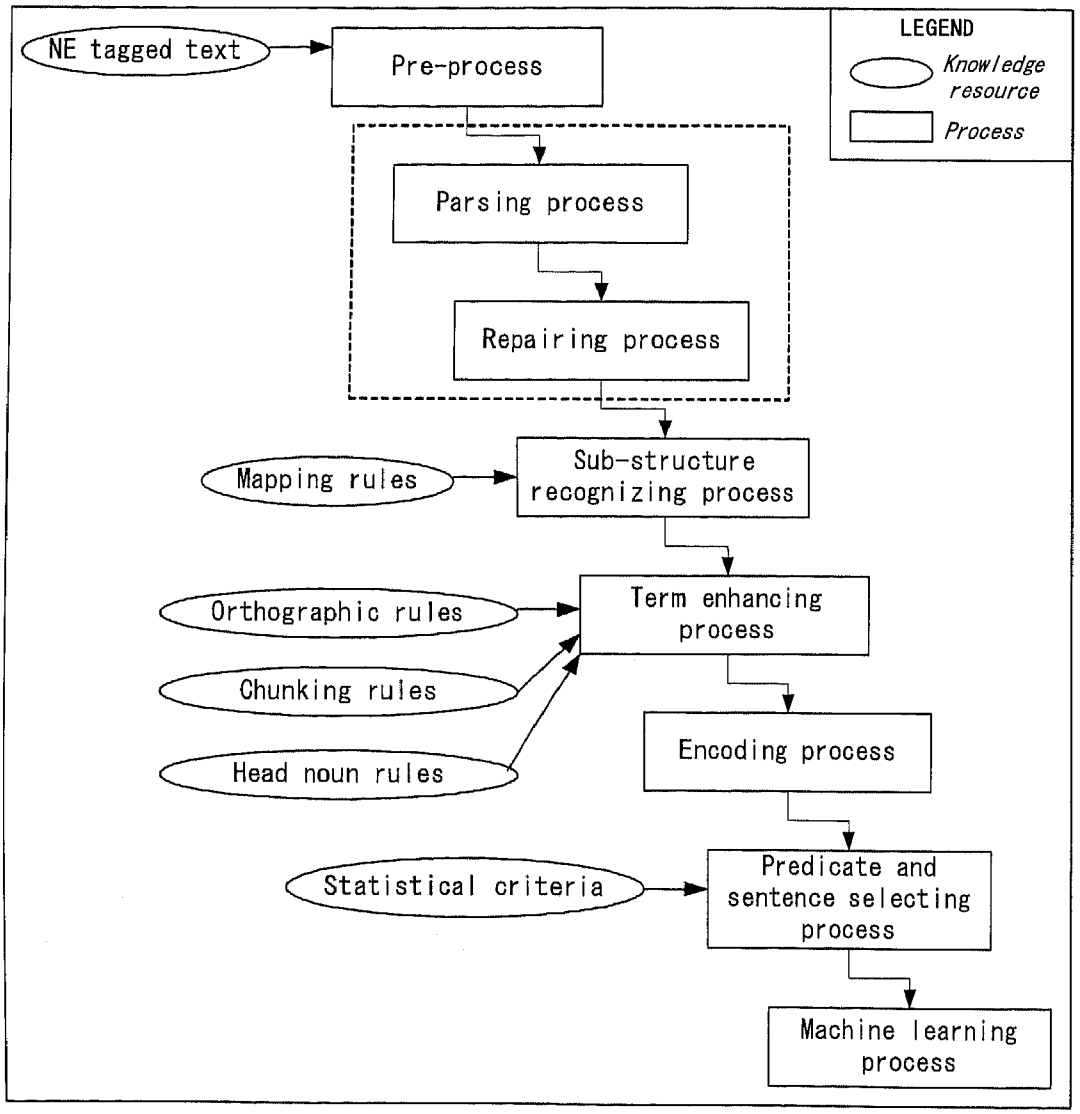


Figure 4-1: Overview of the processes and knowledge components in using system

The overall architecture for the entire system is shown by processes and knowledge components in Figure 4-1. This system is composed of 8 main processes: (1) pre-process explained in section 4.1.1, (2) parsing process and (3) repairing process explained in section 4.1.2, (4) sub-structure recognizing process explained in section 4.1.3, (5) term enhancing process explained in section 4.1.4, (6) encoding process explained in section 4.1.5, (7) predicate and sentence selecting process explained in section 4.1.6 and (8) machine learning process and its results explained in section 4.2-4.4.

4.1.1 Data set and pre-process

The GENIA corpus V3.02 (Ohta et al., 2002; Kim et al., 2003) is used as the data set of named-entity tagged text. It is the largest annotated corpus in the molecular biology domain available publicly[44]. The GENIA corpus contains 2,000 MEDLINE abstracts that were collected using the search terms *human*, *transcription factor* and *blood cell* since the corpus creators aimed their annotation work to converge on biological reactions concerning transcription factors in human blood cells.

In GENIA corpus, each word in the text from totally 490,941 words is annotated with part-of-speech tags according to its syntactic role. In addition, about 100K biological terms (97,876 terms to be precise) are annotated with 36 terminal classes (*thick circle* nodes) from the GENIA ontology shown in Figure 4-4. The GENIA ontology is intended to be a formal model of concepts corresponding cell signaling reactions in human. It is to be used by text processing applications: IE, IR-Information Retrieval, classification and categorization of documents, text summarization, etc.

(1)…we have detected two crossreacting proteins in [activated normal human lymphocytes]$_{DNA}$.

(2)Although these heterodimers do not recognize a classical [thyroid hormone response element]$_{cell\text{-}type}$ (TRE) characterized by direct repeat…

Figure 4-2: Example of qualifier inconsistency in GENIA corpus

[44] http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/topics/Corpus/

The annotated text with 36 classes in GENIA corpus has been done by 2 domain experts. However, no interannotator agreement for term annotation has been published. Through a simple inspection done in this thesis, some inconsistencies are found. From sentence (1) in Figure 4-2, all qualifiers are included in the boundary of the named entity "*activate normal human lymphocytes*" which refers to *DNA*. On the other hand, in sentence (2) the quantifier "*classical*" is not annotated as a part of *cell-type*. As in the GENIA corpus annotation scheme, to include or not include quantifiers in the named entities is left to the expert judgment, thus it is not surprise to have qualifiers inconsistency. Figure 4-3 illustrates another inconsistency. The term "*monoclonal antibody*" is annotated as *protein* in sentence (3), but the same term is not annotated in sentence (4).

(3) Using Western blot analysis with a [monoclonal antibody]$_{protein}$ recognizing 17-amino acid epitope...

(4) We demonstrate that cross-linking CD30 with an anti-CD30-specific monoclonal antibody, which mimics...

Figure 4-3: Example of loss annotation in GENIA corpus

To work with all 36 classes in the corpus may contain quite high inconsistency, therefore in this work only 5 classes as in JNLPBA shared task (Kim et al., 2004) are used. These classes are *protein*, *DNA*, *RNA*, *cell line* and *cell type* (shown as *bold-italic-named-with-yellow-background* nodes in Figure 4-4). Several subclasses of *protein*, *DNA*, and *RNA* from the original taxonomy are simplified into their mother classes corresponding to them. The classes of *cell line* and *cell type* are of interest to be annotated in order to make the task realistic for post-processing by a potential template filling application.

According to the above mentioning, pre-processing is responsible for converting all named entities tagged with classes, other than classes of *cell line* and *cell type* as well as subclasses of *protein*, *DNA*, and *RNA*, into non-tagged entities. Moreover, all subclasses of *protein*, *DNA*, and *RNA* must be converted to the classes of *protein*, *DNA*, and *RNA*, respectively.

Not only is the pre-process stage involved with converting all tagged classes into 5 required classes, but it also removes non-named entity XML elements from the text. In addition, it prepares the XML attributes in a form which is acceptable to the parser.

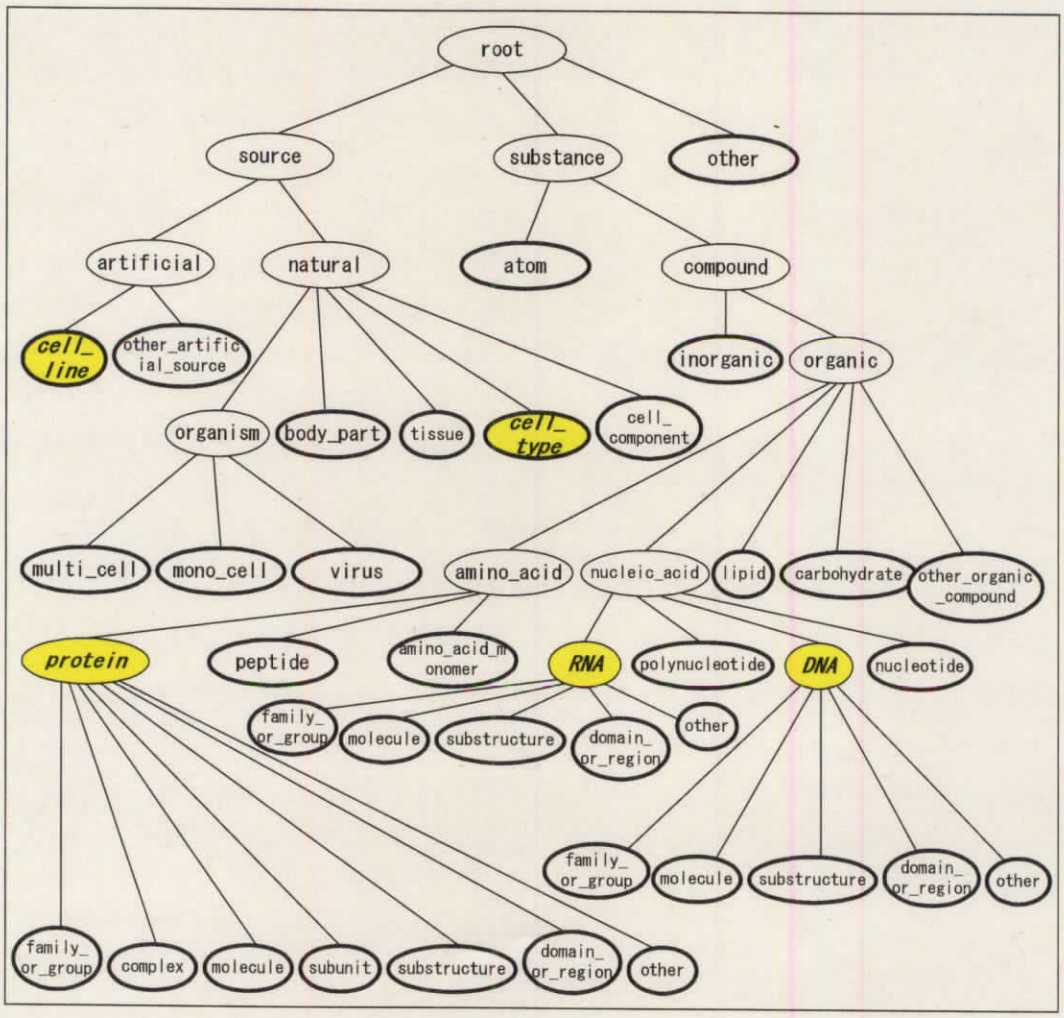


Figure 4-4: The GENIA ontology (36 terminal classes shown as *thick circle* nodes, 5 classes used in the following experiments shown as *bold-italic-named-yellow background* nodes)

4.1.2 Parsing and repairing process

The Conexor FDG parser (Tapanainen and Jarvinen, 1997) is used in the parsing process as it is widely used and is considered to be a state-of-the-art commercial parser. The

grammatical relations connecting one word to another can be obtained from the FDG parser which works based on a dependency grammar. This grammatical relation between words is needed for identifying dependencies between a verb and its arguments. In addition to functional dependency relations between words, the parser gives syntactic tags for phrases and morphological tags as well. After parsing the named-entity tagged sentence "*Both compounds altered the <NAME cl="protein">NFAT-1 transcriptional complex</NAME>, causing its retarded mobility on gels.*", the syntactic information for each token or word comes out in columns as shown in Figure 4-5. These columns are word number or word position indexes in a sentence, surface form of word, lemma form or base form of word, syntactic relation of the focused word to the word on which it depends (e.g., from line 2 in Figure 4-5, *subj:>3* means that the word *compounds* has syntactic relation as the subject of the word number 3 that is *altered* in this sentence), syntactic function (e.g., *@DN>* means functioning as determiner, *@SUBJ* means functioning as subject, *@+FMAINV* means functioning as finite main verb), surface syntactic information of a word (e.g., from line 1 in Figure 4-5, the word *Both* has *%>N* as its surface syntactic information, so this word is known by the parser that it is the determiner or premodifier of a nominal), and the last column is referred to morphological information [45] (e.g., from line 3 in Figure 4-5, *<MORPH>V PAST</MORPH>* indicates that the word *altered* is a verb in past tense form).

The repairing process is a post-process of parsing process. Its main purpose is to correct for wrongly segmented sentences which the parser sometimes outputs. Certain multi-word expressions are also identified by inserting underscores between two separate words tokenized by the parser, such as *in_vivo*, *even_though* and *so_far*.

4.1.3 Sub-structure recognizing process

The sub-structure recognizing process is the process to identify the tokens that constitute arguments of predicates. In this study, a predicate in verbal form but not nominal form is mainly focused. Therefore, for a verb such as *activate*, the surface forms of this predicate to be analyzed include *activate*, *activates*, *activated*, and *activating*, but not *activation*.

[45] Morphological information describes about the structure of word form. The FDG parser gives different aspect of the word form for different part-of-speech. For instance, the degree of comparison is a morphological information given to adjectives, but for verbs the information to indicate if a verb is modal auxiliary, or infinitive, as well as tense (e.g. present or past) and grammatical category of forms that designate a speaker or writer (e.g. singular-first person, singular-third person, etc.) are given.

With regard to arguments, only arguments corresponding to the syntactic relation of either subject or object are bound in this study. At present, it lacks a practical semantic role labeling system to identify arguments of a predicate, especially for molecular biology domain. Thus, this study which is to investigate the contribution of semantic relationship between a predicate and its arguments simplifies its scope by only concerning over arguments being grammatical subject or object. The general algorithm used for finding a subject and object term of each verb from a set of data processed until 4.1.2 is as follows:

(1) Find a position of target predicate which must be in a verbal form only.

(2) Interpreting which voice this verb is in by checking at the column 6 of the verb token. If it is *%VA*, the verb is an active verb. On the other hand, if it is *%VP*, the verb is a passive verb.

(3) Find a subject or object of a verb by traversing through syntactic relations given by the parser at the column 4. Basically, the system will traverse up until *subj:>#* is found in case of subject and traverse down until *obj:>#* is found in case of object (# refers to the word number of the target verb or the verb in focus). From Figure 4-5, the token *compounds* is found to have subject relation to the verb *alter* and the token *complex* is found to be an object.

(4) After the head of surface subject and surface object is found from (3), the full boundary of a subject or an object is identified by propagating to the premodifiers of a noun which is a subject head and an object head. These premodifiers will have *@A>* at the column 5 in parsing data. All modifiers except determiners are included in the surface subject or the surface object boundary as shown in Figure 4-5 that *NFAT-1* and *transcriptional* are included but *the* is not included in the boundary of surface object containing *complex* as the object head. A determiner is not included into a boundary of object or subject because any determiners never ever are parts of biological terms. This rule not to include a determiner is also used by Rindflesch and colleagues in their system to extract binding relationships from text (Rindflesch et al., 2000).

The step (3) explained above is practical for only basic case. However, in some cases, to look for only *subj:>#* or *obj:>#* is not enough to get subject head or object head. The extended criterion as explained in the following section is needed.

1	Both	Both	det:>2	@DN>	%>N	<MORPH> DET - </MORPH>	
2	compounds	compound	subj:>3	@SUBJ	%NH	<MORPH> N NOM_PL </MORPH>	<- Surface Subject
3	altered	alter	main:>0	@+FMAINV	%VA	<MORPH> V PAST </MORPH>	<- Target Verb
4	the	the	det:>7	@DN>	%>N	<MORPH> DET - </MORPH>	
<NAME	cl="protein">						
5	NFAT-1	nfat-1	attr:>6	@A>	%>N	<MORPH> N NOM_SG </MORPH>	<- Surface Object
6	transcriptional	transcriptional	attr:>7	@A>	%>N	<MORPH> A ABS </MORPH>	
7	complex	complex	obj:>3	@OBJ	%NH	<MORPH> A ABS </MORPH>	
</NAME>							
8	.	.					
9	causing	cause	ha:>3	@-FMAINV	%VA	<MORPH> V ING </MORPH>	
10	its	it	attr:>11	@A>	%>N	<MORPH> PRON GEN_SG3 </MORPH>	
11	retarded	retarded	attr:>12	@A>	%>N	<MORPH> A ABS </MORPH>	
12	mobility	mobility	obj:>9	@OBJ	%NH	<MORPH> N NOM_SG </MORPH>	
13	on	on	loc:>9	@ADVL	%EH	<MORPH> PREP - </MORPH>	
14	gels	gel	pcomp:>13	@<P	%NH	<MORPH> N NOM_PL </MORPH>	
15	.	.					

Figure 4-5: A parsing result from FDG parser of a sentence "Both compounds altered the NFAT-1 transcriptional complex, causing its retarded mobility on gels." Boundaries of surface subject and object are shown by *red* squares.

4.1.3.1 Bounding constituent for subject

- **Modal auxiliary verbs, Auxiliary verbs and Verb phrase functioning similar to auxiliary verbs**:

  In many cases, modal auxiliary verbs (e.g., can, must, should, would, could, may, shall, will, might, and ought to) or auxiliary verbs (e.g., have, be, and do) or verb phrases functioning similar to auxiliary verbs (e.g., have been shown to, be needed to, be used to, be known to, and play a role in) are used with a target verb as shown in Figure 4-8. When these verbs are used with a target verb, to traverse up through syntactic relations for *subj:>#*, where # is the word number of target verb will not succeed. In the case of modal auxiliary verb or auxiliary verb, the system has to traverse down starting from the subject which links to another verb at the position #. This subject will be the subject head for the target verb if this linkage is finally chained to the target verb as the main of this auxiliary verb. In the case of a verb phrase such as *have been shown* in Figure 4-6, after traversing through the linkage from the subject (*Mutations*) linking to non-target verb (*have*), the linkage will be ended at the main verb *shown*. Then, the target verb is checked if it links to this main verb or not. In this example, the target verb *block* links to the main verb *shown*. Therefore, the token *Mutations* is identified as a surface subject of the target verb *block*.

- **Sharing subject with other verbs**:

  Sometimes, the target verb shares its subject with other verbs it connects to by coordinating conjunctions (e.g., *and*, and *or*). If a token has syntactic relation as a subject of the verb which the target verb connects, that token will be identified as the subject of the target verb as well. In Figure 4-7, the token *Cytokines* is a subject of the word number 2 (*subj:>2*) and the target verb *activate* is conjunct to the word number 2 (*cc:>2*). Thus, the token *Cytokines* is bound to be a subject head of the target verb *activate*.

1	Mutations	mutation	subj:>6	@SUBJ	%NH	<MORPH> N NOM_PL </MORPH>	<- Surface Subject
2	in	in	mod:>1	@<NOM	%N<	<MORPH> PREP - </MORPH>	
3	the	the	det:>5	@DN>	%>N	<MORPH> DET - </MORPH>	
<NAME cl="DNA">							
4	Tat	tat	attr:>5	@A>	%>N	<MORPH> N NOM_SG </MORPH>	
5	gene	gene	pcomp:>2	@<P	%NH	<MORPH> N NOM_SG </MORPH>	
</NAME>							
6	have	have	v-ch:>7	@+FAUXV	%AUX	<MORPH> V PRES </MORPH>	<- Verb phrase which functions similar to auxiliary verbs
7	been	be	v-ch:>8	@-FAUXV	%AUX	<MORPH> V EN </MORPH>	
8	shown	show	main:>0	@-FMAINV	%VP	<MORPH> V EN </MORPH>	
9	to	to	pm:>10	@INFMARK>	%AUX	<MORPH> INFMARK - </MORPH>	
10	block	block	cnt:>8	@-FMAINV	%VA	<MORPH> V INF </MORPH>	<- Target Verb
11	HIV	hiv	attr:>12	@A>	%>N	<MORPH> ABBR NOM_SG </MORPH>	<- Surface Object
12	replication	replication	obj:>10	@OBJ	%NH	<MORPH> N NOM_SG </MORPH>	
13	in	in	loc:>10	@ADVL	%EH	<MORPH> PREP - </MORPH>	
<NAME cl="cell-type">							
14	human	human	attr:>15	@A>	%>N	<MORPH> A ABS </MORPH>	
15	T	t	attr:>16	@A>	%>N	<MORPH> ABBR NOM_SG </MORPH>	
16	Cells	cell	pcomp:>13	@<P	%NH	<MORPH> N NOM_PL </MORPH>	
</NAME>							
17	.	.					

Figure 4-6: A parsing result in case a target verb is not a main verb of a sentence

<NAME	cl="protein">						
1	Cytokines	cytokine	subj:>2	@SUBJ	%NH	<MORPH> N NOM_PL </MORPH>	<- Surface Subject
</NAME>							
2	bind	bind	main:>0	@+FMAINV	%VA	<MORPH> V PRES </MORPH>	
3	to	to	ha:>2	@ADVL	%EH	<MORPH> V - </MORPH>	
<NAME	cl="protein">						
4	hematopoietin	hematopoietin	attr:>5	@A>	%>N <?>	<MORPH> N NOM_SG </MORPH>	
5	receptors	receptor	pcomp:>3	@<P	%NH	<MORPH> N NOM_PL </MORPH>	
</NAME>							
6	and	and	cc:>2	@CC	%CC	<MORPH> CC - </MORPH>	
7	activate	activate	cc:>2	@+FMAINV	%VA	<MORPH> V PRES </MORPH>	<- Target Verb
<NAME	cl="protein">						
8	JAK	jak	attr:>9	@A>	%>N	<MORPH> N NOM_SG </MORPH>	<- Surface Object in the form
9	kinases	kinase	obj:>7	@OBJ	%NH	<MORPH> N NOM_PL </MORPH>	
</NAME>							
10	.	.					

Figure 4-7: A parsing result in case a target verb shares its subject with another verb

- **Relative pronoun resolution**:

  It is possibly found that a target verb exists as a main verb in a subordinate clause of which the relative pronoun (e.g., *which*, *who*, and *that*) presents as the subject of the clause. In Figure 4-8, a relative pronoun *that* (word number 7) is identified by the parser as a subject of the target verb *mediate*. To identify *that* as a subject of a verb *mediate* would not be useful to investigate the contribution of semantic relations between predicate and its arguments to NER task. The relationship between the noun phrase *cis-acting elements* and the target predicate *mediate*

must be recovered. In other words, the object of the main clause is the subject of the subordinate clause.

	1	Here	here		@ADVL	%EH	<MORPH> ADV – </MORPH>
	2	we	we	subj:>3	@SUBJ	%NH	<MORPH> PRON PERS_NOM_PL1 </MORPH>
	3	map	map	main:>0	@+FMAINV	%VA	<MORPH> V PRES </MORPH>
	4	the	the	det:>6	@DN>	%>N	<MORPH> DET – </MORPH>
<NAME	cl="DNA">						
	5	cis-acting	cis-acting	attr:>6	@A>	%>N	<MORPH> A ABS </MORPH>
	6	elements	element	obj:>3	@OBJ	%NH	<MORPH> N NOM_PL </MORPH>
</NAME>							
	7	that	that	subj:>8	@SUBJ	%NH	<MORPH> PRON – </MORPH>
	8	mediate	mediate	mod:>6	@+FMAINV	%VA	<MORPH> V PRES </MORPH>
	9	interleukin	interleukin	attr:>10	@A>	%>N	<MORPH> N NOM_SG </MORPH>
	10	responsiveness	responsiveness	obj:>8	@OBJ	%NH	<MORPH> N NOM_SG </MORPH>
	11	Of	of	mod:>10	@<NOM-OF	%N<	<MORPH> PREP – </MORPH>
	12	The	the	det:>16	@DN>	%>N	<MORPH> DET – </MORPH>
<NAME	cl="DNA">						
	13	mouse	mouse	attr:>14	@A>	%>N	<MORPH> N NOM_SG </MORPH>
	14	IL-2R	il-2r	attr:>15	@A>	%>N	<MORPH> N NOM_SG </MORPH>
	15	alpha	alpha	attr:>16	@A>	%>N	<MORPH> N NOM_SG </MORPH>

	16	gene	gene	pcomp:>11	@<P	%NH	<MORPH> N NOM_SG </MORPH>
</NAME>							
	17	using	use	man:>8	@-FMAINV	%VA	<MORPH> V ING </MORPH>
	18	A	a	det:>21	@DN>	%>N	<MORPH> DET SG </MORPH>
<NAME	cl="cell-line">						
	19	thymic	thymic	attr:>20	@A>	%>N	<MORPH> A ABS </MORPH>
	20	lymphoma-derived	lymphoma-derived	attr:>21	@A>	%>N	<MORPH> A ABS </MORPH>
	21	hybridoma	hybridoma	obj:>17	@OBJ	%NH	<MORPH> N NOM_SG </MORPH>
</NAME>							
	22	(	(				
<NAME	cl="cell-line">						
	23	PC60	pc60	mod:>21	@NH	%NH	<MORPH> N NOM_SG </MORPH>
</NAME>							
	24	)	)				
	25	.	.				

Figure 4-8: A sentence example showing the case when a target verb is a main verb of a subordinate and relative pronoun such as "that" in this example (the word number 7) posses syntactic relation as a subject of the target verb "mediate" (the word number 8)

4.1.3.2 Bounding constituent for object

In addition to the simple case that the object head of a target verb can be captured by searching for a token with *obj:>#* (where # is a word number of the target verb), there is also a case that the target verb shares its object with other verbs. Similar to sharing subject, if the target verb links to the same word as the object links to, such object would be an object head of the target verb as well.

Furthermore, an object of some verbs not only has syntactic relation directly to the target verb as an object, but also it has a relation to a preposition which follows a target verb as a complement. In Figure 4-9, the token *receptors* has no direct dependency relation to the word number 2 which is the target verb, but it links to the word number 3

as the complement of that word (shown by *pcomp:>3* in column 4). Among all verbs used in the experiments, only prepositional complements of 5 verbs (i.e., bind, associate, interact, result, and lead) are bound as the object. From these 5 verbs, there are 7 types of co-occurrence patterns between verbs and prepositions found from GENIA corpus such as "bind to", "associate to", "associate with", "interact to", "result in", "result from", and "lead to".

The consequences of this sub-structure recognition process are 4 PAS-related features (i.e., predicate surface form, predicate lemma, voice, and surface syntactic role) added for all tokens in both a surface subject boundary and a surface object boundary related to a particular target predicate. These features are detailed as follows.

- **Predicate surface form**: The surface form of the predicate preserving orthographic and morphological information.
- **Predicate lemma**: The lemmatized form of the predicate in lower case and infinitive form. Various inflected forms of the same predicate are mapped to their common root. For example, *activate*, *activates*, *activated*, and *activating* are mapped to *activate* treated as the same thing.
- **Voice**: This is used to distinguish between the *active* and *passive* voice of the predicate. Tags used for this feature are *ACT* for *active* voice and *PAS* for *passive* voice.
- **Surface syntactic role**: This is either *surface subject* or *surface object* which is identified by the method explained above. Tags used are *SSUBJ* for surface subject and *SOBJ* for direct object. Furthermore, the tag used for *surface object* which is found as a prepositional complement (explained in section 4.1.3.2) is *PCOMP*. From preliminary experiments, it is shown that to make distinction between *direct object* and *prepositional complement object* influences the correct interpretation of the objective argument's semantic roles.

4.1.4 Term enhancing process

This process adds in certain features that are specifically designed to reduce known problems of ambiguity for term recognition. This includes adding an orthographic feature for each token by using rules proposed in earlier studies (Collier et al., 2000), identifying

text chunks for noun phrases (NP), verb phrases (VP), adverb phrases (ADVP) and prepositional phrases (PP). Text chunks are found using regular expressions on parts of speech found by the parser and provide a flat representation of syntactically correlated words.

<NAME	cl="protein">						
1	Cytokines	cytokine	subj:>2	@SUBJ	%NH	<MORPH> N NOM_PL </MORPH>	<- Surface Subject
</NAME>							
2	bind	bind	main:>0	@+FMAINV	%VA	<MORPH> V PRES </MORPH>	<- Target Verb
3	to	to	ha:>2	@ADVL	%EH	<MORPH> V - </MORPH>	
<NAME	cl="protein">						
4	hematopoietin	hematopoietin	attr:>5	@A>	%>N <?>	<MORPH> N NOM_SG </MORPH>	<- Surface Object in the form of prepositional complement
5	receptors	receptor	pcomp:>3	@<P	%NH	<MORPH> N NOM_PL </MORPH>	
</NAME>							
6	and	and	cc:>2	@CC	%CC	<MORPH> CC - </MORPH>	
7	activate	activate	cc:>2	@+FMAINV	%VA	<MORPH> V PRES </MORPH>	
<NAME	cl="protein">						
8	JAK	jak	attr:>9	@A>	%>N	<MORPH> N NOM_SG </MORPH>	
9	kinases	kinase	obj:>7	@OBJ	%NH	<MORPH> N NOM_PL </MORPH>	
</NAME>							
10	.	.					

Figure 4-9: A sentence example showing the object of a target verb can be found from the complement of preposition co-occurred with the target verb

Column 1	2	3	4	5	6	7	8	9	10	11
Further-more	Further-more	-	RB	ic	B-ADVP	O	O	O	O	O
,	,	-	,	pu	O	O	O	O	O	O
we	we	we	PRP	lw	B-NP	O	O	O	O	O
show	show	-	NN	lw	B-VP	O	O	O	O	O
that	that	-	DT	lw	B-SBAR	O	O	O	O	O
IL-12	il-12	IL-12	NNP	cd h	B-NP	O	O	O	O	B-protein
stimulates	stimulate	-	VBZ	lw	B-VP	O	O	O	O	O
formation	formation	form-ation	NN	lw	B-NP	O	O	O	O	O
of	of	-	IN	lw	B-PP	O	O	O	O	O
a	a	-	DT	lw	B-NP	O	O	O	O	O
DNA-binding	dna-binding	com-plex	VBG	2c	I-NP	recog-nizes	recog-nize	ACT	SSUBJ	B-protein
complex	complex	com-plex	JJ	lw	I-NP	recog-nizes	recog-nize	ACT	SSUBJ	I-protein
that	that	that	DT	lw	B-SBAR	O	O	O	O	O
recognizes	recognize	-	VBZ	lw	B-VP	O	O	O	O	O
a	a	-	DT	lw	B-NP	O	O	O	O	O
DNA	dna	se-quence	NNP	2c	I-NP	recog-nizes	recog-nize	ACT	SOBJ	O
sequence	sequence	se-quence	NN	lw	I-NP	recog nizes	recog-nize	ACT	SOBJ	O
previously	previously	-	RB	lw	B-ADVP	O	O	O	O	O
shown	show	-	VBN	lw	B-VP	O	O	O	O	O
to	to	-	TO	lw	I-VP	O	O	O	O	O
bind	bind	-	NN	lw	I-VP	O	O	O	O	O
STAT	stat	pro-teins	NNP	2c	B-NP	O	O	O	O	B-protein
proteins	protein	pro-teins	NNS	lw	I-NP	O	O	O	O	I-protein
and	and	-	CC	lw	O	O	O	O	O	O
that	that	-	DT	lw	B-SBAR	O	O	O	O	O
this	this	-	DT	lw	B-NP	O	O	O	O	O
complex	complex	com-plex	JJ	lw	I-NP	O	O	O	O	O
contains	contain	-	VBZ	lw	B-VP	O	O	O	O	O
STAT4	stat4	STAT4	NNP	cd	B-NP	O	O	O	O	B-protein
.	.	-	.	pu	O	O	O	O	O	O

Figure 4-10: Training data is in IOB2 format. Feature columns are separated with spaces.

4.1.5 Encoding process

Encoding process has its function to constitute features derived directly from parsing result in section 4.1.2 and derived through processes in section 4.1.3 and section 4.1.4. In each experiment, a column formatted table of features with the named entity classes provided in IOB2 format[46] is used. Until now, the training data which will be used for machine learning contains 11 columns as shown in Figure 4-10.

All features comprise of 1) surface word, 2) lemma form, 3) head word of NP-chunk, 4) part-of-speech, 5) orthographic feature, 6) phrase-chunk, 7) predicate surface form, 8) predicate lemma, 9) voice, 10) surface syntactic role, and 11) named entity classes which is the answers to be learned by machine.

This whole set of training data will be selected to be trained by machines for each predicate separately. The next section describes what predicates to be studied are and what are the criteria to choose these predicates.

4.1.6 Predicate and sentence selecting process

As the predicate-argument relationship is a specific characteristic for each individual predicate, thus influences of features derived from the knowledge of predicate-argument relations are explored separately for each predicate. The training data for each predicate contains sentences extracted from the whole set of training data, mentioned in section 4.1.1, by using the criteria that these sentences must contain a focus predicate in verbal form at least once.

With respect to what predicates were studied, the predicate selecting process were started by gathering predicates used in earlier works to capture biological events (Blaschke et al., 1999; Ono et al., 2001; Pustejovsky et al., 2002) and gathered from predicates used in my work to construct the PASBio[47] database as explained in chapter 3. Most predicates from the 44 predicates which have been gathered are found rarely in the GENIA corpus (the source of training sets). In order to avoid having too small set of

---

[46] IOB2 format (Ratnaparkhii, 1998) is a standard format for word-chunk used in many gold standard collection and evaluation exercises such as CoNLL's shared tasks or the MUCs. The tag "O" is given to words outside a chunk, "B-*k*" to the first word in a chunk of type *k*, and "I-*k*" to the remaining words.

[47] PASBio contains PAS frames analyzed from the literatures in the molecular biology domain. Available online at http://research.nii.ac.jp/~collier/projects/PASBio/ (Wattarujeekrit et al., 2004).

training data, the predicates containing less than 100 examples[48] are filtered out. This filtering process results in 19 predicates in total as shown in Figure 4-11. From these predicates, 6 predicates were selected to be used for investigating of the effects of various forming of predicate-argument related features. The idea to obtain these 6 predicates is described later in this section and the outcome of the investigation will be discussed in section 4.3. For the other 13 predicates, they will be used to evaluate the overall effectiveness of the best model using predicate-argument related features compared to the lexical-based model in which predicate-argument related features are not employed. This experiment to affirm the best model and a general analysis of the sources of errors as well as the possibilities to the improvement are discussed in section 4.3-4.4.

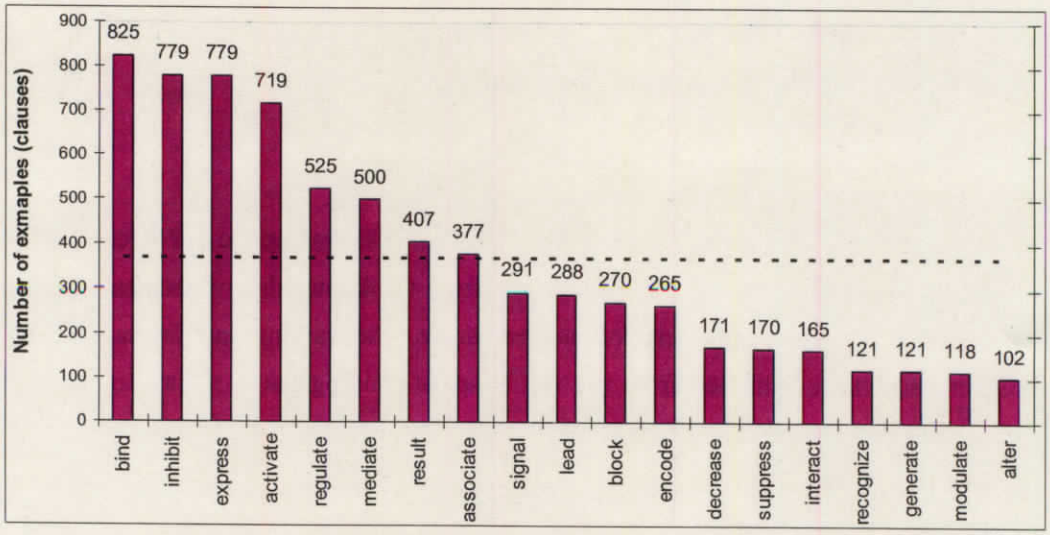


Figure 4-11: Graph showing the number of examples for each of 19 predicates used in the experiments. The dotted line represents the average number of examples for these predicates

Due to the hypothesis that the proportion of named entities as arguments should be a key impact on NER, the statistics expressing type of named entity for *agent* argument and

[48] The number of examples is a number of clauses containing a particular predicate. In a sentence, it is possible to have more than one clause related to the predicate.

theme argument [49] for each predicate is calculated from the corpus and summarized in Table 4-1. From this statistics, the 19 predicates are classified into 3 groups:

- **Group 1 (A group of predicates having both arguments *agent* and *theme* with a higher probability of belonging to a named entity class than non-named entity class)**: The predicates in this group include *bind*, *express*, *activate*, *encode*, *interact* and *recognize*.
- **Group 2 (A group of predicates having both arguments *agent* and *theme* with a lower probability of belonging to a named entity class than non-named entity class)**: The predicates in this group include *inhibit*, *result*, *signal*, *block*, *decrease*, *suppress*, *generate*, *modulate*, *lead* and *alter*.
- **Group 3 (A group of predicates having either arguments *agent* or *theme* with a higher probability of belonging to a named entity class than non-named entity class)**: The predicates in this group include *regulate*, *mediate* and *associate*.

In order to understand whether the probability to belong to a named entity class of an argument affects to the performance of NER or not, the experiments should be done on predicates from every group. However, if the number of examples for predicates from each group is not in balance, the analysis could be inconclusive. Thus, the 6 predicates are selected (2 predicates from each group) on the basis that these predicates have a similar number of training examples. Their numbers of examples are around the average line shown in Figure 4-11. Selected predicates for group 1 are *encode* and *recognize*, group 2 are *block* and *lead*, and group 3 are *regulate* and *associate*.

A set of sentences for each of these predicates will be investigated separately through several sets of features (i.e., models 1-6). To form each set of features are explained in the next section and the experimental results are given in section 4.3.

---

[49] Hence, the *agent* argument refers to the argument which has syntactic role as *subject* in the case of active voice and refers to the argument having syntactic role as *object* introduced by the preposition "by" in the case of passive voice. The *theme* argument refers to the argument which has syntactic role as *object* in the case of active voice and refers to the argument having syntactic role as *subject* in the case of passive voice.

Table 4-1: Proportion of *agent* and *theme* arguments in 5 classes of named entities

Predicate	Agent Argument								Theme Argument							
	Total Agent	Protein%	DNA%	RNA%	Cell line%	Cell type%	Total NE%	Non-NE%	Total Theme	Protein%	DNA%	RNA%	Cell line%	Cell type%	Total NE%	Non-NE%
Bind	399	46.1	10.3	00.0	00.3	01.8	58.5	41.5	662	26.6	31.6	00.2	00.2	01.2	59.8	40.2
Inhibit	561	26.2	01.3	00.0	00.2	00.4	28.1	71.9	710	09.3	01.4	00.6	00.1	00.7	12.1	87.9
Express	338	03.9	03.6	00.6	23.7	30.1	61.9	38.1	620	54.7	07.9	06.8	01.1	02.1	72.6	27.4
Activate	464	49.6	03.0	00.2	01.0	00.7	54.5	45.5	625	35.5	15.2	00.0	02.2	08.3	61.2	38.8
Regulate	381	54.3	08.9	00.0	00.0	01.0	64.2	35.8	482	10.2	10.2	00.2	00.4	00.4	21.4	78.6
Mediate	418	54.6	12.7	00.0	00.0	02.6	69.8	30.1	433	04.2	01.6	00.0	00.0	00.9	06.7	93.3
Result	329	06.1	02.7	00.6	00.9	02.1	12.5	87.5	396	03.3	01.3	00.0	01.0	00.0	05.6	94.4
Associate	39	41.0	00.0	00.0	05.1	05.1	51.2	48.8	614	16.6	05.0	00.0	00.5	02.0	24.1	75.9
Signal	154	20.1	00.0	00.0	00.7	00.0	20.8	79.2	148	11.5	00.0	00.7	00.0	01.4	13.6	86.4
Lead	241	06.6	00.4	00.0	01.2	00.8	09.1	90.9	296	02.4	00.7	00.0	00.0	00.3	03.4	96.6
Block	209	28.7	02.4	00.0	00.5	00.5	32.1	67.9	234	11.6	01.7	00.9	00.0	01.3	15.4	84.6
Encode	228	03.5	47.8	04.8	00.4	00.4	56.9	43.1	234	66.7	00.9	00.9	00.0	00.0	68.5	31.5
Decrease	75	20.0	04.0	01.3	00.0	00.0	25.3	74.7	115	10.4	00.9	00.0	00.0	00.9	12.2	87.8
Suppress	101	20.8	02.0	00.0	02.0	01.0	25.8	74.2	158	07.6	01.3	01.3	00.0	00.7	10.9	89.1
Interact	133	53.3	09.0	00.0	01.5	00.7	64.6	35.3	145	52.4	27.5	00.0	00.0	00.6	80.6	19.3
Recognize	113	53.1	00.0	00.0	07.0	14.2	74.2	25.7	94	25.5	25.5	00.0	00.0	00.0	51.0	49.0
Generate	69	08.7	04.4	00.0	07.3	07.3	27.7	72.3	103	20.3	05.8	02.9	08.7	04.9	42.6	57.4
Modulate	66	28.8	01.5	00.0	00.0	07.6	37.9	62.1	109	05.5	00.0	00.0	00.0	00.0	05.5	94.5
Alter	53	28.3	01.9	00.0	00.0	01.9	32.1	67.9	93	10.8	00.0	00.0	00.0	01.1	11.9	88.1

4.2 Machine learning process

The machine learning process is the last process in this methodology. Not only does this process involve applying a learning algorithm to the training data, but also is responsible for feature engineering (i.e., feature selection and feature design).

Firstly, the prepared training data such as shown in Figure 4-10 are formed into 2 sets of training models. The Model 1 is composed of the training data with a set of features not related to PAS information (i.e., from column 1 to column 6). On the other hand, the Model 2 is composed of the training data with a set of features including PAS-related features (i.e., features from column 7 to column 10 added to features from column 1 to column 6). To compare learning results from these two models would help to test if the intuition (i.e., semantic relations between predicate and its arguments are salient to improve NER) is on the right track. The SVMs learning algorithm is used in this work.

However, the representation of the PAS-related knowledge within 4 features (columns 7-10) as basically given to the word tokens within the boundaries of surface subject and surface object recognized through the process in section 4.1.3 would not be enough for being able to clearly see the impact of semantic knowledge within PAS to the NER. Thus, some extra features in accordance with the subject and object boundaries (e.g., a feature representing the knowledge of transitive and intransitive sense, a feature representing syntactic path from the subject head or the object head to the target verb) are derived and added to the Model 2, leading to 4 more models (the Model 3, 4, 5 and 6) to be explored. The theoretical thought underlying the derivation of these additional features are explained in section 4.2.3.

4.2.1 Support Vector Machines (SVMs)

SVM classification method is known as the stronger learning method in comparison with decision tree learning and other statistical learning methods (Vapnik, 1995; Sholkopf et al., 1997; Vapnik 1998). In computational linguistics domain, SVMs have achieved highest performance in various shared tasks. For instance, the method using SVMs for automatic semantic role labeling of Hacioglu and colleagues (Hacioglu et al., 2004) succeed to be the best model in the CoNLL-2004 shared task [50] or the success of the Kudo

[50] http://www.lsi.upc.edu/~srlconll/st04/st04.html

and Matsumoto's text chunking method (Kudo and Matsumoto, 2000) in the CoNLL-2000 shared task [51]. Also, SVM classifier has been widely recognized to be practical for the molecular biology NER task (Kazama et al., 2002; Lee et al., 2003; Takeuchi and Collier, 2003; Yamamoto et al., 2003).

The main idea of SVMs is to construct a hyperplane to separate the two classes with a maximum margin which is the distance between two hyperplanes. Suppose N training examples (x_i, y_i), where $(1 \leq i \leq N)$ and x_i is a feature vector, y_i is the class label $\{-1, +1\}$ of x_i are given, SVMs find a hyperplane $w.x + b = 0$ which correctly separates the training examples and has a maximum margin by applying the equation $f(x) = sign(\sum_{i=1}^{m} w_i K(x, z_i) + b)$ where $f(x) = +1$ means x is a member of a certain class and $f(x) = -1$ means x is not a member, m is the number of support vectors and $K(x, z_i)$ is a kernel function. According to Takeuchi and Collier's work, the polynomial function degree 2 has been shown to be the best kernel for their NER system (Takeuchi and Collier, 2002), thus the kernel function used in the experiments is $k(x) = (1+x)^2$.

The TinySVM package used in this work is implemented by NAIST Computational Linguistic Laboratory [52]. This package allows setting parameters F and T for specifying the context boundary to be used as features. The parameter F and T are used to specify the window in the form "F:[beginning position of token]..[end position of token]:[beginning position of column]..[end position of column] [53]" and "T:[beginning position of defined class]..[end position of defined class]", respectively. In this work, the context window for all experiments is set as "F:-1..1:0.." and "T:-2..-1". This setting of context window is shown to be the best setting from the preliminary experiment that have been conducted. Figure 4-12 shows the boundary of the context windows used in this work (the *blue* square is the position being estimated class by SVMs, the *green* area represents the context fixed by parameter F and the *light blue* area represents the context fixed by parameter T.

---

[51] http://www.cnts.ua.ac.be/conll2000/chunking/

[52] http://chasen.org/~taku/software/TinySVM/.

[53] In case of parameter F, if [end position of column] is omitted, the last column is set as [end position of column].

Column 1	2	3	4	5	6	7	8	9	10	11
...										
IL-12	il-12	IL-12	NNP	cdh	B-NP	O	O	O	O	B-protein
stimulates	stimulate	-	VBZ	lw	B-VP	O	O	O	O	O
Formation	formation	form-ation	NN	lw	B-NP	O	O	O	O	O
Of	of	-	IN	lw	B-PP	O	O	O	O	O
A	a	-	DT	lw	B-NP	O	O	O	O	O
DNA-binding	dna-binding	com-plex	VBG	2c	I-NP	recog-nizes	recog-nize	ACT	SSUBJ	B-protein
Complex	complex	com-plex	JJ	lw	I-NP	recog-nizes	recog-nize	ACT	SSUBJ	
That	that	that	DT	lw	B-SBAR	O	O	O	O	
recognizes	recognize	-	VBZ	lw	B-VP	O	O	O	O	
A	A	-	DT	lw	B-NP	O	O	O	O	
DNA	dna	se-quence	NNP	2c	I-NP	recog-nizes	recog-nize	ACT	SOBJ	
Sequence	sequence	se-quence	NN	lw	I-NP	recog nizes	recog-nize	ACT	SOBJ	
...										

Figure 4-12: Context windows from setting "F:-1..1:0.. T:-2..-1"

Due to Tiny SVM is a binary classifier, another aspect need to be decided for multi-class classification task of NER is about the strategy for combining several binary classifiers. In this work, all experiments with Tiny SVM use the strategy of one-against-one (also called pairwise) rather than one-against-the rest. The one-against-one strategy will construct $K(K-1)/2$ binary SVMs where K denotes the number of the target classes. Each binary SVM has one vote for its answer class after learning the sample. The final class to be answered from the combination of these several binary classifiers is the class with the maximum votes.

4.2.2 Lexical-based model and PAS-based model

As stated before, the Model 1 and Model 2 are constituted from the 13 features of the already prepared training data from the pre-process (section 4.1.1) to the encoding process (section 4.1.5). The Model 1 is considered to be a base model to evaluate the overall effects of PAS-related features to the performance of NER, whereas the Model 2 is considered to be a base model for comparing the contribution of various forming of PAS-related features to NER. Therefore, the Model 1 will be called lexical-based model and the Model 2 will be called PAS-based model henceforth.

- **Model 1**: Lexical-based model

  A set of features used in this lexical-based model are state-of-the-art features for the NER task. The features such as *surface word*, *lemma form*, and *orthographic feature* are quite cheap to be obtained because they are the information derivable from the lexical appearance of a word by using just linguistic knowledge (e.g., morphological analysis). The domain knowledge is not necessary for extracting these features. A little bit more expensive features are *part-of-speech*, *phrase-chunk*, and *head word of NP-chunk*. These features rely on the capability of parser to investigate the syntactic functions of a word in a sentence.

  As discussed in chapter 2, these cheap features work well for NER in general or newswire domain (more than 90% for the best performance). However, the state-of-the-art performance for NER in molecular biology domain (the performance is less than 80%) declares the need of features representing deeper information than the lexical-based features do. The first effort to employ deeper knowledge for NER is expresses in a set of features added in the Model 2, explained as follows.

- **Model 2**: PAS-based model

  The Model 2 contains all lexical-based features used in Model 1, with an additional set of features representing semantic relations between a predicate and its arguments (also known as arguments' semantic roles). However, these 4 additional features (i.e., *predicate surface form*, *predicate lemma*, *voice* and *surface syntactic role*) can be correctly determined an argument's semantic role (*agent* or *theme*) for merely simple cases. If both *surface subject* and *surface object* co-occur with a target verb, it can surely determine that the argument functioning as *subject* plays the role of *agent* and the argument functioning as *object* plays role of *theme* when the predicate is used in active voice, and vice versa in passive voice. The correct determination of semantic role would lead to the correct named entity classification, underlying the hypothesis that semantic relationships in PAS (arguments' semantic roles) for each predicate confine classes of named entities participating in the event indicated by the predicate. However, as the arguments with the same semantic role possibly belong to

different named entity classes, the lexical-based features and PAS semantic based-features are required altogether to solve this ambiguity.

In the next section, several kinds of syntactic information are derived to be features supporting the original PAS-based features with the aim to diminish the ambiguity in determining semantic roles.

4.2.3 Additional predicate-argument related features

As stated in the previous section, the more semantic role of each argument can be correctly interpreted, the more semantic knowledge in PAS can take effect to improve NER systems. This claim will be confirmed by the experimental results provided in section 4.3.

It is worth to note that the problem placing underneath the problem of employing PAS-knowledge for enhancing NER task is to find the choice of syntactic features to support identification of semantic structure.

- **Model 3**: Path

  Path feature represents the syntactic path from the subject argument to the related predicate and from the related predicate to the object argument. The partial parser based on dependency grammar is used in our work, so the path is derived from the flat structure of dependency tree. For example, the path between the subject constituent and the predicate is "NP_VP_ADVP_VP" and the path between the object constituent and the predicate is "VP_PP_NP" for the sentence "[Increased cytokine secretion]$_{NP}$ [was]$_{VP}$ [specifically]$_{ADVP}$ [**inhibited**]$_{VP}$ [by]$_{PP}$ [G1]$_{NP}$". As can be seen, the information that in a sentence the subject posits before the predicate but the object posits after the predicate is also embedded in the representation of path as shown.

- **Model 4**: A pair composed of the subject and object's heads

  A pair of subject and object's head feature is designed following the intuition that a named entity class of an agent should restrict a possible type of a named entity playing role as theme and vice versa. Moreover, the head words of subject and object are used as these heads may be considered as subtypes of any named

entity classes. The using of a pair of lemma forms of subject-object head words would help to reduce data sparseness problem compared to the using of a pair of surface forms of subject-object head words.

- **Model 5**: Transitive and intransitive sense

  A column is added to be a feature representing if a predicate is used in transitive or intransitive sense. For each surface subject's constituent, a tag "fobj" is set if the surface object is found in the current clause. A tag "O" is set if the surface object is not found. However, this feature helps just in part to correctly determine transitive or intransitive sense implicit in the usage of a predicate as the object argument can be omit in a clause although a predicate is used in transitive sense. For instance, the predicate "eat" is used in transitive sense without mentioning any objects in the sentence "Yesterday, John ate at ABC restaurant".

- **Model 6**: Joining of a subject-object's head pair and transitive-intransitive feature

  A pair of subject-object head is used to be assigned to a column of transitive-intransitive feature instead of "fobj" when the object is found in the corresponding clause.

4.2.4 Assessment

In this work, all prepared data set is divided into 9 parts of training set and 1 part of testing set due to 10-fold cross validation is using. The testing set has the named entity class hidden from the learner model and results from the learner model are then compared against the correct class to determine F1 scores (van Rijsbergen, 1979). The F1 score is calculated from the equation $F1 = (2PR)/(P+R)$, where P denotes Precision and R Recall. P is the ratio of the number of correctly found named entity chunks to the number of found named entity chunks, and R is the ratio of the number of correctly found named entity chunks to the number of true named entity chunks.

4.3 Experimental results and discussion

Firstly, all models of different feature sets described in section 4.2.3 are applied to 6 predicates. As shown in Table 4-1, these predicates include *encode* and *recognize*

(predicates from group 1), *block* and *lead* (predicates from group 2), as well as *regulate* and *associate* (predicates from group 3). The F1-scores resulted from this experiment are shown in Table 4-2.

In each record, the F1-score of a corresponding predicate is given for Model 1 (Lexical-based model), Model 2 (PAS-related model), Model 3 (the Model 2 with added Path feature), Model 4 (the Model 2 with added Pair of subject and object's heads feature), Model 5 (the Model 2 with added Transitive/Intransitive feature) and the Model 6 (the Model 4 is embodied in Model 5). Compared to the F1-score in Model 1, the higher F1-scores from any other models are shown in *bold* number. Moreover, if the F1 scores in any models among Models 3-6 are higher than in Model 2, the scores will be highlighted with a *gray* background. The number of examples for each predicate is given in a bracket next to a predicate's name.

Table 4-2: F1-scores of representative predicates trained with features in Models 1-6

Predicate \ Model		**M 1** *Lexical-based*	**M 2** *PAS-based*	**M 3** *Path*	**M 4** *Pair of Head*	**M 5** *Trans/ Intrans*	**M 6** *M4+M5*
Group 1: *both high*	Encode (265)	56.60	**57.56**	**58.38**	**57.16**	**57.69**	**57.64**
	Recognize (121)	47.24	**49.39**	**48.47**	**49.54**	**49.16**	**49.39**
Group 2: *both low*	Block (270)	51.19	**51.47**	**52.23**	**51.85**	**52.02**	**51.95**
	Lead (288)	57.01	**57.40**	56.70	**57.12**	**57.53**	**57.49**
Group 3: *high/ low*	Regulate (525)	61.87	60.48	60.13	60.72	60.01	60.37
	Associate (377)	52.09	51.48	51.29	50.43	51.40	50.97

As can be observed from Table 4-2, the simple representation of PAS related knowledge such as in Model 2 improves the performance for all predicates except the predicates *regulate* and *associate* which only have either argument *agent* or *theme* with a higher possibility to belong to a named entity class than non-named entity class, compared to lexical-based features (Model 1). On the other hand, predicates in groups 3 do not show any improvement in any models using PAS-related features (Models 3-6). Thus, in the following, only predicates in group 1 and group 2 will be discussed in terms of the effectiveness of each type of the extra PAS-related features used in Models 3-6,

compared to the model using PAS-based feature set (Model 2) which influences to the improvement in performance for all predicates of group 1 and 2. The reason of the performance degradation of the predicates in group 3 will be discussed in section 4.4.

Surface Word	Phrase-chunk	Syntactic roles	Path pattern between SSUBJ or SOBJ and the predicate	
the	B-NP	O	O	
proteins	I-NP	SSUBJ	NP_VP	
encoded	B-VP	O	O	<- target predicate "*encode*"
by	B-ADVP	O	O	
these	B-NP	O	O	
two	I-NP	O	O	
latter	I-NP	SOBJ	VP_ADVP_NP	
genes	I-NP	SOBJ	VP_ADVP_NP	
is	B-VP	O	O	
approximately	B-ADVP	O	O	
65%	B-NP	O	O	

Figure 4-13: Examples of simple Path patterns between arguments and the predicates found in the data set of *encode*

With regard to Path feature (used in Model 3), the performance is improved from Model 2 for only the model training on data set of predicates *encode* and *block*. Empirically, one reason for this should be that the surface subject and surface object of these two predicates are located close to the predicate in most cases. For example, as shown in Figure 4-13, the path patterns between arguments and the predicate *encode* of "...[protein]$_{NP}$ [encoded]$_{VP}$ [by]$_{ADVP}$ [these two latter genes]$_{NP}$..." are "NP_VP" for the subject argument and "VP_ADVP_NP" for the object argument. Owing to the short path patterns, the path patterns can be generalized throughout the data sets. On the contrary, long path patterns are mostly found in the examples for other 2 predicates such as *recognize* and *lead*. For example, as shown in Figure 4-14, from the sentence "[Control peptides]$_{NP}$ [corresponding]$_{VP}$ [to]$_{ADVP}$ [the normal pml]$_{NP}$ [and]$_{O}$ [RAR alpha proteins]$_{NP}$ [were]$_{VP}$ [not]$_{ADVP}$ [recognized]$_{VP}$.", the path from the subject argument "Control peptides" to the predicate *recognize* is "NP_VP_ADVP_NP_O_NP_VP_ADVP_VP". This long path pattern would cause data-sparseness problems for the path feature. Similarly, the long path pattern of the predicate *lead* is given an example in Figure 4-15.

Surface Word	Phrase-chunk	Syntactic roles	Path between SSUBJ or SOBJ and the predicate	
Control	B-NP	SSUBJ	NP_VP_ADVP_NP_O_NP_VP_ADVP_VP	
peptides	I-NP	SSUBJ	NP_VP_ADVP_NP_O_NP_VP_ADVP_VP	
corresponding	B-VP	O	O	
to	B-ADVP	O	O	
the	B-NP	O	O	
normal	I-NP	O	O	
pml	I-NP	O	O	
and	O	O	O	
RAR	B-NP	O	O	
alpha	I-NP	O	O	
proteins	I-NP	O	O	
were	B-VP	O	O	
not	B-ADVP	O	O	
recognized	B-VP	O	O	<- target predicate "*recognize*"
.	O	O	O	

Figure 4-14: An example showing the long Path patterns between arguments and the predicates found in the data set of *recognize*. The subject argument is always followed by some modification before reaching its predicate.

The next feature, the Head-Pair feature, which aims to use the named entity type of subject as a restriction for the named entity type of object and vice versa, does not show its usefulness for predicate *encode* and *lead*. In case of *lead*, its arguments (i.e., both *agent* and *theme*) are prone to be non-named entity rather than to belong to a named entity class (shown in Table 4-1), so the pair of the head words of its arguments can have many variants such as shown in Figure 4-16. This causes this feature to be ineffective in constraining a named entity functioning as a subject with named entity functioning as an object and vice versa. In case of the predicate *encode*, although both its arguments are prone to belong to named entity classes rather than to be non-named entity, the Head-Pair feature is not helpful for the predicate *encode*. As the predicate *encode* used in the molecular biology domain has a specific semantic to describe relationships between genes and gene products, the head pair of arguments for this predicate is mostly found as *gene_protein*. Therefore, this feature contains too general information to be helpful for *encode*. One of the sentences contain *gene_protein* is shown in Figure 4-13.

Surface Word	Phrase-chunk	Syntactic roles	Path between SSUBJ or SOBJ and the predicate
The	B-NP	O	O
elevation	I-NP	SSUBJ	NP_O_NP_SBAR_VP_ADVP_NP_O_NP_VP
in	O	O	O
intracellular	B-NP	O	O
calcium	I-NP	O	O
that	B-SBAR	O	O
is	B-VP	O	O
induced	I-VP	O	O
by	B-ADVP	O	O
interactions	B-NP	O	O
at	O	O	O
the	B-NP	O	O
antigen	I-NP	O	O
receptor	I-NP	O	O
leads	B-VP	O	O <- target predicate "*lead*"
to	B-ADVP	O	O
the	B-NP	O	O
activation	I-NP	PCOMP	VP_ADVP_NP
of	B-PP	O	O
the	B-NP	O	O
calcium-dependent	I-NP	O	O
phosphatase	I-NP	O	O
calcineurin	I-NP	O	O
.	O	O	O

Figure 4-15: An example showing the long Path patterns between arguments and the predicates found in the data set of *lead*. The subject argument is always followed by some modification before reaching its predicate.

Surface Word	Phrase-chunk	Syntactic roles	Head Pair of arguments
The	B-NP	O	O
generation	I-NP	SSUBJ	generation_downstream
of	B-PP	O	O
second	B-NP	O	O
messengers	I-NP	O	O
In	O	O	O
T	B-NP	O	O
cells	I-NP	O	O
normally	B-ADVP	O	O
leads	B-VP	O	O <- target predicate "*lead*"
to	B-ADVP	O	O
downstream	B-NP	PCOMP	generation_downstream

signaling	B-VP	O	O	
that	B-NP	O	O	
results	B-VP	O	O	
In	B-ADVP	O	O	
transcriptional	B-NP	O	O	
activation	I-NP	O	O	
of	B-PP	O	O	
the	B-NP	O	O	
IL-2	I-NP	O	O	
gene	I-NP	O	O	
.	O	O	O	
The	B-NP	O	O	
consequences	**I-NP**	**SSUBJ**	**consequence_failure**	
of	B-PP	O	O	
EBV	B-NP	O	O	
infection	I-NP	O	O	
of	B-PP	O	O	
T	B-NP	O	O	
cells	I-NP	O	O	
at	O	O	O	
an	B-NP	O	O	
early	I-NP	O	O	
stage	I-NP	O	O	
Of	B-PP	O	O	
differentiation	B-NP	O	O	
may	B-VP	O	O	
lead	**I-VP**	**O**	**O**	**<- target predicate "*lead*"**
to	B-ADVP	O	O	
failure	**B-NP**	**PCOMP**	**consequence_failure**	
of	B-PP	O	O	
normal	B-NP	O	O	
T	I-NP	O	O	
Cell	I-NP	O	O	
repertoire	I-NP	O	O	
development	I-NP	O	O	
,	O	O	O	
autoimmunity	B-NP	O	O	
,	O	O	O	
Or	O	O	O	
malignancy	B-NP	O	O	
.	O	O	O	

Figure 4-16: Examples of the Subject-Object Head Pair for *lead*. The head pairs are *generation_downstream* and *consequence_failure* for sentence 1 and 2

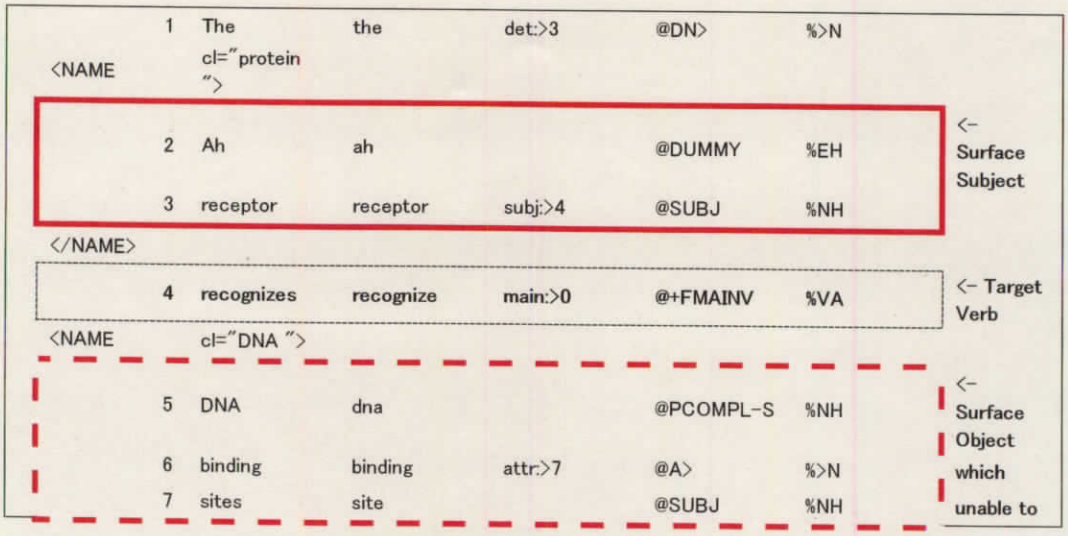
In case of Transitive/Intransitive feature, it is surprised that for some predicates this feature is not useful in improving the performances though it is important in correctly interpreting the semantic role of an argument. From Figure 4-17, the subject argument "John" has the semantic role of *agent* in sentence (1) but the subject argument "the window" has the semantic role of *theme* in sentence (2). These two sentences illustrate that to know only the syntactic function as a subject or object cannot have a correct determination on the semantic role. To give information stating if the object is found in a sentence or not would therefore help to some extent to imply the sense in which the predicate is used. The performance of the model having this feature (Model 5) should outperform the PAS-based model (Model2).

(1) [John]$_{agent}$ **broke** [the window]$_{theme}$. ***<- Transitive sense***

(2) [The window]$_{theme}$ **broke.** ***<- Intransitive sense***

Figure 4-17: Sentences show the use of the predicate *broke* in the transitive sense (sentence 1) and in the intransitive sense (sentence 2)

However, the performance for *recognize* decreased when this feature was applied. From the result analysis, the parsing error accounts for this unexpected result. Some linkages between words are lost as shown in Figure 4-18.



</NAME>							capture
	8	for	for	mod:>7	@<NOM	%N<	as no
	9	the	the	det:>13	@DN>	%>N	relation
<NAME		cl="protein ">					to the
	10	B	B	attr:>11	@A>	%>N	predicate
	11	cell	cell	attr:>12	@A>	%>N	is found
	12	transcription	transcription	attr:>13	@A>	%>N	
	13	factor	factor	pcomp:>8	@<P	%NH	
</NAME>							
	14	,	,				
<NAME		cl="protein ">					
	15	BSAP	bsap		@SUBJ	%NH	
</NAME>							
	16	:	:				
	17	a	a	det:>19	@DN>	%>N	
	18	possible	possible	attr:>19	@A>	%>N	
	19	mechanism	mechanism	mod:>15	@NH	%NH	
	20	for	for	mod:>19	@<NOM	%N<	
	21	dioxin-mediated	dioxin-mediated	attr:>22	@A>	%>N	
	22	alteration	alteration	pcomp:>20	@<P	%NH	
	23	of	of	mod:>22	@<NOM-OF	%N<	
	24	CD19	cd19	attr:>25	@A>	%>N	
	25	gene	gene	attr:>26	@A>	%>N	
	26	expression	expression	pcomp:>23	@<P	%NH	
	27	in	in	mod:>26	@<NOM	%N<	
<NAME		cl="cell-type">					
	28	human	human	attr:>29	@A>	%>N	
	29	B	b	attr:>30	@A>	%>N	
	30	lymphocytes	lymphocyte	pcomp:>27	@<P	%NH	
</NAME>							
	31	.	.				

Figure 4-18: Incomplete parsing results for the predicate *recognize*

As the linkage from token *sites* (the word number 7) to the predicate is lost, then a constituent for surface object of *recognize* cannot be captured. This causes a subsequent problem for the Transitive/Intransitive feature; i.e., this feature is set to "O" to represent that the predicate *recognize* is used in the transitive sense, whereas it does not. Thus, this incomplete parsing result accounts for decreasing F1-score of *recognize* when using the Transitive/Intransitive feature (Model 5) compared to when not using it (Model 2).

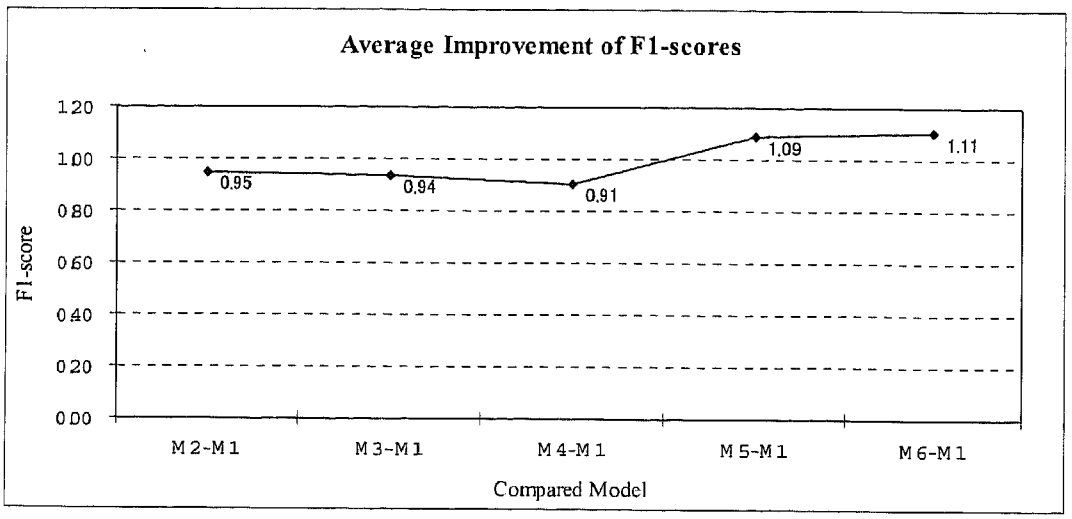


Figure 4-19: Average improvement of F1-scores for each of PAS-related models (Models 2-6) compared to the lexical-based model (Model 1)

A graph of the average improvement of the F1-score from each PAS-related model compared to the lexical-based model is shown in Figure 4-19. All predicates in group 3 did not show the performance improvement when a feature set of every PAS-related model was applied to. So, the average F1-score of each model is calculated from the performances resulted from the applying of the model to the predicates in only group 1 and group 2. In the case that the mix model (Model 6) is not considered, the results show that the Transitive/Intransitive feature (Model 5) gives the highest contribution as expected. Some more performance improvement can be obtained in Model 6 where the Head-Pair feature (Model 4) is embedded in the Transitive/Intransitive feature (Model 5). In this, the average F1-score rises up from 1.09 to 1.11. Prior to concluding that the Model 6 is the best model in this work, the other ways of joining features from different models have been explored. However, their performances get worse compared to Model 6 and to the use of each model separately.

In order to affirm the effectiveness of PAS-related features to NER, the Model 1, 2, and 6 are applied to the rest of selected predicates. The F1-scores of all 19 predicates are summarized in Table 4-3. The predicates that have either F1-score in Model 2 or in Model 6 higher than the F1-score resulted from Model 1 are marked in *bold-faced* characters. A group of each predicate corresponding to the proportion of its arguments to belong to named entity classes (as explained in section 4.1.6) is also given in this table, as

well as a comparison of the F1-score from Model 2 and from Model 6 with the F1-score from Model 1 for every predicate.

Table 4-3: F1-scores of all 19 predicates trained with features in the Model 1, 2 and 6

Group	Model / Predicate	Size of Examples	Model 1 *(Lexical-based)*	Model 2 *(PAS-based)*	Model 6 *(Model 4 + Model 5)*	M2-M1	M6-M1
1	**Recognize**	**121**	**47.24**	**49.39**	**49.39**	**2.15**	**2.15**
	Encode	**265**	**56.60**	**57.56**	**57.64**	**0.96**	**1.04**
	Interact	**165**	**56.46**	**57.02**	**57.46**	**0.56**	**1.00**
	Bind	**825**	**62.74**	**62.76**	**62.78**	**0.02**	**0.04**
	Express	779	60.43	59.64	59.71	-0.79	-0.72
	Activate	719	66.18	65.23	65.12	-0.95	-1.06
2	**Result**	**407**	**55.04**	**55.76**	**56.12**	**0.72**	**1.08**
	Block	**270**	**51.19**	**51.47**	**51.95**	**0.28**	**0.76**
	Inhibit	**779**	**60.70**	**61.80**	**61.23**	**1.10**	**0.53**
	Lead	**288**	**57.01**	**57.40**	**57.49**	**0.39**	**0.48**
	Signal	**291**	**54.68**	**54.86**	**54.74**	**0.18**	**0.06**
	Suppress	170	51.32	50.82	50.76	-0.50	-0.56
	Modulate	118	44.68	44.67	43.99	-0.01	-0.69
	Alter	102	50.12	48.91	48.91	-1.21	-1.21
	Generate	121	40.46	37.90	39.25	-2.56	-1.21
	Decrease	171	48.75	47.87	47.51	-0.88	-1.24
3	Mediate	500	60.22	60.01	60.18	-0.21	-0.04
	Associate	377	52.09	51.48	50.97	-0.61	-1.12
	Regulate	525	61.87	60.48	60.37	-1.39	-1.50

From empirical results in Table 4-3, what can be concluded is that by applying the PAS-related features in addition to the lexical-based features, not only the performance of NER system was improved in case of the predicates in group 1, but also it was improved in case of the predicates in group 2. As the predicates in group 1 have both arguments *agent* and *theme* with a higher probability of belonging to a named entity class than non-named entity class, the application of PAS-related features for the predicates in this group

was expected to allow NER system to have higher performance than what it got from lexical-based features. The predicates such as *recognize*, *encode*, *interact* and *bind* conformed to this expectation, but not *express* and *activate*. From the analysis through the experimental results, one reason of the performance degradation of the predicates *express* and *activate* was from the incomplete parsing results. The boundaries of *agent* and *theme* arguments can not be identified correctly if the incomplete parsing results are conditioned. Without the impact from incorrect boundary identification, all predicates in group 1 have shown the positive effect of using PAS-related features. This claim is affirmed in further experiments explained in detail in section 4.4.

As described in section 4.1.6, the property of predicates in group 2 is opposite to that of predicates in group 1. However, the effectiveness of using PAS-related features for the predicates in this group is not contrast to group 1. Some predicates such as *inhibit*, *result*, *signal*, *lead* and *block* show the performance improvement for using PAS-related features. On the contrary, there are also some predicates showing negative effects such as *decrease*, *suppress*, *generate*, *modulate* and *alter*. This phenomenon can be explained as that the lower probability of belonging to a named entity class of both arguments *agent* and *theme* of predicates in group 2 helps the machine learning model to identify the arguments as non-named entity. An example of the predicate *inhibit* is shown in Figure 4-20. In this figure, the constituent "B95-8 cytosol" seems lexically a kind of molecular entity as it starts with capital letter, number, hyphen and so on, but it is correctly classified not to be named entity as it is a surface subject of *inhibit*.

However, some predicates in this group show the negative effects. Their small numbers of examples are the reason for inability of machine learning model to take advantage of PAS-related features. A predicate showing positive effect such as *inhibit*, *result*, *signal*, *lead* and *block* has the number of training example as 779, 407, 291, 288 and 270, respectively; whereas a negative effect predicate such as *decrease*, *suppress*, *generate*, *modulate* and *alter* has the number of training examples as 171, 170, 121, 118, and 102, respectively. This statistical information affirms previous explanation.

Surf.	Lem.	Head	POS	Orth.	Phr.	PSurf	PLem	Voice	SynR	Path	H_NE	M_NE
B95-8	b95-8	Cytosol		cdh	B-NP	Inhibited	inhibit	ACT	SSUBJ	cytosol_binding	O	O
cytosol	cytosol	Cytosol		lw	I-NP	Inhibited	inhibit	ACT	SSUBJ	cytosol_binding	O	O
inhibited	**inhibit**	**-**		**lw**	**B-VP**	**O**	**O**	**O**	**O**	**O**	**O**	**O**
specific	specific	Binding		lw	B-NP	Inhibited	inhibit	ACT	SOBJ	O	O	O
binding	binding	Binding		lw	I-NP	Inhibited	inhibit	ACT	SOBJ	O	O	O
of	of	-		lw	B-PP	O	O	O	O	O	O	O
[	[	-		pu	O	O	O	O	O	O	O	O
3H	3h	3H		cd	B-NP	O	O	O	O	O	O	O
]	]	-		pu	O	O	O	O	O	O	O	O
dexamethasone	dexamethasone	dexamethasone		lw	B-NP	O	O	O	O	O	O	O
(	(	-		pu	B-NP	O	O	O	O	O	O	O
P	p	P		sc	I-NP	O	O	O	O	O	O	O
<	<	-		ot	O	O	O	O	O	O	O	O
0.01	0.01	0.01		ot	B-NP	O	O	O	O	O	O	O
)	)	-		pu	O	O	O	O	O	O	O	O
when	when	-		lw	B-ADVP	O	O	O	O	O	O	O
mixed	mix	-		lw	B-VP	O	O	O	O	O	O	O
with	with	-		lw	B-ADVP	O	O	O	O	O	O	O
cytosol	cytosol	Cytosol		lw	B-NP	O	O	O	O	O	O	O
prepared	prepare	-		lw	B-VP	O	O	O	O	O	O	O
from	from	-		lw	B-ADVP	O	O	O	O	O	O	O
either	either	-		lw	I-ADVP	O	O	O	O	O	O	O
a	a	-		lw	B-NP	O	O	O	O	O	O	O
human	human	Line		lw	I-NP	O	O	O	O	O	B-cell_line	B-cell_line
lymphoid	lymphoid	Line		lw	I-NP	O	O	O	O	O	I-cell_line	I-cell_line
cell	cell	Line		lw	I-NP	O	O	O	O	O	I-cell_line	I-cell_line
line	line	Line		lw	I-NP	O	O	O	O	O	I-cell_li	I-cell_li

										ne	ne
(	(	-	pu	O	O	O	O	O	O	O	O
HL	hl	HL	2c	B-NP	O	O	O	O	O	B-cell_line	O
)	)	-	pu	O	O	O	O	O	O	O	O
Or	or	-	lw	O	O	O	O	O	O	O	O
Rat	rat	Thymus	lw	B-NP	O	O	O	O	O	O	O
thymus	thymus	Thymus	lw	I-NP	O	O	O	O	O	O	O
.	.	-	pu	O	O	O	O	O	O	O	O

Figure 4-20: An example of the classification results from SVMs using PAS-related features for the predicate *inhibit*. The result from SVMs is shown in *blue*, the boundary of surface subject is shown in *pink*, and the surface object is shown in *yellow*[54]

In case of group 3, the performance degradation is found for all predicates in this group (i.e., *mediate*, *associate* and *regulate*) when the PAS-related features are employed. As predicates in this group have only argument *agent* or *theme* with a higher probability of belonging to a named entity class than non-named entity, these predicates lacks of the restriction between named entity types of arguments *agent* and *theme*. This seems to be the reason of performance degradation when applying PAS-related features to the predicates in group 3. The machine learning model is likely to fail in identifying the arguments at the surface subject and surface object boundaries to the correct named entity class. An example of the predicate *associate* is shown in Figure 4-21. The statistical evidence observed from GENIA corpus expresses that only the *agent* argument of the predicate *associate* is prone to be a kind of named entity. As there is no restriction pattern as *DNA-associate-DNA*, as well as lexical information in the following sentence is not an obvious guiding for being *DNA*, thus SVMs classify "transcriptional regulatory element" and "nuclease-hypersensitive site" as non-named entities.

[54] Column names: Surf=Surface word, Lem=Lemma form, Head=Head word of NP-chunk, POS=Part-of-speech, Orth=Orthographic feature, Phr=Phrase-chunk, PSurf=Predicate surface form, PLem=Predicate Lemma, Voice=voice, SynR=Surface syntactic role, Head=Pair of subject and object's heads, H_NE=Human annotated class of named entity and M_NE=Machine annotated class of named entity.

Surf	Lem	Head	POS	Orth	Phr	PSurf	PLem	Voice	SynR	Head	H_NE	M_NE
A	A	-		sc	B-NP	O	O	O	O	O	O	O
transcriptional	Transcriptional	element		lw	I-NP	Associated	associate	PAS	SSUBJ	element_site	B-DNA	O
regulatory	regulatory	element		lw	I-NP	Associated	associate	PAS	SSUBJ	element_site	I-DNA	O
element	element	element		lw	I-NP	Associated	associate	PAS	SSUBJ	element_site	I-DNA	O
is	be	-		lw	B-VP	O	O	O	O	O	O	O
associated	**associate**	**-**		**lw**	**I-VP**	**O**	**O**	**O**	**O**	**O**	**O**	**O**
with	with	-		lw	B-ADVP	O	O	O	O	O	O	O
a	A	-		lw	B-NP	O	O	O	O	O	O	O
nuclease-hypersensitive	nuclease-hypersensitive	site		ot	I-NP	Associated	associate	PAS	PCOMP	O	B-DNA	O
site	site	site		lw	I-NP	associated	associate	PAS	PCOMP	O	I-DNA	O
in	in	-		lw	O	O	O	O	O	O	O	O
the	the	-		lw	B-NP	O	O	O	O	O	O	O
pol	pol	gene		lw	I-NP	O	O	O	O	O	B-DNA	B-DNA
gene	gene	gene		lw	I-NP	O	O	O	O	O	I-DNA	I-DNA
of	of	-		lw	B-PP	O	O	O	O	O	O	O
human	human	virus		lw	B-NP	O	O	O	O	O	O	O
immunodeficiency	immunodeficiency	virus		lw	I-NP	O	O	O	O	O	O	O
virus	virus	virus		lw	I-NP	O	O	O	O	O	O	O
type	type	type		lw	I-NP	O	O	O	O	O	O	O
1	1	1		1ds	I-NP	O	O	O	O	O	O	O
.	.	-		pu	O	O	O	O	O	O	O	O

Figure 4-21: An example of the classification results from SVMs using PAS-related features for the predicate *associate*. The result from SVMs is shown in *blue*, the boundary of surface subject is shown in *pink*, and the surface object is shown in *yellow*

Summarily, the PAS-related features will help to improve NER using lexical-based features for a predicate conforming to the following criteria:

- a predicate in group 1 (arguments both *agent* and *theme* with a higher probability of belonging to a named entity class than non-named entity class)[55]
- a predicate in group 2 (arguments both *agent* and *theme* with a lower probability of belonging to a named entity class than non-named entity class) plus enough examples (with regard to the empirical evidence, at least 270 examples should be enough[56])

Moreover, PAS-related features can give more contribution to improve the lexical-based NER system when it is applied to the predicates in group 1 than in group 2.

4.4 Impediments to high performance improvement

In Table 4-3, the experimental results show that the PAS-related features only slightly improve NER. The highest performance improvement was only 2.15 resulted from applying the PAS-related features to the predicate *recognize*. However, this is not because the semantic relationship between a predicate and its argument is an insignificant knowledge for NER. According to the analysis on the experimental results, it was found that there were several factors impede the NER system in taking full benefits from the PAS-related features. These factors can be classified into 3 main groups: (1) boundary of arguments (2) semantic role representation and (3) named entities outside argument boundaries.

4.4.1 Boundary of arguments

This impediment factor involves the incorrect identification of an argument boundary. It is the consequence from the problems pertaining to the parsing results and the quantifiers.

4.4.1.1 Problem from the parsing result

There are 2 kinds of parsing errors which affect the NER system to incorrectly identify the argument boundary. The first error is the lost of linkage between words. By the nature of the sentences in bio-molecular literature, multiple ideas are conveyed in a single

[55] In section 4.4, the experiment to prove that the performance improvement from using PAS-related features is also obtainable in case of the predicates *express* and *activate*, which are 2 predicates in group 1 showing performance degradation in Table 4-3.

[56] This remark is just a preliminary conclusion. Further detailed investigation will be performed in future work.

sentence. Thus, sentences from the bio-molecular literature are rather the complex sentences than simple ones. The sentence "*HTLV-1 encodes an essential 40-kDa protein termed Tax that not only transactivates the long terminal repeat of this retrovirus but also induces an array of cellular genes*" shown in Figure 4-22 is an example of complex sentences which are often found in the bio-molecular literature. In this sentence, the parser failed to give syntactic relation "*object:>2*" to the word number 6 "*protein*" (i.e. the linkage between the argument at the surface object position and its predicate is lost). Thus, the argument shown in the *red-dotted* squares cannot be captured by the sub-structure recognizing process of the NER system used in this work.

	1	HTLV-I	htlv-i	subj:>2	@SUBJ	%NH	<MORPH> N NOM_SG </MORPH>	<- Surface Subject
	2	encodes	encode		@+FMAINV	%VA	<MORPH> V PRES_SG3 </MORPH>	<- Target Verb
	3	an	an		@DN>	%>N	<MORPH> DET SG </MORPH>	
	4	essential	essential	attr:>5	@A>	%>N	<MORPH> A ABS </MORPH>	<- Surface Object which unable to capture as no relation to the predicate is found
<NAME	cl="protein">							
	5	40-kDa	40-kda	attr:>6	@A>	%>N<?>	<MORPH> N NOM_SG </MORPH>	
	6	protein	protein		@SUBJ	%NH	<MORPH> N NOM_SG </MORPH>	
</NAME>								
	7	termed	term		@+FMAINV	%VA	<MORPH> V PAST </MORPH>	
<NAME	cl="protein">							
	8	Tax	tax		@NH	%NH	<MORPH> N NOM_SG </MORPH>	
</NAME>								
	9	that	that	subj:>12	@SUBJ	%NH<Rel>	<MORPH> PRON - </MORPH>	
	10	not	not	neg:>11	@ADVL	%EH	<MORPH> NEG-PART - </MORPH>	

11	only	only	meta:>12	@ADVL	%EH	<MORPH> ADV – </MORPH>
12	transactivates	transactivat*	mod:>8	@+FMAINV	%VA<?>	<MORPH> V PRES_SG3 </MORPH>
13	the	The	det:>16	@DN>	%>N	<MORPH> DET – </MORPH>
14	long	Long	attr:>15	@A>	%>N	<MORPH> A ABS </MORPH>
15	terminal	Terminal	attr:>16	@A>	%>N	<MORPH> A ABS </MORPH>
16	repeat	Repeat	obj:>12	@OBJ	%NH	<MORPH> N NOM_SG </MORPH>
17	of	Of	mod:>16	@<NOM-OF	%N<	<MORPH> PREP – </MORPH>
18	this	This	det:>19	@DN>	%>N	<MORPH> DET– </MORPH>
19	retrovirus	Retrovirus	pcomp:>17	@<P	%NH	<MORPH> N NOM_SG </MORPH>
20	but	But	cc:>12	@CC	%CC	<MORPH> CC – </MORPH>
21	also	Also	meta:>22	@ADVL	%EH	<MORPH> ADV – </MORPH>
22	induces	Induce	cc:>12	@+FMAINV	%VA	<MORPH> V PRES_SG3 </MORPH>
23	an	An	det:>24	@DN>	%>N	<MORPH> DET SG </MORPH>
24	array	Array	obj:>22	@OBJ	%NH	<MORPH> N NOM_SG </MORPH>
25	of	Of	mod:>24	@<NOM-OF	%N<	<MORPH> PREP – </MORPH>
<NAME	cl="DNA">					
26	cellular	Cellular	attr:>27	@A>	%>N	<MORPH> A ABS </MORPH>
27	genes	Gene	pcomp:>25	@<P	%NH	<MORPH> N NOM_PL </MORPH>
</NAME>						
28	.	.				

Figure 4-22: An example of incomplete parsing results for the predicate *encode*

Another parsing error is that the information is given, but it is incorrect. From Figure 4-23, the token "*bZIP*" highlighted in *pink* is suggested by the parser as a noun head (%NH), thus this token is not included in the same NP-chunk as the tokens "*transcriptional*" and "*activator*". The boundary of the subject argument of the predicate *bind* is incorrectly identified as shown in the *light-green* square.

	1	ZEBRA	zebra	subj:>2	@SUBJ	%NH	<MORPH> N NOM_SG </MORPH>	
</NAME>								
	2	is	be		@+FMAINV	%VA	<MORPH> V PRES_SG3 </MORPH>	
	3	a	a	det:>4	@DN>	%>N	<MORPH> DET SG </MORPH>	
<NAME	cl="protein">							
	4	bZIP	bzip	subj:>20	@SUBJ	%NH<?>	<MORPH> N NOM_SG </MORPH>	
	5	transcriptional	transcriptional	attr:>6	@A>	%>N	<MORPH> A ABS </MORPH>	<- Surface Subject
	6	activator	activator	mod:>4	@APP	%NH	<MORPH> N NOM_SG </MORPH>	
</NAME>								
	7	which	which	subj:>8	@SUBJ	%NH<Rel>	<MORPH> PRON WH_NOM </MORPH>	
	8	binds	bind	mod:>6	@+FMAINV	%VA	<MORPH> V PRES_SG3 </MORPH>	<-Target Verb
	9	as	as	copred:>8	@ADVL	%EH	<MORPH> PREP - </MORPH>	
	...	...	...	...	...	...	...	

Figure 4-23: An example of the incorrect parsing results for the predicate *bind*

These types of parsing error influence the system to incorrectly interpret semantic roles. As the system cannot capture the object argument of the sentence in Figure 4-22, the system would wrongly interpret that this sentence is intransitive sentence and the subject

constituent "*HTLV-I*" plays the role of *theme* of the predicate "*encodes*". In fact, the constituent that plays the role of *theme* is "*essential 40-kDa protein*".

To investigate the contribution of PAS-related features without the impact from parsing error, the manual identification of the boundaries of surface subject and surface object was done instead of using the information of syntactic relations resulted from the parser to identify these boundaries. The argument boundaries corresponding to the predicates such as *express*, *activate*, *decrease*, *associate* and *recognize* were identified manually on a set of 100 examples of each predicate, as well as a full set of examples (265 examples) of the predicate *encode*. The reasons why these predicates were selected to show the effects from the parsing error were as follows.

The predicates *express* and *activate* were chosen to confirm the claim in the previous section that these 2 predicates should not result in performance degradation in case of using PAS-related features. The predicate *decrease*, a representative of group 2, was selected to test if its performance degradation found in the previous section was still remain in non-parsing error condition. Similarly, the predicate *associate* was selected for group 3. With regard to *recognize* and *encode*, as they gave the two best performance improvements shown in the previous section, it is interesting to see the upper bound of the performance improvements if there is no impact from parsing errors. The F1-scores resulted from SVMs training on these data sets are shown in Table 4-4.

Table 4-4: F1-scores obtained from the training sets containing manually identified the surface subject and surface object boundaries

Model	Group	Number of Examples	**Model 1** *(Lexical-based)*	**Model 6** *(Model 4 + Model 5)*	**Performance Improvement (Manual)**
Express	1	100	49.83	50.82	0.99
Activate	1	100	58.65	59.27	0.62
Decrease	2	100	51.87	51.73	-0.14
Associate	3	100	43.28	39.91	-3.37
Recognize	1	100	46.31	52.43	6.12
Encode	1	100	56.89	59.29	2.40
Encode	1	265 (Full)	56.60	59.87	3.27

In Table 4-3, the using of PAS-related features in addition to lexical-based features for parsing-examples[57] of *express* and *activate* resulted in the performance degradation compared to the using of only lexical-based features. On the contrary, the performance improvement is obtained in case of manual-examples. In this regard, the F1-scores of using PAS-related features are 0.99 (for *express*) and 0.62 (for *activate*) higher than of using only lexical-based features as shown in Table 4-4. This result supports the conclusion mentioned at the end of section 4.3 that the positive effects of PAS-related features on NER system should be obtained for all predicates in group 1.

For the predicates *decrease* (group 2) and *associate* (group 3), NER systems still result in performance degradation although manual-examples are used instead of parsing-examples. This conforms to the conclusion mentioned at the end of section 4.3 that the positive effects of PAS-related features to NER system will be obtained from a predicate in group 2 in the condition that its training data must be large. Furthermore, the positive effects of PAS-related features to NER system will not be obtained from applying these features to the predicates in group 3.

Considering *recognize*, from training on 100 manual-examples, the performance improvement increase to 6.12—which is about 3 times of 2.15 that is obtained from 121 parsing-examples. The size of the manual-examples of *recognize* is nearly equal to that of the parsing-examples; thus, it can be implied that the parsing error affects in decreasing the power of PAS-related features to enhance NER system for at least 3 times. In accordance to this, the performance improvement of *encode* also increase about 3 times in case 256 manual-examples are used, compared to 256 parsing-examples.

Moreover, the results in Table 4-4 show that the more training data, the better performance. The performance improvement of 2.40 acquired from the NER system training on just 100 manual-examples of *encode* increase to 3.27 from training on 256 manual-examples affirms what is just stated.

To be summarized, the parsing errors (i.e., the lost of syntactic relation and the incorrect information for a word) must be handled to allow the NER system using PAS-

[57] Hence, the training examples are called manual-examples when argument boundaries are identified manually and are called parsing-examples when argument boundaries are identified automatically based on syntactic relation information given by the parser.

related features to get a maximum bound of performance at 3 times higher than what can be achieved under the parsing error condition.

4.4.1.2 Problem from the quantifier

The use of quantifier in a sentence is a problematic factor in identifying a boundary of an argument. From Figure 4-24, the constituent "*multiple isotypes*" in the sentence "*T cells express multiple isotypes of protein kinase C*" will be bounded to be the argument *theme* of the predicate *express* because the general algorithm for sub-structure recognition uses raw information given by the parser that "*multiple isotypes*" has syntactic relation as the object of "*express*".

```
T cells express [multiple isotypes] of protein kinase C...
                [     protein     ]
                [        O        ]   [    protein    ]
```

Figure 4-24: A sentence showing human annotation in GENIA corpus (*green part*) and the answer from the NER system using PAS-related features (*blue part*)

However, the substantial argument that should be identified as a *theme* argument is "*protein kinase C*" (a protein-entity that can be a participant of the expression event). The constituent "*multiple isotypes*" is merely a quantifier. This is analogous to a sentence in general language; for instance, in the sentence "*David drinks a cup of coffee*", the entity that is drunk is "*coffee*" but not "*a cup*". If the sub-structure recognition process can distinguish a substantial argument from a quantifier, the semantic role will be assigned to the correct constituent.

```
...TCF-1 specifically recognizes [T beta 5 element] of the TCR
beta enhancer
                                 [        DNA       ]
                                 [        DNA       ]
```

Figure 4-25: A sentence showing human annotation in GENIA corpus (*green line*) and the answer from the NER system using PAS-related features (*blue line*)

To distinguish between a substantial argument and a quantifier is a nontrivial task. In Figure 4-25, the constituent "*T beta 5 element*" of the sentence "*...TCF-1 specifically recognizes the T beta 5 element of the TCR beta enhancer...*" is a substantial argument; even though, this substantial argument is mentioned in this sentence with the same surface syntactic pattern (i.e. the pattern *X* of *Y*) as the quantifier "*multiple isotypes*" in the previous sentence.

4.4.2 Semantic role representation

This impediment factor involves the imperfect identification of a semantic role. The more efficient representation of semantic roles *agent* and *theme*, as well as the efficient representation for semantic roles other than *agent* and *theme* are required.

So far, the best set of features used in this work is capable of representing only semantic roles *agent* and *theme*. Furthermore, the syntactic features used in the current system are still not sufficient even to make distinction between these two semantic roles.

(1) [Donazol]$_{agent}$ **decreases** [transcription]$_{theme}$ of estrogen receptor gene in human monocytes. <- ***Transitive sense***

(2) [CD40 cells]$_{theme}$ in a resting B-cell preparation from hepatitis B patients **decreased** to zero after a 5-day culture with sequence I. <- ***Intransitive sense***

Figure 4-26: Sentences showing the using of the predicate *decrease* in the transitive sense (sentence 1) and in the intransitive sense (sentence 2)

Analogous to the examples of the predicate *break* in Figure 4-17, from sentences: (1) "*Donazol* ***decreases*** *transcription of estrogen receptor gene in human monocytes*" and (2) "*CD40 cells in a resting B-cell preparation from hepatitis B patients* ***decreased*** *to zero after a 5-day culture with sequence I*" as shown in Figure 4-26, the system can correctly interpret that the constituent "*transcription*" functioning as the surface object in sentence (1) and "*CD40 cells*" functioning as the surface subject in sentence (2) have the same semantic role (i.e. *theme*). Current set of features (PAS-related features) in representing semantic roles consists of *predicate's surface form*, *predicate's lemma*, *voice*, *surface*

syntactic role and the feature combined from a *subject-object head pair* and *transitive-intransitive* sense. As mentioned in section 4.2.3, from using this feature set the system incorrectly interpret that "*John*" in the sentence "*John ate at the Royal-Host restaurant yesterday*" plays the semantic role of *theme*. As the surface object is not mentioned, thus the system identify that the corresponding predicate is used in intransitive sense. The surface subject will play the semantic role of *theme* if the predicate is used in intransitive sense. However, the predicate *eat* is used in the transitive sense because the event "eating" implies that there must be something is eaten although it is not mentioned. In this regard, a predefined PAS frame for each predicate seems to be very important. As a PAS frame can clue that whether the predicate can be used in intransitive sense or not. Subsequently, the predefined PAS-frame should be represented in a feature set of the NER system as well in order to increase the system's capability in representing the semantic roles *agent* and *theme*.

In addition to the insufficiency of features for representing semantic roles *agent* and *theme*, the representation of other semantic roles is necessary as well.

(3) [The fNF-E2 isoform]$_{Arg3:RNA}$ **is transcribed** from an alternative promoter.

(4) [The plastid genome]$_{Arg1:DNA}$ **is known to be transcribed** by a plasmid-encoded prokaryoric.

Figure 4-27: Sentences showing the using of the predicate transcribe in 2 different senses

From sentences: (3) "*The fNF-E2 isoform is transcribed from an alternative promoter*" and (4) "*The plastid genome is known to be transcribed by a plasmid-encoded prokaryotic*" as shown in Figure 4-27, the constituents "*The fNF-E2 isoform*" in sentence (1) and "*The plastid genome*" in sentence (2) will be identified by the system as *theme* argument if the system aims to cover only the semantic roles *agent* and *theme*[58]. In this case, the using of semantic roles cannot help the NER system to classify "*fNF-E2 isoform*" as RNA and "*plastid genome*" as DNA due to the same semantic role they have.

[58] This is based on the simple rule that if a sentence is in passive voice, the *subject* constituent will have semantic role as *theme* and the *object* constituent will have semantic role as *agent*.

From the analysis in the process of constructing PASBio database (Wattarujeekrit et al., 2004), the PAS frame of the predicate *transcribe* consists of 5 arguments (i.e., Arg0-Arg4). Each argument has its semantic role: *agent*, *entity transcribed (theme)*, *transcription site*, *transcription product* and *location (organ or tissue where the transcription happen)* as shown in Figure 4-28.

```
Predicate: transcribe = convert DNA into RNA
Argument-Semantic role set:
        Arg0-causer agent
        Arg1-entity to be transcribed (transcription source)
        Arg2-transcription site
        Arg3-entity after transcription (transcription product)
        Arg4-location as organ/tissue expressing transcription
```

Figure 4-28: The PAS frame of the predicate *transcribe* defined in PASBio database

According to the semantic role set in the PAS frame of the predicate *transcribe*, the constituents "*The fNF-E2 isoform*" plays the semantic role of *transcription product* and "*The plastid genome*" plays the semantic role of *entity to be transcribed (theme)*. According to the domain knowledge, DNA is the only molecular entity that can be transcribed and the molecular entity that is a product of transcription process is RNA. Therefore, if the system can cover the semantic roles other than *agent* and *theme* such as *Arg1-entity to be transcribed* and *Arg3-entity after transcription*, the semantic roles will show more positive effect to the system.

4.4.3 Named entities outside the argument boundaries

This impediment factor concerns the named entities found in a sentence at the constituents outside a focus predicate's argument boundaries. So far, a semantic role represented by syntactic features (PAS-related features) is assigned for only the argument boundaries of one predicate at a time. Thus, in Figure 4-29, the NER system can take advantage of semantic roles to classify named entity types of only "*gene*" and "*transcription factors*", if the focus predicate is "*encode*". As most of sentences found in molecular biology literature are not simple sentences, two or more predicates are likely to

be found in the same sentence. To apply the semantic roles of all predicate's arguments in a sentence at the same time should help the NER system get better results than to apply the semantic roles of the arguments of only one predicate at a time. Concerning to this, the critical point to be considered is about the weight or priority of each predicate in a sentence.

... [gene]$_{DNA}$ **encoding** [transcription factors]$_{protein}$ that **bind** to the [canonical DNA sequence]$_{DNA}$ ...

Figure 4-29: An example of sentences containing more than one predicates (hence, *encode* and *bind*)

For instance, there are two predicates (i.e. *encode* and *bind*) found in the sentence "*...gene encoding transcription factors that bind to the canonical DNA sequence...*" as shown in Figure 4-29. In this sentence, the constituent "*transcription factors*" is an object of "*encoding*" and at the same time it is a subject of "*bind*". Between to assign semantic role as the *theme* of the predicate *encode* or to assign semantic role as the *agent* of the predicate *bind*, which choice will result in better performance must be investigated.

... we **identify** the [gene]$_{DNA}$ **encoding** the [lymphocyte homing and migration protein]$_{DNA}$ ...

Figure 4-30: An example of sentences containing more than one predicates (hence, *identify* and *encode*)

In Figure 4-30, it is easy to decide that the constituent "*gene*" should be assigned the semantic role of *agent* of the predicate "*encoding*" instead of *theme* of the predicate "*identify*". Roughly, the lowest priority should be given to the predicates used to write a paper. For each rhetorical zone, the predicates used in composing a paper are, for example, Background—*suggest* and *discover*; Problem setting—*test*, *evaluate* and *address*; Outline—*report*, *ask*, *find* and *show* (Mizuta and Collier, 2004).

4.5 The effectiveness of an argument's semantic role

In this work, the use of semantic relationship between a predicate and its argument for enhancing NER has been explored. This semantic relationship, the semantic role to be precise, is employed in terms of a feature set for the SVM-based NER system. This feature set is composed of PAS-related features which are the features pertaining to syntactic information capable of semantic role representation. So far, the PAS-related features are assigned to only tokens in a boundary of a predicate's argument functioning as a surface *subject* or a surface *object*. Therefore, in this work, the use of PAS-related features in addition to the lexical-based features directly helps NER system to identify the type of a named entity locating at the surface *subject* or surface *object* position.

Surf	Lem	Head	Orth	Phr	PSurf	PLem	Voice	SynR.	Head	H_NE	M_NE
More over	more over	-	lc	B-ADVP	O	O	O	O	O	O	O
,	,	-	Pu	O	O	O	O	O	O	O	O
mono cytes	mono cyte	Mono cytes	Lw	B-NP	O	O	O	O	O	B-cell_type	O
expre ss	expre ss	-	Lw	B-VP	O	O	O	O	O	O	O
a	a	-	Lw	B-NP	O	O	O	O	O	O	O
novel	novel	Protei n	Lw	I-NP	O	O	O	O	O	O	O
IL-10-stimul ated	il-10-stimul ated	Protei n	Cdh	I-NP	O	O	O	O	O	B-protein	B-protein
STAT	stat	Protei n	2c	I-NP	O	O	O	O	O	I-protein	I-protein
protei n	protei n	Protei n	Lw	I-NP	O	O	O	O	O	I-protein	I-protein
with	with	-	Lw	B-ADVP	O	O	O	O	O	O	O
an	an	-	Lw	B-NP	O	O	O	O	O	O	O
M	m	M	Sc	I-NP	O	O	O	O	O	O	O
(	(	-	Pu	O	O	O	O	O	O	O	O
r	r	R	lw	B-NP	O	O	O	O	O	O	O
)	)	-	pu	B-NP	O	O	O	O	O	O	O
of	of	-	lw	B-PP	O	O	O	O	O	O	O
70	70	-	2ds	B-NP	recog nized	recog nize	PAS	SSUB J	kda_a b	B-protein	B-protein
kDa	kda	kDa	lc	I-NP	recog nized	recog nize	PAS	SSUB J	kda_a b	I-protein	I-protein
that	that	That	lw	B-SBAR	O	O	O	O	O	O	O
is	be	-	lw	B-VP	O	O	O	O	O	O	O

recog nized	recog nize	-	lw	I-VP	O	O	O	O	O	O	O
by	by	-	lw	B-ADVP	O	O	O	O	O	O	O
the	the	-	lw	B-NP	O	O	O	O	O	O	O
anti-STAT 3	anti-stat3	anti-STAT 3	cdh	I-NP	recog nized	recog nize	PAS	SOBJ	O	B-protein	B-protein
Ab	ab	-	ic	O	recog nized	recog nize	PAS	SOBJ	O	I-protein	I-protein
but	but	-	lw	O	O	O	O	O	O	O	O
is	be	-	lw	B-VP	O	O	O	O	O	O	O
not	not	-	lw	B-ADVP	O	O	O	O	O	O	O
obser ved	obser ve	-	lw	B-VP	O	O	O	O	O	O	O
in	in	-	lw	B-ADVP	O	O	O	O	O	O	O
T	t	Cells	sc	B-NP	O	O	O	O	O	B-cell_type	B-cell_type
cells	cell	Cells	lw	I-NP	O	O	O	O	O	I-cell_type	I-cell_type
.	.	-	pu	O	O	O	O	O	O	O	O

Figure 4-31: An example of classification results from SVMs using PAS-related features for a predicate *recognize*. The result from SVMs is shown in *blue*, a human-annotated named entity is shown in column H_NE, the surface subject argument is shown in *pink*, and the surface object argument is shown in *yellow*

With regards to the difficulties of NER in the molecular biology domain, the PAS-related features have their usefulness mainly for the case of polysemy or sharing names among entities (i.e., both systematic polysemy and homonymy). This is because the lexical knowledge is totally unlikely to be able to deal with the case of polysemy. Also, the PAS-related features can partly handle the lexical difficulty (i.e., the various patterns of terminology problem).

In Figure 4-31, the constituent "*70 kDa*" is correctly classified as protein name (*Red square*) due to the clue of being an argument *theme* of the predicate *recognize*. By using only lexical knowledge, the model may interpret "*70 kDa*" as a mentioning of a particular quantity. This misleads the NER system to recognize "*70 kDa*" as non-NE (shown in Figure 4-32, *Red square*). Similarly, without PAS-related features for semantic relation between a predicate *recognize* and a term "*anti-STAT3 Ab*" shown in Figure 4-32 (*Green square*), only the word "*anti-STAT3*" is correctly classified as a part of a protein name

Surf	Lem	Head	Orth	Phr	H_NE	M_NE
Moreover	moreover	-	ic	B-ADVP	O	O
,	,	-	pu	O	O	O
monocytes	monocyte	Monocytes	lw	B-NP	B-cell_type	O
express	express	-	lw	B-VP	O	O
a	A	-	lw	B-NP	O	O
novel	Novel	Protein	lw	I-NP	O	O
IL-10-stimulated	il-10-stimulated	Protein	cdh	I-NP	B-protein	B-protein
STAT	Stat	Protein	2c	I-NP	I-protein	I-protein
protein	protein	Protein	lw	I-NP	I-protein	I-protein
with	With	-	lw	B-ADVP	O	O
an	An	-	lw	B-NP	O	O
M	M	M	sc	I-NP	O	O
(	(	-	pu	O	O	O
r	R	R	lw	B-NP	O	O
)	)	-	pu	B-NP	O	O
of	Of	-	lw	B-PP	O	O
70	70	-	2ds	B-NP	B-protein	O
kDa	Kda	kDa	lc	I-NP	I-protein	O
that	That	That	lw	B-SBAR	O	O
is	Be	-	lw	B-VP	O	O
recognized	**recognize**	**-**	**lw**	**I-VP**	**O**	**O**
by	By	-	lw	B-ADVP	O	O
the	The	-	lw	B-NP	O	O
anti-STAT3	anti-stat3	anti-STAT3	cdh	I-NP	B-protein	B-protein
Ab	Ab	-	ic	O	I-protein	O
but	But	-	lw	O	O	O
is	Be	-	lw	B-VP	O	O
not	Not	-	lw	B-ADVP	O	O
observed	observe	-	lw	B-VP	O	O
in	in	-	Lw	B-ADVP	O	O
T	T	Cells	Sc	B-NP	B-cell_type	B-cell_type
cells	cell	Cells	Lw	I-NP	I-cell_type	I-cell_type
.	.	-	Pu	O	O	O

Figure 4-32: An example of classification results of the SVM-based NER system using only lexical features for a predicate *recognize*. The result from the SVM-based NER system is shown in *blue*, a human-annotated named entity is shown in column H_NE, the surface subject argument is shown in *pink*, and the surface object argument is shown in *yellow*

because its lexical information is obvious hint. However, by using only lexical features, the following word "*Ab*" is incorrectly classified to non-NE. In contrast, the word "*Ab*" is correctly classified as a part of protein name when PAS-related features are used. This confirms the importance of semantic relation between a predicate and its arguments for NER.

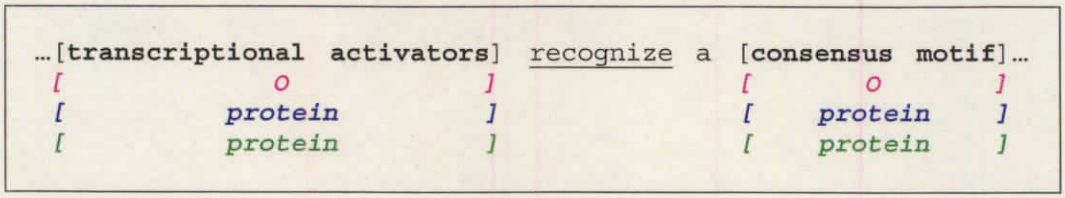


Figure 4-33: A sentence showing the answer from the NER system using only lexical features (*pink line*) and using also PAS-related features (*blue line)*, as well as named entity class annotated by GENIA corpus's annotators (*green line*)

Figure 4-33 is another example to show that the using of PAS-related features can handle the lexical difficulty. As discussed in the introduction of Chapter 4, a molecular entity is not only in the form of a proper noun, but also a descriptive term. In Figure 4-33, the constituents "*transcriptional activators*" and "*consensus motif*" are the examples of named entities in the forms of the descriptive terms. They do not contain any capital letters for expressing themselves as the proper nouns. By using only lexical features, the NER system incorrectly recognizes them as non-NE (*pink line*). On the contrary, the using of PAS-related features helps the NER system to correctly recognize them as protein entities (*blue line*).

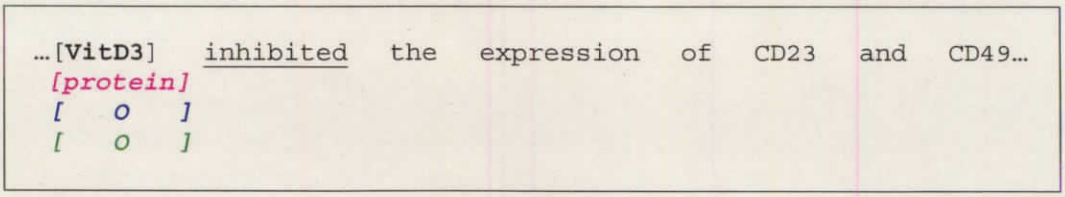


Figure 4-34: A sentence showing the answer from the NER system using only lexical features (*pink line*) and using also PAS-related features (*blue line)*, as well as named entity class annotated by GENIA corpus's annotators (*green line*)

In Figure 4-34, because the constituent "*VitD3*" is lexically a proper noun, the NER system using only lexical features recognizes this constituent as protein (*pink line*). But, the term "*VitD3*" actually stands for *Vitamin D3* which is not a named entity of interest in this work[59]. Due to the lack of naming conventions in the molecular biology domain, the lexical patterns of different types of named entities are overlap across each other. The lexical information only is too weak to hint the NER system to correctly identify if a term is a named entity of interest. In addition to lexical information, the semantic role information will be another important clue. The constituent "*VitD3*" plays the role of *agent* in the clause "*...VitD3 inhibited the expression of CD23 and CD49...*". In accordance with the restriction between a named entity type and a semantic role, the argument *agent* of the predicate *inhibit* is unlikely to be protein. By using together both lexical information and semantic role information, the NER system can correctly identify "*VitD3*" as non-NE as shown in the *blue line* in the Figure 4-34.

```
...a cDNA encoding the [murine B-cell specific coactivator]...
                       [                DNA                 ]
                       [                protein             ]
                       [                protein             ]
```

Figure 4-35: A sentence showing the answer from the NER system using only lexical features (*pink line*) and using also PAS-related features (*blue line)*, as well as named entity class annotated by GENIA corpus's annotators (*green line*)

An example in Figure 4-35 shows that the semantic role has a potential to handle polysemy problem. The entity named as "*murine B-cell specific coactivator*" can refer to either DNA or protein. However, in the example, this entity plays the role of *theme* in the encoding event, so this entity is correctly classified as protein by the NER system using PAS-related features (*blue line*). This is because the entity being encoded in the molecular event can be only a protein.

[59] Please see section 4.1.1 for more details.

Chapter 5
Conclusions and Future works

The aim of this thesis is to prove the hypothesis that the semantic role, the semantic information describing the relationships between a predicate and its arguments, can enhance molecular NER system. The semantic role seems to be useful for molecular NER system due to a key idea that each argument's semantic role should impose type restrictions on the entities within the argument.

In order to investigate this hypothesis, two main subsidiary tasks were done: (1) the construction of PAS frames from analyzing the sublanguage used in the bio-molecular literature and (2) the employment of semantic roles in machine-learning based NER system. The consequences of doing these tasks are discussed as follows.

5.1 Concluding Remarks

5.1.1 Construction of PASBio resource

The first subtask in this thesis is to analyze the molecular biology sublanguage presented in the literature (e.g., MEDLINE, EMBO, PNAS, NAR and JV) for constructing PAS frames in this scientific domain. As a consequence, 34 PAS frames for 29 predicates are included in PASBio resource[60] as well as about 300 annotated example sentences corresponding to the frames.

Through the analysis, the predicates can be categorized into 4 groups grounded on their defined PAS frames in the molecular biology domain (PASBio project) compared to their PAS frames in newswire domain (PropBank project). The characteristic of predicates in each group is as follows: Group A—predicates have similar senses in the molecular biology domain and in newswire domain, but more arguments are required to completely express the event in the molecular biology domain; Group B—predicates have similar senses in the molecular biology domain and in newswire domain, but less arguments are

[60] As each PAS frame is defined for a sense of predicate, the predicates with more than one sense will have more than one PAS frames.

required to completely express the event in the molecular biology domain because unnecessary arguments are counted as the basic knowledge that domain experts must have already known; Group C—predicates have the same senses in the molecular biology domain and in newswire domain and required the same set of arguments in both domains; Group D—predicates are rarely used in newswire domain or have special meaning in the molecular biology domain. The percentage of predicates for each group is 30%, 20%, 20% and 30%, respectively[61]. From this information, it is worth to note that about 30% of predicates (Group D's predicates) are totally in need to be analyzed specifically for the molecular biology domain. On the other hand, only 20% of predicates (Group C's predicates) do not required the construction of PAS frames specifically for bio-molecular sublanguage. The PAS frames can be naturally borrowed from what are already defined in newswire domain. This result supports the importance of the analysis of PAS frames for the molecular biology domain.

Because in this thesis semantic information between a predicate and its arguments is simply applied to NER system (i.e. only a semantic role of an argument with its syntactic function as either *subject* or *object* is considered), a defined PAS frame showing all arguments seem unnecessary thus far. However, the construction of PASBio is important for the further step of NER system in future works. In addition, the PASBio frames can currently be used as follows[62]:

- a reference knowledge in case the annotation of semantic information at the PAS level is required
- a reference knowledge for designing extraction patterns, i.e. a biological information extraction system can use PAS frames as a guideline concerning the types and numbers of arguments that can be expected from the extraction process
- a reference knowledge to explain the empirical results

5.1.2 Employment of semantic roles in machine-learning based NER

The semantic role is applied to NER system in terms of features for SVM-based NER system. In this thesis, only the semantic role of an argument functioning as either *subject*

[61] Please see section 3.4 for the full list of predicates in each group.
[62] Please see section 3.5 for detailed explanation of the usefulness of PASBio's frames.

or *object* of a target predicate is involved. The syntactic features used to represent the arguments' semantic roles are named ad PAS-related features. With reference to two types of evidences for NER (i.e., term internal and term external), the semantic role is considered as the term external evidence as same as co-occurrence information. Nevertheless, to use semantic role is more consistent than co-occurrence[63]. The base model for evaluating the importance of the semantic roles for molecular NER consists of six lexical features: *surface word*, *lemma form*, *head word of noun phrase*, *part-of-speech*, *orthographic feature* and *phrase-chunk*. The GENIA corpus V3.02 with five classes (e.g., protein, DNA, RNA, cell line and cell type) after conflation is a set of training data used in this thesis[64].

As a result, the set of PAS-related features that allows the use of semantic role to show its highest positive effect on the NER system is composed of *predicate's surface form*, *predicate's lemma*, *voice* and *the united feature of subject-object head's lemma and transitive-intransitive sense*. Moreover, the set of PAS-related features takes effect to each predicate differently. Their positive effects can be shown for only the predicates conforming to the criteria as follows. A predicate must have its arguments as both *agent* and *theme* with a higher probability of belonging to a named entity class than non-named entity class; otherwise, a predicate must have its arguments as both *agent* and *theme* with a lower probability of belonging to a named entity class than non-named entity class and the number of training examples for this predicate should be large enough. In addition, the use of PAS-related features for the predicates fitting into the first criterion gives higher improvement in performance.

5.2 Future directions

As to apply semantic roles for NER systems is quite new in bio-text mining community, this thesis is a kind of pioneer research. Therefore, various problematic aspects still remain for being investigation in the future. These aspects are explained separately for each subtask as follows.

- **Construction of PASBio resource**: In order to broaden usage scope of PASBio frame, PAS frames for other predicates besides 30 predicates that have

[63] Please see section 2.3 for detailed discussion on the using of semantic role compared to co-occurrence.
[64] Please see section 4.1.1 for detailed discussion about the data set in GENIA corpus.

been analyzed should be focused. Also, the number of annotated sentences should be increased because so far there are too small to be trained for automatic semantic role labeling.

- **Employment of semantic roles in machine-learning based NER**: In this thesis, the upper bound of the performance of the NER system employing semantic roles have not been obtained yet; thus, the future directions for this subtask are planed according to the 3 main impediments illustrated in section 4.4.1.1. First, the sophisticated rules for the NER system to have better ability in identifying the argument of a predicate should be done. Second, the set of features for representing an argument's semantic role should be extended to increase both correctness and coverage of semantic role representation. Third, the semantic roles of multiple predicates found in a sentence should be employed at the same time.

About Author

Name	Tuangthong Wattarujeekrit
Birth Date	July 18, 1975
Birth Place	Songkhla, Thailand
Nationality	Thai
Status	Single
Educations	2000-2002 Master of Computer Engineering, Department of Computer Engineering, Kasetsart University, Bangkok, Thailand 1992-1996 Bachelor of Computer Engineering (First Class Honors), Department of Computer Engineering, Chiangmai University, Chiangmai, Thailand
Experiences	Jun 2000-Sep 2002 Teacher Assistant, Department of Computer Engineering, Kasetsart University, Bangkok, Thailand Apr 1996-Oct 1999 Computer Engineer (Project Leader), Hoya Glass Disk (Thailand) Ltd., Northern Region Industrial Estate, Lamphun, Thailand

Related Publications

Wattarujeekrit T. and Collier N. (2005). **Exploring Predicate-Argument Relations for Named Entity Recognition in the Molecular Biology Domain.** *Proceedings of the 8th International Conference on Discovery Science (DS'2005)*, pp. 267-280. (Springer-Verlag, Lecture Notes in Artificial Intelligence, Vol. 3735)

Wattarujeekrit T., Shah P. K., and Collier N. (2004). **PASBio: predicate-argument structures for event extraction in molecular biology.** *BMC Bioinformatics* 5:155 (Published 19th October 2004)

Wattarujeekrit T. and Collier N. (2004). **Integrating Event Frame Annotation into the Open Ontology Forge Annotation Tool.** *Proceedings of the 4th International Workshop on Knowledge Markup and Semantic Annotation (at ISWC'2004)*, pp. 119-122.

Collier N., Takeuchi K., Kawazoe A., Mullen T., and Wattarujeekrit T. (2003). **A Framework for Integrating Deep and Shallow Semantic Structures in Text Mining.** *Proceedings of the 7th International Conference on Knowledge-based Intelligent Information and Engineering Systems (KES'2003)*, pp. 824-834. (Springer-Verlag, Lecture Notes in Computer Science, Vol. 2773)

Bibliography

Andrade M. A. and Valencia A. (1998). **Automatic Extraction of Keywords from Scientific Text: Application to the Knowledge Domain of Protein Families**. *Bioinformatics* 14(1):600-607.

Blaschke C., Andrade M. A., Ouzounis C. and Valencia A. (1999) **Automatic extraction of biological information from scientific text: Protein-protein interactions**. Proc. of *Int. Conf. Intelligent System Molecular Biology*, pp. 60-67.

Bock J. and Gough D. (2001). **Predicting protein-protein interactions from primary structure**. *Bioinformatics* 17:455-460.

Bodenreider O., Mitchell J. A. and McCray A. T. (2002). **Evaluation of the UMLS as a terminology and knowledge resource for biomedical informatics**. Proc. of *the AMIA Symposium*, pp. 61-65.

Brill E. (1992). **A simple rule-based part of speech tagger**. Proc. of *the 3rd Conf. on Applied Natural Language Processing (ANLP'92)*, pp. 152-155.

Brown M., Grundy W., Lin D., Cristianini N., Sugnet C., Furey T., Ares M. and Hassler D. (2000). **Knowledge-based analysis of microarray gene expression data by using support vector machines**. Proc. of *the National Academy of Science*, pp. 262-267.

Buitelaar P. (1998). **CoreLex: Systematic Polysemy and Underspecification**. *Ph.D. thesis*, Computer Science Department, Brandeis University, Febuary.

Bunescu R. C., Ge R., Kate R. J., Marcotte E. M., Mooney R. J., Ramani A. K., and Wond Y. W. (2005). **Comparative Experiments on Learning Information Extractors for Proteins and their Interactions**. *Artificial Intelligence in Medicine* 33(2): 139-155.

Cohen K. B., Dolbey A. E., Acquaash-Mensah G. K. and Hunter L. (2002). **Contrast and variability in gene names**. Proc. of *the Workshop on NLP in the Biomedical Domain (at ACL'02)*, pp. 14-20.

Cohen K. B. and Hunter L. (2004). **Natural Language Processing and Systems Biology**. *Artificial Intelligence methods and tools for systems biology*, Dubitzky and Pereira eds., Springer Verlag.

Collier N. and Takeuchi K. (2004). **Comparison of character-level and part of speech features for name recognition in bio-medical texts**. *Journal of Biomedical Informatics* 37(6): 423-425.

Collier N., Nobata C. and Tsujii J. (2000). **Extracting the names of genes and gene products with a Hidden Markov Model**. Proc. of *the 18th Int. Conf. on Computational Linguistics (COLING'00)*, pp. 201-207.

Corney D. P. A., Buxton B. F., Langdon W. B. and Jones D. T. (2004). **BioRAT: extracting biological information from full-length papers**. *Bioinformatics* 20(17): 3206-3213.

Couto F., Silva M. and Coutinho P. (2005). **Finding Genomic Ontology Terms in Unstructured Text**. *BMC Bioinformatics* 6(Suppl 1): S21.

Cowie J. and Lehnert W. (1996). **Information extraction**. *Communications of the ACM* 39(1): 80-91.

Craven M. and Kumlein J. (1999). **Constructing biological knowledge bases by extracting information from text sources**. Proc. of *the 7th Int. Conf. on Intelligent Systems for Molecular Biology (ISMB'99)*, pp. 77-86.

Cristianini N. and Shawe-Taylor J. (2000). *An introduction to support vector machines and other kernel-based learning methods*, Cambridge University Press, ISBN 0521780195.

DARPA. (1995). Proc. of *the 6th Message Understanding Conference (MUC-6)*, Columbia, Marryland.

DARPA. (1998). Proc. of *the 7th Message Understanding Conference (MUC-7)*, Fairfax, VA, USA.

Dickman S. (2003). **Tough Mining: The challenges of searching the scientific literature.** *PLoS Biology* 1(2):e8.

Ehrler F., Jimeno A. and Ruch P. (2005). **Data-poor categorization and passage retrieval for Gene Ontology annotation in Swiss-Prot.** *BMC Bioinformatics* 6(Suppl 1): S23.

Erhardt R. A. and Blaschke C. (2005). **Identifying biomedical information in the scientific knowledge.** *White paper*, almaKnowledgeServer (AKS).

Franzen K., Eriksson G., Olsson F., Asker L., Liden P. and Coster J. (2002). **Protein names and how to find them.** *International Journal of Medical Informatics* 67(1-3):49-61.

Friedman C., Kra P., Yu H., Krauthammer M. and Rzhetsky A. (2001). **GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles.** *Bioinformatics* 1(1):1-9.

Fukuda K., Tamura A., Tsunoda T. and Takagi T. (1998). **Toward information extraction: identifying protein names from biological papers.** Proc. of *the Pacific Symposium on Biocomputing (PSB '98)*, pp. 705-716.

Gaizauskas R., Demetriou G. and Humphreys K. (2000). **Term Recognition and Classification in Biological Science Journal Articles**. Proc. of *the Workshop on Computational Terminology for Medical and Biological Applications*, pp. 37-44.

Gaizauskas R., Demetriou G., Artymiuk P. J., Willett P. (2003). **Protein structures and information extraction from biological texts: the PASTA system**. *Bioinformatics* 19:135-143.

Guyon I., Weston J., Barnhill S. and Vapnik V. (2002). **Gene selection for cancer classification using support vector machines**. *Machine Learning* 46:389-422.

Hacioglu K., Pradhan S., Ward W., Marting J. H. and Jurafsky D. (2004). **Semantic role labeling by tagging syntactic chunks**. Proc. of *the 8th Conf. on Computational Natural Language Learning (CoNLL'04)*, pp. 110-113.

Hajic J., Cmejrek M., Dorr B., Ding Y., Eisner J., Gildea D., Koo T., Parton K., Penn G., Redev D. and Rambow O. (2002). **Natural Language Generation in the Context of Machine Translation**. Technical Report, the Center for Language and Speech Processing, Johns Hopkins University.

Han C., Lavoie B., Palmer M., Rambow O., Kittredge R., Korelsky T., Kim N. and Kim M. (2000). **Handling Structural Divergences and Recovering Deropped Arguments in a Korean/English Machine Translation System**. Proc. of *the Conf. on Association for Machine Translation in the Americas (AMTM'00)*, pp. 40-53.

Harris Z. S. (1968). *Mathematical Structures of Language*. Wiley-Interscience.

Harris Z. S. (2002). **The structure of science information**. *Biomedical Informatics* 35:215-221.

Hatzivassiloglou V., Duboue P. A. and Rzhetsky A. (2001). **Disambiguating proteins, genes, and RNA in text: a machine learning approach**. *Bioinformatics* 17(1):97-106.

Hirschman L., Colosimo M., Morgan A. A. and Yeh A. S. (2005). **Overview of BioCreAtIvE task 1B: normalized gene lists.** *BMC Bioinformatics* 6(Suppl 1):S11.

Hirschman L., Morgan A. A. and Yeh A. S. (2002). **Tutabaga by any other name: extracting biological names**. *Journal of Biomed Inform* 35(4):247-259.

Hirschman L. and Sager N. (1982). **Automatic Information Formatting of a Medical Sublanguage**. *Sublanguage: Studies of Language in Restricted Semantic Domains*, Kittredge and Lehrberger eds., Walter de Gruyter, pp. 27-80.

Hua S. J. and Sun Z. R. (2001). **Support vector machine approach for protein subcellular localization prediction.** *Bioinformatics* 17:721-728.

Humphreys B. L., Lindberg D. A. B., Schollman H. M. and Barnett G. O. (1998). **The Unified Medical Language System: An informatics research collaboration**. *Journal of the American Medical Informatics Association* 5:1.

Iliopoulos I., Enright A. J. and Ouzounis C. A. (2001). **Textquest: document clustering of Medline abstracts for concept discovery in molecular biology**. Proc. of *the Pacific Symposium on Biocomputing (PSB '01)*, pp. 384-395.

Jaakkola T., Diekhans M. and Haussler D. (1999). **Using the Fisher kernel method to detect remote protein homologies**. Proc. of *the 7th Int. Conf. on Intelligent Systems for Molecular Biology*, pp. 149-158.

Jenssen T. K., Laegreid A., Komorowski J. and Hovig E. (2001). **A Literature Network of Human Genes for High-Throughput Analysis of Gene Expression**. *Nature Genetics* 28:21-28.

Joachims T. (1998). **Text Categorization with Support Vector Machines: Learning with Many Relevant Features**. Proc. of *the 10th European Conference on Machine Learning (ECML '98)*, pp. 137-142.

Joachims T. (1999). **Making large-scale SVM learning practical.** *Advances in Kernel Methods – Support Vector Learning*, MIT Press.

Kazama J., Makino T., Ohta Y. and Tsujii J. (2002). **Tuning Support Vector Machines for Biomedical Named Entity Recognition**. Proc. of *the Workshop on NLP in the Biomedical Domain (at ACL '02)*, pp. 1-8.

Kim J. D., Ohta T., Tateisi Y. and Tsujii J. (2003). **GENIA corpus-semantically annotated corpus for bio-textmining**. *Bioinformatics* 19(1):180-182.

Kim J. D., Ohta T., Tsuruoka Y. and Tateisi Y. (2004). **Introduction to the Bio-Entity Recognition Task at JNLPBA**. Proc. of *the Int. Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA '04)*, pp. 70-75.

Krauthammer M. and Nenadic G. (2004). **Term identification in the biomedical literature**. *Journal Biomedical Informatics*. 37(6):512-526.

Krauthammer M., Rzhetsky A., Morozov P. and Friedman C. (2000). **Using BLAST for identifying gene and protein names in journal articles**. *Gene* 259(1-2):245-252.

Kudo T. and Matsumoto Y. (2000). **Use of support vector learning for chunk identification**. Proc. of *the 4th Conference on Computational Natural Language Learning (CoNLL '00)*, pp. 142-144.

Lee K. J., Hwang Y. S. and Rim H. C. (2003). **Two-phase biomedical NE Recognition based on SVMs**. Proc. of *the Workshop on Natural Language Processing in Biomedicine (at ACL '03)*, pp. 33-40.

Levin B. (1993). *English Verb Classes and Alternations: A preliminary Investigation*, University of Chicago Press.

Lin Y. F., Tsai T. H., Chou W. C., Wu K. P., Sung T. Y. and Hsu W. L. (2004). **A maximum entropy approach to biomedical named entity recognition**. Proc. of *the Workshop on Data Mining in Bioinformatics (BioKDD'04)*, pp. 56-61.

Lui H., Lussier Y. A. and Friedman C. (2001). **Disambiguating ambiguous biomedical terms in biomedical narrative text: an unsupervised method**. *Journal of Biomed Informatics* 34:249-261.

Marcu D. (2000). *The Theory and Practice of Discourse Parsing and Summarization.* MIT Press, ISBN 0262133725.

Meyers A., Macleod C. and Grishman R. (1996). **Standardization of the Complement/ Adjunct Distinction**. Proc. of *the 7th Euralex International Congress.*

Miller G. A. (1990). **WordNet: An on-line lexical database**. *International Journal of Lexicography* 3: 235-312.

Mizuta Y. and Collier N. (2004). **An Annotation Scheme for a Rhetorical Analysis of Biology Articles**. Proc. of *the 4th Int. Conf. on Language Resources and Evaluation (LREC'04)*, pp. 1737-1740.

Morgan A., Yeh A., Hirschman L. and Colosimo M. (2003) **Gene Name Extraction Using FlyBase resources**. Proc. of *the Workshop for NLP I Biomedicine (at ACL'03)*, pp. 1-8.

Nakagawa T., Kudoh T. and Matsumoto Y. (2001). **Unknown word guessing and part-of-speech tagging using support vector machines**. Proc. of *the 6th Natural Language Processing Pacific Rim Symposium (NL-PRS'01)*, pp. 325-331.

Narayanaswamy M., Ravikumar K. E. and Vijay-Shanker K. (2003). **A Biological Named Entity Recognizer**. Proc. of *the Pacific Symposium on Biocomputing (PSB'03)*, pp. 427-438.

Nenadic G., Ananiadou S. and McNaught J. (2004). **Enhancing automatic term recognition through recognition of variation.** Proc. of *the 20th Conference on Computational Linguistics (COLING'04)*, pp. 604-610.

Ng S. K. and Wong M. (1999). **Toward routine automatic pathway discovery from on-line scientific text abstracts.** *Genome Inform Ser. Workshop Genome Inform*, pp. 104-112.

Nobata C., Collier N. and Tsujii J. (1999). **Automatic term identification and classification in biology texts.** Proc. of *the Natural Language Pacific Rim Symposium (NL-PRS'99)*, pp. 369-374.

Noble W. S. (2004). **Support vector machine applications in computational biology.** *Kernel methods in computational biology*, The MIT Press, ISBN 0-262-19509-7.

Novichkova S., Egorov S. and Daraselia N. (2003). **MedScan, a natural language processing engine for MEDLINE abstracts.** *Bioinformatics* 19(13):1699-1706.

Ohta T., Tateishi Y., Mima H. and Tsujii J. (2002). **The GENIA Corpus: An Annotated Research Abstract Corpus in the Molecular Biology Domain.** Proc. of *the Human Language Technologies Conference (HLT'02)*, pp. 73-77.

Ono T., Hishigaki H., Tanigami A. and Takagi T. (2001). **Automated extraction of information on protein-protein interactions from the biological literature.** *Bioinformatics* 17(2):155-161.

Park K. M., Kim S. H., Lee K. J., Lee D. G. and Rim H. C. (2004). **Incorporating Lexical Knowledge into Biomedical NE Recognition.** Proc. of *the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA'04)*, pp. 76-79.

Proux D., Rechenmann F., Julliard L., Piller V. and Jacq B. (1998). **Detecting Gene Symbols and Names in Biological Texts: A First Step toward Pertinent Information Extraction**. Proc. of *the 9th Workshop Genome Informatics*, pp. 72-80.

Pruitt K. D. and Maglott D. R. (2001). **RefSeq and LocusLink: NCBI gene-centered resources**. *Nucleic Acids Res.* 29:137-140.

Pustejovsky J., Castano J. and Zhang J. (2002). **Robust Relational parsing over Biomedical Literature: Extracting Inhibit Relations**. Proc. of *the Pacific Symposium on Biocomputing (PSB '02)*, pp.505-516.

Quinlan J. R. (1990). **Learning logical definitions from relations**. *Machine Learning* 5:239-266.

Ratnaparkhi A. (1998). **Maximum Entropy Models for Natural Language Ambiguity Resolution**. *PhD thesis*, Computer and Information Science, University of Pennsylvania.

Rindflesch T. C., Hunter L. and Aronson A. R. (1999). **Mining molecular binding terminology from biomedical text**. Proc. of *AMIA Symposium*, pp. 127-131.

Rindflesch T. C., Rajan J. V. and Hunter L. (2000). **Extracting Molecular Binding Relationships from Biomedical Text**. Proc. of *the 6th Conf. on Applied Natural Language Processing (ANLP '00)*, pp. 188-195.

Rindflesch T. C., Tanabe L., Weinstein J. N. and Hunter L. (2000). **Edgar: extraction of drugs, genes, and relations from the biomedical literature**. Proc. of *the Pacific Symposium on Biocomputing (PSB '00)*, pp. 514-525.

Rossler M. (2004). **Adapting an NER-System for German to the Biomedical Domain**. Proc. of *the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA '04)*, pp. 92-95.

Rzhetsky A., Iossifov I., Koike T., Krauthammer M., Kra P., Morris M., Yu H., Duboue P. A., Weng W., Wilbur W. J., Hatzivassiloglou V. and Friedman C. (2004). **Geneways: a system for extracting, analyzing, visualizing, and integrating molecular pathway data**. *Journal of Biomedical Informatics* 37(1):43-53.

Scolkopf B., Sung K., Burges C., Girosi F., Niyogi P., Poggio T. and Vapnik V. (1997). **Comparing Support Vector Machines with Gaussian Kernels to Radial Basis Function Classifiers**. *IEEE Transaction of Signal Processing* 45:2758-2765.

Schuemie M. J., Weeber M., Schijvennaars B. J. A., van Mulligen E. M., van der Eijk C. C., Jelier R., Mons B. and Kors J. A. (2004). **Distribution of information in biomedical abstracts and full-text publications**. *Bioinformatics* 20(16):2597-2604.

Settles B. (2004). **Biomedical Named Entity Recognition Using Conditional Random Fields and Novel Feature Sets**. Proc. of *the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA'04)*, pp. 104-107.

Shah P. K., Perez-Iratxeta C., Bork P. and Andrade M. A. (2003). **Information extraction from full text scientific articles: Where are the keywords?** *BMC Bioinformatics* 4(20).

Shen D., Zhang J., Zhou G., Su J. and Tan C. (2003) **Effective Adaptation of Hidden Markov Model based Named Entity Recognizer for Biomedical Domain**. Proc. of *the Workshop on NLP in Biomedicine (at ACL'03)*, pp. 49-56.

Shi L. and Campagne F. (2005). **Building a protein name dictionary from full text: a machine learning term extraction approach**. *BMC Bioinformatics* 6(88).

Song Y., Kim E., Lee G. G. and Yi B. K. (2004). **POSBIOTM-NER in the shared task of BioNLP/NLPBA 2000**. Proc. of *the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA-2004)*, pp. 100-103.

Spasi I., Nenadic G., and Ananiadou S. (2003). **Using domain-Specific Verbs for Term Classification**. Proc. of *the Workshop on NLP in Biomedicine (at ACL '03)*, pp. 17-24.

Stapley B. J. and Benoit G. (2000). **Biobibliometrics: Information retrieval and visualization from co-occurrences of gene names in Medline abstracts**. Proc. of *the Pacific Symposium on Biocomputing (PSB '00)*, pp. 529-540.

Takeuchi K. and Collier N. (2002). **Use of Support Vector Machines in Extended Named Entity Recognition**. Proc. of *the 6th Conf. on Natural Language Learning (CoNLL '02)*, pp. 119-125.

Takeuchi K. and Collier N. (2003). **Bio-medical Entity Extraction using Support Vector Machines**. Proc. of *the Workshop on NLP in Biomedicine (at ACL '2003)*, pp. 57-64.

Tanabe L. and Wilbur W. J. (2002). **Tagging gene and protein names in biomedical text**. *Bioinformatics* 17: 1053-1057.

Tapanainen P. and Jarvinen T. (1997). **A non-projective dependency parser**. Proc. of *the 5th Conf. on Applied Natural Language Processing (ANLP'97)*, pp. 64-71.

The Gene Ontology Consortium. (2000). **Gene ontology: Tool for the unification of biology**. *Nature Genetics* 25:25-29.

Thomas J., Milward D., Ouzounis C., Pulman S. and Carroll M. (2000). **Automatic extraction of protein interactions from scientific abstracts**. Proc. of *the Pacific Symposium on Biocomputing (PSB '00)*, pp. 541-552.

Tsuruoka Y. and Tsujii J. (2003). **Probabilistic Term Variant Generator for Biomedical Terms**. Proc. of *the Workshop on NLP in Biomedicine (at ACL '03)*, pp. 41-48.

Tuason O., Chen L., Lui H., Blake J. A. and Friedman C. (2004). **Biological Nomenclature: A Source of Lexical Knowledge and Ambiguity**. Proc. of *the Pacific Symposium on Biocomputing (PSB'04)*, pp. 238-249.

van Rijsbergen C. J. (1979). *Information Retrieval*, Butterworths, London.

Vapnik V. (1995). *The Natural of Statistical Learning* Theory, Springer-Verlag, New York.

Vapnik V. (1998). *The Statistical Learning Theory*, Wiley, New York.

Wheeler D. L., Chappey C., Lash A. E., Leipe D. D., Madden T. L., Schuler G. D., Tatusova T. A. and Rapp B. A. (2000). **Database resources of the National Center for Biotechnology Information**. *Nucleic Acids Res.* 28:10-14.

Willett R. (1988). **Recent trends in hierarchic document clustering: a critical review**. *Information Processing & Management*, 25:577.

Yakushiji A., Tateisi Y., Miyao Y. and Tsujii J. (2001). **Event extraction from biomedical papers using a full parser**. Proc. of *the Pacific Symposium on Biocomputing (PSB'98)*, pp. 408-419.

Yamamoto K., Kudo T., Konagaya A. and Matsumoto Y. (2003). **Protein Name Tagging for Biomedical Annotation in Text**. Proc. of *the Workshop on NLP in Biomedicine (at ACL' 03)*, pp. 65-72.

Yamamoto K., Kudo T., Konagaya A. and Matsumoto Y. (2004). **Use of morphological analysis in protein name recognition**. *Journal of Biomedical Informatics* 37(6):471-482.

Yu H., Hatzivassiloglou V., Friedman C., Rzhetsky A. and Wilbur W. J. (2002). **Automatic Extraction of Gene and Protein Synonyms from MEDLINE and Journal**

Articles. Proc. of *American Medical Informatics Association Symposium (AMIA '02)*, pp. 919-923.

Zhang J., Shen D. Zhou G., Jian S. and Tan C. L. (2004). **Enhancing HMM-based Biomedical Named Entity Recognition by Studying Special Phenomena**. *Journal of Biomedical Informatics (special issue on Natural Language Processing in Biomedicine:Aims, Achievements and Challenge)* 37(6): 411-422.

Zhao S. (2004). **Named Entity Recognition in Biomedical Text using an HMM model**. Proc. of *the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA '04)*, pp. 84-87.

Zhou G. D. and Su J. (2004). **Exploring Deep Knowledge Resources in Biomedical Name Recognition**. In Proc. of *the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA '04)*, pp. 96-99.

Zhou G., Zhang J., Su J., Shen D. and Tan C. L. (2004). **Recognizing Names in Biomedical Texts: a Machine Learning Approach**. *Bioinformatics* 20(7):1178-1190.

Zien A., Ratsch G., Mika S., Scholkopf B., Lemmen C., Smola A., Lengauer T. and Muller K. (2003). **Engineering support vector machine kernels that recognize translation initiation sites**. *Bioinformatics* 16(9):799-807.

Appendix A – a list of all acronyms

The following is a list of acronyms used in this thesis.

ADVP: Adverb Phrase

EMBO: European Molecular Biology Organization

FDG: Functional Dependency Grammar

GO: Gene Ontology

IE: Information Extraction

JV: Journal of Virology

MUC: Message Understanding Conference

NAR: Nucleic Acids Research

NER: Named Entity Recognition

NP: Noun Phrase

PAS: Predicate-Argument Structure

PNAS: Proceedings of the National Academy of Sciences of the United States of America

PP: Prepositional Phrase

VP: Verb Phrase

Appendix B – PASBio tagging labels

The following is a list of tagging labels used for annotating predicate-argument structure in PASBio project.

ADV: An adverb of any types except a manner adverb

ArgR: An argument giving information about consequences after the occurrence of the event denoted by the predicate in focus

Arg*X*: An argument number *X*

MAN: A manner adverb (e.g., normally, specifically)

MOD: A modal verb (e.g., will, may, can, shall, must)

NEG: A negator word (e.g., not or n't)

Appendix C – PASBio frames

All PASBio frames and annotated sentences corresponding to each frame are given below and are accessible publicly from http://research.nii.ac.jp/~collier/projects/PASBio/.

Declaration

I declare that the work presented in this thesis is, to the best of my knowledge and belief, original and my own work. The material has not been submitted, either in whole or part, for a degree at this or any other universities.

Tuangthong Wattarujeekrit

Tuangthong Wattarujeekrit