A Semi-Automatic Method for Annotating a Biomedical Proposition Bank

Wen-Chi Chou[1], Richard Tzong-Han Tsai[1,2], Ying-Shan Su[1], Wei Ku[1,3], Ting-Yi Sung[1] and Wen-Lian Hsu[1]

[1]Institute of Information Science, Academia Sinica, Taiwan, ROC.

[2]Dept. of Computer Science and Information Engineering, National Taiwan University, Taiwan, ROC.

[3]Institute of Molecular Medicine, National Taiwan University, Taiwan, ROC.

{jacky957,thtsai,qnn,wilmaku,tsung,hsu}@iis.sinica.edu.tw

Abstract

In this paper, we present a semi-automatic approach for annotating semantic information in biomedical texts. The information is used to construct a biomedical proposition bank called BioProp. Like PropBank in the newswire domain, BioProp contains annotations of predicate argument structures and semantic roles in a treebank schema. To construct BioProp, a semantic role labeling (SRL) system trained on PropBank is used to annotate BioProp. Incorrect tagging results are then corrected by human annotators. To suit the needs in the biomedical domain, we modify the PropBank annotation guidelines and characterize semantic roles as components of biological events. The method can substantially reduce annotation efforts, and we introduce a measure of an upper bound for the saving of annotation efforts. Thus far, the method has been applied experimentally to a 4,389-sentence treebank corpus for the construction of BioProp. Inter-annotator agreement measured by kappa statistic reaches .95 for combined decision of role identification and classification when all argument labels are considered. In addition, we show that, when trained on BioProp, our biomedical SRL system called BIOSMILE achieves an F-score of 87%.

1 Introduction

The volume of biomedical literature available on the Web has grown enormously in recent years, a trend that will probably continue indefinitely. Thus, the ability to process literature automatically would be invaluable for both the design and interpretation of large-scale experiments. To this end, several information extraction (IE) systems using natural language processing techniques have been developed for use in the biomedical field. Currently, the focus of IE is shifting from the extraction of nominal information, such as named entities (NEs) to verbal information that represents the relations between NEs, e.g., events and function (Tateisi et al., 2004; Wattarujeekrit et al., 2004). In the IE of relations, the roles of NEs participating in a relation must be identified along with a verb of interest. This task involves identifying main roles, such as agents and objects, and adjunct roles (ArgM), such as location, manner, timing, condition, and extent. This identification task is called *semantic role labeling* (SRL). The corresponding roles of the verb (*predicate*) are called *predicate arguments*, and the whole proposition is known as a *predicate argument structure* (PAS).

To develop an automatic SRL system for the biomedical domain, it is necessary to train the system with an annotated corpus, called *proposition bank* (Palmer et al., 2005). This corpus contains annotations of semantic PAS's superimposed on the Penn Treebank (PTB) (Marcus et al., 1993; Marcus et al., 1994). However, the process of manually annotating the PAS's to construct a proposition bank is quite time-consuming. In addition, due to the complexity of proposition bank annotation, inconsistent annotation may occur frequently and further complicate

Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora 2006, pages 5–12,
Sydney, July 2006.

the annotation task. In spite of the above difficulties, there are proposition banks in the newswire domain that are adequate for training SRL systems (Xue and Palmer, 2004; Palmer et al., 2005). In addition, according to the CoNLL-2005 shared task (Carreras and Màrquez, 2005), the performance of SRL systems in general does not decline significantly when tagging out-of-domain corpora. For example, when SRL systems trained on the Wall Street Journal (WSJ) corpus were used to tag the Brown corpus, the performance only dropped by 15%, on average. In comparison to annotating from scratch, annotation efforts based on the results of an available SRL system are much reduced. Thus, we plan to use a newswire SRL system to tag a biomedical corpus and then manually revise the tagging results. This semi-automatic procedure could expedite the construction of a biomedical proposition bank for use in training a biomedical SRL system in the future.

2 The Biomedical Proposition Bank - BioProp

As proposition banks are semantically annotated versions of a Penn-style treebank, they provide consistent semantic role labels across different syntactic realizations of the same verb. The annotation captures predicate-argument structures based on the sense tags of polysemous verbs (called *framesets*) and semantic role labels for each argument of the verb. Figure 1 shows the annotation of semantic roles, exemplified by the following sentence: "IL4 and IL13 receptors activate STAT6, STAT3 and STAT5 proteins in normal human B cells." The chosen predicate is the word "activate"; its arguments and their associated word groups are illustrated in the figure.

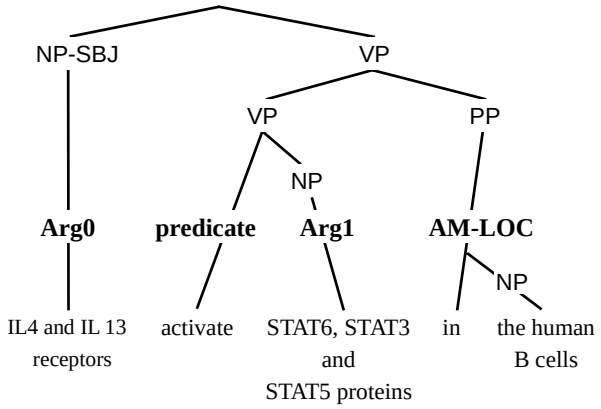

Figure 1. A treebank annotated with semantic role labels

Since proposition banks are annotated on top of a Penn-style treebank, we selected a biomedical corpus that has a Penn-style treebank as our corpus. We chose the GENIA corpus (Kim et al., 2003), a collection of MEDLINE abstracts selected from the search results with the following keywords: human, blood cells, and transcription factors. In the GENIA corpus, the abstracts are encoded in XML format, where each abstract also contains a MEDLINE UID, and the title and content of the abstract. The text of the title and content is segmented into sentences, in which biological terms are annotated with their semantic classes. The GENIA corpus is also annotated with part-of-speech (POS) tags (Tateisi and Tsujii, 2004), and co-references are added to part of the GENIA corpus by the MedCo project at the Institute for Infocomm Research, Singapore (Yang et al., 2004).

The Penn-style treebank for GENIA, created by Tateisi et al. (2005), currently contains 500 abstracts. The annotation scheme of the GENIA Treebank (GTB), which basically follows the Penn Treebank II (PTB) scheme (Bies et al., 1995), is encoded in XML. However, in contrast to the WSJ corpus, GENIA lacks a proposition bank. We therefore use its 500 abstracts with GTB as our corpus. To develop our biomedical proposition bank, BioProp, we add the proposition bank annotation on top of the GTB annotation.

In the following, we report on the selection of biomedical verbs, and explain the difference between their meaning in PropBank (Palmer et al., 2005), developed by the University of Pennsylvania, and their meaning in BioProp (a biomedical proposition bank). We then introduce BioProp's annotation scheme, including how we modify a verb's framesets and how we define framesets for biomedical verbs not defined in VerbNet (Kipper et al., 2000; Kipper et al., 2002).

2.1 Selection of Biomedical Verbs

We selected 30 verbs according to their frequency of use or importance in biomedical texts. Since our targets in IE are the relations of NEs, only sentences containing protein or gene names are used to count each verb's frequency. Verbs that have general usage are filtered out in order to ensure the focus is on biomedical verbs. Some verbs that do not have a high frequency, but play important roles in describing biomedical relations, such as "phosphorylate" and "transactivate", are also selected. The selected verbs are listed in Table 1.

Predicate	Frameset	Example
express (VerbNet)	**Arg0:** agent **Arg1:** theme **Arg2:** recipient or destination	[Some legislators$_{Arg0}$][expressed$_{predicate}$] [concern that a gas-tax increase would take too long and possibly damage chances of a major gas-tax-increasing ballot initiative that voters will consider next June$_{Arg1}$].
translate (VerbNet)	**Arg0:** causer of transformation **Arg1:** thing changing **Arg2:** end state **Arg3:** start state	But some cosmetics-industry executives wonder whether [techniques honed in packaged goods$_{Arg1}$] [will$_{AM-MOD}$] [translate$_{predicate}$] [to the cosmetics business$_{Arg2}$].
express (BioProp)	**Arg0:** causer of expression **Arg1:** thing expressing	[B lymphocytes and macrophages$_{Arg0}$] [express$_{predicate}$] [closely related immunoglobulin G (IgG) Fc receptors (Fc gamma RII) that differ only in the structures of their cytoplasmic domains$_{Arg1}$].

Table 2. Framesets and examples of "express" and "translate"

Type	Verb list
1	encode, interact, phosphorylate, transactivate
2	express, modulate
3	bind
4	activate, affect, alter, associate, block, decrease differentiate, encode, enhance, increase, induce, inhibit, mediate, mutate, prevent, promote, reduce, regulate, repress, signal, stimulate, suppress, transform, trigger

Table 1. Selected biomedical verbs and their types

2.2 Framesets of Biomedical Verbs

Annotation of BioProp is mainly based on Levin's verb classes, as defined in the VerbNet lexicon (Kipper et al., 2000). In VerbNet, the arguments of each verb are represented at the semantic level, and thus have associated semantic roles. However, since some verbs may have different usages in biomedical and newswire texts, it is necessary to customize the framesets of biomedical verbs. The 30 verbs in Table 1 are categorized into four types according to the degree of difference in usage: (1) verbs that do not appear in VerbNet due to their low frequency in the newswire domain; (2) verbs that do appear in VerbNet, but whose biomedical meanings and framesets are undefined; (3) verbs that do appear in VerbNet, but whose primary newswire and biomedical usage differ; (4) verbs that have the same usage in both domains.

Verbs of the first type play important roles in biomedical texts, but rarely appear in newswire texts and thus are not defined in VerbNet. For example, "phosphorylate" increasingly appears in the fast-growing PubMed abstracts that report experimental results on phosphorylated events; therefore, it is included in our verb list. However, since VerbNet does not define the frameset for "phosphorylate", we must define it after analyzing all the sentences in our corpus that contain the verb. Other type 1 verbs may correspond to verbs in VerbNet; in such cases, we can borrow the VerbNet definitions and framesets. For example, "transactivate" is not found in VerbNet, but we can adopt the frameset of "activate" for this verb.

Verbs of the second type appear in VerbNet, but have unique biomedical meanings that are undefined. Therefore, the framesets corresponding to their biomedical meanings must be added. In most cases, we can adopt framesets from VerbNet synonyms. For example, "express" is defined as "say" and "send very quickly" in VerbNet. However, in the biomedical domain, its usage is very similar to "translate". Thus, we can use the frameset of "translate" for "express". Table 2 shows the framesets and corresponding examples of "express" in the newswire domain and biomedical domain, as well as that of "translate" in VerbNet.

Verbs of the third type also appear in VerbNet. Although the newswire and biological senses are defined therein, their primary newswire sense is not the same as their primary biomedical sense. "Bind," for example, is common in the newswire domain, and it usually means "to tie" or "restrain with bonds." However, in the biomedical domain, its intransitive use- "attach or stick to"- is far more common. For example, a Google search for the phrase "glue binds to" only returned 21 results, while the same search replacing "glue" with "protein" yields 197,000 hits. For such verbs, we only need select the appropriate alternative meanings and corresponding framesets. Lastly, for verbs of the fourth type, we can di-

rectly adopt the newswire definitions and framesets, since they are identical.

2.3 Distribution of Selected Verbs

There is a significant difference between the occurrence of the 30 selected verbs in biomedical texts and their occurrence in newswire texts. The verbs appearing in verb phrases constitute only 1,297 PAS's, i.e., 1% of all PAS's, in PropBank (shown in Figure 2), compared to 2,382 PAS's, i.e., 16% of all PAS's, in BioProp (shown in Figure 3). Furthermore, some biomedical verbs have very few PAS's in PropBank, as shown in Table 3. The above observations indicate that it is necessary to annotate a biomedical proposition bank for training a biomedical SRL system.

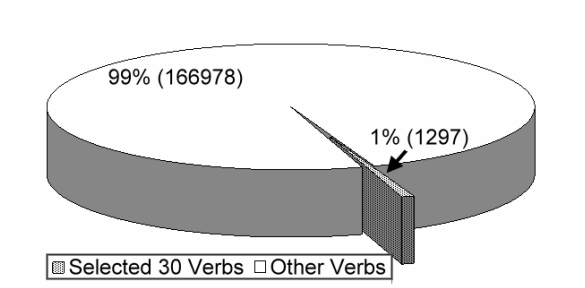

Figure 2. The percentage of the 30 biomedical verbs and other verbs in PropBank

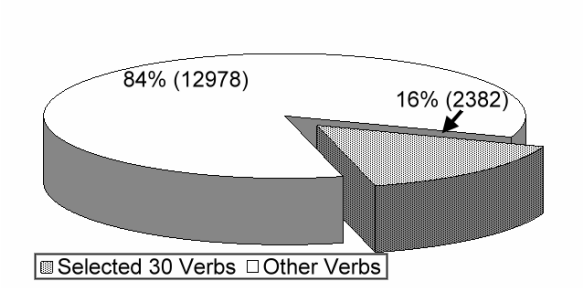

Figure 3. The percentage of the 30 biomedical verbs and other verbs in BioProp

3 Annotation of BioProp

3.1 Annotation Process

After choosing 30 verbs as predicates, we adopted a semi-automatic method to annotate BioProp. The annotation process consists of the following steps: (1) identifying predicate candidates; (2) automatically annotating the biomedical semantic roles with our WSJ SRL system; (3) transforming the automatic tagging results into *WordFreak* (Morton and LaCivita, 2003) format; and (4) manually correcting the annotation results with the *WordFreak* annotation tool. We now describe these steps in detail:

Verbs	# in BioProp	Ratio(%)	# in PropBank	Ratio(%)
induce	290	1.89	16	0.01
bind	252	1.64	0	0
activate	235	1.53	2	0
express	194	1.26	53	0.03
inhibit	184	1.20	6	0
increase	166	1.08	396	0.24
regulate	122	0.79	23	0.01
mediate	104	0.68	1	0
stimulate	93	0.61	11	0.01
associate	82	0.53	51	0.03
encode	79	0.51	0	0
affect	60	0.39	119	0.07
enhance	60	0.39	28	0.02
block	58	0.38	71	0.04
reduce	55	0.36	241	0.14
decrease	54	0.35	16	0.01
suppress	38	0.25	4	0
interact	36	0.23	0	0
alter	27	0.18	17	0.01
transactivate	24	0.16	0	0
modulate	22	0.14	1	0
phosphorylate	21	0.14	0	0
transform	21	0.14	22	0.01
differentiate	21	0.14	2	0
repress	17	0.11	1	0
prevent	15	0.10	92	0.05
promote	14	0.09	52	0.03
trigger	14	0.09	40	0.02
mutate	14	0.09	1	0
signal	10	0.07	31	0.02

Table 3. The number and percentage of PAS's for each verb in BioProp and PropBank

1. Each word with a VB POS tag in a verb phrase that matches any lexical variant of the 30 verbs is treated as a predicate candidate. The automatically selected targets are then double-checked by human annotators. As a result, 2,382 predicates were identified in BioProp.

2. Sentences containing the above 2,382 predicates were extracted and labeled automatically by our WSJ SRL system. In total, 7,764 arguments were identified.

3. In this step, sentences with PAS annotations are transformed into *WordFreak* format (an XML format), which allows annotators to view a sentence in a tree-like fashion. In addition, users can customize the tag set of arguments. Other linguistic information can also be integrated and displayed in

WordFreak, which is a convenient annotation tool.

4. In the last step, annotators check the predicted semantic roles using *WordFreak* and then correct or add semantic roles if the predicted arguments are incorrect or missing, respectively. Three biologists with sufficient biological knowledge in our laboratory performed the annotation task after receiving computational linguistic training for approximately three months.

Figure 4 illustrates an example of BioProp annotation displayed in *WordFreak* format, using the frameset of "phophorylate" listed in Table 4.

This annotation process can be used to construct a domain-specific corpus when a general-purpose tagging system is available. In our experience, this semi-automatic annotation scheme saves annotation efforts and improves the annotation consistency.

Predicate	Frameset
phosphorylate	**Arg0:** causer of phosphorylation **Arg1:** thing being phosphorylated **Arg2:** end state **Arg3:** start state

Table 4. The frameset of "phosphorylate"

3.2 Inter-annotation Agreement

We conducted preliminary consistency tests on 2,382 instances of biomedical propositions. The inter-annotation agreement was measured by the kappa statistic (Siegel and Castellan, 1988), the definition of which is based on the probability of inter-annotation agreement, denoted by $P(A)$, and the agreement expected by chance, denoted by $P(E)$. The kappa statistics for inter-annotation agreement were .94 for semantic role identification and .95 for semantic role classification when ArgM labels were included for evaluation. When ArgM labels were omitted, kappa statistics were .94 and .98 for identification and classification, respectively. We also calculated the results of combined decisions, i.e., identification and classification. (See Table 5.)

3.3 Annotation Efforts

Since we employ a WSJ SRL system that labels semantic roles automatically, human annotators can quickly browse and determine correct tagging results; thus, they do not have to examine

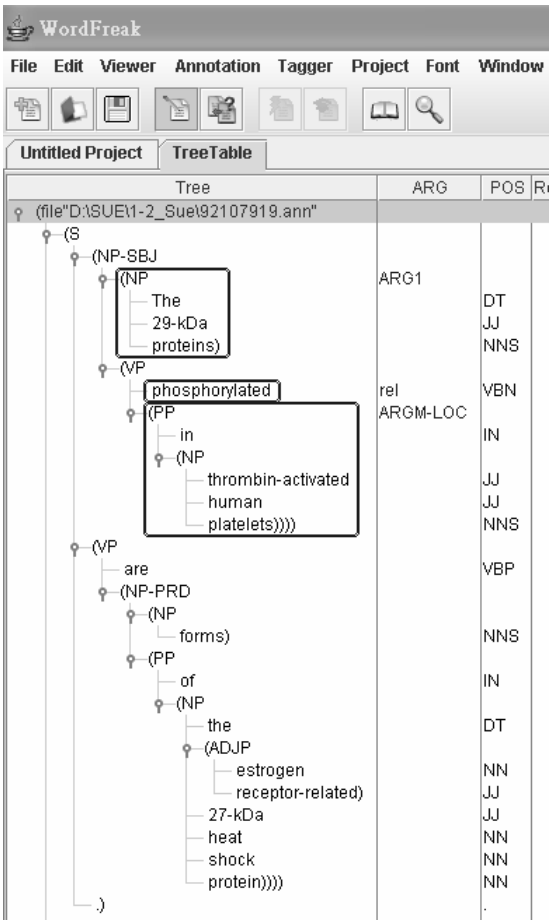

Figure 4. An example of BioProp displayed with *WordFreak*

		$P(A)$	$P(E)$	Kappa score
including ArgM	role identification	.97	.52	.94
	role classification	.96	.18	.95
	combined decision	.96	.18	.95
excluding ArgM	role identification	.97	.26	.94
	role classification	.99	.28	.98
	combined decision	.99	.28	.98

Table 5. Inter-annotator agreement

all tags during the annotation process, as in the full manual annotation approach. Only incorrectly predicted tags need to be modified, and missed tags need to be added. Therefore, annotation efforts can be substantially reduced. To quantify the reduction in annotation efforts, we define the saving of annotation effort, ρ, as:

$$\rho = \frac{\#\text{ of correctly labeled nodes}}{\#\text{ of all nodes}} < \frac{\#\text{ of correctly labeled nodes}}{\#\text{ of correct} + \#\text{ of incorrect} + \#\text{ of missed nodes}} \quad (1)$$

In Equation (1), since the number of nodes that need to be examined is usually unknown, we

use an easy approximation to obtain an upper bound for ρ. This is based on the extremely optimistic assumption that annotators should be able to recover a missed or incorrect label by only checking one node. However, in reality, this would be impossible. In our annotation process, the upper bound of ρ for BioProp is given by:

$$\rho < \frac{18932}{18932 + 6682 + 15316} = \frac{18932}{40975} = 46\%,$$

which means that, at most, the annotation effort could be reduced by 46%.

A more accurate tagging system is preferred because the more accurate the tagging system, the higher the upper bound ρ will be.

4 Disambiguation of Argument Annotation

During the annotation process, we encountered a number of problems resulting from different usage of vocabulary and writing styles in general English and the biomedical domain. In this section, we describe three major problems and propose our solutions.

4.1 Cue Words for Role Classification

PropBank annotation guidelines provide a list of words that can help annotators decide an argument's type. Similarly, we add some rules to our BioProp annotation guideline. For example, "in vivo" and "in vitro" are used frequently in biomedical literature; however, they seldom appear in general English articles. According to their meanings, we classify them as location argument (AM-LOC).

In addition, some words occur frequently in both general English and in biomedical domains but have different meanings/usages. For instance, "development" is often tagged as Arg0 or Arg1 in general English, as shown by the following sentence:

Despite the strong case for stocks, however, most pros warn that [individuals$_{\text{Arg0}}$] shouldn't try to [profit$_{\text{predicate}}$] [from short-term developments$_{\text{Arg1}}$].

However, in the biomedical domain, "development" always means the stage of a disease, cell, etc. Therefore, we tag it as temporal argument (AM-TMP), as shown in the following sentence:

[Rhom-2 mRNA$_{\text{Arg1}}$] is [expressed$_{\text{predicate}}$] [in early mouse development$_{\text{AM-TMP}}$] [in central nervous system, lung, kidney, liver, and spleen but only very low levels occur in thymus$_{\text{AM-LOC}}$].

4.2 Additional Argument Types

In PropBank, the negative argument (AM-NEG) usually contains explicit negative words such as "not". However, in the biomedical domain, researchers usually express negative meaning implicitly by using "fail", "unable", "inability", "neither", "nor", "failure", etc. Take "fail" as an example. It is tagged as a verb in general English, as shown in the following sentence:

But [the new pact$_{\text{Arg1}}$] will force huge debt on the new firm and [could$_{\text{AM-MOD}}$] [still$_{\text{AM-TMP}}$] [fail$_{\text{predicate}}$] [to thwart rival suitor McCaw Cellular$_{\text{Arg2}}$].

Negative results are important in the biomedical domain. Thus, for annotation purposes, we create additional negation tag (AM-NEG1) that does not exist in PropBank. The following sentence is an example showing the use of AM-NEG1:

[They$_{\text{Arg0}}$] [fail$_{\text{AM-NEG1}}$] to [induce$_{\text{predicate}}$] [mRNA of TNF-alpha$_{\text{Arg1}}$] [after 3 h of culture $_{\text{AM-TMP}}$].

In this example, if we do not introduce the AM-NEG1, "fail" is considered as a verb like in PropBank, not as a negative argument, and it will not be included in the proposition for the predicate "induce". Thus, BioProp requires the "AM-NEG1" tag to precisely express the corresponding proposition.

4.3 Essentiality of Biomedical Knowledge

Since PAS's contain more semantic information, proposition bank annotators require more domain knowledge than annotators of other corpora. In BioProp, many ambiguous expressions require biomedical knowledge to correctly annotate them, as exemplified by the following sentence in BioProp:

In the cell types tested, the LS mutations indicated an apparent requirement not only for the intact NF-kappa B and SP1-binding sites but also for [several regions between -201 and -130$_{\text{Arg1}}$] [not$_{\text{AM-NEG}}$] [previously$_{\text{AM-MNR}}$] [associated$_{\text{predicate}}$][with viral infectivity$_{\text{Arg2}}$].

Annotators without biomedical knowledge may consider [between -201 and -130] as extent argument (AM-EXT), because the PropBank guidelines define numerical adjuncts as AM-

EXT. However, it means a segment of DNA. It is an appositive of [several regions]; therefore, it should be annotated as part of Arg1 in this case.

5 Effect of Training Corpora on SRL Systems

To examine the possibility that BioProp can improve the training of SRL systems used for automatic tagging of biomedical texts, we compare the performance of systems trained on BioProp and PropBank in different domains. We construct a new SRL system (called a BIOmedical SeMantIc roLe labEler, BIOSMILE) that is trained on BioProp and employs all the features used in our WSJ SRL system (Tsai et al., 2006).

As with POS tagging, chunking, and named entity recognition, SRL can also be formulated as a sentence tagging problem. A sentence can be represented by a sequence of words, a sequence of phrases, or a parsing tree; the basic units of a sentence in these representations are words, phrases, and constituents, respectively. Hacioglu et al. (2004) showed that tagging phrase-by-phrase (P-by-P) is better than word-by-word (W-by-W). However, Punyakanok et al. (2004) showed that constituent-by-constituent (C-by-C) tagging is better than P-by-P. Therefore, we use C-by-C tagging for SRL in our BIOSMILE.

SRL can be divided into two steps. First, we identify all the predicates. This can be easily accomplished by finding all instances of verbs of interest and checking their part-of-speech (POS) tags. Second, we label all arguments corresponding to each predicate. This is a difficult problem, since the number of arguments and their positions vary according to a verb's voice (active/passive) and sense, along with many other factors.

In BIOSMILE, we employ the maximum entropy (ME) model for argument classification. We use Zhang's MaxEnt toolkit (http://www.nlplab.cn/zhangle/maxent_toolkit.html) and the L-BFGS (Nocedal and Wright, 1999) method of parameter estimation for our ME model. Table 6 shows the features we employ in BIOSMILE and our WSJ SRL system.

To compare the effects of using biomedical training data versus using general English data, we train BIOSMILE on 30 randomly selected training sets from BioProp (g_1,.., g_{30}), and WSJ SRL system on 30 from PropBank (w_1,.., w_{30}), each of which has 1,200 training PAS's.

BASIC FEATURES
- **Predicate** – The predicate lemma
- **Path** – The syntactic path through the parsing tree from the parse constituent being classified to the predicate
- **Constituent type**
- **Position** – Whether the phrase is located before or after the predicate
- **Voice** – passive: If the predicate has a POS tag VBN, and its chunk is not a VP, or it is preceded by a form of "to be" or "to get" within its chunk; otherwise, it is active
- **Head word** – Calculated using the head word table described by Collins (1999)
- **Head POS** – The POS of the Head Word
- **Sub-categorization** – The phrase structure rule that expands the predicate's parent node in the parsing tree
- **First and last Word and their POS tags**
- **Level** – The level in the parsing tree

PREDICATE FEATURES
- **Predicate's verb class**
- **Predicate POS tag**
- **Predicate frequency**
- **Predicate's context POS**
- **Number of predicates**

FULL PARSING FEATURES
- **Parent's, left sibling's, and right sibling's paths, constituent types, positions, head words and head POS tags**
- **Head of PP parent –** If the parent is a PP, then the head of this PP is also used as a feature

COMBINATION FEATURES
- **Predicate distance combination**
- **Predicate phrase type combination**
- **Head word and predicate combination**
- **Voice position combination**

OTHERS
- **Syntactic frame of predicate/NP**
- **Headword suffixes of lengths 2, 3, and 4**
- **Number of words in the phrase**
- **Context words & POS tags**

Table 6. The features used in our argument classification model

We then test both systems on 30 400-PAS test sets from BioProp, with g_1 and w_1 being tested on test set 1, g_2 and w_2 on set 2, and so on. Then we generate the scores for g_1-g_{30} and w_1-w_{30}, and compare their averages.

Table 7 shows the experimental results. When tested on the biomedical corpus, BIOSMILE outperforms the WSJ SRL system by 22.9%. This result is statistically significant as expected.

Training	Test	Precision	Recall	F-score
PropBank	BioProp	74.78	56.25	64.20
BioProp	BioProp	88.65	85.61	87.10

Table 7. Performance comparison of SRL systems trained on BioProp and PropBank

6 Conclusion & Future Work

The primary contribution of this study is the annotation of a biomedical proposition bank that incorporates the following features. First, the choice of 30 representative biomedical verbs is made according to their frequency and importance in the biomedical domain. Second, since some of the verbs have different usages and others do not appear in the WSJ proposition bank, we redefine their framesets and add some new argument types. Third, the annotation guidelines in PropBank are slightly modified to suit the needs of the biomedical domain. Fourth, using appropriate argument types, framesets and annotation guidelines, we construct a biomedical proposition bank, BioProp, on top of the popular biomedical GENIA Treebank. Finally, we employ a semi-automatic annotation approach that uses an SRL system trained on the WSJ PropBank. Incorrect tagging results are then corrected by human annotators. This approach reduces annotation efforts significantly. For example, in BioProp, the annotation efforts can be reduced by, at most, 46%. In addition, trained on BioProp, BIOSMILE's F-score increases by 22.9% compared to the SRL system trained on the PropBank.

In our future work, we will investigate more biomedical verbs. Besides, since there are few biomedical treebanks, we plan to integrate full parsers in order to annotate syntactic and semantic information simultaneously. It will then be possible to apply the SRL techniques more extensively to biomedical relation extraction.

References

Ann Bies, Mark Ferguson, Karen Katz, and Robert MacIntyre. 1995. Bracketing Guidelines for Treebank II Style Penn Treebank Project. *Technical report, University of Pennsylvania*.

Xavier Carreras and Lluís Màrquez. 2005. Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling. *In Proceedings of CoNLL-2005*.

Michael Collins. 1999. Head-driven Statistical Models for Natural Language Parsing. Ph.D. thesis, University of Pennsylvania.

Kadri Hacioglu, Sameer Pradhan, Wayne Ward, James H. Martin, and Daniel Jurafsky. 2004. Semantic Role Labeling by Tagging Syntactic Chunks. *In Proceedings of CoNLL-2004*.

Jin-Dong Kim, Tomoko Ohta, Yuka Teteisi, and Jun'-ichi Tsujii. 2003. GENIA corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(Suppl. 1): i180-i182.

Karin Kipper, Hoa Trang Dang, and Martha Palmer. 2000. Class-based construction of a verb lexicon. *In Proceedings of AAAI-2000.*

Karin Kipper, Martha Palmer, and Owen Rambow. 2002. Extending PropBank with VerbNet semantic predicates. *In Proceedings of AMTA-2002.*

Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. 1994. The Penn Treebank: Annotating predicate argument structure. *In Proceedings of ARPA Human Language Technology Workshop.*

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19(2): 313-330.

Thomas Morton and Jeremy LaCivita. 2003. WordFreak: an open tool for linguistic annotation. *In Proceedings of HLT/NAACL-2003.*

Jorge Nocedal and Stephen J Wright. 1999. *Numerical Optimization*, Springer.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1).

Vasin Punyakanok, Dan Roth, Wen-tau Yih, and Dav Zimak. 2004. Semantic Role Labeling via Integer Linear Programming Inference. *In Proceedings of COLING-2004*.

Sidney Siegel and N. John Castellan. 1988. *Nonparametric Statistics for the Behavioral Sciences*. New York, McGraw-Hill.

Richard Tzong-Han Tsai, Wen-Chi Chou, Yu-Chun Lin, Cheng-Lung Sung, Wei Ku, Ying-Shan Su, Ting-Yi Sung, and Wen-Lian Hsu. 2006. BIOSMILE: Adapting Semantic Role Labeling for Biomedical Verbs: An Exponential Model Coupled with Automatically Generated Template Features. *In Proceedings of BioNLP'06*.

Yuka Tateisi, Tomoko Ohta, and Jun-ichi Tsujii. 2004. Annotation of Predicate-argument Structure of Molecular Biology Text. *In Proceedings of the IJCNLP-04 workshop on Beyond Shallow Analyses*.

Yuka Tateisi and Jun-ichi Tsujii. 2004. Part-of-Speech Annotation of Biology Research Abstracts. *In Proceedings of the 4th International Conference on Language Resource and Evaluation*.

Yuka Tateisi, Akane Yakushiji, Tomoko Ohta, and Jun-ichi Tsujii. 2005. Syntax Annotation for the GENIA corpus. *In Proceedings of IJCNLP-2005*.

Tuangthong Wattarujeekrit, Parantu K Shah, and Nigel Collier1. 2004. PASBio: predicate-argument structures for event extraction in molecular biology. *BMC Bioinformatics*, 5(155).

Nianwen Xue and Martha Palmer. 2004. Calibrating Features for Semantic Role Labeling. *In Proceedings of the EMNLP-2004*.

Xiaofeng Yang, Guodong Zhou, Jian Su, and Chew Lim Tan. 2004. Improving Noun Phrase Coreference Resolution by Matching Strings. *In Proceedings of 1st International Joint Conference on Natural Language Processing*: 226-233.