BIOSMILE: Adapting Semantic Role Labeling for Biomedical Verbs: An Exponential Model Coupled with Automatically Generated Template Features

Richard Tzong-Han Tsai[1,2], **Wen-Chi Chou**[1], **Yu-Chun Lin**[1,2], **Cheng-Lung Sung**[1], **Wei Ku**[1,3], **Ying-Shan Su**[1,4], **Ting-Yi Sung**[1] **and Wen-Lian Hsu**[1]

[1]Institute of Information Science, Academia Sinica

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

[3]Institute of Molecular Medicine, National Taiwan University

[4]Dept. of Biochemical Science and Technology, National Taiwan University

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

Abstract

In this paper, we construct a biomedical semantic role labeling (SRL) system that can be used to facilitate relation extraction. First, we construct a proposition bank on top of the popular biomedical GENIA treebank following the PropBank annotation scheme. We only annotate the predicate-argument structures (PAS's) of thirty frequently used biomedical predicates and their corresponding arguments. Second, we use our proposition bank to train a biomedical SRL system, which uses a maximum entropy (ME) model. Thirdly, we automatically generate argument-type templates which can be used to improve classification of biomedical argument types. Our experimental results show that a newswire SRL system that achieves an F-score of 86.29% in the newswire domain can maintain an F-score of 64.64% when ported to the biomedical domain. By using our annotated biomedical corpus, we can increase that F-score by 22.9%. Adding automatically generated template features further increases overall F-score by 0.47% and adjunct arguments (AM) F-score by 1.57%, respectively.

1 Introduction

The volume of biomedical literature available has experienced unprecedented growth in recent years. The ability to automatically process this literature would be an invaluable tool for both the design and interpretation of large-scale experiments. To this end, more and more information extraction (IE) systems using natural language processing (NLP) have been developed for use in the biomedical field. A key IE task in the biomedical field is extraction of relations, such as protein-protein and gene-gene interactions.

Currently, most biomedical relation-extraction systems fall under one of the following three approaches: cooccurence-based (Leroy et al., 2005), pattern-based (Huang et al., 2004), and machine-learning-based. All three, however, share the same limitation when extracting relations from complex natural language. They only extract the relation targets (e.g., proteins, genes) and the verbs representing those relations, overlooking the many adverbial and prepositional phrases and words that describe location, manner, timing, condition, and extent. The information in such phrases may be important for precise definition and clarification of complex biological relations.

The above problem can be tackled by using semantic role labeling (SRL) because it not only recognizes main roles, such as agents and objects, but also extracts adjunct roles such as location, manner,

timing, condition, and extent. The goal of SRL is to group sequences of words together and classify them with semantic labels. In the newswire domain, Morarescu et al. (2005) have demonstrated that full-parsing and SRL can improve the performance of relation extraction, resulting in an F-score increase of 15% (from 67% to 82%). This significant result leads us to surmise that SRL may also have potential for relation extraction in the biomedical domain. Unfortunately, no SRL system for the biomedical domain exists.

In this paper, we aim to build such a biomedical SRL system. To achieve this goal we roughly implement the following three steps as proposed by Wattarujeekrit et al., (2004): (1) create semantic roles for each biomedical verb; (2) construct a biomedical corpus annotated with verbs and their corresponding semantic roles (following definitions created in (1) as a reference resource;) (3) build an automatic semantic interpretation model using the annotated text as a training corpus for machine learning. In the first step, we adopt the definitions found in PropBank (Palmer et al., 2005), defining our own framesets for verbs not in PropBank, such as "phosphorylate". In the second step, we first use an SRL system (Tsai et al., 2005) trained on the Wall Street Journal (WSJ) to automatically tag our corpus. We then have the results double-checked by human annotators. Finally, we add automatically-generated template features to our SRL system to identify adjunct (modifier) arguments, especially those highly relevant to the biomedical domain.

2 Biomedical Proposition Bank

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 (Palmer et al., 2005). 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 the human B cells." The chosen predicate is the word "activate"; its arguments and their associated word groups are illustrated in the figure.

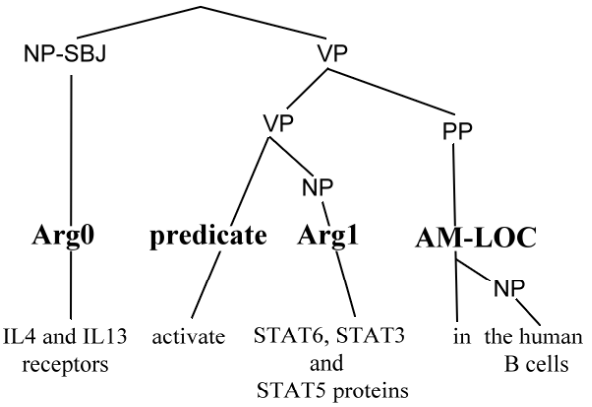

Figure 1. A Treebank Annotated with Semantic Role Labels

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 et al., 2004), and coreferences (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.

2.1 Important Argument Types

In the biomedical domain, relations are often dependent upon locative and temporal factors (Kholodenko, 2006). Therefore, locative (AM-LOC) and temporal modifiers (AM-TMP) are particularly important as they tell us where and when biomedical events take place. Additionally, nega-

tive modifiers (AM-NEG) are also vital to correctly extracting relations. Without AM-NEG, we may interpret a negative relation as a positive one or vice versa. In total, we use thirteen modifiers in our biomedical proposition bank.

2.2 Verb Selection

We select 30 frequently used verbs from the molecular biology domain given in Table 1.

express	trigger	encode
associate	repress	enhance
interact	signal	increase
suppress	activate	induce
prevent	alter	Inhibit
modulate	affect	Mediate
phosphorylate	bind	Mutated
transactivate	block	Reduce
transform	decrease	Regulate
differentiated	promote	Stimulate

Table 1. 30 Frequently Biomedical Verbs

Let us examine a representative verb, "activate". Its most frequent usage in molecular biology is the same as that in newswire. Generally speaking, "activate" means, "to start a process" or "to turn on." Many instances of this verb express the action of waking genes, proteins, or cells up. The following sentence shows a typical usage of the verb "activate."

[NF-kappaB $_{\text{Arg1}}$] is [not $_{\text{AM-NEG}}$] [activated $_{\text{predicate}}$] [upon tetracycline removal $_{\text{AM-TMP}}$] [in the NIH3T3 cell line $_{\text{AM-LOC}}$].

3 Semantic Role Labeling on BioProp

In this section, we introduce our BIOmedical SeMantIc roLe labEler, BIOSMILE. Like POS tagging, chunking, and named entity recognition, SRL can 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 are words, phrases, and constituents arranged in the above representations, respectively. Hacioglu et al. (2004) showed that tagging phrase by phrase (P-by-P) is better than word by word (W-by-W). Punyakanok et al., (2004) further showed that constituent-by-constituent (C-by-C) tagging is better than P-by-P. Therefore, we choose C-by-C tagging for SRL. The gold standard SRL corpus, PropBank, was designed as an additional layer of annotation on top of the syntactic structures of the Penn Treebank.

SRL can be broken into two steps. First, we must identify all the predicates. This can be easily accomplished by finding all instances of verbs of interest and checking their POS's.

Second, for each predicate, we need to label all arguments corresponding to the predicate. It is a complicated problem since the number of arguments and their positions vary depending on a verb's voice (active/passive) and sense, along with many other factors.

In this section, we first describe the maximum entropy model used for argument classification. Then, we illustrate basic features as well as specialized features such as biomedical named entities and argument templates.

3.1 Maximum Entropy Model

The maximum entropy model (ME) is a flexible statistical model that assigns an outcome for each instance based on the instance's history, which is all the conditioning data that enables one to assign probabilities to the space of all outcomes. In SRL, a history can be viewed as all the information related to the current token that is derivable from the training corpus. ME computes the probability, $p(o|h)$, for any o from the space of all possible outcomes, O, and for every h from the space of all possible histories, H.

The computation of $p(o|h)$ in ME depends on a set of binary features, which are helpful in making predictions about the outcome. For instance, the node in question ends in "cell", it is likely to be AM-LOC. Formally, we can represent this feature as follows:

$$f(h,o) = \begin{cases} 1 : \text{if current\_node\_ends\_in\_cell}(h) = \text{true} \\ \quad \text{and o} = \text{AM - LOC} \\ 0 : \text{otherwise} \end{cases}$$

Here, current_node_ends_in_cell(h) is a binary function that returns a true value if the current node in the history, h, ends in "cell". Given a set of features and a training corpus, the ME estimation process produces a model in which every feature f_i has a weight α_i. Following Bies et al. (1995), we can compute the conditional probability as:

$$p(o \mid h) = \frac{1}{Z(h)} \prod_i \alpha_i^{f_i(h,o)}$$

$$Z(h) = \sum_o \prod_i \alpha_i^{f_i(h,o)}$$

The probability is calculated by multiplying the weights of the active features (i.e., those of $f_i(h,o) = 1$). α_i is estimated by a procedure called Generalized Iterative Scaling (GIS) (Darroch et al., 1972). The ME estimation technique guarantees that, for every feature, f_i, the expected value of α_i equals the empirical expectation of α_i in the training corpus. We use Zhang's MaxEnt toolkit and the L-BFGS (Nocedal et al., 1999) method of parameter estimation for our ME model.

BASIC FEATURES • **Predicate** – The predicate lemma • **Path** – The syntactic path through the parsing tree from the parse constituent be-ing 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 2. The Features Used in the Baseline Argument Classification Model

3.2 Basic Features

Table 2 shows the features that are used in our baseline argument classification model. Their effectiveness has been previously shown by (Pradhan et al., 2004; Surdeanu et al., 2003; Xue et al., 2004). Detailed descriptions of these features can be found in (Tsai et al., 2005).

3.3 Named Entity Features

In the newswire domain, Surdeanu et al. (2003) used named entity (NE) features that indicate whether a constituent contains NEs, such as personal names, organization names, location names, time expressions, and quantities of money. Using these NE features, they increased their system's F-score by 2.12%. However, because NEs in the biomedical domain are quite different from newswire NEs, we create bio-specific NE features using the five primary NE categories found in the GENIA ontology[1]: protein, nucleotide, other organic compounds, source and others. Table 3 illustrates the definitions of these five categories. When a constituent exactly matches an NE, the corresponding NE feature is enabled.

NE	Definition
Protein	Proteins include protein groups, families, molecules, complexes, and substructures.
Nucleotide	A nucleic acid molecule or the compounds that consist of nucleic acids.
Other organic compounds	Organic compounds exclude protein and nucleotide.
Source	Sources are biological locations where substances are found and their reactions take place.
Others	The terms that are not categorized as sources or substances may be marked up, with

Table 3. Five GENIA Ontology NE Categories

3.4 Biomedical Template Features

Although a few NEs tend to belong almost exclusively to certain argument types (such as "…cell" being mainly AM-LOC), this information alone is not sufficient for argument-type classification. For one, most NEs appear in a variety of argument types. For another, many appear in more than one constituent (node in a parsing tree) in the same sentence. Take the sentence "IL4 and IL13 receptors activate STAT6, STAT3 and STAT5 proteins in the human B cells," for example. The NE "the human B cells" is found in two constituents ("the

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

human B cells" and "in the human B cells") as shown in figure 1. Yet only "in the human B cells" is an AM-LOC because here "human B cells" is preceded by the preposition "in" and the determiner "the". Another way to express this would be as a template—<prep> the <cell>." We believe such templates composed of NEs, real words, and POS tags may be helpful in identifying constituents' argument types. In this section, we first describe our template generation algorithm, and then explain how we use the generated templates to improve SRL performance.

Template Generation (TG)

Our template generation (TG) algorithm extracts general patterns for all argument types using the local alignment algorithm. We begin by pairing all arguments belonging to the same type according to their similarity. Closely matching pairs are then aligned word by word and a template that fits both is created. Each slot in the template is given constraint information in the form of either a word, NE type, or POS. The hierarchy of this constraint information is word > NE type > POS. If the arguments share nothing in common for a given slot, the TG algorithm will put a wildcard in that position. Figure 2 shows an aligned pair arguments. For this pair, the TG algorithm generated the template "AP-1 CC PTN" (PTN: protein name) because in the first position, both arguments have "AP-1;" in the second position, they have the same POS "CC;" and in the third position, they share a common NE type, "PTN." The complete TG algorithm is described in Algorithm 1.

AP-1/PTN/NN and/O/CC NF-AT/PTN/NN
AP-1/PTN/NN or/O/CC NFIL-2A/PTN/NN

Figure 2. Aligned Argument Pair

Applying Generated Templates

The generated templates may match exactly or partially with constituents. According to our observations, the former is more useful for argument classification. For example, constituents that perfectly match the template "IN a * <cell>" are overwhelmingly AM-LOCs. Therefore, we only accept exact template matches. That is, if a constituent exactly matches a template *t*, then the feature corresponding to *t* will be enabled.

Algorithm 1 Template Generation

Input: Sentences set $S = \{s_1, \ldots, s_n\}$,
Output: A set of template $T = \{t_1, \ldots, t_k\}$.

1: $T = \{\}$;
2: **for** each sentence s_i from s_1 to s_{n-1} **do**
3: **for** each sentence s_j from s_i to s_n **do**
4: perform ***alignment*** on s_i and s_j, then
5: pair arguments according to similarity;
6: generate common template ***t*** from argument pairs;
7: $T \leftarrow T \cup \boldsymbol{t}$;
8: **end**;
9: **end**;
10: return T;

4 Experiments

4.1 Datasets

In this paper, we extracted all our datasets from two corpora, the Wall Street Journal (WSJ) corpus and the BioProp, which respectively represent the newswire and biomedical domains. The Wall Street Journal corpus has 39,892 sentences, and 950,028 words. It contains full-parsing information, first annotated by Marcus et al. (1997), and is the most famous treebank (WSJ treebank). In addition to these syntactic structures, it was also annotated with predicate-argument structures (WSJ proposition bank) by Palmer et al. (2005).

In biomedical domain, there is one available treebank for GENIA, created by Yuka Tateshi et al. (2005), who has so far added full-parsing information to 500 abstracts. In contrast to WSJ, however, GENIA lacks any proposition bank.

Since predicate-argument annotation is essential for training and evaluating statistical SRL systems, to make up for GENIA's lack of a proposition bank, we constructed BioProp. Two biologists with masters degrees in our laboratory undertook the annotation task after receiving computational linguistic training for approximately three months.

We adopted a semi-automatic strategy to annotate BioProp. First, we used the PropBank to train a statistical SRL system which achieves an F-score of over 86% on section 24 of the PropBank. Next, we used this SRL system to annotate the GENIA treebank automatically. Table 4 shows the amounts of all adjunct argument types (AMs) in BioProp. The detail description of can be found in (Babko-Malaya, 2005).

Type	Description	#	Type	Description	#
NEG	negation marker	103	ADV	general purpose	307
LOC	location	389	PNC	purpose	3
TMP	time	145	CAU	cause	15
MNR	manner	489	DIR	direction	22
EXT	extent	23	DIS	discourse connectives	179
			MOD	modal verb	121

Table 4. Subtypes of the AM Modifier Tag

4.2 Experiment Design

Experiment 1: Portability

Ideally, an SRL system should be adaptable to the task of information extraction in various domains with minimal effort. That is, we should be able to port it from one domain to another. In this experiment, we evaluate the cross-domain portability of our SRL system. We use Sections 2 to 21 of the PropBank to train our SRL system. Then, we use our system to annotate Section 24 of the PropBank (denoted by Exp 1a) and all of BioProp (denoted by Exp 1b).

Experiment 2: The Necessity of BioProp

To compare the effects of using biomedical training data vs. using newswire data, we train our SRL system on 30 randomly selected training sets from BioProp (g_1,.., g_{30}) and 30 from PropBank (w_1,.., w_{30}), each having 1200 training PAS's. We then test our system 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 add up the scores for w_1-w_{30} and g_1-g_{30}, and compare their averages.

Experiment 3: The Effect of Using Biomedical-Specific Features

In order to improve SRL performance, we add domain specific features. In Experiment 3, we investigate the effects of adding biomedical NE features and argument template features composed of words, NEs, and POSs. The dataset selection procedure is the same as in Experiment 2.

5 Results and Discussion

All experimental results are summarized in Table 5. For argument classification, we report the precision (P), recall (R) and F-scores (F). The details are illustrated in the following paragraphs.

Configuration	Training	Test	P	R	F
Exp 1a	PropBank	PropBank	90.47	82.48	86.29
Exp 1b	PropBank	BioProp	75.28	56.64	64.64
Exp 2a	PropBank	BioProp	74.78	56.25	64.20
Exp 2b	BioProp	BioProp	88.65	85.61	87.10
Exp 3a	BioProp	BioProp	88.67	85.59	87.11
Exp 3b	BioProp	BioProp	89.13	86.07	87.57

Table 5. Summary of All Experiments

Role	Exp 1a			Exp 1b			+/-(%)
	P	R	F	P	R	F	
Overall	90.47	82.48	86.29	75.28	56.64	64.64	-21.65
ArgX	91.46	86.39	88.85	78.92	67.82	72.95	-15.90
Arg0	86.36	78.01	81.97	85.56	64.41	73.49	-8.48
Arg1	95.52	92.11	93.78	82.56	75.75	79.01	-14.77
Arg2	87.19	84.53	85.84	32.76	31.59	32.16	-53.68
AM	86.76	70.02	77.50	62.70	32.98	43.22	-34.28
-ADV	73.44	52.32	61.11	39.27	26.34	31.53	-29.58
-DIS	81.71	48.18	60.62	67.12	48.18	56.09	-4.53
-LOC	89.19	57.02	69.57	68.54	2.67	5.14	-64.43
-MNR	67.93	57.86	62.49	46.55	22.97	30.76	-31.73
-MOD	99.42	92.5	95.84	99.05	88.01	93.2	-2.64
-NEG	100	91.21	95.40	99.61	80.13	88.81	-6.59
-TMP	88.15	72.83	79.76	70.97	60.36	65.24	-14.52

Table 6. Performance of Exp 1a and Exp 1b

Experiment 1

Table 6 shows the results of Experiment 1. The SRL system trained on the WSJ corpus obtains an F-score of 64.64% when used in the biomedical domain. Compared to traditional rule-based or template-based approaches, our approach suffers acceptable decrease in overall performance when recognizing ArgX arguments. However, Table 6 also shows significant decreases in F-scores from other argument types. AM-LOC drops 64.43% and AM-MNR falls 31.73%. This may be due to the fact that the head words in PropBank are quite different from those in BioProp. Therefore, to achieve better performance, we believe it will be necessary to annotate biomedical corpora for training biomedical SRL systems.

Experiment 2

Table 7 shows the results of Experiment 2. When tested on BioProp, BIOSMILE (Exp 2b) outperforms the newswire SRL system (Exp 2a) by 22.9% since the two systems are trained on different domains. This result is statistically significant.

Furthermore, Table 7 shows that BIOSMILE outperforms the newswire SRL system in most

argument types, especially Arg0, Arg2, AM-ADV, AM-LOC, AM-MNR.

Role	Exp 2a			Exp 2b			+/-(%)
	P	R	F	P	R	F	
Overall	74.78	56.25	64.20	88.65	85.61	87.10	22.90
ArgX	78.40	67.32	72.44	91.96	89.73	90.83	18.39
Arg0	85.55	64.40	73.48	92.24	90.59	91.41	17.93
Arg1	81.41	75.11	78.13	92.54	90.49	91.50	13.37
Arg2	34.42	31.56	32.93	86.89	81.35	84.03	51.10
AM	61.96	32.38	42.53	81.27	76.72	78.93	36.40
-ADV	36.00	23.26	28.26	64.02	52.12	57.46	29.20
-DIS	69.55	51.29	59.04	82.71	75.60	79.00	19.96
-LOC	75.51	3.23	6.20	80.05	85.00	82.45	76.25
-MNR	44.67	21.66	29.17	83.44	82.23	82.83	53.66
-MOD	99.38	88.89	93.84	98.00	95.28	96.62	2.78
-NEG	99.80	79.55	88.53	97.82	94.81	96.29	7.76
-TMP	67.95	60.40	63.95	80.96	61.82	70.11	6.16

Table 7. Performance of Exp 2a and Exp 2b

The performance of Arg0 and Arg2 in our system increases considerably because biomedical verbs can be successfully identified by BIOSMILE but not by the newswire SRL system. For AM-LOC, the newswire SRL system scored as low as 76.25% lower than BIOSMILE. This is likely due to the reason that in the biomedical domain, many biomedical nouns, e.g., organisms and cells, function as locations, while in the newswire domain, they do not. In newswire, the word "cell" seldom appears. However, in biomedical texts, cells represent the location of many biological reactions, and, therefore, if a constituent node on a parsing tree contains "cell", this node is very likely an AM-LOC. If we use only newswire texts, the SRL system will not learn to recognize this pattern. In the biomedical domain, arguments of manner (AM-MNR) usually describe how to conduct an experiment or how an interaction arises or occurs, while in newswire they are extremely broad in scope. Without adequate biomedical domain training corpora, systems will easily confuse adverbs of manner (AM-MNR), which are differentiated from general adverbials in semantic role labeling, with general adverbials (AM-ADV). In addition, the performance of the referential arguments of Arg0, Arg1, and Arg2 increases significantly.

Experiment 3

Table 8 shows the results of Experiment 3. The performance does not significantly improve after adding NE features. We originally expected that NE features would improve recognition of AM arguments such as AM-LOC. However, they failed to ameliorate the results since in the biomedical domain most NEs are just matched parts of a constituent. This results in fewer exact matches. Furthermore, in matched cases, NE information alone is insufficient to distinguish argument types. For example, even if a constituent exactly matches a protein name, we still cannot be sure whether it belongs to the subject (Arg0) or object (Arg1). Therefore, NE features were not as effective as we had expected.

Role	NE (Exp 3a)			Template (Exp 3b)			+/-(%)
	P	R	F	P	R	F	
Overall	88.67	85.59	87.11	89.13	86.07	87.57	0.46
ArgX	91.99	89.70	90.83	91.89	89.73	90.80	-0.03
Arg0	92.41	90.57	91.48	92.19	90.59	91.38	-0.1
Arg1	92.47	90.45	91.45	92.42	90.44	91.42	-0.03
Arg2	86.93	81.3	84.02	87.08	81.66	84.28	0.26
AM	81.30	76.75	78.96	82.96	78.18	80.50	1.54
-ADV	64.11	52.23	57.56	65.66	55.60	60.21	2.65
-DIS	82.51	75.42	78.81	83.00	75.79	79.23	0.42
-LOC	80.07	85.09	82.50	84.24	85.48	84.86	2.36
-MNR	83.50	82.19	82.84	84.56	84.14	84.35	1.51
-MOD	98.14	95.28	96.69	98.00	95.28	96.62	-0.07
-NEG	97.66	94.81	96.21	97.82	94.81	96.29	0.08
-TMP	81.14	62.06	70.33	83.10	63.95	72.28	1.95

Table 8. Performance of Exp 3a and Exp 3b

6 Conclusions and Future Work

In Experiment 3b, we used the argument templates as features. Since ArgX's F-score is close to 90%, adding the template features does not improve its score. However, AM's F-score increases by 1.54%. For AM-ADV, AM-LOC, and AM-TMP, the increase is greater because the automatically generated templates effectively extract these AMs.

In Figure 3, we compare the performance of argument classification models with and without argument template features. The overall F-score improves only slightly. However, the F-scores of main adjunct arguments increase significantly.

The contribution of this paper is threefold. First, we construct a biomedical proposition bank, BioProp, on top of the popular biomedical GENIA treebank following the PropBank annotation scheme. We employ semi-automatic annotation using an SRL system trained on PropBank, thereby significantly reducing annotation effort. Second, we create BIOSMILE, a biomedical SRL system, which uses BioProp as its training corpus. Thirdly, we develop a method to automatically generate templates that can boost overall performance, es-

pecially on location, manner, adverb, and temporal arguments. In the future, we will expand BioProp to include more verbs and will also integrate an automatic parser into BIOSMILE.

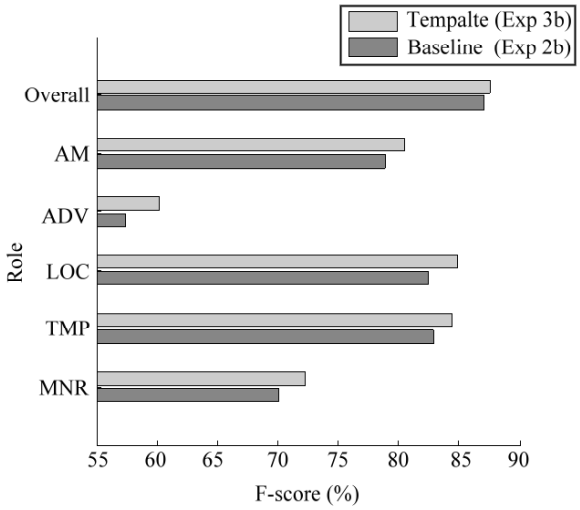

Figure 3. Improvement of Template Features Overall and on Several Adjunct Types

Acknowledgement

We would like to thank Dr. Nianwen Xue for his instruction of using the WordFreak annotation tool. This research was supported in part by the National Science Council under grant NSC94-2752-E-001-001 and the thematic program of Academia Sinica under grant AS94B003. Editing services were provided by Dorion Berg.

References

Babko-Malaya, O. (2005). *Propbank Annotation Guidelines*.

Bies, A., Ferguson, M., Katz, K., MacIntyre, R., Tredinnick, V., Kim, G., et al. (1995). *Bracketing Guidelines for Treebank II Style Penn Treebank Project*

Collins, M. J. (1999). *Head-driven Statistical Models for Natural Language Parsing*. Unpublished Ph.D. thesis, University of Pennsylvania.

Darroch, J. N., & Ratcliff, D. (1972). Generalized Iterative Scaling for Log-Linear Models. *The Annals of Mathematical Statistics*.

Hacioglu, K., Pradhan, S., Ward, W., Martin, J. H., & Jurafsky, D. (2004). *Semantic Role Labeling by Tagging Syntactic Chunks.* Paper presented at the CONLL-04.

Huang, M., Zhu, X., Hao, Y., Payan, D. G., Qu, K., & Li, M. (2004). Discovering patterns to extract protein-protein interactions from full texts. *Bioinformatics, 20*(18), 3604-3612.

Kholodenko, B. N. (2006). Cell-signalling dynamics in time and space. *Nat Rev Mol Cell Biol, 7*(3), 165-176.

Kim, J. D., Ohta, T., Tateisi, Y., & Tsujii, J. (2003). GENIA corpus--semantically annotated corpus for bio-textmining. *Bioinformatics, 19 Suppl 1*, i180-182.

Leroy, G., Chen, H., & Genescene. (2005). An ontology-enhanced integration of linguistic and co-occurrence based relations in biomedical texts. *Journal of the American Society for Information Science and Technology, 56*(5), 457-468.

Marcus, M. P., Santorini, B., & Marcinkiewicz, M. A. (1997). Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics, 19*.

Morarescu, P., Bejan, C., & Harabagiu, S. (2005). *Shallow Semantics for Relation Extraction.* Paper presented at the IJCAI-05.

Nocedal, J., & Wright, S. J. (1999). *Numerical Optimization*: Springer.

Palmer, M., Gildea, D., & Kingsbury, P. (2005). The proposition bank: an annotated corpus of semantic roles. *Computational Linguistics, 31*(1).

Pradhan, S., Hacioglu, K., Kruglery, V., Ward, W., Martin, J. H., & Jurafsky, D. (2004). Support vector learning for semantic argument classification. *Journal of Machine Learning*

Punyakanok, V., Roth, D., Yih, W., & Zimak, D. (2004). *Semantic Role Labeling via Integer Linear Programming Inference.* Paper presented at the COLING-04.

Surdeanu, M., Harabagiu, S. M., Williams, J., & Aarseth, P. (2003). *Using Predicate-Argument Structures for Information Extraction.* Paper presented at the ACL-03.

Tateisi, Y., & Tsujii, J. (2004). *Part-of-Speech Annotation of Biology Research Abstracts.* Paper presented at the LREC-04.

Tateisi, Y., Yakushiji, A., Ohta, T., & Tsujii, J. (2005). *Syntax Annotation for the GENIA corpus*.

Tsai, T.-H., Wu, C.-W., Lin, Y.-C., & Hsu, W.-L. (2005). *Exploiting Full Parsing Information to Label Semantic Roles Using an Ensemble of ME and SVM via Integer Linear Programming.* . Paper presented at the CoNLL-05.

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

Xue, N., & Palmer, M. (2004). *Calibrating Features for Semantic Role Labeling.* Paper presented at the EMNLP-04.

Yang, X., Zhou, G., Su, J., & Tan., C. (2004). *Improving Noun Phrase Coreference Resolution by Matching Strings.* Paper presented at the IJCNLP-04.