A Comparative Study on Generalization of Semantic Roles in FrameNet

Yuichiroh Matsubayashi[†] **Naoaki Okazaki**[†] **Jun'ichi Tsujii**[†‡*]
[†]Department of Computer Science, University of Tokyo, Japan
[‡]School of Computer Science, University of Manchester, UK
[*]National Centre for Text Mining, UK
{y-matsu,okazaki,tsujii}@is.s.u-tokyo.ac.jp

Abstract

A number of studies have presented machine-learning approaches to semantic role labeling with availability of corpora such as FrameNet and PropBank. These corpora define the semantic roles of predicates for each frame independently. Thus, it is crucial for the machine-learning approach to generalize semantic roles across different frames, and to increase the size of training instances. This paper explores several criteria for generalizing semantic roles in FrameNet: role hierarchy, human-understandable descriptors of roles, semantic types of filler phrases, and mappings from FrameNet roles to thematic roles of VerbNet. We also propose feature functions that naturally combine and weight these criteria, based on the training data. The experimental result of the role classification shows 19.16% and 7.42% improvements in error reduction rate and macro-averaged F1 score, respectively. We also provide in-depth analyses of the proposed criteria.

1 Introduction

Semantic Role Labeling (SRL) is a task of analyzing predicate-argument structures in texts. More specifically, SRL identifies predicates and their arguments with appropriate semantic roles. Resolving surface divergence of texts (e.g., voice of verbs and nominalizations) into unified semantic representations, SRL has attracted much attention from researchers into various NLP applications including question answering (Narayanan and Harabagiu, 2004; Shen and Lapata, 2007; Moschitti et al., 2007), and information extraction (Surdeanu et al., 2003).

buy.v	PropBank	FrameNet
Frame	buy.01	Commerce_buy
Roles	ARG0: buyer	Buyer
	ARG1: thing bought	Goods
	ARG2: seller	Seller
	ARG3: paid	Money
	ARG4: benefactive	Recipient
	...	...

Figure 1: A comparison of frames for *buy.v* defined in PropBank and FrameNet

In recent years, with the wide availability of corpora such as PropBank (Palmer et al., 2005) and FrameNet (Baker et al., 1998), a number of studies have presented statistical approaches to SRL (Màrquez et al., 2008). Figure 1 shows an example of the frame definitions for a verb *buy* in PropBank and FrameNet. These corpora define a large number of frames and define the semantic roles for each frame independently. This fact is problematic in terms of the performance of the machine-learning approach, because these definitions produce many roles that have few training instances.

PropBank defines a frame for each sense of predicates (e.g., *buy.01*), and semantic roles are defined in a frame-specific manner (e.g., *buyer* and *seller* for *buy.01*). In addition, these roles are associated with tags such as *ARG0-5* and *AM-**, which are commonly used in different frames. Most SRL studies on PropBank have used these tags in order to gather a sufficient amount of training data, and to generalize semantic-role classifiers across different frames. However, Yi et al. (2007) reported that tags *ARG2–ARG5* were inconsistent and not that suitable as training instances. Some recent studies have addressed alternative approaches to generalizing semantic roles across different frames (Gordon and Swanson, 2007; Zapi-

Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pages 19–27,
Suntec, Singapore, 2-7 August 2009. ©2009 ACL and AFNLP

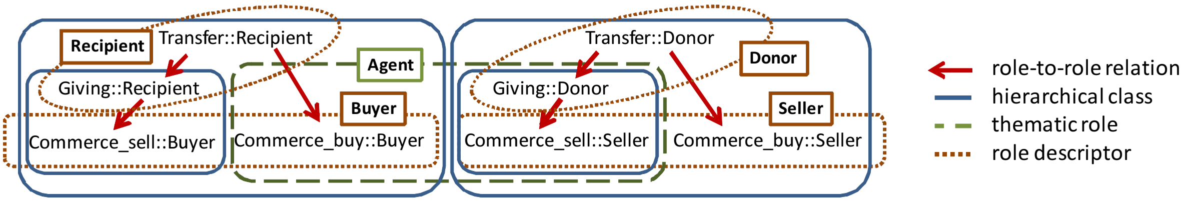

Figure 2: An example of role groupings using different criteria.

rain et al., 2008).

FrameNet designs semantic roles as frame specific, but also defines hierarchical relations of semantic roles among frames. Figure 2 illustrates an excerpt of the role hierarchy in FrameNet; this figure indicates that the *Buyer* role for the *Commerce_buy* frame (Commerce_buy::Buyer hereafter) and the Commerce_sell::Buyer role are inherited from the Transfer::Recipient role. Although the role hierarchy was expected to generalize semantic roles, no positive results for role classification have been reported (Baldewein et al., 2004). Therefore, the generalization of semantic roles across different frames has been brought up as a critical issue for FrameNet (Gildea and Jurafsky, 2002; Shi and Mihalcea, 2005; Giuglea and Moschitti, 2006)

In this paper, we explore several criteria for generalizing semantic roles in FrameNet. In addition to the FrameNet hierarchy, we use various pieces of information: human-understandable descriptors of roles, semantic types of filler phrases, and mappings from FrameNet roles to the thematic roles of VerbNet. We also propose feature functions that naturally combines these criteria in a machine-learning framework. Using the proposed method, the experimental result of the role classification shows 19.16% and 7.42% improvements in error reduction rate and macro-averaged F1, respectively. We provide in-depth analyses with respect to these criteria, and state our conclusions.

2 Related Work

Moschitti et al. (2005) first classified roles by using four coarse-grained classes (Core Roles, Adjuncts, Continuation Arguments and Co-referring Arguments), and built a classifier for each coarse-grained class to tag PropBank *ARG* tags. Even though the initial classifiers could perform rough estimations of semantic roles, this step was not able to solve the ambiguity problem in PropBank *ARG2-5*. When training a classifier for a semantic role, Baldewein et al. (2004) re-used the training instances of other roles that were similar to the target role. As similarity measures, they used the FrameNet hierarchy, peripheral roles of FrameNet, and clusters constructed by a EM-based method. Gordon and Swanson (2007) proposed a generalization method for the PropBank roles based on syntactic similarity in frames.

Many previous studies assumed that thematic roles bridged semantic roles in different frames. Gildea and Jurafsky (2002) showed that classification accuracy was improved by manually replacing FrameNet roles into 18 thematic roles. Shi and Mihalcea (2005) and Giuglea and Moschitti (2006) employed VerbNet thematic roles as the target of mappings from the roles defined by the different semantic corpora. Using the thematic roles as alternatives of *ARG* tags, Loper et al. (2007) and Yi et al. (2007) demonstrated that the classification accuracy of PropBank roles was improved for *ARG2* roles, but that it was diminished for *ARG1*. Yi et al. (2007) also described that *ARG2–5* were mapped to a variety of thematic roles. Zapirain et al. (2008) evaluated PropBank ARG tags and VerbNet thematic roles in a state-of-the-art SRL system, and concluded that PropBank *ARG* tags achieved a more robust generalization of the roles than did VerbNet thematic roles.

3 Role Classification

SRL is a complex task wherein several problems are intertwined: *frame-evoking word identification*, *frame disambiguation* (selecting a correct frame from candidates for the evoking word), *role-phrase identification* (identifying phrases that fill semantic roles), and *role classification* (assigning correct roles to the phrases). In this paper, we focus on role classification, in which the role generalization is particularly critical to the machine learning approach.

In the role classification task, we are given a sentence, a frame evoking word, a frame, and

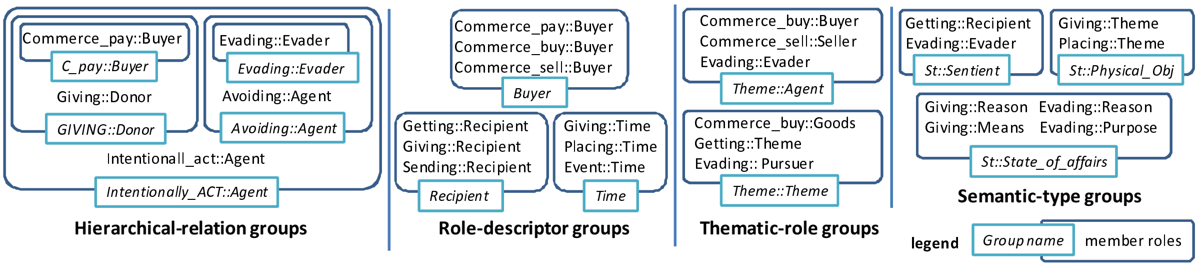

Figure 4: Examples for each type of role group.

INPUT:
frame = Commerce_sell
candidate roles ={Seller, Buyer, Goods, Reason, Time, ... , Place}
sentence = Can't [you] [**sell** $_{\text{Commerce\_sell}}$] [the factory] [to some other company] ?

OUTPUT:
sentence = Can't [you $_{\text{Seller}}$] [**sell** $_{\text{Commerce\_sell}}$] [the factory $_{\text{Goods}}$] [to some other company $_{\text{Buyer}}$] ?

Figure 3: An example of input and output of role classification.

phrases that take semantic roles. We are interested in choosing the correct role from the candidate roles for each phrase in the frame. Figure 3 shows a concrete example of input and output; the semantic roles for the phrases are chosen from the candidate roles: Seller, Buyer, Goods, Reason, ... , and Place.

4 Design of Role Groups

We formalize the generalization of semantic roles as the act of grouping several roles into a class. We define a *role group* as a set of role labels grouped by a criterion. Figure 4 shows examples of role groups; a group *Giving::Donor* (in the hierarchical-relation groups) contains the roles Giving::Donor and Commerce_pay::Buyer. The remainder of this section describes the grouping criteria in detail.

4.1 Hierarchical relations among roles

FrameNet defines hierarchical relations among frames (frame-to-frame relations). Each relation is assigned one of the seven types of directional relationships (*Inheritance*, *Using*, *Perspective_on*, *Causative_of*, *Inchoative_of*, *Subframe*, and *Precedes*). Some roles in two related frames are also connected with role-to-role relations. We assume that this hierarchy is a promising resource for generalizing the semantic roles; the idea is that the role at a node in the hierarchy inherits the characteristics of the roles of its ancestor nodes. For example, Commerce_sell::Seller in Figure 2 inherits the property of Giving::Donor.

For *Inheritance*, *Using*, *Perspective_on*, and *Subframe* relations, we assume that descendant roles in these relations have the same or specialized properties of their ancestors. Hence, for each role y_i, we define the following two role groups,

$$\begin{aligned} H^{\text{child}}_{y_i} &= \{y \mid y = y_i \lor y \text{ is a child of } y_i\}, \\ H^{\text{desc}}_{y_i} &= \{y \mid y = y_i \lor y \text{ is a descendant of } y_i\}. \end{aligned}$$

The hierarchical-relation groups in Figure 4 are the illustrations of $H^{\text{desc}}_{y_i}$.

For the relation types *Inchoative_of* and *Causative_of*, we define role groups in the opposite direction of the hierarchy,

$$\begin{aligned} H^{\text{parent}}_{y_i} &= \{y \mid y = y_i \lor y \text{ is a parent of } y_i\}, \\ H^{\text{ance}}_{y_i} &= \{y \mid y = y_i \lor y \text{ is an ancestor of } y_i\}. \end{aligned}$$

This is because lower roles of *Inchoative_of* and *Causative_of* relations represent more neutral stances or consequential states; for example, Killing::Victim is a parent of Death::Protagonist in the *Causative_of* relation.

Finally, the *Precedes* relation describes the sequence of states and events, but does not specify the direction of semantic inclusion relations. Therefore, we simply try $H^{\text{child}}_{y_i}$, $H^{\text{desc}}_{y_i}$, $H^{\text{parent}}_{y_i}$, and $H^{\text{ance}}_{y_i}$ for this relation type.

4.2 Human-understandable role descriptor

FrameNet defines each role as frame-specific; in other words, the same identifier does not appear in different frames. However, in FrameNet, human experts assign a human-understandable name to each role in a rather systematic manner. Some names are shared by the roles in different frames, whose identifiers are different. Therefore, we examine the semantic commonality of these names; we construct an equivalence class of the roles sharing the same name. We call these human-understandable names *role descriptors*. In Figure 4, the role-descriptor group *Buyer* collects the roles Commerce_pay::Buyer, Commerce_buy::Buyer, and Commerce_sell::Buyer.

This criterion may be effective in collecting similar roles since the descriptors have been annotated by intuition of human experts. As illustrated in Figure 2, the role descriptors group the semantic roles which are similar to the roles that the FrameNet hierarchy connects as sister or parent-child relations. However, role-descriptor groups cannot express the relations between the roles as inclusions since they are equivalence classes. For example, the roles Commerce_sell::Buyer and Commerce_buy::Buyer are included in the role descriptor group *Buyer* in Figure 2; however, it is difficult to merge Giving::Recipient and Commerce_sell::Buyer because the Commerce_sell::Buyer has the extra property that one gives something of value in exchange and a human assigns different descriptors to them. We expect that the most effective weighting of these two criteria will be determined from the training data.

4.3 Semantic type of phrases

We consider that the selectional restriction is helpful in detecting the semantic roles. FrameNet provides information concerning the semantic types of role phrases (fillers); phrases that play specific roles in a sentence should fulfill the semantic constraint from this information. For instance, FrameNet specifies the constraint that Self_motion::Area should be filled by phrases whose semantic type is *Location*. Since these types suggest a coarse-grained categorization of semantic roles, we construct role groups that contain roles whose semantic types are identical.

4.4 Thematic roles of VerbNet

VerbNet thematic roles are 23 frame-independent semantic categories for arguments of verbs, such as *Agent*, *Patient*, *Theme* and *Source*. These categories have been used as consistent labels across verbs. We use a partial mapping between FrameNet roles and VerbNet thematic roles provided by SemLink. [1] Each group is constructed as a set $T_{t_i} = \{y | \text{SemLink maps } y \text{ into the thematic role } t_i\}$.

SemLink currently maps 1,726 FrameNet roles into VerbNet thematic roles, which are 37.61% of roles appearing at least once in the FrameNet corpus. This may diminish the effect of thematic-role groups than its potential.

5 Role classification method

5.1 Traditional approach

We are given a frame-evoking word e, a frame f and a role phrase x detected by a human or some automatic process in a sentence s. Let Y_f be the set of semantic roles that FrameNet defines as being possible role assignments for the frame f, and let $\mathbf{x} = \{x_1, \ldots, x_n\}$ be observed features for x from s, e and f. The task of semantic role classification can be formalized as the problem of choosing the most suitable role $\tilde{y}$ from Y_f. Suppose we have a model $P(y|f, \mathbf{x})$ which yields the conditional probability of the semantic role y for given f and $\mathbf{x}$. Then we can choose $\tilde{y}$ as follows:

$$\tilde{y} = \underset{y \in Y_f}{\operatorname{argmax}} P(y|f, \mathbf{x}). \tag{1}$$

A traditional way to incorporate role groups into this formalization is to overwrite each role y in the training and test data with its role group $m(y)$ according to the memberships of the group. For example, semantic roles Commerce_sell::Seller and Giving::Donor can be replaced by their thematic-role group *Theme::Agent* in this approach. We determine the most suitable role group $\tilde{c}$ as follows:

$$\tilde{c} = \underset{c \in \{m(y) | y \in Y_f\}}{\operatorname{argmax}} P_m(c|f, \mathbf{x}). \tag{2}$$

Here, $P_m(c|f, \mathbf{x})$ presents the probability of the role group c for f and $\mathbf{x}$. The role $\tilde{y}$ is determined uniquely iff a single role $y \in Y_f$ is associated with $\tilde{c}$. Some previous studies have employed this idea to remedy the data sparseness problem in the training data (Gildea and Jurafsky, 2002). However, we cannot apply this approach when multiple roles in Y_f are contained in the same class. For example, we can construct a semantic-type group *St::State_of_affairs* in which Giving::Reason and Giving::Means are included, as illustrated in Figure 4. If $\tilde{c}$ = *St::State_of_affairs*, we cannot disambiguate which original role is correct. In addition, it may be more effective to use various

[1] http://verbs.colorado.edu/semlink/

groupings of roles together in the model. For instance, the model could predict the correct role Commerce_sell::Seller for the phrase "you" in Figure 3 more confidently, if it could infer its thematic-role group as *Theme::Agent* and its parent group *Giving::Donor* correctly. Although the ensemble of various groupings seems promising, we need an additional procedure to prioritize the groupings for the case where the models for multiple role groupings disagree; for example, it is unsatisfactory if two models assign the groups *Giving::Theme* and *Theme::Agent* to the same phrase.

5.2 Role groups as feature functions

We thus propose another approach that incorporates group information as feature functions. We model the conditional probability $P(y|f, \mathbf{x})$ by using the maximum entropy framework,

$$p(y|f,\mathbf{x}) = \frac{\exp(\sum_i \lambda_i g_i(\mathbf{x}, y))}{\sum_{y \in Y_f} \exp(\sum_i \lambda_i g_i(\mathbf{x}, y))}. \quad (3)$$

Here, $G = \{g_i\}$ denotes a set of n feature functions, and $\Lambda = \{\lambda_i\}$ denotes a weight vector for the feature functions.

In general, feature functions for the maximum entropy model are designed as indicator functions for possible pairs of x_j and y. For example, the event where the head word of x is "you" ($x_1 = 1$) and x plays the role Commerce_sell::Seller in a sentence is expressed by the indicator function,

$$g_1^{\textit{role}}(\mathbf{x}, y) = \begin{cases} 1 & (x_1 = 1 \wedge \\ & \ y = \text{Commerce\_sell::Seller}) \\ 0 & \text{(otherwise)} \end{cases}. \quad (4)$$

We call this kind of feature function an *x-role*.

In order to incorporate role groups into the model, we also include all feature functions for possible pairs of x_j and role groups. Equation 5 is an example of a feature function for instances where the head word of x is "you" and y is in the role group *Theme::Agent*,

$$g_2^{\textit{theme}}(\mathbf{x}, y) = \begin{cases} 1 & (x_1 = 1 \wedge \\ & \ y \in \textit{Theme::Agent}) \\ 0 & \text{(otherwise)} \end{cases}. \quad (5)$$

Thus, this feature function fires for the roles wherever the head word "you" plays *Agent* (e.g., Commerce_sell::Seller, Commerce_buy::Buyer and Giving::Donor). We call this kind of feature function an *x-group* function.

In this way, we obtain *x-group* functions for all grouping methods, e.g., $g_k^{\textit{theme}}$, $g_k^{\textit{hierarchy}}$. The role-group features will receive more training instances by collecting instances for fine-grained roles. Thus, semantic roles with few training instances are expected to receive additional clues from other training instances via role-group features. Another advantage of this approach is that the usefulness of the different role groups is determined by the training processes in terms of weights of feature functions. Thus, we do not need to assume that we have found the best criterion for grouping roles; we can allow a training process to choose the criterion. We will discuss the contributions of different groupings in the experiments.

5.3 Comparison with related work

Baldewein et al. (2004) suggested an approach that uses role descriptors and hierarchical relations as criteria for generalizing semantic roles in FrameNet. They created a classifier for each frame, additionally using training instances for the role *A* to train the classifier for the role *B*, if the roles *A* and *B* were judged as similar by a criterion. This approach performs similarly to the overwriting approach, and it may obscure the differences among roles. Therefore, they only re-used the descriptors as a similarity measure for the roles whose *coreness* was *peripheral*. [2]

In contrast, we use all kinds of role descriptors to construct groups. Since we use the feature functions for both the original roles and their groups, appropriate units for classification are determined automatically in the training process.

6 Experiment and Discussion

We used the training set of the Semeval-2007 Shared task (Baker et al., 2007) in order to ascertain the contributions of role groups. This dataset consists of the corpus of FrameNet release 1.3 (containing roughly 150,000 annotations), and an additional full-text annotation dataset. We randomly extracted 10% of the dataset for testing, and used the remainder (90%) for training.

Performance was measured by micro- and macro-averaged F1 (Chang and Zheng, 2008) with respect to a variety of roles. The micro average biases each F1 score by the frequencies of the roles,

[2] In FrameNet, each role is assigned one of four different types of *coreness* (*core*, *core-unexpressed*, *peripheral*, *extra-thematic*) It represents the conceptual necessity of the roles in the frame to which it belongs.

and the average is equal to the classification accuracy when we calculate it with all of the roles in the test set. In contrast, the macro average does not bias the scores, thus the roles having a small number of instances affect the average more than the micro average.

6.1 Experimental settings

We constructed a baseline classifier that uses only the *x-role* features. The feature design is similar to that of the previous studies (Màrquez et al., 2008). The characteristics of *x* are: **frame, frame evoking word, head word, content word** (Surdeanu et al., 2003), **first/last word, head word of left/right sister, phrase type, position, voice, syntactic path (directed/undirected/partial), governing category** (Gildea and Jurafsky, 2002), **WordNet supersense in the phrase**, combination features of **frame evoking word & headword**, combination features of **frame evoking word & phrase type**, and combination features of **voice & phrase type**. We also used **PoS tags** and **stem forms** as extra features of any word-features.

We employed Charniak and Johnson's reranking parser (Charniak and Johnson, 2005) to analyze syntactic trees. As an alternative for the traditional named-entity features, we used WordNet supersenses: 41 coarse-grained semantic categories of words such as person, plant, state, event, time, location. We used Ciaramita and Altun's Super Sense Tagger (Ciaramita and Altun, 2006) to tag the supersenses. The baseline system achieved 89.00% with respect to the micro-averaged F1.

The *x-group* features were instantiated similarly to the *x-role* features; the *x-group* features combined the characteristics of *x* with the role groups presented in this paper. The total number of features generated for all *x-roles* and *x-groups* was 74,873,602. The optimal weights Λ of the features were obtained by the maximum a posterior (MAP) estimation. We maximized an L_2-regularized log-likelihood of the training set using the Limited-memory BFGS (L-BFGS) method (Nocedal, 1980).

6.2 Effect of role groups

Table 1 shows the micro and macro averages of F1 scores. Each role group type improved the micro average by 0.5 to 1.7 points. The best result was obtained by using all types of groups together. The result indicates that different kinds of group complement each other with respect to semantic role generalization. Baldewein et al. (2004) reported that hierarchical relations did not perform well for their method and experimental setting; however, we found that significant improvements could also be achieved with hierarchical relations. We also tried a traditional label-replacing approach with role descriptors (in the third row of Table 1). The comparison between the second and third rows indicates that mixing the original fine-grained roles and the role groups does result in a more accurate classification.

Feature	Micro	Macro	−Err.
Baseline	89.00	68.50	0.00
role descriptor	90.78	**76.58**	16.17
role descriptor (replace)	90.23	76.19	11.23
hierarchical relation	90.25	72.41	11.40
semantic type	90.36	74.51	12.38
VN thematic role	89.50	69.21	4.52
All	**91.10**	75.92	19.16

Table 1: The accuracy and error reduction rate of role classification for each type of role group.

Feature	#instances	Pre.	Rec.	Micro
baseline	≤ 10	63.89	38.00	47.66
	≤ 20	69.01	51.26	58.83
	≤ 50	75.84	65.85	70.50
+ all groups	≤ 10	72.57	55.85	63.12
	≤ 20	76.30	65.41	70.43
	≤ 50	80.86	74.59	77.60

Table 2: The effect of role groups on the roles with few instances.

By using all types of groups together, the model reduced 19.16 % of the classification errors from the baseline. Moreover, the macro-averaged F1 scores clearly showed improvements resulting from using role groups. In order to determine the reason for the improvements, we measured the precision, recall, and F1-scores with respect to roles for which the number of training instances was at most 10, 20, and 50. In Table 2, we show that the micro-averaged F1 score for roles having 10 instances or less was improved (by 15.46 points) when all role groups were used. This result suggests the reason for the effect of role groups; by bridging similar semantic roles, they supply roles having a small number of instances with the information from other roles.

6.3 Analyses of role descriptors

In Table 1, the largest improvement was obtained by the use of role descriptors. We analyze the effect of role descriptors in detail in Tables 3 and 4. Table 3 shows the micro-averaged F1 scores of all

Coreness	#roles	#instances/#role	#groups	#instances/#group	#roles/#group
Core	1902	122.06	655	354.4	2.9
Peripheral	1924	25.24	250	194.3	**7.7**
Extra-thematic	763	13.90	171	62.02	4.5

Table 4: The analysis of the numbers of roles, instances, and role-descriptor groups, for each type of coreness.

Coreness	Micro
Baseline	89.00
Core	89.51
Peripheral	**90.12**
Extra-thematic	89.09
All	90.77

Table 3: The effect of employing role-descriptor groups of each type of coreness.

No.	Relation Type	Micro
-	baseline	89.00
1	+ Inheritance (children)	89.52
2	+ Inheritance (descendants)	**89.70**
3	+ Using (children)	89.35
4	+ Using (descendants)	**89.37**
5	+ Perspective on (children)	89.01
6	+ Perspective on (descendants)	89.01
7	+ Subframe (children)	89.04
8	+ Subframe (descendants)	89.05
9	+ Causative of (parents)	89.03
10	+ Causative of (ancestors)	89.03
11	+ Inchoative of (parents)	89.02
12	+ Inchoative of (ancestors)	89.02
13	+ Precedes (children)	89.01
14	+ Precedes (descendants)	89.03
15	+ Precedes (parents)	89.00
16	+ Precedes (ancestors)	89.00
18	+ all relations (2,4,6,8,10,12,14)	**90.25**

Table 5: Comparison of the accuracy with different types of hierarchical relations.

semantic roles when we use role-descriptor groups constructed from each type of coreness (core[3], peripheral, and extra-thematic) individually. The *peripheral* type generated the largest improvements.

Table 4 shows the number of roles associated with each type of coreness (#roles), the number of instances for the original roles (#instances/#role), the number of groups for each type of coreness (#groups), the number of instances for each group (#instances/#group), and the number of roles per each group (#roles/#group). In the *peripheral* type, the role descriptors subdivided 1,924 distinct roles into 250 groups, each of which contained 7.7 roles on average. The *peripheral* type included semantic roles such as *place*, *time*, *reason*, *duration*. These semantic roles appear in many frames, because they have general meanings that can be shared by different frames. Moreover, the semantic roles of *peripheral* type originally occurred in only a small number (25.24) of training instances on average. Thus, we infer that the *peripheral* type generated the largest improvement because semantic roles in this type acquired the greatest benefit from the generalization.

6.4 Hierarchical relations and relation types

We analyzed the contributions of the FrameNet hierarchy for each type of role-to-role relations and for different depths of grouping. Table 5 shows the micro-averaged F1 scores obtained from various relation types and depths. The *Inheritance* and *Using* relations resulted in a slightly better accuracy than the other types. We did not observe any real differences among the remaining five relation types, possibly because there were few semantic roles associated with these types. We obtained better results by using not only groups for parent roles, but also groups for all ancestors. The best result was obtained by using all relations in the hierarchy.

6.5 Analyses of different grouping criteria

Table 6 reports the precision, recall, and micro-averaged F1 scores of semantic roles with respect to each coreness type.[4] In general, semantic roles of the *core* coreness were easily identified by all of the grouping criteria; even the baseline system obtained an F1 score of 91.93. For identifying semantic roles of the *peripheral* and *extra-thematic* types of coreness, the simplest solution, the descriptor criterion, outperformed other criteria.

In Table 7, we categorize feature functions whose weights are in the top 1000 in terms of greatest absolute value. The behaviors of the role groups can be distinguished by the following two characteristics. Groups of role descriptors and semantic types have large weight values for the first word and supersense features, which capture the characteristics of adjunctive phrases. The original roles and hierarchical-relation groups have strong

[3] We include *Core-unexpressed* in *core*, because it has a property of *core* inside one frame.

[4] The figures of role descriptors in Tables 4 and 6 differ. In Table 4, we measured the performance when we used one or all types of coreness for training. In contrast, in Table 6, we used all types of coreness for training, but computed the performance of semantic roles for each coreness separately.

Feature	Type	Pre.	Rec.	Micro
baseline	c	91.07	92.83	91.93
	p	81.05	76.03	78.46
	e	78.17	66.51	71.87
+ descriptor group	c	92.50	93.41	92.95
	p	84.32	82.72	83.51
	e	80.91	69.59	74.82
+ hierarchical relation class	c	92.10	93.28	92.68
	p	82.23	79.84	81.01
	e	77.94	65.58	71.23
+ semantic type group	c	92.23	93.31	92.77
	p	83.66	81.76	82.70
	e	80.29	67.26	73.20
+ VN thematic role group	c	91.57	93.06	92.31
	p	80.66	76.95	78.76
	e	78.12	66.60	71.90
+ all group	c	92.66	93.61	93.13
	p	84.13	82.51	83.31
	e	80.77	68.56	74.17

Table 6: The precision and recall of each type of coreness with role groups. Type represents the type of coreness; c denotes core, p denotes peripheral, and e denotes extra-thematic.

features of x	class type				
	or	hr	rl	st	vn
frame	0	4	0	1	0
evoking word	3	4	7	3	0
ew & hw stem	9	34	20	8	0
ew & phrase type	11	7	11	3	1
head word	13	19	8	3	1
hw stem	11	17	8	8	1
content word	7	19	12	3	0
cw stem	11	26	13	5	0
cw PoS	4	5	14	15	2
directed path	19	27	24	6	7
undirected path	21	35	17	2	6
partial path	15	18	16	13	5
last word	15	18	12	3	2
first word	11	23	53	26	10
supersense	7	7	35	25	4
position	4	6	30	9	5
others	27	29	33	19	6
total	188	298	313	152	50

Table 7: The analysis of the top 1000 feature functions. Each number denotes the number of feature functions categorized in the corresponding cell. Notations for the columns are as follows. ‘or’: original role, ‘hr’: hierarchical relation, ‘rd’: role descriptor, ‘st’: semantic type, and ‘vn’: VerbNet thematic role.

associations with lexical and structural characteristics such as the syntactic path, content word, and head word. Table 7 suggests that role-descriptor groups and semantic-type groups are effective for *peripheral* or adjunctive roles, and hierarchical relation groups are effective for *core* roles.

7 Conclusion

We have described different criteria for generalizing semantic roles in FrameNet. They were: role hierarchy, human-understandable descriptors of roles, semantic types of filler phrases, and mappings from FrameNet roles to thematic roles of VerbNet. We also proposed a feature design that combines and weights these criteria using the training data. The experimental result of the role classification task showed a 19.16% of the error reduction and a 7.42% improvement in the macro-averaged F1 score. In particular, the method we have presented was able to classify roles having few instances. We confirmed that modeling the role generalization at feature level was better than the conventional approach that replaces semantic role labels.

Each criterion presented in this paper improved the accuracy of classification. The most successful criterion was the use of human-understandable role descriptors. Unfortunately, the FrameNet hierarchy did not outperform the role descriptors, contrary to our expectations. A future direction of this study would be to analyze the weakness of the FrameNet hierarchy in order to discuss possible improvement of the usage and annotations of the hierarchy.

Since we used the latest release of FrameNet in order to use a greater number of hierarchical role-to-role relations, we could not make a direct comparison of performance with that of existing systems; however we may say that the 89.00% F1 micro-average of our baseline system is roughly comparable to the 88.93% value of Bejan and Hathaway (2007) for SemEval-2007 (Baker et al., 2007). [5] In addition, the methodology presented in this paper applies generally to any SRL resources; we are planning to determine several grouping criteria from existing linguistic resources and to apply the methodology to the PropBank corpus.

Acknowledgments

The authors thank Sebastian Riedel for his useful comments on our work. This work was partially supported by Grant-in-Aid for Specially Promoted Research (MEXT, Japan).

References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The berkeley framenet project. In *Proceedings of Coling-ACL 1998*, pages 86–90.

Collin Baker, Michael Ellsworth, and Katrin Erk. 2007. Semeval-2007 task 19: Frame semantic struc-

[5] There were two participants that performed whole SRL in SemEval-2007. Bejan and Hathaway (2007) evaluated role classification accuracy separately for the training data.

ture extraction. In *Proceedings of SemEval-2007*, pages 99–104.

Ulrike Baldewein, Katrin Erk, Sebastian Padó, and Detlef Prescher. 2004. Semantic role labeling with similarity based generalization using EM-based clustering. In *Proceedings of Senseval-3*, pages 64–68.

Cosmin Adrian Bejan and Chris Hathaway. 2007. UTD-SRL: A Pipeline Architecture for Extracting Frame Semantic Structures. In *Proceedings of SemEval-2007*, pages 460–463. Association for Computational Linguistics.

X. Chang and Q. Zheng. 2008. Knowledge Element Extraction for Knowledge-Based Learning Resources Organization. *Lecture Notes in Computer Science*, 4823:102–113.

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 173–180.

Massimiliano Ciaramita and Yasemin Altun. 2006. Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In *Proceedings of EMNLP-2006*, pages 594–602.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

Ana-Maria Giuglea and Alessandro Moschitti. 2006. Semantic role labeling via FrameNet, VerbNet and PropBank. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the ACL*, pages 929–936.

Andrew Gordon and Reid Swanson. 2007. Generalizing semantic role annotations across syntactically similar verbs. In *Proceedings of ACL-2007*, pages 192–199.

Edward Loper, Szu-ting Yi, and Martha Palmer. 2007. Combining lexical resources: Mapping between propbank and verbnet. In *Proceedings of the 7th International Workshop on Computational Semantics*, pages 118–128.

Lluís Màrquez, Xavier Carreras, Kenneth C. Litkowski, and Suzanne Stevenson. 2008. Semantic role labeling: an introduction to the special issue. *Computational linguistics*, 34(2):145–159.

Alessandro Moschitti, Ana-Maria Giuglea, Bonaventura Coppola, and Roberto Basili. 2005. Hierarchical semantic role labeling. In *Proceedings of CoNLL-2005*, pages 201–204.

Alessandro Moschitti, Silvia Quarteroni, Roberto Basili, and Suresh Manandhar. 2007. Exploiting syntactic and shallow semantic kernels for question answer classification. In *Proceedings of ACL-07*, pages 776–783.

Srini Narayanan and Sanda Harabagiu. 2004. Question answering based on semantic structures. In *Proceedings of Coling-2004*, pages 693–701.

Jorge Nocedal. 1980. Updating quasi-newton matrices with limited storage. *Mathematics of Computation*, 35(151):773–782.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Dan Shen and Mirella Lapata. 2007. Using semantic roles to improve question answering. In *Proceedings of EMNLP-CoNLL 2007*, pages 12–21.

Lei Shi and Rada Mihalcea. 2005. Putting Pieces Together: Combining FrameNet, VerbNet and WordNet for Robust Semantic Parsing. In *Proceedings of CICLing-2005*, pages 100–111.

Mihai Surdeanu, Sanda Harabagiu, John Williams, and Paul Aarseth. 2003. Using predicate-argument structures for information extraction. In *Proceedings of ACL-2003*, pages 8–15.

Szu-ting Yi, Edward Loper, and Martha Palmer. 2007. Can semantic roles generalize across genres? In *Proceedings of HLT-NAACL 2007*, pages 548–555.

Beñat Zapirain, Eneko Agirre, and Lluís Màrquez. 2008. Robustness and generalization of role sets: PropBank vs. VerbNet. In *Proceedings of ACL-08: HLT*, pages 550–558.