Chapter 9

Application to a New Product Release Domain

9.1 Applying RHB$^+$

This section describes results of applying RHB$^+$ to the learning of IE rules in a new product release domain [53].

9.1.1 Illustration of a Learning Process

Table 9.1: Sample Sentences

Article id	Sample Article
#1	"ABC Corp. this week announced that it will release a color printer on Jan. 20."
#2	"XYZ Corp. released a color scanner last month."

Now, we examine the two short notices of new products release in Table 9.1. The following table shows a sample template for articles reporting a new product release.

Template
1. article id:
2. company:
3. product:
4. release date:

Preparation

Suppose that the following semantic representation is obtained from Article 1.

```
(c1) announce( 1,                              % article number
               past,                           % tense
               "this week",                    % tn1
               "ABC Corp.",                    % n1
               (c2) ).                         % n10
(c2) release( 1,                               % article number
              future,                          % tense
              "Jan. 20",                       % tn1
              "ABC Corp.",                     % n1
              "a color printer" ).             % n2
```

The filled template for Article 1 is as follows.

Template 1
1. article id: 1
2. company: ABC Corp.
3. product: a color printer
4. release date: Jan. 20

Suppose that the following semantic representation is obtained from Article 2.

```
(c3) release( 2,                               % article number
              past,                            % tense
              "last month",                    % tn1
              "XYZ Corp.",                     % n1
              "a color scanner" ).             % n2
```

The filled template for Article 2 is as follows.

Template 2
1. article id: 2
2. company: XYZ Corp.
3. product: a color scanner
4. release date: last month

Head Construction

Two positive examples are selected for the template slot "company".

```
company( 1,                 % article number
         "ABC Corp").       % name
company( 2,                 % article number
         "XYZ Corp").       % name
```

By computing the lgg, the following head is obtained:

```
company( Art: number                 % article number
         Co: organization).          % name
```

Body Construction

Generate possible literals[1] by combining predicate names and variables, then check the PWI values of clauses to which one of the literal added. In this case, suppose that adding the following literal with predicate *release* is the best one. After the dynamic sort restriction, the current clause satisfies the stopping condition. Finally, the rule for extracting "company name" is returned. Extraction rules for other slots "product" and "release date" can be learned in the same manner. Note that several literals may be needed in the body of the clause to satisfy the stopping condition.

```
company( Art:number,                 % article number
         Co: organization    )       % name
  :- release( Art,                   % article number
              T: tense,              % tense
              D: time,               % tn1
              Co,                    % n1
              P: product     ).      % n2
```

[1] "literals" here means atomic formulae or negated ones.

Extraction

Now, we have the following semantic representation extracted from the new article:
Article 3: "JPN Corp. has released a new CD player."[2]

```
(c4) release( 3,                      % article number
              perfect_present,        % tense
              nil,                    % tn1
              "JPN Corp.",            % n1
              "a new CD player" ).    % n2
```

Applying the learned IE rules and other rules, we can obtain the filled template for Article 3.

Template 3
1. article id: 3
2. company: JPN Corp.
3. product: CD player
4. release date:

9.1.2 Setting of Experiments

We extracted articles related to the release of new products from a one-year newspaper corpus written in Japanese [3] . One-hundred articles were randomly selected from 362 relevant articles. The template we used consisted of five slots: company name, product name, release date, announce date, and price. We also filled one template for each article. Only the sentences including words in the slots of the filled templates were chosen (which eliminates irrelevant sentences) and analyzed. After that, case frames were converted into atomic formulae representing semantic representation. The semantic representations were given to the learner as background knowledge, and the filled templates were given as positive examples. Precision, recall and F-measure, the standard metrics for IE tasks, are counted by using the remove-one-out cross validation on the 100 examples for each item.

[2] We always assume *nil* for the case that is not included in the sentence.

[3] We used articles from the Mainichi Newspapers of 1994 with permission.

9.1.3 Results

Table 9.2: Learning Results of New Product Release

	company name	product name	release date	announce date	price
precision (all)	82.8%	90.9%	98.9%	100.0%	91.3%
precision (correct)	90.0%	100.0%	98.9%	100.0%	92.6%
recall (all)	80.9%	70.0%	84.2%	88.2%	76.8%
recall (correct)	90.0%	88.6%	93.8%	88.2%	87.5%
F-measure (all)	81.8%	79.1%	91.0%	93.7%	83.4%
F-measure (correct)	90.0%	94.0%	96.3%	93.7%	90.0%
ave. time (sec.)	302.0	396.4	466.9	46.1	202.4

Table 9.2 shows the results of our experiment. Overall, precision was very high. 70-88% recall was achieved with all semantic representations, including errors in case role selection and semantic category selection. With only correct semantic representations, 88-94% recall was achieved.

It is important that the extraction of five different pieces of information showed good results. This indicates that the ILP system RHB$^+$ has a high potential in IE tasks.

9.2 Applying ψ-RHB$^+$

This section describes results of applying ψ-RHB$^+$ to the learning of IE rules in a new product release domain [52].

9.2.1 Setting of Experiments

We extracted articles related to the release of new products from a one-year newspaper corpus written in Japanese [4] . As a preliminary experiment, twenty articles were randomly selected from 362 articles related to new product release. The template we used consisted of four slots: company name, product name, release date and price. We also filled one template for each article. Only the sentences including words in the slots of the filled templates were chosen to remove irrelevant sentences and then analyzed. After that, tagged parse trees were converted into atomic formulae which represent case frames. Those case frames were given to our learner as background knowledge and filled templates as positive examples. Accuracy and recall, the standard measures for IE tasks, are counted by using the remove-one-out cross validation on the twenty examples.

9.2.2 Results

Table 9.3 shows the results of our experiment. Overall, very high accuracy was recored. 40-75% recall was achieved with all case frames including errors in case selection and semantic tag selection. Those selections had errors in a range of 5-35%. With only correct case frames, 60-85% recall were achieved.

It is important to see that the extraction of four different pieces of information showed good results. This indicates that the preliminary version of our learner has high potential in IE tasks.

9.3 Discussion

Because of the limited robustness and performance of current natural language processing techniques (for Japanese texts), errors in semantic tag selection and parsing were found in case frames and they were not negligible. We think, however, that our approach, *i.e.*applying ILP to the learning from case frames, will become more practical as natural language processing techniques make a progress. Just like the

[4] We used the Mainichi Newspapers articles of 1994 with permission.

Table 9.3: Learning Results of New Product Release

	company	product	release date	price
Accuracy	100.0%	92.3%	100.0%	100.0%
Recall (all case frames)	60.0%	40.0%	57.1%	75.0%
Recall (correct case frames)	75.0%	61.5%	61.5%	85.0%
F-measure (all case frames)	75.0%	55.8%	72.7%	85.7%
F-measure (correct case frames)	85.7%	73.8%	76.2%	91.9%
Time (sec.)	1043.1	1119.7	2082.7	958.3

rapid progress seen in part-of-speech taggers, other natural language processing techniques may be improved faster than expected. Along this line, extending ILP to ψ-term capable ILP creates opportunities to apply rich ILP techniques to IE tasks. Therefore, the extension is indispensable.

9.4 Summary

This chapter has illustrated the advantages of applying the hierarchically sorted ILP approach to IE in an new product release domain and described experimental evaluations. The results showed the potential of our approach in IE tasks, extracting information from real-world texts.