Re-interpretation of Rules for Named Entity Task by Probability Assignment

Tatsunori Mori
Graduate School of Environment and Information Sciences
Yokohama National University
79-5 Tokiwadai, Hodogaya, Yokohama 240-8501, JAPAN
mori@forest.eis.ynu.ac.jp

Abstract

This paper proposes a scheme of re-interpreting rules for recognizing named entities(NE) by estimating probability of NE-tag assignment about every sub-pattern in each rule. An NE rule is usually consist of a matching pattern and an NE assignment. However, it is not easy to create such NE rules by paying attention to both of them. Especially, it is hard for humans to take account of all of possible NE assignment for a pattern. Thus, in our re-interpretation, we firstly discard all of NE assignment information and use matching patterns only as detectors of word sequence. The assignment of NE tags for each matching pattern is newly obtained in probabilistic form by examining every portion of training corpora which matches with the pattern. Since probability of NE-tag assignment is provided for every sub-pattern in each original pattern, all of sub-patterns may contribute to finding the globally optimum assignment of NE tags.

1 Introduction

Information extraction is one of the important technologies to obtain useful information from a huge amount of documents available in machine readable form. Since identification of named entities(NEs) plays a fundamental role in information extraction, many researchers have been engaged in the development of high-performance NE recognizers. NE recognition task is also one of the main tasks in Message Understanding Conferences(MUC-6 program committee, 1996; MUC-7 program committee, 1998), MET and IREX-NE task(IREX Committee, 1999). Through those conferences, it has been shown that NE recognizers based on handcrafted rules are superior to other types of systems.

An NE rule is usually consist of a matching pattern and an NE assignment. However, it is not easy to create such NE rules by paying attention to both of them. Especially, it is hard for humans to take account of all of possible NE assignment for a pattern.

As a way to cope with the problem, in this paper, we propose a scheme of re-interpretation of NE rules by estimating probability of NE-tag assignment about every sub-pattern in each rule. Altough the main target of our scheme are hand-crafted rules, it can be apply to other types of NE rules which comes from other sources, like the result of automated rule generation.

The main idea of our scheme is that each NE rule can be regarded as the combination of the following two independent functions, and thus, mechanisms which realize those functions can be derived from different sources.

Word sequence detection: Detecting some useful sequences of words by matching them with a pattern, which is usually a sequence of sub-patterns.

NE tag assignment: Assigning NE tag(s) to a part/whole of the detected word sequences.

In the process of hand-crafting NE rules, humans usually assign at most one NE tags to (sub-)patters, namely in deterministic way, because we do not take account of all possibility of NE-tag assignment in large number of examples. Thus, such NE tag assignment is not comprehensive and has potentially a lot of exceptions.

Accordingly, in our re-interpretation, we firstly discard all of NE assignment information. Only patterns are used as word sequence detectors. The assignment of NE tags for each matching pattern is newly obtained in probabilistic form by thoroughly examining every portion of training corpora, which are annotated with NE-tags by hand. Since probability of NE-tag assignment is provided for every sub-pattern in each original pattern, all of sub-patterns may contribute to finding the globally optimum assignment of NE tags.

Therefore, our scheme can be regarded as the integration of 1) humans' speciality, namely, complicated pattern generation, and 2) computers' speciality, namely, probabilistic estimation with large corpora.

2 General Scheme of Probabilistic NE Recognition

In this section, we describe a general scheme of probabilistic NE recognizer.

2.1 General Scheme of NE Recognition

As shown in Figure 1, the general function of an NE recognizer is to assign a pair of NE tags (the right column of Figure 1) to each word in the input text (the left column of Figure 1). By tokenizer, morphological analyzer and part-of-speech (POS) tagger, the text is tokenized into a sequence of words, or morphemes, and each morpheme is attached some helpful information for the NE recognizer, such as a POS information. The sequence of tuples of morpheme and other information is an input of the NE recognizer (the left column of Figure 1). The NE recognizer selects one pair of NE tags for each word depending on its process. The pair consists of the NE tag at the start point of the word and the NE tag at the end point of the word. The sequence of pairs is the output of NE recognizer(the right column of Figure 1). Since one NE can consist of more than one word, a tag represents the role of a word in an NE. Namely, tags with suffixes '`-st`','`-md`' and '`-ed`' represent the start, middle and end point of an NE. In Figure 1, the output shows that the sequence of `Word4`, `Word5` and `Word6` makes one NE `PER`(`PERSON`).

```
     Input                                   Output
Morpheme Info                            Start    End
------------                            -------------
Word1   Art                             None    None
Word2   Noun                            None    None
Word3   Verb                            None    None
Word4   Noun      ==process=>           PER-st  PER-md
Word5   Noun                            PER-md  PER-md
Word6   Noun                            PER-md  PER-ed
Word7   Prep                            None    None
  :       :                                   :
```

Figure 1: General scheme of NE system

In the rest of this section, we will describe two types of process for NE recognition. First one is a traditional rule-based system. Second one is a system based on probability, which we adopted.

Note that this type of probability-based systems also have been proposed by several researchers such as (Sekine et al., 1998; Borthwic, 1999). The scheme described in this section can be considered as a generalization of schemes proposed by those former researches.

To the scheme of probability-based systems, our main contribution is that we propose a re-interpretation of NE rules and make all sub-patterns of NE rules contribute to 'local assignment of probability of NE' described in Section 2.3.

2.2 Traditional rule-based NE system

In the traditional rule-based NE systems, each NE rule performs as a small NE recognizer. For example, let us consider the NE rule shown in Figure 2. The first column contains the serial numbers of sub-patterns, which are used for explanation only. The second column contains the sequence of sub-patterns which represents the sequence we want to detect in the input. Each sub-pattern is usually one or more iterations of a morpheme pattern which matches with a pair of word and supplemental information like POS in the input. In description of patterns, we use the Perl-like notation, which is a variant of the regular expression, such as '`*`' for zero-or-more time iteration, '`+`' for one-or-more time iteration, '?' for zero-or-one time iteration and '[...]' for a character class, which matches with one character in the bracket(Wall et al., 1996). The third column contains the NE tags which are associated with the sub-patterns in the same row. When some sequence in input matches with the sequence of sub-patterns, the rule fires and outputs corresponding NE tags for the morphemes. In this example, when the sequence of all sub-patterns 1 to 7 is matched with some sequence of the input pairs, the sequence of input pairs matched with the sub-pattern 1 are assigned the NE tags '`PER`-'(`PERSON`), similarly the sequence matched with the sub-pattern 6 and 7 are assigned the NE tags '`ORG`-'(`ORGANIZATION`).

We call this type of system 'deterministic NE rule system', because NE tag assignment is deterministically performed only when an NE rules is applicable.

```
-- ------------------- -----------------
No.       Sub-pattern    NE tag assignment
     Morpheme     Info    Start       End
-- ------------------- -----------------
1  ([A-Z][a-z]* Noun)+  PER-st    PER-ed
2  (,           Punc)
3  ((a|the)     Art)?
4  ([a-z]+      Noun)+
5  (of          Prep)
6  ([A-Z][a-z]* Noun)+  ORG-st    ORG-md
7  (Corp.       Noun)   ORG-md    ORG-ed
-- ------------------- -----------------
```

Figure 2: An example of NE rule for English text, which detects the word sequence '$\boxed{1}$, a ... of $\boxed{2}$ Corp.' and assigns the NE tags `PER` and `ORG` to the sequences corresponding to $\boxed{1}$ and '$\boxed{2}$ Corp.', respectively.

2.3 NE system based on Probability

The difference between probability-based NE systems and other types of systems is found in NE tag assignment. NE tag assignment of probability-based systems consists of two stages: Local assignment of probability of NE tags and Global selection of most plausible sequence of NE tags.

In the first stage, a system of this type assigns each morpheme, not one plausible NE tag, but the probabilities of all of NE tag candidates. Each probability is associated with one tag, and it represents how often the tag appears at the position. The second column of Figure 3 is an example of the local assignment of probability. Since we have to take account of all possibility for tag assignment, the tag '`NONE`' is introduced to represent that any NE tags are assigned to the position.

In the second stage, to generate the final output like the third column of Figure 3, the system globally finds the most plausible sequence of NE tags under the adjacency constraint. Usually, Viterbi algorithm is used to find the optimum path. The adjacency constraint consists of some rules to maintain the consistency of NE sequence. For example, the next NE tag of the tag '`PER-st`'(PERSON start) must be one of the tags '`PER-ed`' (PERSON end) or '`PER-md`'(PERSON middle).

One of the preferable features of this scheme is that the result of several different types of NE recognizers can be locally integrated into the probability of NE tags at each morpheme. Therefore, a multi-strategic system can be constructed in the probabilistic way. Although there are several ways to integrate probabilities derived from different evidences, our system uses the average of probabilities because of its simplicity.

```
-------------   -------------------------------   -------------
     Input       Probability for each NE tag         Output
Morpheme Info         Start           End         Start   End
-------------   -------------------------------   -------------
  :      :              :              :                :
Word1  Art     (.0 .0 .0 .0 1.)(.0 .0 .0 .0 1.)  None   None
Word2  Noun    (.1 .0 .1 .0 .8)(.0 .1 .0 .1 .8)  None   None
Word3  Verb    (.0 .0 .0 .0 1.)(.0 .0 .0 .0 1.)  None   None
Word4  Noun    (.5 .1 .2 .1 .1)(.1 .5 .1 .2 .1)  PER-st PER-md
Word5  Noun    (.3 .5 .0 .0 .2)(.3 .5 .0 .0 .2)  PER-md PER-md
Word6  Noun    (.4 .4 .0 .0 .2)(.6 .3 .0 .0 .2)  PER-md PER-ed
Word7  Prep    (.0 .0 .0 .0 1.)(.0 .0 .0 .0 1.)  None   None
  :      :              :              :
  :      :              :              :
------------   -------------------------------   ------------
               Start: (p(PER-st) p(PER-md) p(ORG-st) p(ORG-md) p(None))
               End  : (p(PER-ed) p(PER-md) p(ORG-ed) p(ORG-md) p(None))
```

Figure 3: General scheme of Probability-based system

3 Probabilistic Re-interpretation of NE Rules for Probabilistic NE Rule Systems

In this section, we describe a probabilistic re-interpretation of NE rules. With the re-interpretation we can not only weave all possibility of NE assignment into NE rules in probabilistic way, but also harmonize NE rules with other types of probabilistic NE recognizers, which are mainstream of recognizer based on machine learning. This is the main contribution of us.

3.1 Re-interpretation of NE Rules

As described before, the function of NE rule can be decomposed into two sub-functions: word sequence detection and NE tag assignment. For example, in Figure 2, the first function 'word sequence detection' is found in the sequence of sub-patterns in second column. The second function 'NE tag assignment' can be seen in the relation between each sub-pattern and the NE tag in the last column.

When we make such rules by hand, we consider both of those functions simultaneously. Since we cannot take account of the probability of the NE tag assignment in large number of example, we have no choice but to select, at most, one plausible NE tag for each sub-pattern. That is, NE assignment made by hand unwillingly becomes deterministic. Consequently, the important part in making NE rules is the accurate description of sequence detectors, or patterns, to exclude matching with undesirable word sequence. However, it is hard for human beings to keep the pattern from exceptions by seeing huge number of examples. Moreover, we usually gives NE tag assignments to not all of sub-patterns but only a few of them, because we cannot pick up uncertain assignments for the same reason described above.

The problem comes from deterministic assignment of NE tags, especially, in the case of maintaining rules fully by hand. When we use NE rules made by hand with deterministic assignment of NE tags, the examples which are not consistent with the NE assignment of a rule must be the exeptions, and do not contribute to NE assignment. However, such instances show other possibilities of NE assignment, even though their probability is not so high. Therefore, the NE tag assignment should be expressed in some manner in which all possible candidates of assignment are maintained.

Conversely, if we introduce the way to deal with the non-deterministic assignment of NE tags, the following things are expected because we can weave all possibility of NE assignment into NE rules.

- With other resources, we can give the information about NE assignment for all sub-patterns even if some of them originally do not have NE assignment. Thus, we would make full use of patterns.
- We can use patterns which are not so accurate in NE detection.
- We are able to filter out rules having poor-quality for deterministic use, if we can obtain probabilistic distribution of NE assignment.

As non-deterministic assignment, we use probabilistic NE assignment described in Section 2.3. Since humans cannot estimate probability for NEs with a large corpus, we firstly discard all of NE assignment information on NE rules. Only patterns are used as word sequence detectors. Therefore, users do not have to describe the NE assignment anymore in our scheme. What remains for users is to describe the structure of sequence detectors, or patterns, to find some peculiar contexts.

Probabilistic assignment of NE for each patterns are estimated with a given corpus. The process is described in the next section.

3.2 Estimate of Probability of NE assignment

Fortunately, there are several corpora annotated with NE-tags by hand. Using such corpora as training data, we can estimate probabilities of NE tag assignment at every sub-pattern in the pattern.

The algorithm to estimate probabilities of NE tag assignment of a pattern is straightforward as described bellow:

- Preparation

  Prepare a set of counters 'NE counters' for each sub-patterns. Each counter corresponds to one NE tag.

- Counting

  1. Find the next portion of corpus which is matched with the pattern. If there is not such portion, then end the counting.

2. If there is an NE tag at the position of a sub-pattern, increase the corresponding NE counter of the sub-pattern.
3. If there is no NE tag at the position of a sub-pattern, increase the special NE counter 'NONE' of the sub-pattern.
4. Goto Step 1.

- Deriving Probability

  The probability that an NE tag appears at the position of a sub-pattern is simply estimated as $\frac{\text{Count of a NE tag}}{\text{Total count of all NE tags}}$.

The actual data structure is more complicated than described above. Since one sub-pattern may be an iteration of a basic pattern and may be matched with more-than-one words, we have to maintain the probabilities of NE tags appearing not only at the beginning and ending of one word, but also at the middle words of word sequence. Thus we describe the probability information for each sub-pattern as a quadruplet of list of probabilities. In our system, members in a quadruplet are labeled with the position names: `ST`(start position of the word sequence), `ED`(end position), `MD-ST`(start of a middle word in the sequence) and `MD-ED`(end of a middle word in the sequence). Those labels correspond to position names of NE tags in a word sequence, as shown in Figure 4. Each list of probabilities in the quadruplet represents probabilities of NE tag candidates at the corresponding position described above.

| ST MD-ED | | MD-ST MD-ED | ... | MD-ST MD-ED | | MD-ST ED |

Figure 4: Position names of NE tags in a word sequence (each box represents one word)

Another technical issue is the smoothing. In the estimation of probability, we are usually suffering from the sparseness of data. For example, there are some cases in which the probabilities of almost all candidates of NE tags become zero. Such zero probability is the main cause of problem that we can not find the globally optimum assignment of NE tags. In order to avoid this problem, we use a smoothing method, called 'M-estimate' defined as follows(Mitchell, 1997):

$$\begin{aligned} p'(T) &= \frac{freq(T) + M * p(T)}{n + M}, \qquad (1) \\ M &= m * n, \end{aligned}$$

where $p'(T)$ is the estimated probability that the NE tag T appears at the position, n is the frequency that the rule is applicable, $freq(T)$ is the frequency of the NE tag T appearing at the position, $p(T)$ is the prior probability that the tag T appears in corpora, and m is the parameter to balance two factors $freq(T)$ and $p(T)$.

Figure 5 shows an example of probabilistic NE assignment. Since the combination of a pattern and probabilistic NE assignment can be regarded as a new NE rule, we call this type of data structure 'probabilistic NE rule.'

3.3 Application of Probabilistic NE Rules to Unseen texts

We can use probabilistic NE rules to locally assign probabilistic distribution of NE tags to a certain word in unseen texts. In general, more than one NE rules may be applied to one word, that is, there may be several evidences of NE tags for one word. Therefore, the final local probability of NE tag candidates for a word should be calculated by combining those plural distributions of probability. Among various ways to make a combination, we use an arithmetical mean, which is the simplest method.

Once all of local probabilities of NE tag candidates are fixed, Viterbi algorithm will be used to find the globally optimum assignment of NE tags.

4 Features of our Scheme

Our scheme described above has, at least, two new and strong features that have never found in the NE recognizers proposed so far:

- Usually, in handcrafted rules, only one NE tag can be assigned to one sub-pattern. Sometimes, however, individual NE rules generate inconsistent results. To remedy this situation, in our scheme, the system can offer multiple NE tag candidates having their own probability for one word, and those candidates can be used to find a globally optimum assignment of NE tags through Viterbi algorithm.

- The system can estimate the probability of NE tag candidates at the position to which the rule developer does not assign any tags. That is, every sub-pattern in NE rules contributes to finding a globally optimum assignment of NE tags.

Based on those features, we expect the following points:

Expectation 1: Our probability-based scheme of NE rules will have the same or better effectiveness than traditional rule-based deterministic NE systems.

Expectation 2: In the deterministic use of NE rules, if the quality of some of NE rules is not good, those rules deteriorate the total effectiveness of the NE recognizer. However, in our scheme, the quality of each rule is evaluated with NE-tagged corpora and the result of evaluation is preserved as probability of assignment of NE tags. The sub-pattern in low quality will be assigned discounted probability. Therefore, we will be able to use automatically generated NE rules, which are usually made from corpora, as well as handcrafted rules.

Expectation 3: Based on the probability distribution of NE assignment, we will be able to filter out poor-quality NE rules to improve the effectiveness of the whole NE rule set. After that, if we prefer deterministic rules only, we may remove probability information from each rules by selecting most plausible NE tags. NE rules in deterministic style do not require special processing like the best path search by Viterbi algorithm, and can be translated into more efficient form like finite state transducers(Roche and Schabes, 1997).

```
--- --------------------  -------------------------------------------------------
No.     Sub-pattern               Probabilities of NE tag assignment
     Morpheme    Info
--- --------------------  -------------------------------------------------------
1   ([A-Z][a-z]*  Noun)+  ST:(.9 .1 .0 .0 .0), ED:(1. .0 .0 .0 .0),
                            MD-ST:(.2 .8 .0 .0 .0), MD-ED:(.3 .7 .0 .0 .0)
2   (,            Punc)   ST:(.0 .0 .0 .0 1.), ED:(.0 .0 .0 .0 1.), MD-ST:, MD-ED:
3   ((a|the)      Art)?            :
4   (.*           Noun)+           :
5   (of           Prep)            :
6   ([A-Z][a-z]*  Noun)+  ST:(.0 .0 .9 .1 .0), ED:(.0 .0 .0 1. .0),
                            MD-ST:(.0 .2 .0 .8 .0), MD-ED:(.0 .2 .0 .8 .0)
7   (Corp.        Noun)   ST:(.0 .0 .0 1. .0), ED:(.0 .0 1. .0 .0), MD-ST:, MD-ED:
--- --------------------  -------------------------------------------------------
                          ST   : (p(PER-st) p(PER-md) p(ORG-st) p(ORG-md) p(None))
                          ED   : (p(PER-ed) p(PER-md) p(ORG-ed) p(ORG-md) p(None))
                          MD-ST: (p(PER-st) p(PER-md) p(ORG-st) p(ORG-md) p(None))
                          MD-ED: (p(PER-ed) p(PER-md) p(ORG-ed) p(ORG-md) p(None))
```

Figure 5: An example of probabilistic NE assignment

Expectation 4: The result of recognizing NEs will be able to be easily integrated with results of other probability-based schemes in probabilistic way, such as an arithmetical mean of probability. The effectiveness of the integrated system is expected to be higher than original ones.

In the following sections, we will examine the above expectations through several experiments.

5 Condition of Experiments

We use the following three Japanese corpora with NE tags, which are distributed in IREX-NE task(IREX Committee, 1999). Note that their contents are mutually disjunctive.

- CRL NE data set(CRL corpus, hereafter): 1174 articles of Mainichi-shinbun Newspaper.
- IREX-NE training data set in topic of arrest(ARREST corpus): 23 articles of Mainichi-shinbun Newspaper.
- IREX-NE formal-run data set of general topics (GENERAL corpus): 72 articles of Mainichi-shinbun Newspaper, which contain 361 `ORGANIZATION`s, 338 `PERSON`s, 413 `LOCATION`s, 48 `ARTIFACT`s, 260 `DATE`s, 54 `TIME`s, 15 `MONEY`s and 21 `PERCENT`s.

In the following examinations, we use CRL corpus and ARREST corpus as the training data set, and GENERAL corpus as the test data set. Each corpus is segmented and pos-tagged by JUMAN 3.6(Kurohashi and Nagao, 1998) which is a Japanese morphological analyzer. In smoothing, we use 0.2 for the parameter m of (1).

6 General Effectiveness of our Scheme

In order to examine Expectation 1, we compare the effectiveness of our probability-based system with the normal deterministic NE rule system. In this comparison, both of those systems use the same set of NE rules, which we handcrafted for IREX-NE task by examining mainly CRL corpus. Hereafter we refer to the set of NE rules as 'IREX-NE rule set'. The condition of this comparison is as follows:

Deterministic NE rule system:
Deterministic Scheme and IREX-NE rule set.

Our probabilistic NE rule system:
Probability-Based scheme, IREX-NE rule set and the probabilistic NE assignment derived form CRL corpus and ARREST corpus.

Test set: GENERAL corpus.

Table 1 and Table 2 are the results of the deterministic NE rule system and our probabilistic NE rule system, respectively. The label 'All' in those tables corresponds to the case that all of NE tags are considered in evaluation. Those tables show that, in the case that the NE rules are constructed by hand, our probability-based scheme of NE rules has the almost same effectiveness as the deterministic NE rule system. This result supports our Expectation 1.

Table 1: Effectiveness of Deterministic NE rule system

	Recall	Precision	F-measure
`ORGANIZATION`	44.99	76.75	56.73
`PERSON`	49.86	56.91	53.15
`LOCATION`	61.06	72.36	66.23
`ARTIFACT`	4.08	15.38	6.45
`DATE`	88.04	83.79	85.86
`TIME`	96.55	76.71	85.49
`MONEY`	86.67	86.67	86.67
`PERCENT`	80.95	100	89.47
All	59.34	72.19	65.14

7 Effectiveness for Poor-quality NE Rules

In order to examine Expectation 2, we prepared poor-quality NE rules for '`PERSON`', which are semi-automatically and exhaustively derived form the CRL corpus as described in Section 7.1. We compare our probabilistic NE rule system with the deterministic NE rule system on the same condition that the set of poor-quality NE rules is used.

7.1 Exhaustive NE Rules for '`PERSON`'

We consider three types of NE rules for '`PERSON`.' One rule and two templates for rules are shown in

Table 2: Effectiveness of Probabilistic NE rule system

	Recall	Precision	F-measure
ORGANIZATION	41.39	74.19	53.14
PERSON	52.68	57.72	55.08
LOCATION	65.38	71.77	68.43
ARTIFACT	4.08	33.33	7.27
DATE	87.68	78.57	82.88
TIME	87.93	96.23	91.89
MONEY	86.67	86.67	86.67
PERCENT	80.95	100	89.47
All	59.85	71.65	65.22

Table 3: POS marks used in NE rules for PERSON

P:	Noun which is analyzed as a part of PERSON by the morphological analyzer
H:	Noun which appears in the prefix dictionary
T:	Noun which appears in the suffix dictionary
N:	Noun which does not appear in either the prefix or suffix dictionaries
S:	One of other suffixes of Nouns
O:	Word other than Noun

Figure 6,7 and 8. The templates generate many NE rules for 'PERSON.' The rule No.1 in Figure 6 finds the word sequence of PERSON by expanding the extent of PERSON, from one word which is analyzed as a part of PERSON by the morphological analyzer, into a possible series of words. The rule generated by the template No.1 in Figure 7 finds PERSON by spotting a suffix expression of PERSON and expanding the extent of word sequence backward. In the contrast to that, the rule generated by the template No.2 in Figure 8 finds PERSON by spotting a prefix expression of PERSON and expanding the extent of word sequence forward.

Since prefix and suffix expressions of PERSON are used in those rules, those rules need some preprocessing which spots the expressions in the target text. To do that, we make a prefix dictionary and a suffix dictionary of PERSON. The prefix dictionary of PERSON contains all of words which appear just before NEs of PERSON in the training corpora. The suffix dictionary of PERSON contains all of words which appear just after NEs of PERSON in the training corpora. With the dictionaries and POS information given by the morphological analyzer, the preprocessing program add one of POS marks described in Table 3 to each morpheme. In Figure 7 and 8, `Prefix` or `Suffix` is one of the expressions in the prefix or suffix dictionaries, respectively.

From those rules and templates, we obtained 3115 NE rules for 'PERSON'. Hereafter, we refer to this rule set as 'PERSON rule set'. Note that some of expressions in the prefix and suffix dictionaries may appear around word sequences of other NEs or non-NE words. Therefore, the quality of the rule set derived by this method is not so high.

7.2 NE Results of both Methods

On the condition described below, we compare the effectiveness of both methods.

Deterministic NE rule system:
Deterministic Scheme and PERSON rule set.

```
---  ---------------  ----------------
No.    Sub-pattern    NE tag assignment
     Morpheme Info     Start       End
---  ---------------  ----------------
1    (.+          N)+  PER-st    PER-md
2     .+          P    PER-md    PER-md
3    (.+          N)+  PER-md    PER-ed
---  ---------------  ----------------
```

Figure 6: NE rule No.1 for 'PERSON'

```
---  ---------------  ----------------
No.    Sub-pattern    NE tag assignment
     Morpheme Info     Start       End
---  ---------------  ----------------
1     .+       [OH]
2    (.+         N )+ PER-ST    PER-ED
3     Suffix   [TS]
---  ---------------  ----------------
```

Figure 7: NE rule template No.1 for 'PERSON'

Our probabilistic NE rule system
Probability-Based scheme, PERSON rule set and the probabilistic NE assignment derived form CRL corpus and ARREST corpus.

Test set: GENERAL corpus.

Table 4 shows that the result of the deterministic NE rule system has high recall rate but precision rate is very low. On the other hand, in the result of our probabilistic NE rule system shown in Table 5, the precision rate is improved and the F-measure increases by about 10 point to 65.08. This result supports our Expectation 2.

8 Filtering out Poor-quality NE Rules

In order to examine Expectation 3, we conduct an experiment about filtering NE rules based on the entropy of probabilistic distribution of NE tags on the patterns.

8.1 Entropy of Probabilistic Distribution of NE Tags on Patterns

As we described in Section 3.2, each sub-pattern has a quadruplet of list of probability, each member of which is labeled with one of position names, ST(start position of the word sequence), ED(end position), MD-ST(start of a middle word in the sub-pattern) and MD-ED (end of a middle word in the sub-pattern). Each information labeled with a position name contains a list of probability of NE tag candidates. Thus, we can easily compute the entropy of the probability distribution for each position. Since the entropy shows the uniformity of the distribution, a low entropy means that the sub-pattern contributes to identifying the

```
---  ---------------  ----------------
No.    Sub-pattern    NE tag assignment
     Morpheme Info     Start       End
---  ---------------  ----------------
1     Prefix      H
2    (.+          N)+ PER-st    PER-ed
3     .+       [STO]
---  ---------------  ----------------
```

Figure 8: NE rule template No.2 for 'PERSON'

Table 4: Result for PERSON with the deterministic NE rules

Recall	Precision	F-measure
71.27	45.67	55.67

Table 5: Result for PERSON with the probabilistic NE rules

Recall	Precision	F-measure
61.41	69.21	65.08

most likely NE tag. Note that each NE rule has several sub-patterns and each sub-pattern has a quadruplet of probability information. Therefore, we define 'entropy of NE rule' as the average of the entropy values of all positions in the rule. We expect that by removing high entropy NE rules we can refine the set of NE rules and the refined rule set achieves good effectiveness even if the rule set is used in traditional deterministic NE rule system.

8.2 Filtering based on Entropy of NE Rules

We performed an experiment of filtering with PERSON rule set in Section 7.1.

Firstly, the entropy of each NE rule is calculated, then we remove NE rules whose entropy is more than several predetermined threshold values. Secondly, each of filtered set of rules is examined on the condition as follows:

System: Deterministic Scheme and filtered PERSON rule set. In filtering, the probabilistic NE assignment derived form CRL corpus and ARREST corpus is used.

Test set: GENERAL corpus.

As known from the result shown in Table 6, the F-measure is improved by about 11 point after filtering out NE rules whose entropy is more than 0.2. This result supports our Expectation 3.

9 Combining Probabilistic NE Rules with other Schemes

Since, as described in Section 3.3, the probabilistic NE rule system usually utilizes several evidences, it is easy to add new source of evidence. We can use the output of any types of NE recognizers, as long as the output is in the form of probability.

In this section, in order to examine Expectation 4, we combine our probabilistic NE rule system with a machine-leaning-based NE recognizer we developed(Matsuo and Mori, 1999).

9.1 NE recognizer based on Decision Tree Learning

The principle of the machine-leaning-based system is the same as the system proposed in Sekine(Sekine et al., 1998). Those systems firstly make a decision tree(Quinlan, 1993) from NE-tagged corpora. The decision tree classifies a trigram of words with POS information into several predetermined 'NE classes', which correspond to leaves of the tree. The 'NE class', in this scheme, is not an NE tag but a list of probability of NE tags, which is the same type of information as the probabilistic NE assignment in this paper.

The machine-leaning-based system is similar to Sekine's system, but improved. It uses the EDR concept dictionary(Japan Electrronic Dictionary Research Institute, 1995), which is one of the largest thesauri for Japanese, to generalize words into more general concepts in order to improve the recall.

Table 7 shows the result of recognizing PERSONs only with the system based on decision tree learning.

Table 7: NE recognizer based on decision tree

Recall	Precision	F-measure
70.12	71.17	70.64

9.2 Combining Probabilities of NE Tags

We made an experiment to combine the probabilistic NE rule systems and the decision-tree based system on the following condition.

Decision-tree based system: Decision-tree made from CRL corpus and ARREST corpus.

Probabilistic NE rule system Probability-Based scheme, PERSON rule set and the probabilistic NE assignment derived form CRL corpus and ARREST corpus.

Test set: GENERAL corpus.

Combination method in the experiment is very simple. We regard the decision-tree as one of NE rules in the probabilistic NE rule system, and combine probability by an arithmetical mean as described in Section 3.3.

As shown in Table 8, the combination improves the F-measure in comparison with the results of both component systems shown in Table 5 and Table 7. Although precisions was little bit degraded, recalls increase significantly from 8 to 16 point.

Table 8: Result of Combination

Recall	Precision	F-measure
77.81	67.96	72.25

10 Related Works

There has been amount of research concerning with NE task in conferences like MUC and MET in TIPSTER Project and IREX-NE in Japan. Generally speaking, there are two types of approaches, namely, handcrafted NE rules and machine learning.

The approach based on handcrafted NE rules is widely used and acheves high performance. In some systems, those NE rules are converted into the finite state transducers in order to achieve higher efficiency(Roche and Schabes, 1997). From the viewpoint of refinement of handcrafted NE rules, Nobata(Nobata and Sekine, 1998) proposes the system which has an interface for users to edit

Table 6: Filtering based on Entropy of NE rules

Threshold	0.10	0.15	0.20	0.25	0.30	No filtering
Recall	64.23	64.51	67.89	67.89	67.89	71.27
Precision	65.71	65.43	64.96	60.86	58.50	45.67
F-measure	64.96	64.97	66.39	64.18	62.85	55.67

NE rules interactively according to the result of pattern matching with an input example sentence.

As for machine learning, there are systems based on different types of learning method, for example, decision tree learning (Sekine et al., 1998; Matsuo and Mori, 1999), maximum entropy method(Borthwic, 1999), hidden Markov model(Freitag and McCallum, 1999). Freitag(Freitag, 1998) also developed a multistrategic learning model of NE which integrates three different learning method in terms of regression models.

In contrast, our scheme can be used in many ways to contribute to those approaches. As for rule based approach, our method would provide with one measure of effectiveness of NE rules. It may be helpful not only in the case of filtering out poor-quality NE rules as described in this paper, but also in the case of revision of NE rules with some user interfaces. As for machine learning approaches, our scheme offers a way to integrate NE rules into other schemes of probability-based machine learning.

11 Conclusion

In this paper, we described a probabilistic reinterpretation of NE rules based on NE-tagged corpora. We also showed that our scheme has some good features including filtering NE rules and provides the easy way to integrate it with other schemes.

The following points should be parts of our future works.

- Experiments of integration with other types of leaning schemes, e.g. learning based on maximum entropy method.
- Detailed experiments with NE rules for not only '`PERSON`' but also other NE tags.

References

Andrew Borthwic. 1999. A Japanese named entity recognizer constructed by a non-speaker of Japanese. In *Proceedings of IREX workshop*, pages 187–194, September.

Dayne Freitag and Andrew Kachites McCallum. 1999. Information extraction with HMMs and shrinkage. In *Proceedings of AAAI-99 workshop: Machine Learning for Information Extraction*.

Dayne Freitag. 1998. Multistrategy learning for information extraction. In *Proceedings of The Fifteenth International Conference on Machine Learning*. Morgan Kaufmann Publishers, July.

IREX Committee, editor. 1999. *Proceedings of IREX workshop*. IREX Committee. (in Japanese).

Japan Electrronic Dictionary Research Institute, 1995. *EDR Electronic Dictionary Specification.*

Tadao Kurohashi and Makoto Nagao, 1998. *Japanese Morphological Analysis System JUMAN version 3.6 Manual.* Kyoto University. (in Japanese).

Mamoru Matsuo and Tatsunori Mori. 1999. Named entity system based on machine learning and edr concept dictionary. In *Symposium on Natural Language Processing for Knowledge Discovery.* (URL http://www.pluto.ai.kyutech.ac.jp/plt/inuilab/pub/NLP_Sympo99/).

Tom M. Mitchell. 1997. *Machine Learning.* McGraw-Hill.

MUC-6 program committee, editor. 1996. *Proceedings of the sixth Message Understanding Conference (MUC-6).* MUC-6 program committee, Morgan Kaufmann Publishers Inc.

MUC-7 program committee, editor. 1998. *Proceedings of the seventh Message Understanding Conference (MUC-7).* MUC-7 program committee. (URL http://www.muc.saic.com/).

Chikashi Nobata and Satoshi Sekine. 1998. Development and evaluation of Japanese information extraction system. SIG Notes 98-NL-127-16, Information Processing Society of Japan, September. (in Japanese).

J. Ross Quinlan. 1993. *C4.5: progarms for machine learning.* Morgan Kaufmann Publishers, May.

Emmanuel Roche and Yves Schabes, editors. 1997. *Finite-State Language Processing.* MIT Press, Cambridge.

Satoshi Sekine, Ralph Grishman, and Hiroyuki Shinnou. 1998. A decision tree method for finding and classifying names in Japanese texts. In *Proceedings of the Sixth Workshop on Very Large Corpra.*

Larry Wall, Tom Christiansen, and Randal L. Schwartz. 1996. *Programming Perl.* O'Reilly and Associates,Inc.