Minimally Supervised Japanese Named Entity Recognition: Resources and Evaluation

Takehito Utsuro*, **Manabu Sassano****

*Department of Information and Computer Sciences, Toyohashi University of Technology
Tenpaku-cho, Toyohashi, 441-8580, Japan
utsuro@ics.tut.ac.jp

**Fujitsu Laboratories, Ltd.
4-4-1, Kamikodanaka, Nakahara-ku, Kawasaki 211-8588, Japan
sassano@flab.fujitsu.co.jp

Abstract

Approaches to named entity recognition that rely on hand-crafted rules and/or supervised learning techniques have limitations in terms of their portability into new domains as well as in the robustness over time. For the purpose of overcoming those limitations, this paper evaluates named entity chunking and classification techniques in Japanese named entity recognition in the context of minimally supervised learning. This experimental evaluation demonstrates that the minimally supervised learning method proposed here improved the performance of the seed knowledge on named entity chunking and classification. We also investigated the correlation between performance of the minimally supervised learning and the sizes of the training resources such as the seed set as well as the unlabeled training data.

1. Introduction

It is widely agreed that named entity recognition is an important step for various applications of natural language processing such as information retrieval, machine translation, information extraction and natural language understanding. In the English language, the task of named entity recognition was one of the tasks of the Message Understanding Conference (MUC) (e.g., MUC-7 (MUC, 1998)) and has been studied intensively. In the Japanese language, several recent conferences, such as MET (Multilingual Entity Task, MET-1 (Maiorano, 1996) and MET-2 (MUC, 1998)) and IREX (Information Retrieval and Extraction Exercise) Workshop (IREX Committee, 1999), focused on named entity recognition as one of their contest tasks, thus promoting research on Japanese named entity recognition.

In Japanese named entity recognition, it is quite common to apply morphological analysis as a preprocessing stage and to segment the sentence string into a sequence of morphemes. Then, hand-crafted pattern matching rules and/or statistical named entity recognizer are applied to recognize named entities. However, in named entity recognition, it is often the case that new entities are introduced as the domain of the texts changes, and even in the same domain, as for newspaper articles for example, where it is quite natural to assume that entities which have never been encountered in past articles are newly introduced in future articles. Therefore, approaches to named entity recognition that rely on hand-crafted rules and/or supervised learning techniques have limitations in terms of their portability into new domains as well as in the robustness over time.

For the purpose of overcoming those limitations, the minimally supervised approach to named entity recognition proposed by (Collins and Singer, 1999; Cucerzan and Yarowsky, 1999) is more promising. The central idea of the minimally supervised approach relates to bootstrapping utilizing redundancy in unlabeled data, with the help of a minimal number of labeled data as initial seeds. The idea of utilizing redundancy in the unlabeled data for entity tagging was proposed by (Yarowsky, 1995) in the context of word sense disambiguation, in which cross-model redundancy in contextual features of word sense disambiguation is exploited. This idea was then theoretically formalized in the context of computational learning theory by (Blum and Mitchell, 1998) and has been termed as "co-training". (Collins and Singer, 1999; Cucerzan and Yarowsky, 1999) independently applied co-training algorithms to named entity classification/ recognition, both of which utilize the redundancy of morphological and contextual evidence of named entity classification/recognition.

Following this prior research which employed the minimally supervised approach to named entity classification/recognition, in this paper we focus on a minimally supervised approach to Japanese named entity recognition. In Japanese named entity recognition, named entities to be recognized may have different segmentation boundaries from those of morphemes obtained by the morphological analysis. For example, in our analysis of the IREX workshop's training corpus of named entities, 44% of the named entities have segmentation boundaries that differ from boundaries obtained through morphological analysis by a Japanese morphological analyzer BREAKFAST (Sassano et al., 1997) (section 2.). Thus, in Japanese named entity recognition, the most difficult problems include this issue of how to recognize such named entities that have a segmentation boundary mismatch in terms of the morphemes obtained by morphological analysis. Furthermore, in almost 90% of cases of segmentation boundary mismatches, named entities to be recognized can be decomposed into several morphemes as their constituents. This means that the problem of recognizing named entities in those cases may be solved by incorporating techniques of base noun phrase chunking (boundary detection) (Ramshaw and Marcus, 1995; Muñoz et al., 1999).

In this paper, we focus on both the issues of named entity chunking and classification in Japanese named entity recognition, and evaluate named entity chunking techniques in the context of minimally supervised learning. First, as a supervised learning method, we employ the supervised decision list learning method of (Yarowsky, 1994), into which we incorporate several noun phrase chunking techniques (sections 3. and 4.). We chose the decision list learning method as the supervised learning technique because it is easy to implement and quite straightforward to extend a supervised learning version to a minimally supervised version (Yarowsky, 1994; Yarowsky, 1995; Collins and Singer, 1999) (section 5.). Then, we applied a minimally supervised learning algorithm to Japanese named entity recognition, where a list of frequent named entities are extracted from unlabeled data by a human and fed to the learning algorithm as *seeds*. The minimally supervised learning method improves the performance of the seed knowledge on named entity recognition by iteratively applying it to unseen data unlabeled with named entity tags and effectively utilizing redundancy in the unlabeled data.

The minimally supervised learning method employed in this paper is a variant of the ones based on the supervised decision list learning (Yarowsky, 1995; Collins and Singer, 1999), which, as discussed above, have been applied to classification tasks such as sense disambiguation and named entity classification, but not to the full named entity chunking and classification task. The results of the experimental evaluation shows that our minimally supervised learning method improves the performance of the seed knowledge on named entity recognition. We also investigate the correlation between performance of the minimally supervised learning and the sizes of the resources such as the seed as well as the unlabeled training data. (section 6.).

2. Japanese Named Entity Recognition

2.1. Task of the IREX Workshop

The task of named entity recognition of the IREX workshop is to recognize eight named entity types in Table 1 (IREX Committee, 1999). The organizer of the IREX workshop provided 1,174 newspaper articles which include 18,677 named entities as the training data. In the formal run (general domain) of the workshop, the participating systems were requested to recognize 1,510 named entities included in the held-out 71 newspaper articles.

Table 1: Statistics of NE Types of IREX

NE Type	frequency (%) Training	frequency (%) Test
ORGANIZATION	3676 (19.7)	361 (23.9)
PERSON	3840 (20.6)	338 (22.4)
LOCATION	5463 (29.2)	413 (27.4)
ARTIFACT	747 (4.0)	48 (3.2)
DATE	3567 (19.1)	260 (17.2)
TIME	502 (2.7)	54 (3.5)
MONEY	390 (2.1)	15 (1.0)
PERCENT	492 (2.6)	21 (1.4)
Total	18677	1510

Table 2: Statistics of Boundary Match/Mismatch of Morphemes (M) and Named Entities (NE)

Match/Mismatch		freq. of NE Tags (%)	
1 M to 1 NE		10480 (56.1)	
$n(\geq 2)$ Ms to 1 NE	$n = 2$	4557 (24.4)	7175 (38.4)
	$n = 3$	1658 (8.9)	
	$n \geq 4$	960 (5.1)	
other boundary mismatch		1022 (5.5)	
Total		18677	

2.2. Segmentation Boundaries of Morphemes and Named Entities

In the work presented here, we compare the segmentation boundaries of named entities in the IREX workshop's training corpus with those of morphemes which were obtained through morphological analysis by a Japanese morphological analyzer BREAKFAST (Sassano et al., 1997).[1] Detailed statistics of the comparison are provided in Table 2. Nearly half of the named entities have boundary mismatches against the morphemes and also almost 90% of the named entities with boundary mismatches can be decomposed into more than one morpheme. Figure 1 shows some examples of those cases.[2]

3. Chunking and Tagging Named Entities

In this section, we formalize the problem of named entity chunking in Japanese named entity recognition.

3.1. Task Definition

First, we will provide our definition of the task of Japanese named entity chunking. Suppose that a se-

[1]The set of part-of-speech tags of BREAKFAST consists of about 300 tags. BREAKFAST achieves 99.6% part-of-speech accuracy against newspaper articles.

[2]In most cases of the "other boundary mismatch" in Table 2, one or more named entities have to be recognized as a part of a correctly analyzed morpheme and those cases are not caused by errors of morphological analysis. One frequent example of this type is a Japanese verbal noun "hou-bei (*visiting United States*)" which consists of two characters "hou (*visiting*)" and "bei (*United States*)", where "bei (*United States*)" have to be recognized as `<LOCATION>`. We believe that boundary mismatch of this type can be easily solved by employing a supervised learning technique such as the decision list learning method.

Table 3: Encoding Schemes of Named Entity Chunking States

Named Entity Tag			<ORG>			<LOC>			<LOC>		
Morpheme Sequence	$\cdots$	M	M	M	M	M	M	M	M	$\cdots$
Inside/Outside Encoding		O	ORG_I	O	LOC_I	LOC_I	LOC_I	LOC_B	O	
Open/Close Encoding		O	ORG_U	O	LOC_S	LOC_C	LOC_E	LOC_U	O	

2 Morphemes to 1 Named Entity

<ORGANIZATION>
$\cdots$ Roshia (*Russian*) gun (*army*) $\cdots$

<PERSON>
$\cdots$ Murayama (last name) Tomiichi (first name) shushou (*prime minister*) $\cdots$

3 Morphemes to 1 Named Entity

<TIME>
$\cdots$ gozen (*AM*) ku (*nine*) ji (*o'clock*) $\cdots$

<ARTIFACT>
$\cdots$ hokubei (*North America*) jiyuu-boueki (*free trade*) kyoutei (*treaty*) $\cdots$

Figure 1: Examples of Boundary Mismatch of Morphemes and Named Entities

quence of morphemes are given as below:

$$
\begin{array}{ccc}
(\text{Left Context}) & (\text{Named Entity}) & (\text{Right Context}) \\
\cdots M_{-k}^{L} \cdots M_{-1}^{L} & M_{1}^{NE} \cdots M_{i}^{NE} \cdots M_{m}^{NE} & M_{1}^{R} \cdots M_{l}^{R} \cdots \\
 & \uparrow & \\
 & (\text{Current Position}) &
\end{array}
$$

Then, given that the current position is at the morpheme M_i^{NE}, the task of named entity chunking is to assign a chunking state (to be described in section 3.2.) as well as a named entity type to the morpheme M_i^{NE} at the current position, considering the patterns of surrounding morphemes. Note that in the supervised learning phase we can use the chunking information on which morphemes constitute a named entity, and which morphemes are in the left/right contexts of the named entity.

3.2. Encoding Schemes of Named Entity Chunking States

In this paper, we use the following two methods for encoding chunking states of named entities. Examples of those encoding schemes are shown in Table 3.

3.2.1. Inside/Outside Encoding

The Inside/Outside scheme of encoding chunking states of base noun phrases was studied by (Ramshaw and Marcus, 1995; Muñoz et al., 1999). Their scheme distinguishes the following three states: O – the word at the current position is outside any base noun phrase. I – the word at the current position is inside some base noun phrase. B – the word at the current position marks the beginning of a base noun phrase that immediately follows another base noun phrase. We extend this scheme to named entity chunking by further distinguishing each of the states I and B into eight named entity types.[3] Thus, this scheme distinguishes $2 \times 8 + 1 = 17$ states.

3.2.2. Open/Close Encoding

The Open/Close scheme of encoding chunking states of named entities was employed in (Sekine et al., 1998; Borthwick et al., 1998). A similar scheme is also studied in (Muñoz et al., 1999) in the context of base noun phrases chunking. This scheme distinguishes the following four states for each named entity type: S – the morpheme at the current position marks the beginning of a named entity consisting of more than one morpheme. C – the morpheme at the current position marks the middle of a named entity consisting of more than one morpheme. E – the morpheme at the current position marks the ending of a named entity consisting of more than one morpheme. U – the morpheme at the current position is a named entity consisting of only one morpheme. The scheme also considers one additional state for the position outside any named entity: O – the morpheme at the current position is outside any named entity. Thus, in this setting, our scheme distinguishes $4 \times 8 + 1 = 33$ states.

3.3. Preceding/Subsequent Morphemes as Contextual Clues

In this paper, we employ the model used in (Sekine et al., 1998; Borthwick et al., 1998) as that of considering preceding/subsequent morphemes as contextual clues to named entity chunking/tagging.[4] Here we provide a basic outline of the model, and the details of how to incorporate it into the decision list learning framework will be described in section 4.2.2..

Suppose that the current position is at the morpheme M_0, as illustrated below. Then, when assigning a chunking state as well as a named entity type to the morpheme M_0, the model considers the preceding single morpheme M_{-1} as well as the subsequent one

[3] We allow the state x_B for a named entity type x only when the morpheme at the current position marks the beginning of a named entity of the type x that immediately follows a named entity of the same type x.

[4] In (Sassano and Utsuro, 2000), we proposed and evaluated a novel model that incorporates richer contextual information as well as patterns of constituent morphemes within a named entity.

morpheme M_1 as the contextual clue.

$$\begin{array}{ccc} \left(\begin{array}{c}\text{Left}\\ \text{Context}\end{array}\right) & \left(\begin{array}{c}\text{Current}\\ \text{Position}\end{array}\right) & \left(\begin{array}{c}\text{Right}\\ \text{Context}\end{array}\right) \\ \cdots\ M_{-1} & M_0 & M_1\ \cdots \end{array} \quad (1)$$

4. Supervised Learning for Japanese Named Entity Recognition

This section describes how to apply the decision list learning method to chunking/tagging named entities.

4.1. Decision List Learning

A decision list (Rivest, 1987; Yarowsky, 1994) is a sorted list of the decision rules each of which decides the value of a *decision* D given some *evidence* E. Each decision rule in a decision list is sorted in descending order with respect to some preference value, and rules with higher preference values are applied first when applying the decision list to some new test data.

First, the random variable D representing a decision varies over several possible values, and the random variable E representing some evidence varies over '1' and '0' (where '1' denotes the presence of the corresponding piece of evidence, '0' its absence). Then, given some training data in which the correct value of the decision D is annotated to each instance, the conditional probabilities $P(D=x \mid E=1)$ of observing the decision $D=x$ under the condition of the presence of the evidence E $(E=1)$ are calculated and the decision list is constructed by the following procedure.

1. For each piece of evidence, calculate the *log of likelihood ratio* of the largest conditional probability of the decision $D=x_1$ (given the presence of that piece of evidence) to the second largest conditional probability of the decision $D=x_2$:

   $$\log_2 \frac{P(D=x_1 \mid E=1)}{P(D=x_2 \mid E=1)}$$

   Then, a decision list is constructed with pieces of evidence sorted in descending order with respect to their log of likelihood ratios, where the decision of the rule at each line is $D=x_1$ with the largest conditional probability.[5]

2. The final line of a decision list is defined as 'a default', where the log of likelihood ratio is calculated as the ratio of the largest marginal probability of the decision $D=x_1$ to the second largest marginal probability of the decision $D=x_2$:

   $$\log_2 \frac{P(D=x_1)}{P(D=x_2)}$$

   The 'default' decision of this final line is $D=x_1$ with the largest marginal probability.

[5](Yarowsky, 1994) discusses several techniques for avoiding problems which arise when an observed count is 0. From among those techniques, we employ the simplest one, i.e., adding a small constant α $(0.1 \le \alpha \le 0.25)$ to the numerator and denominator. With this modification, more frequent evidence is preferred when several evidence candidates exist with the same unsmoothed conditional probability $P(D=x \mid E=1)$. Yarowsky's training algorithm also differs somewhat in his use of the ratio $\frac{P(D=x_i|E=j)}{P(\neg D=x_i|E=j)}$, which is equivalent in the case of binary classifications, and also by the interpolation between the global probabilities (used here) and the residual probabilities further conditional on higher-ranked patterns failing to match in the list.

4.2. Decision List Learning for Chunking/Tagging Named Entities

4.2.1. Decision

For each of the two schemes of encoding chunking states of named entities described in section 3.2., as the possible values of the decision D, we consider exactly the same categories of chunking states as those described in section 3.2..

4.2.2. Evidence

The evidence E of the decision list learning is a combination of the features of preceding/subsequent morphemes as well as the morpheme in the current position. The evidence E represents a tuple (F_{-1}, F_0, F_1), where F_{-1} and F_1 denote the features of immediately preceding/subsequent morphemes M_{-1} and M_1, respectively, F_0 the feature of the morpheme M_0 at the current position (see Formula (1) in section 3.3.). The definition of the possible values of those features F_{-1}, F_0, and F_1 are given below, where M_i denotes the morpheme itself (i.e., including its lexical form as well as part-of-speech), C_i the character type (i.e., Japanese (hiragana or katakana), Chinese (kanji), numbers, English alphabets, symbols, and all possible combinations of these) of M_i, T_i the part-of-speech of M_i:

$$\begin{array}{lcl} F_{-1} & ::= & M_{-1} \mid (C_{-1}, T_{-1}) \mid T_{-1} \mid \text{null} \\ F_1 & ::= & M_1 \mid (C_1, T_1) \mid T_1 \mid \text{null} \\ F_0 & ::= & M_0 \mid (C_0, T_0) \mid T_0 \end{array}$$

As the evidence E, we consider each possible combination of the values of those three features.

4.3. Procedures for Training and Testing

Next we will briefly describe the entire process of supervised learning the decision list for chunking/tagging named entities as well as applying it to chunking/tagging unseen named entities.

4.3.1. Training

In the training phase, at each morpheme position, as described in the section 4.2., each *allowable* combination of features is considered as the evidence E. Then, the frequency of each decision D and evidence E is counted and the decision list is learned as described in section 4.1..

4.3.2. Testing

When applying the decision list to chunking/tagging unseen named entities, first, in each morpheme position, the combination of features is considered as in the case of the non-entity position in the training phase. Then, the decision list is consulted and all the decisions of the rules with a log of likelihood ratio above a certain threshold are recorded. Finally, as in the case of previous research (Sekine et al.,

	<PERSON>				<PERSON>			
···	Murayama	Tomiichi	shushou	ha	Kooru	shushou	to	···
	(last name)	(first name)	(*prime minister*)	(Topic)	(last name)	(*prime minister*)	(*with*)	

Figure 2: The Basic Idea of Utilizing Redundancy in the Minimally Supervised Learning

1998; Borthwick et al., 1998), the most appropriate sequence of the decisions that are consistent throughout the whole sequence is searched for. By consistency of the decisions, we mean requirements such as that the decision representing the beginning of some named entity type has to be followed by that representing the middle of the same entity type (in the case of the Open/Close encoding). Also, in our case, the appropriateness of the sequence of the decisions is measured by the sum of the log of likelihood ratios of the corresponding decision rules.

5. Minimally Supervised Learning

5.1. The Basic Idea of Utilizing Redundancy

The following example describes how to utilize redundancy in the unlabeled data in the minimally supervised learning.

Suppose that the sequence of morphemes in Figure 2 is included in an unlabeled data set, where the names of the prime ministers of Germany and Japan need to be recognized as <PERSON>s. Both of the two names are followed by a noun representing the social position, i.e., in this case, "shushou" (*prime minister*). In the Japanese language, the name of a person at a certain social position is followed by a noun representing the position, as in the case of "President Clinton" in English. Also suppose that, as a seed, the learner is given the name of Japanese prime minister "Murayama Tomiichi" as a <PERSON> named entity. Then, the algorithm can learn the following decision rule from this seed and unlabeled data in Figure 2.

> If X is followed by "shushou" (*prime minister*),
> then X is <PERSON>.

By applying this decision rule to "Kooru", the algorithm can easily annotate "Kooru" as a <PERSON> named entity. In this case, the redundancy of the seed knowledge and a reliable contextual clue at the named entity position "Murayama Tomiichi" is the key to bootstrapping in minimally supervised learning.

5.2. Learning Algorithm

In following section we provide our minimally supervised learning algorithm, which uses a list of named entities as seeds and a large amount of unlabeled data, but not use any labeled data.

1. Initialization

Seed Selection

First, a list of frequent named entities are extracted from the result of the morphological analysis of unlabeled data. This is done by a human who consults the definition of the named entity tags if necessary.

Converting Seeds into the Initial Decision List

Then, the list of frequent named entities is converted into a decision list which is to be used as an initial decision list in the minimally supervised learning.

2. Minimally Supervised Learning

The following application and training iteration is iterated and the changes in precision, recall, and f-measure against held-out data labeled with named entity tags are observed.

Applying Decision List to Unlabeled Data

Usually, in the case of supervised learning, the best performance against test data is obtained when only those rules with a log of likelihood ratio above a certain threshold are considered. However, in our experimental results, the best performance in the minimally supervised learning was obtained when all the rules in the decision list were considered at each step of the training iteration. The results of minimally supervised learning reported in this paper, then, are those obtained without a threshold in the decision lists.

Learning Decision List

A new decision list is learned from the data with named entity tags annotated by the decision list learned in the previous step. The data may contain errors of named entity recognition.

6. Experimental Evaluation

We next experimentally evaluate the performance of the supervised and minimally supervised learning for Japanese named entity recognition on the IREX workshop's training and test data. In this evaluation, we exclude the named entities with "other boundary mismatch" in Table 2.

6.1. Comparison of Named Entity Chunking Methods

First, we compare the performance of the two encoding schemes of named entity chunking states (the Inside/Outside and the Open/Close encoding schemes) in full supervised and minimally supervised learning.

6.1.1. Supervised Learning

For each of those encoding schemes, a decision list is learned by full supervised learning from the IREX workshop's training data and evaluated against the IREX workshop's test data. We searched for an optimal threshold of the log of likelihood ratio in the decision list. The performance of each encoding scheme measured by f-measure ($\beta = 1$)/precision/recall is given in Table 4. We classified the system output according to the number of constituent morphemes of each named entity and evaluate the performance for each subset of the system output. For each subset and

Table 4: Evaluation Results of Supervised Learning

		n Morphemes to 1 Named Entity				
		$n \geq 1$	$n = 1$	$n = 2$	$n = 3$	$n \geq 4$
Inside/Outside	F-measure ($\beta = 1$)	**72.9**	75.4	**79.7**	**51.4**	29.2
	(Precision)	(72.7)	(68.0)	(78.2)	(74.7)	(87.5)
	(Recall)	(73.1)	(84.7)	(81.1)	(39.2)	(17.5)
Open/Close	F-measure ($\beta = 1$)	72.7	**76.1**	79.5	43.7	**29.6**
	(Precision)	(76.7)	(71.8)	(82.2)	(79.6)	(95.5)
	(Recall)	(69.0)	(81.0)	(77.0)	(30.1)	(17.5)

the whole set, we show the higher performance in f-measure with bold-faced font.

The Inside/Outside encoding scheme achieves slightly better performance in total f-measure and significantly better in f-measure of the "$n = 3$" subset, while the Open/Close encoding scheme achieves slightly better performance in f-measure of "$n = 1$" and "$n \geq 4$" subsets. The Inside/Outside encoding scheme achieves better performance in recall, while the Open/Close encoding scheme achieves better performance in precision. This difference in precision/recall is consistent with the degree of generalization of the two encoding schemes.

6.1.2. Minimally Supervised Learning

Next, we simulate the minimally supervised learning algorithm of section 5.2. by taking a small portion of highly ranked decision rules learned by supervised learning as seeds. First, from the first half of the IREX workshop's training data, a decision list is learned by supervised learning. Then, top 100 and 500 rules are extracted from the decision list as seeds of minimally supervised learning, and the minimally supervised learning algorithm is run with the other half of the IREX workshop's training data (containing about 9,300 named entities) as the unlabeled training data by ignoring the annotated named entity tags. The whole decision list learned by supervised learning has approximately 60,000 rules for both the Inside/Outside and the Open/Close encoding schemes and the top 500 rules are those without ambiguities of decisions (i.e., the unsmoothed conditional probability $P(D=x \mid E=1)$ equals to 1). Figure 3 shows the results of comparing the performance changes (in f-measure ($\beta = 1$)) of the two encoding schemes in the minimally supervised learning.

This results clearly shows that the initial performance of the Inside/Outside encoding scheme is better than that of the Open/Close encoding scheme with the same number of decision rules. Also with respect to the maximum f-measure throughout the training iteration, the Inside/Outside encoding scheme achieves better performance than the Open/Close encoding scheme. One of the obvious advantages of the Inside/Outside encoding scheme over the Open/Close encoding scheme is that the former can generalize decision rules among named entities consisting of different number of morphemes, while the latter can not generalize decision rules between those consisting of one morpheme and those consisting of more than one morpheme. The generalization ability of the Inside/Outside encoding scheme is particularly effective in minimally supervised learning where the data sparseness problem easily occurs. From this result, we conclude that the Inside/Outside encoding scheme is more suitable for minimally supervised learning and the rest of our evaluation is carried out with the Inside/Outside encoding scheme.

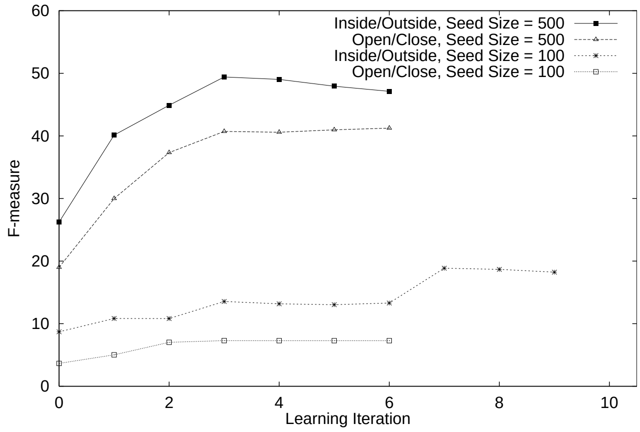

Figure 3: Comparison of Inside/Outside and Open/Close Encoding Schemes in Minimally Supervised Learning

6.2. Minimally Supervised Learning with Inside/Outside Encoding Scheme

Japanese named entity chunking and tagging with the Inside/Outside encoding scheme was evaluated in the minimally supervised learning of section 5.2.. In this evaluation, a list of frequent named entities consisting of one morpheme is extracted as seeds by a human from the first half of the IREX workshop's training data without referring to the named entity tags annotated to the original training data. Then, the minimally supervised learning algorithm is run with the other half of the IREX workshop's training data as the unlabeled training data. The performance of the learned decision list is evaluated against the IREX workshop's test data.

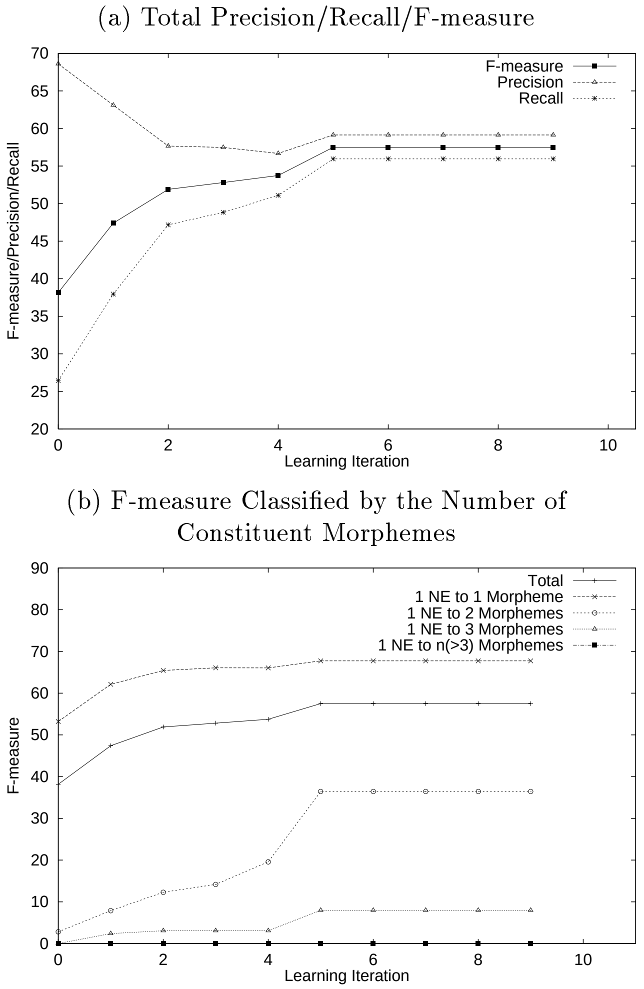

Figure 4: Performance Changes across Minimally Supervised Training Iteration

6.2.1. Changes of Precision/Recall/F-measure in Training Iteration

First, for the case where the most frequent 200 seed named entities are used, Figure 4 shows the changes of precision/recall/f-measure in the iteration of minimally supervised learning (Figure 4 (a)) as well as the changes of f-measure classified according to the number of constituent morphemes of each named entity (Figure 4 (b)). These performance are measured against the unlabeled training data. The curve of Figure 4 (a) represents a typical performance changes in the minimally supervised learning, in that the recall and the f-measure dramatically increases, while the precision decreases. One of the remarkable results of Figure 4 (b) which has to be pointed out is that the performance for the "1 named entity to 2 morphemes" and "1 named entity to 3 morphemes" increases even though each named entity in the seed list consists of exactly one morpheme. This result clearly supports the claim that the Inside/Outside encoding scheme can generalize decision rules among named entities consisting of different number of morphemes.

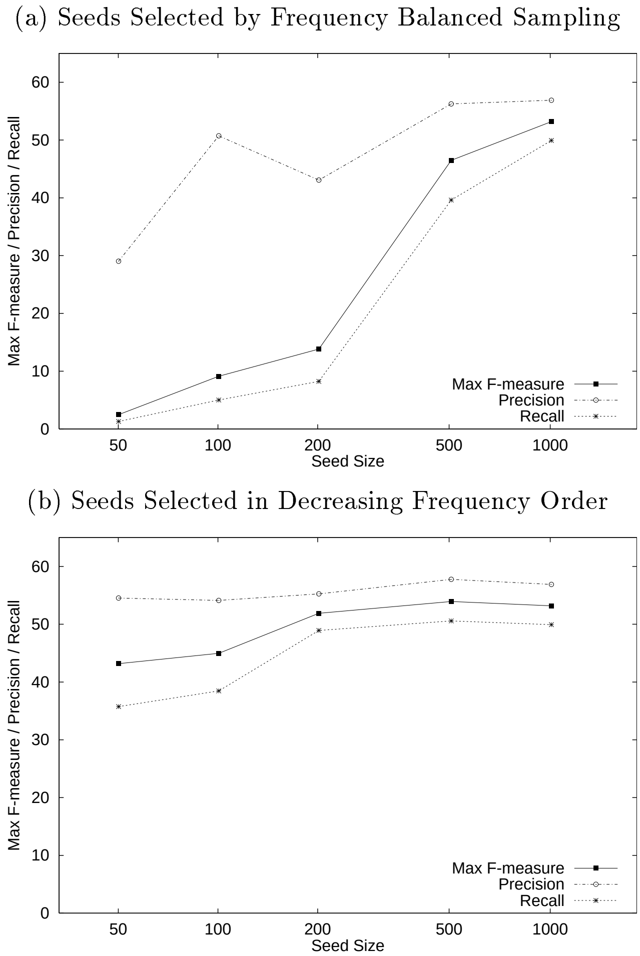

Figure 5: Performance at Different Seed Sizes

6.2.2. Evaluation at Different Resource Sizes

Next, we evaluate minimally supervised learning at different sizes of seed named entities as well as of unlabeled training data.

First, we change the number of seed named entities as 50, 100, 200, 500, and 1000, and run the minimally supervised learning algorithm and plot the maximum f-measure value throughout the training iteration for each seed size. Figure 5 (a) shows the result when the seed named entities are selected so that their frequency counts are balanced among different seed sizes, while Figure 5 (b) shows that when the most frequent named entities are selected as seeds. In Figure 5 (a), the maximum f-measure value increases roughly loglinearly with the number the seed named entities, which is consistent with the results reported in (Cucerzan and Yarowsky, 1999). In Figure 5 (b), on the other hand, the maximum f-measure value stops to increase when the number of the seed named entities is 500, simply because the larger seed lists are constructed by adding less frequent named entities to the smaller seed lists.

Second, we change the size of the unlabeled training data and run the minimally supervised learning algorithm with 200 most frequent named entities as seeds and plot the maximum f-measure value throughout the training iteration for each unlabeled training data size. Figure 6 (a) shows the result when the maximum f-

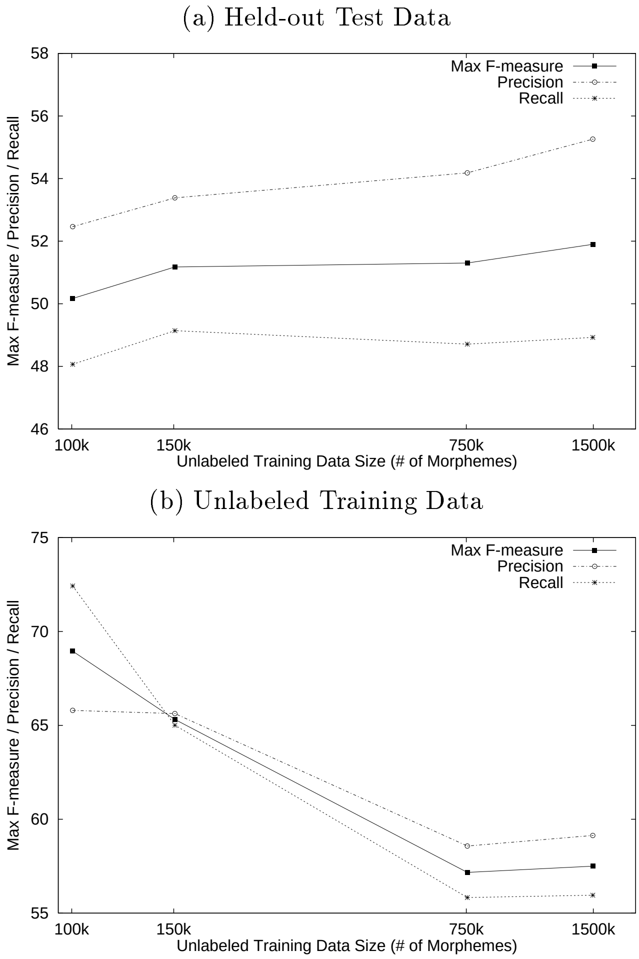

Figure 6: Performance at Different Unlabeled Data Sizes

measure value is measured against the held-out IREX workshop's test data, while Figure 6 (b) shows that when the the maximum f-measure value is measured against the training unlabeled data. In Figure 6 (a), the maximum f-measure value slightly increases due to increases in precision, as the size of the unlabeled training data increases, which is consistent with the results reported in (Cucerzan and Yarowsky, 1999). This paper claims that if more unlabeled data are available, more accurate rules for named entity recognition will be learned by minimally supervised learning. In Figure 6 (b), on the other hand, the maximum f-measure value against the unlabeled training data decreases with the size of the unlabeled training data. This is because as the unlabeled training data set becomes larger, the unlabeled training data tends to include more infrequent named entities that are more difficult to recognize than frequent named entities.

7. Conclusion

This paper evaluated named entity chunking and classification techniques in Japanese named entity recognition in the context of minimally supervised learning. The experimental evaluation showed that our minimally supervised learning method improves the performance of the seed knowledge on named entity recognition. We also investigated the correlation between performance of the minimally supervised learning and the sizes of the resources such as the seed as well as the unlabeled training data. We are now working on incorporating the model of (Sassano and Utsuro, 2000) into minimally supervised learning environments, where richer contextual information as well as patterns of constituent morphemes within a named entity are considered. The results of this experimental evaluation will be reported in the near future.

8. Acknowledgments

This research was carried out while the authors were visiting scholars at Department of Computer Science, Johns Hopkins University. The authors would like to thank Prof. David Yarowsky of Johns Hopkins University for valuable supports to this research.

9. References

Blum, A. and T. Mitchell, 1998. Combining labeled and unlabeled data with co-training. In *Proc. 11th COLT*.

Borthwick, A., J. Sterling, E. Agichtein, and R. Grishman, 1998. Exploiting diverse knowledge sources via maximum entropy in named entity recognition. In *Proc. 6th WVLC*.

Collins, M. and Y. Singer, 1999. Unsupervised models of named entity classification. In *Proc. 1999 Joint SIGDAT Conference on EMNLP and VLC*.

Cucerzan, S. and D. Yarowsky, 1999. Language independent named entity recognition combining morphological and contextual evidence. In *Proc. 1999 Joint SIGDAT Conference on EMNLP and VLC*.

IREX Committee (ed.), 1999. *Proc. the IREX Workshop*. (in Japanese).

Maiorano, S., 1996. The multilingual entity task (MET): Japanese results. In *Proc. TIPSTER PROGRAM PHASE II*.

MUC, 1998. *Proc. 7th Message Understanding Conference (MUC-7)*. MUC.

Muñoz, M., V. Punyakanok, D. Roth, and D. Zimak, 1999. A learning approach to shallow parsing. In *Proc. 1999 Joint SIGDAT Conference on EMNLP and VLC*.

Ramshaw, L. and M. Marcus, 1995. Text chunking using transformation-based learning. In *Proc. 3rd WVLC*.

Rivest, R. L., 1987. Learning decision lists. *Machine Learning*, 2:229–246.

Sassano, M., Y. Saito, and K. Matsui, 1997. Japanese morphological analyzer for NLP applications. In *Proc. 3rd Annual Meeting of the Association for Natural Language Processing*. (in Japanese).

Sassano, M. and T. Utsuro, 2000. Named entity chunking techniques in supervised learning for Japanese named entity recognition. In *Proc. 18th COLING*. (to appear).

Sekine, S., R. Grishman, and H. Shinnou, 1998. A decision tree method for finding and classifying names in Japanese texts. In *Proc. 6th WVLC*.

Yarowsky, D., 1994. Decision lists for lexical ambiguity resolution: Application to accent restoration in Spanish and French. In *Proc. 32nd ACL*.

Yarowsky, D., 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proc. 33rd ACL*.