Named Entity Extraction Based on A Maximum Entropy Model and Transformation Rules

Kiyotaka Uchimoto, Qing Ma, Masaki Murata,
Hiromi Ozaku and Hitoshi Isahara
Communications Research Laboratory
Ministry of Posts and Telecommunications
588-2, Iwaoka, Iwaoka-cho, Nishi-ku
Kobe, Hyogo, 651-2492, Japan
[uchimoto,qma,murata,romi,isahara]@crl.go.jp

Abstract

This paper describes named entity (NE) extraction based on a maximum entropy (M.E.) model and transformation rules. There are two types of named entities when focusing on the relationship between morphemes and NEs as defined in the NE task of the IREX competition held in 1999. Each NE consists of one or more morphemes, or includes a substring of a morpheme. We extract the former type of NE by using the M.E. model. We then extract the latter type of NE by applying transformation rules to the text.

1 Introduction

Named entity (NE) extraction is one basic technique used in information extraction. It can also help to improve the accuracy of morphological and syntactic analysis. The competition held at the MUC (Message Understanding Conference) (SAIC, 1998) since 1980 in the U.S. has helped to improve the technique. In Japan, the "IREX (Information Retrieval and Extraction Exercise)" project began sponsoring a similar competition in 1998. NE extraction is one of two tasks in the competition. The targets for extraction in this task are the names of organizations such as "郵政省 (the Ministry of Posts and Telecommunications)," people's names such as "川端康成 (Yasunari Kawabata)," names of locations such as "神戸 (Kobe)," names of artifacts such as "カローラ (Toyota's Corolla car)," and expressions which represent dates, times, sums of money, and percentages, such as "9月 28日 (September 28th)," "午後 3時 (3 p. m.)," "100万円 (one million yen)," and "10%." There are many and various NEs, and new ones are produced all of the time, so it is impossible to add all of them to a dictionary. There are also ambiguities in usage so that a given expression may be used as a location name in one context and as a person's name in another context. Therefore, it is not easy to identify NEs, and to identify the type of each NE, in a given sentence.

There are two main approaches to extracting NEs, one based on hand-crafted rules and the other based on a machine-learning. The former approach is costly because definitions differ across applications, and the rules have to be changed according to the application. The machine-learning approach requires a training corpus, but a high accuracy can be achieved without requiring a large amount of data if we use a learning model which includes ways of overcoming the data sparseness problem. Therefore, we have taken the latter approach. Many methods based on maximum entropy (M.E.) models have been very accurate (Ratnaparkhi, 1996; Ratnaparkhi, 1997; Borthwick et al., 1998a; Uchimoto et al., 1999), and the M.E. model can be adapted to deal with the data sparseness problem effectively. We have thus used the M.E. model to extract NEs. After identifying NEs in a given text by applying our model, we apply transformation rules which have been acquired by an error-driven learning method to the text.

2 Named Entity Extraction Algorithm

We have used the definition of NEs which is used in the IREX-NE task (IREX Executive Committee, 1999). Eight types of NE, "ORGANIZATION", "PERSON", "LOCATION", "ARTIFACT", "DATE", "TIME", "MONEY", and "PERCENT" are defined. This section describes the method of identifying NEs in a given text and assigning one of eight SGML tags which represent the type of NE to each one.

Each NE consists of one or more morphemes, or includes a substring of a morpheme. We define 40 NE labels, as explained below, and extract an NE which consists of one or more morphemes by estimating the appropriate NE labels according to an M.E. model. The trained M.E. model detects the relationship between features and the NE labels assigned to morphemes. The features are clues used for estimating the labels. After estimating the NE labels according to the M.E. model, we extract an NE, which includes a substring of a morpheme, by using transformation rules that will be explained later.

In detail, the following steps are used to extract NEs.

1. Morphological analysis of a given text.

   We used JUMAN (Kurohashi and Nagao, 1998) for morphological analysis. For example, the phrase "在米女性を中心に「人権を考える会」ができ，…," is divided into the morphemes shown in the first line of Table 1, and morphological information as shown in the second and third lines of Table 1 is assigned to each morpheme.

2. Assigning NE labels to each morpheme.

   We defined the following 40 NE labels, and the rules for connectivity between the labels, which we call *connectivity rules*, as shown in Table 2.

   (a) We added an "OPTIONAL" tag to the eight NE tags, then divided each into four types of sub labels which represented the beginning, middle, and end of an NE, or an NE which consisted of a single morpheme. We thus defined 9×4=36 NE labels. For example, the "PERSON" tag was divided into "PERSON:BEGIN", "PERSON:MIDDLE", "PERSON:END", and "PERSON:SINGLE". We divided the NE tags into four types because several morphemes can constitute a single NE.

   The "OPTIONAL" tag was defined because, in some cases, even a human judge would find it difficult to decide which tag should be assigned to a string, or whether a string is or is not an NE. For example, should "東京高裁 (The Tokyo high court)" be tagged as "LOCATION" or "ORGANIZATION"? Should "日経 (Nikkei, the abbreviation of the name of a newspaper publishing company in Japan)" in the expression "日経平均株価 (Nikkei stock average)" be tagged as "ORGANIZATION" or not? In these cases, "東京高裁" and "日経" are tagged as "OPTIONAL", and are not extracted as NEs. The definition of the "OPTIONAL" tag is also the same as that which is used in the IREX-NE task. We defined the tag to learn its characteristics and to avoid assigning NE tags to strings in such difficult cases as those explained above.

   (b) We defined three more NE labels, "PRE", "POST", and "MID", to distinguish morphemes to the left and right, and between NEs, respectively, from the other morphemes. For example, "大阪 (Osaka)" and "神戸 (Kobe)" in the phrase "昨日大阪と神戸で… (Yesterday in Osaka and Kobe...)" are the names of locations, so the whole phrase is tagged in the following way.

   "昨日 (PRE) ／大阪 (LOCATION:SINGLE) ／と (MID) ／神戸 (LOCATION:SINGLE)

Table 1: Example of the assignment of NE labels by the M.E. model.

Entry		在米 (*zaibei*, staying in the U.S.)	女性 (*josei*, women)	を (*wo*)	中心 (*chuushin*, center)	に (*ni*)	「	人権 (*jinken*, human rights)
POS (major)		noun	noun	post positional particle	noun	post positional particle	special	noun
POS (minor)		common noun	common noun	case marker	common noun	case marker	beginning of brackets	common noun
Candidate of label	1	OTHER	OTHER	OTHER	OTHER	OTHER	PRE	ORG:BEGIN
	2	OTHER	OTHER	OTHER	OTHER	OTHER	PRE	ART:SINGLE
	...	...	...	...	...	...	...	...

を (*wo*)	考える (*kangaeru*, to think)	会 (*kai*, meeting)	」	が (*ga*)	でき (*deki*, organized)	、	Score
post positional particle case marker	verb	noun common noun	special end of brackets	post positional particle case marker	verb	special comma	
ORG:MIDDLE	ORG:MIDDLE	ORG:END	POST	OTHER	OTHER	OTHER	0.8
POST	OTHER	OTHER	OTHER	OTHER	OTHER	OTHER	0.7
	...	...	...	...	.....	...	...

(In this table, "ORG" and "ART" indicate "ORGANIZATION" and "ARTIFACT," respectively.)

／で (POST) ···"
(The labels in parentheses indicate the candidate NE labels assigned to the strings to their left.)

These three labels are used in particular to distinguish morphemes to the left or right of an NE from those to which the "OTHER" tag, explained immediately below, is assigned, because morphemes such as suffixes can be clues which assist in finding NEs.

(c) We defined the label "OTHER" to designate morphemes to which none of the labels defined above can be assigned.

Given tokenization of a test corpus, the extraction of named entities can be reduced to the problem of assigning one NE label to each morpheme in each sentence. The 40 NE labels form the space of "futures" in the M.E. formulation of our problem of extracting named entities. The M.E. model, as well as other similar models, allows the computation of $P(f|h)$ for any f in the space of possible futures, F, and for every h in the space of possible histories, H. A "history" in maximum entropy is all of the conditioning data that enable us to make a decision in the space of futures. In the problem of extracting named entities, we could reformulate this in terms of finding the probability of f associated with the relationship at index t in the test corpus as:

$$P(f|h_t) = P(f|\text{Information derivable from the test corpus related to relationship } t)$$

The computation of $P(f|h)$ in M.E. is dependent on a set of "features" which should be helpful in making a prediction about the future. Like most current M.E. models in computational linguistics, our model is restricted to the features which are binary functions of the history and the future. For instance, one of our features is

$$g(h,f) = \begin{cases} 1 : \text{if has}(h,x) = \text{true}, \\ \quad x = \text{"POS(major)(0) : verb"} \\ \quad \& \ f = 1 \\ 0 : \text{otherwise.} \end{cases} \quad (1)$$

Here "has(h,x)" is a binary function which returns true if the history h has the attribute x. $g(h,f)$ in Eq. (1) can return 1 when the major part-of-speech of the target morpheme is verb. We use the following information as features on the target morpheme: a lexical item and the parts-of-speech it belongs to, and the same information on the four closest morphemes, the two on the left and the two on the right of the target morpheme. In our experiments, we used 12,368 lexical items that appeared five times or more in the training corpus. The part-of-speech

Table 2: Connectivity rules

NE label	labels connectable to the left	labels connectable to the right
x	#(BOS), x, y, x:END, y:END, PRE, MID	\$(EOS), x, y, x:BEGIN, y:BEGIN, POST, MID
x:BEGIN	#(BOS), x, y, x:END, y:END, PRE, MID	x:MIDDLE, x:END
x:MIDDLE	x:BEGIN, x:MIDDLE	x:MIDDLE, x:END
x:END	x:BEGIN, x:MIDDLE	\$(EOS), x, y, x:BEGIN, y:BEGIN, POST, MID
MID	x, x:END	x, x:BEGIN
PRE	#(BOS), POST, OTHER	x, x:BEGIN
POST	x, x:END	\$(EOS), PRE, OTHER
OTHER	#(BOS), POST, OTHER	\$(EOS), PRE, OTHER
\$(EOS)	x, x:END, POST, OTHER	
#(BOS)		x, x:BEGIN, PRE, OTHER

(BOS and EOS indicate "beginning of sentence" and "end of sentence," respectively. x and y correspond to "OPTIONAL" or the other eight tags defined for the IREX-NE task.)

categories are the same as those used by JUMAN. We used 27,370 features that were found three times or more in the training corpus.

Given a set of features and some training data, the maximum entropy estimation process produces a model in which every feature g_i has associated with it a parameter α_i. This allows us to compute the conditional probability as follows (Berger et al., 1996):

$$P(f|h) = \frac{\prod_i \alpha_i^{g_i(h,f)}}{Z_\lambda(h)} \quad (2)$$

$$Z_\lambda(h) = \sum_f \prod_i \alpha_i^{g_i(h,f)}. \quad (3)$$

The maximum entropy estimation technique guarantees that for every feature g_i, the expected value of g_i according to the M.E. model will equal the empirical expectation of g_i in the training corpus. In other words:

$$\sum_{h,f} \tilde{P}(h,f) \cdot g_i(h,f) = \sum_h \tilde{P}(h) \cdot \sum_f P_{M.E.}(f|h) \cdot g_i(h,f). \quad (4)$$

Here $\tilde{P}$ is an empirical probability and $P_{M.E.}$ is the probability assigned by the M.E. model.

Let us assume that a given sentence consists of n morphemes. One of the NE labels as defined above is assigned to each morpheme $m_i (1 \leq i \leq n)$ by using the morphological information acquired in the first step of the process we are describing. The NE label assigned to the i-th morpheme m_i is selected according to probabilities estimated by a trained M.E. model. We call the probability of a particular NE label being assigned to a morpheme, the *labeling probability*. The labeling probability is represented by Eq. (2). We assume that a labeling probability for a whole sentence can be determined as the product of all labeling probabilities in the sentence. We employ the Viterbi algorithm to find the optimal set of assigned NE labels in a sentence with the condition that the placement of labels satisfies connectivity rules shown in Table 2.

3. Post-processing by using transformation rules.

   The boundaries between morphemes which result from analysis by JUMAN do not always correspond to the boundaries between named entities as defined in the IREX-NE task. So after the NEs have been labeled in the second step, we use transformation rules which are automatically determined to extract NEs with boundaries that are not same as those between morphemes. Transformation rules are acquired by an error-driven learning method which is similar to that used by Brill (Brill, 1995) for POS tagging. The difference between our method of rule acquisition and Brill's is that Brill

uses templates to acquire rules and we do not. In our method, rules are automatically acquired by investigating the difference between two sets of data, NE labels in a tagged corpus and those extracted during the previous step from the same corpus without tags. We extract all of the differences in places where the two data sets are broken up into a different number of units or morphemes even though the strings are the same, and use them as transformation rules. For example, the rule shown in Figure 1 was acquired. The antecedent and consequent interpretations are from the result of the previous step and a tagged corpus, respectively. If several different rules have the same antecedent part, only the rule with the highest frequency is chosen. If several rules share the highest frequency, all of the rules are removed from transformation rules. Furthermore, if there are rules which decrease the accuracy of the method on the training corpus, they are removed.

4. Transforming NE labels to NE tags.

   After transforming NE labels to NE tags, the "OPTIONAL" tag is removed because it is not a target of the task.

For example, "在米 (OTHER)" on the first candidate in Table 1 is transformed to "在 (PRE) 米 (LOCATION:SINGLE)" in the third step. We get the following output after transforming NE labels to NE tags.

"在 <LOCATION> 米 < /LOCATION> 女性を中心に
「<ORGANIZATION> 人権を考える会 < /ORGANIZATION>」
ができ，"

3 Experiments and Discussion

3.1 Data Used in Our Experiments

For training, we used the CRL (Communications Research Laboratory) NE data, IREX-NE dry-run training data, IREX-NE dry-run data, and IREX-NE formal-run domain-specific data. The total number of sentences is about 12,000, and the total number of morphemes is about 303,200. All data consist of articles from the Mainichi newspaper, and are tagged with the nine NE tags in SGML format. We used these data after morphologically analyzing the text and transforming the NE tags into our new NE labels. For testing, we used the IREX-NE formal-run data, which consists of articles of two kinds, 71 (about 400 sentences) in a general domain and 20 (about 100 sentences) in a specific domain, the topic being an arrest. They were selected from the Mainichi newspaper articles which appeared from April 14th to May 13th in 1999, and were also tagged with NE tags [1]. The definition of tags is that of the IREX-NE task.

3.2 Experimental Results

The results are shown in Table 3. The first and second columns show the results for the specific domain (ARREST) and the general domain (GENERAL), respectively. We did not tune our model to either domain. Comparing the results with those of experiments carried out without transformation rules, we found the accuracy for the formal experiments had an F-measure, for both domains, one or two points better than those without transformation rules, as shown in Table 3.

In the IREX-NE formal-run, any tags assigned by a system within the region tagged "OPTIONAL" in the formal-run data are ignored in the evaluation. When a region tagged by a system and the region tagged "OPTIONAL" overlap, it is counted as an error. Our evaluation followed this standard.

3.3 Transformation Rules and Accuracy

We applied the transformation rules to NEs which included a substring of a morpheme. The rules were applied to 18 such NEs in the specific domain, and 79 in the general domain. Each of the figures represents about 5% of the NEs in the formal-run data, for each domain. 362 rules were automatically acquired from the training corpus. Nine rules were applied eleven times in processing of the specific domain data, with one error. The re-

[1] All data are available on the IREX web site (IREX Executive Committee, 1999).

	Antecedent part	⇒	Consequent part	
Entry	在日 (*zainichi*, staying in Japan)		在 (*zai*, staying)	日 (*nichi*, Japan)
POS (major)	noun		noun	noun
POS (minor)	SAHEN noun		common noun	common noun
Label	OTHER		PRE	LOCATION:SINGLE

Figure 1: Example of transformation rules.

Table 3: Results for extraction of named entities.

	With transformation rules				Without transformation rules			
	ARREST		GENERAL		ARREST		GENERAL	
Named entity	Recall (%)	Precision (%)	Recall (%)	Precision (%)	Recall (%)	Precision (%)	Recall (%)	Precision (%)
ORGANIZATION	59.46	81.48	59.28	79.55	59.46	81.48	58.73	81.85
PERSON	84.54	84.54	76.92	83.87	84.54	84.54	76.92	83.87
LOCATION	83.02	81.48	76.27	84.45	73.58	77.23	69.73	82.52
ARTIFACT	61.54	66.67	35.42	50.00	61.54	66.67	35.42	50.00
DATE	97.22	97.22	91.15	94.80	97.22	97.22	90.38	94.76
TIME	94.74	100.00	87.04	94.00	94.74	100.00	87.04	94.00
MONEY	100.00	100.00	93.33	93.33	100.00	100.00	93.33	93.33
PERCENT	-	-	100.00	95.45	-	-	80.95	94.44
Total	81.75	86.18	74.50	85.03	79.18	85.08	72.19	84.96
F-measure	83.91		79.42		82.02		78.05	

call and precision were 56% (10/18) and 91% (10/11), respectively. Twelve rules were applied 42 times in processing of the general domain data. There were 10 errors. The recall and precision were 41% (32/79) and 76% (32/42), respectively. We found the following two types of errors.

- A substring of an NE was extracted as an NE by mistake in one case.

  The substring "日 (*nichi*, Japan)" was extracted as LOCATION from the name of a location "在日米軍横田基地 (*zainichi_beigun_yokota_kichi*, an American military base in Yokota)." The whole of "在日米軍横田基地" should have been extracted as LOCATION according to the IREX-NE definition. The M.E. model was not able, however, to achieve this. Consequently, a transformation rule was applied to the whole string, and the substring was extracted by mistake. To reduce such errors, the M.E. model needs to be improved.

- Definitions assigned in the test data differed from those in the training data (10 cases).

  "邦" in the word "邦人 (*houjin*, Japanese)" and "外" in the word "外相会談 (*gaisou_kaidan*, Foreign Office Minister conference)" were defined as LOCATION and ORGANIZATION, respectively, in the training corpus while they were not NEs in the test data. To reduce such errors, maintenance of the training corpus is essential.

We obtained an improvement of about two points in the F-measure for the specific domain, and about 1.5 points in the F-measure for the general domain, by applying transformation rules. In our experiments, the system automatically acquired rules with consequent parts that always have NEs which include a substring of a morpheme, but did not acquire rules with consequent parts that do not have NEs which include a substring of a morpheme. So we carried out the experiments with all of the rules. We then obtained F-measures of 72.23 for the specific domain and 73.12 for the general domain. For the specific domain the results were ten points worse, and for the general domain five points worse, than the accuracies of the experimental results obtained without transformation rules. This result shows that the transformation rules acquired for any types of NEs do not have the ability to correctly revise NE labels assigned by our M.E. model. However, our rule ac-

Table 4: Accuracy with all feature sets, single feature sets, and one set omitted (with transformation rules).

Feature set	ARREST				GENERAL			
	Recall (%)	Precision (%)	F	Difference	Recall (%)	Precision (%)	F	Difference
All	81.75	86.18	83.91	0	74.50	85.03	79.42	0
Lexical items alone	73.26	80.97	76.92	-6.99	62.58	74.29	67.94	-11.48
POS (major) alone	5.40	70.00	10.02	-73.89	2.85	42.16	5.33	-74.09
POS (minor) alone	51.41	62.50	56.42	-27.49	45.23	61.31	52.06	-27.36
No lexical items	51.41	63.49	56.82	-27.09	46.16	65.45	54.14	-25.28
No POS (major)	80.46	85.99	83.13	-0.78	72.91	82.29	77.32	-2.10
No POS (minor)	76.09	87.57	81.43	-2.48	66.89	82.72	73.97	-5.45

Table 5: Accuracy with features of the target morpheme plus those of additional surrounding morphemes (with transformation rules).

Feature set	ARREST				GENERAL			
	Recall (%)	Precision (%)	F	Difference	Recall (%)	Precision (%)	F	Difference
On only (0)	31.11	48.79	37.99	-45.92	35.56	70.57	47.29	-32.13
On (-1)(0)(1)	76.86	84.46	80.48	-3.43	72.32	85.11	78.20	-1.22
On (-2) to (2)	81.75	86.18	83.91	0	74.50	85.03	79.42	0
On (-3) to (3)	80.72	85.09	82.85	-1.06	73.38	84.19	78.41	-1.01

quisition method is simple and we obtained good results with the rules acquired for NEs which include a substring of a morpheme. So we can conclude that the transformation rules acquired by our method are effective in extracting NEs which include a substring of a morpheme, which cannot be extracted by our M.E. model.

3.4 Features and Accuracy

This section describes how much each feature set contributes to improving the accuracy.

We carried out the experiments with each feature set alone, and with all feature sets but one, omitting each in turn. We used transformation rules in those experiments. Table 4 shows the performance under these conditions. In this table, "F" indicates the F-measure and "Difference" indicates the degradation from the results for the formal experiment. We achieved high accuracy with lexical items, and the accuracy decreased significantly when lexical items were not used. This result shows that the lexical items are the most important features for improving the accuracy.

Table 5 is a comparison with performance of the analysis for features of the target morpheme alone, and for performance with the features of surrounding morphemes as well. In this table, "On only (0)" indicates that we used features of the target morpheme alone, "On (-1) to (1)" indicates that we used features of the target morpheme and two adjacent morphemes. "On (-2) to (2)" indicates that we used features of the target morpheme and four other morphemes, the two on the left and the two on the right of the target. "On (-3) to (3)" indicates that we used features of the target morpheme and the six nearest morphemes, i.e., the three on the left and the three on the right. The best accuracy was achieved when we used the features of the target morpheme and the four nearest morphemes. The accuracy decreased when we used the features of the target morpheme and the six nearest morphemes. We believe that it is due to the data sparseness problem.

3.5 Amount of Training Data and Accuracy

Figure 2 shows the relationship between the amount of training data (the number of sentences) and accuracy. The horizontal axis indicates the number of sentences in training data, and the vertical axis indicates the F-measure. In this figure, the notation "arrest" and "general" are used to indicate the results

Table 6: Results of named entity extraction.

Named entity	ARREST		GENERAL	
	Recall (%)	Precision (%)	Recall (%)	Precision (%)
ORGANIZATION	68.92	85.00	62.33	79.79
PERSON	83.51	85.26	77.22	83.92
LOCATION	83.96	84.76	76.76	86.38
ARTIFACT	61.54	80.00	35.42	48.57
DATE	97.22	97.22	90.77	94.78
TIME	94.74	100.00	90.74	94.23
MONEY	100.00	88.89	93.33	82.35
PERCENT	-	-	100.00	100.00
Total	83.55 (+1.80)	88.08 (+1.90)	75.89 (+1.39)	85.20 (+0.17)
F-measure	85.75 (+1.84)		80.17 (+0.75)	

in the specific and general domains, respectively, and "with_rules" and "without_rules" are used to indicate the results obtained with and without transformation rules, respectively. These learning curves suggest that we can expect a certain amount of improvement with the use of more training data.

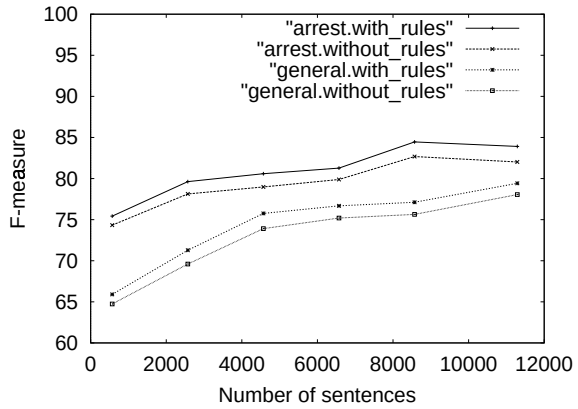

Figure 2: Relationship between the amount of training data and accuracy.

3.6 Use of an NE dictionary

Borthwick (Borthwick, 1999) and Nobata (Nobata, 1999) have developed other systems for extracting NEs. They have obtained improved accuracy by using an NE dictionary. We carried out an experiment with an NE dictionary. We used whether or not the target morpheme is in the NE dictionary as a feature.

We used the same dictionary as used by Borthwick and Nobata, available on Sekine's web site (Sekine, 1999) This is an NE dictionary of the names of organizations and locations, with about 1,000 entries. We also extracted NEs which appeared three or more times in a training corpus and added them to the NE dictionary. About 1,400 NEs were extracted (ORGANIZATION: 272, PERSON: 336, LOCATION: 339, ARTIFACT: 45, DATE: 233, TIME: 31, MONEY: 21, PERCENT: 45, OPTIONAL: 56). The total number of NEs in the NE dictionary was then about 2,400. We used JUMAN to morphologically analyze the NEs in the dictionary, and assigned one of the NE labels that we defined in Section 2 to each morpheme. There was a total of about 10,000 morphemes in the NE dictionary. When a string for a target morpheme was found in the dictionary, we used the NE label assigned to the corresponding morpheme in the dictionary as a feature value.

Table 6 shows the result obtained with the NE dictionary. The accuracy as expressed by the F-measure improved by about two points in the specific domain and about one point in the general domain, over the accuracy obtained without the NE dictionary. If we had an NE dictionary with more entries, we could achieve yet higher accuracies.

3.7 Related Works

With regard to named entity extraction from English sentences, statistical methods based on a hidden Markov model (HMM) (Bikel et al., 1997; Miller et al., 1998), a decision tree model (Cowie, 1995), an M.E. model

(Borthwick et al., 1998a), collocation statistics (Lin, 1998), and a transformation-based error-driven learning model (Aberdeen et al., 1995) have been proposed so far. In the MUC competition, the highest accuracy has been achieved by a system called Nymble (Bikel et al., 1997) which is based on an HMM. This system extracts NEs by applying the following procedure. A finite-state transition network is prepared. Each state of the network represents an NE defined in the MUC-NE task, such as PERSON or ORGANIZATION, or represents NOT-A-NAME which means the word is not a defined NE. Each transition has a transition probability, which represents the transition's conditional probability for a given input word. The analysis is a search for the optimal path in the network which uses the Viterbi algorithm. The states in the optimal path give us NEs. In the other systems, named entities are extracted by a similar procedure, except that the way of estimating the probability varies. Borthwick and his coworkers selected several systems which obtained a high accuracy in the MUC-NE task from among those based on statistical methods and those based on hand-crafted rules, and obtained better results than any of the individual systems by integrating them on the basis of the M.E. model (Borthwick et al., 1998a). They reported that a good accuracy which surpassed human performance could be obtained for a certain data set by integrating several systems (Borthwick et al., 1998b).

With regard to named entity extraction from Japanese sentences, similar statistical methods have been proposed, including methods based on an HMM (Shinnou, 1999), a decision tree model (Sekine et al., 1998; Nobata, 1999), and an M.E. model (Borthwick, 1999). Borthwick's approach is similar to ours except that he used hand-crafted transformation rules while we use automatically acquired rules alone. The accuracy we reported in Section 3.6 is better than that which Borthwick obtained. Our method is more accurate than any other system based on a statistical method that participated in the last IREX-NE workshop, and is close to that obtained by the system which obtained the highest accuracy for the IREX-NE task.

4 Conclusion

This paper described the extraction of named entities on the basis of an M.E. (maximum entropy) model and transformation rules. Eight types of NE are defined by IREX-NE, and each NE consists of one or more morphemes, or includes a substring of a morpheme. We defined 40 NE labels to indicate the beginning, middle, and end of NEs, and extract NEs which consist of one or more morphemes by estimating the labels according to an M.E. model. After this estimation, we extract NEs, which include a substring of a morpheme, by using transformation rules. These rules are automatically acquired by investigating the difference between NE labels in a tagged corpus and those extracted from the same corpus without tags by our system.

Through our experiments, we found that the transformation rules contribute to an improved accuracy, lexical items are the most important features, and the best accuracy was achieved when we used the features of the target morpheme and the four morphemes closest to it, i.e., the two on the left and the two on the right, when a training corpus with 12,000 sentences was used. These results were obtained with the information in the training corpus alone. When we used an NE dictionary which is available on the web as well, we achieved an F-measure of 85.75 for a specific domain, and 80.17 for a general domain, for IREX-NE formal-run data.

There are several possible future directions. In particular, we are interested in the following issues.

- Finding effective features

  We expect that we can achieve higher accuracy by using information that we are not using at the moment, such as information on dependencies between phrasal units called 'bunsetsu', anaphoric relations, and the information given in the process of analyzing text.

- Corpus revision and an NE dictionary

  We found that errors in a training corpus will lead to a lower accuracy, and that dictionary information helps to improve the accuracy. Therefore, corpus revision should be actively studied, and larger NE dictionaries will also be helpful.

  We may be able to tune the model to a particular domain by preparing an NE dictionary adapted to the domain. We would like to try this, and see how well an adapted dictionary works.

Acknowledgments

The authors would like to thank Satoshi Sekine and Andrew Borthwic for fruitful comments and helpful discussions during the progress of this work.

References

John Aberdeen, John Burger, David Day, Lynette Hirschman, Patricia Robinson, and Marc Vilain. 1995. MITRE: Description of the ALEMBIC System used for MUC-6. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pages 141–155.

Adam L. Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra. 1996. A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics*, 22(1):39–71.

Daniel M. Bikel, Scott Miller, Richard Schwartz, and Ralph Weischedel. 1997. Nymble: a High-Performance Learning Name-finder. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pages 194–201.

Andrew Borthwick, John Sterling, Eugene Agichtein, and Ralph Grishman. 1998a. Exploiting Diverse Knowledge Sources via Maximum Entropy in Named Entity Recognition. In *Proceedings of the Sixth Workshop on Very Large Corpora*, pages 152–160.

Andrew Borthwick, John Sterling, Eugene Agichtein, and Ralph Grishman. 1998b. NYU: Description of the MENE Named Entity System as Used in MUC-7. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*. http://www.muc.saic.com/.

Andrew Borthwick. 1999. A Japanese Named Entity Recognizer Constructed by a Non-Speaker of Japanese. In *Proceedings of the IREX Workshop*, pages 187–193.

Eric Brill. 1995. Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging. *Computational Linguistics*, 21(4):543–565.

Jim Cowie. 1995. CRL/NMSU Description of the CRL/NMSU System Used for MUC-6. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pages 157–166.

IREX Executive Committee. 1999. IREX homepage. http://cs.nyu.edu/cs/projects/proteus/irex/.

Sadao Kurohashi and Makoto Nagao, 1998. *Japanese Morphological Analysis System JUMAN Version 3.6.* Department of Informatics, Kyoto University.

Dekang Lin. 1998. Using Collocation Statistics in Information Extraction. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*. http://www.muc.saic.com/.

Scott Miller, Michael Crystal, Heidi Fox, Lance Ramshaw, Richard Schwartz, Rebecca Stone, Ralph Weischedel, and the Annotation Group. 1998. Algorithms that Learn to Extract Information BBN: Description of the Sift System as Used for MUC-7. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*. http://www.muc.saic.com/.

Chikashi Nobata. 1999. Named Entity Tagging System Based on a Decision Tree Model. In *Proceedings of the IREX Workshop*, pages 201–206. (in Japanese).

Adwait Ratnaparkhi. 1996. A Maximum Entropy Model for Part-Of-Speech Tagging. In *Conference on Empirical Methods in Natural Language Processing*, pages 133–142.

Adwait Ratnaparkhi. 1997. A Linear Observed Time Statistical Parser Based on Maximum Entropy Models. In *Conference on Empirical Methods in Natural Language Processing.*

SAIC. 1998. MUC homepage. http://www.muc.saic.com/.

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 Corpora*, pages 171–178.

Satoshi Sekine. 1999. Satoshi Sekine homepage. http://www.cs.nyu.edu/cs/projects/proteus/sekine/.

Hiroyuki Shinnou. 1999. Extraction of Proper Nouns through Extended Character Based HMM. In *Proceedings of the IREX Workshop*, pages 151–157. (in Japanese).

Kiyotaka Uchimoto, Satoshi Sekine, and Hitoshi Isahara. 1999. Japanese Dependency Structure Analysis Based on Maximum Entropy Models. In *Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics (EACL'99)*, pages 196–203.