Bottom-up Named Entity Recognition using a Two-stage Machine Learning Method

Hirotaka Funayama Tomohide Shibata Sadao Kurohashi
Kyoto University, Yoshida-honmachi,
Sakyo-ku, Kyoto, 606-8501, Japan
{funayama, shibata, kuro}@nlp.kuee.kyoto-u.ac.jp

Abstract

This paper proposes Japanese bottom-up named entity recognition using a two-stage machine learning method. Most work has formalized Named Entity Recognition as a sequential labeling problem, in which only local information is utilized for the label estimation, and thus a long named entity consisting of several morphemes tends to be wrongly recognized. Our proposed method regards a compound noun (*chunk*) as a labeling unit, and first estimates the labels of all the chunks in a phrasal unit (*bunsetsu*) using a machine learning method. Then, the best label assignment in the *bunsetsu* is determined from bottom up as the CKY parsing algorithm using a machine learning method. We conducted an experimental on CRL NE data, and achieved an F measure of 89.79, which is higher than previous work.

1 Introduction

Named Entity Recognition (NER) is a task of recognizing named entities such as person names, organization names, and location. It is used for several NLP applications such as Information Extraction (IE) and Question Answering (QA). Most work uses machine learning methods such as Support Vector Machines (SVMs) (Vapnik, 1995) and Conditional Random Field (CRF) (Lafferty et al., 2001) using a hand-annotated corpus (Krishnan and D.Manning, 2006; Kazama and Torisawa, 2008; Sasano and Kurohashi, 2008; Fukushima et al., 2008; Nakano and Hirai, 2004; Masayuki and Matsumoto, 2003).

In general, NER is formalized as a sequential labeling problem. For example, regarding a morpheme as a basic unit, it is first labeled as S-PERSON, B-PERSON, I-PERSON, E-PERSON, S-ORGANIZATION, etc. Then, considering the labeling results of morphemes, the best NE label sequence is recognized.

When the label of each morpheme is estimated, only local information around the morpheme (e.g., the morpheme, the two preceding morphemes, and the two following morphemes) is utilized. Therefore, a long named entity consisting of several morphemes tends to be wrongly recognized. Let us consider the example sentences shown in Figure 1.

In sentence (1), the label of "*Kazama*" can be recognized to be S-PERSON (PERSON consisting of one morpheme) by utilizing the surrounding information such as the suffix "*san*" (Mr.) and the verb "*kikoku shita*" (return home).

On the other hand, in sentence (2), when the label of "*shinyou*" (credit) is recognized to be B-ORGANIZATION (the beginning of ORGANIZATION), only information from "*hatsudou*" (invoke) to "*kyusai*" (relief) can be utilized, and thus the information of the morpheme "*ginkou*" (bank) that is apart from "*shinyou*" by three morphemes cannot be utilized. To cope with this problem, Nakano et al. (Nakano and Hirai, 2004) and Sasano et al. (Sasano and Kurohashi, 2008) utilized information of the head of *bunsetsu*[1]. In their methods, when the label of "*shinyou*" is recognized, the information of the morpheme "*ginkou*" can be utilized.

However, these methods do not work when the morpheme that we want to refer to is not a head of *bunsetsu* as in sentence (3). In this example, when "*gaikoku*" (foreign) is recognized to be B-ARTIFACT (the beginning of ARTIFACT), we want to refer to "*hou*" (law), not "*ihan*" (violation), which is the head of the *bunsetsu*.

This paper proposes Japanese bottom-up named

[1]Bunsetsu is the smallest coherent phrasal unit in Japanese. It consists of one or more content words followed by zero or more function words.

Proceedings of the 2009 Workshop on Multiword Expressions, ACL-IJCNLP 2009, pages 55–62,
Suntec, Singapore, 6 August 2009.

(1) *kikoku-shita Kazama-san-wa* ...
return home Mr.Kazama TOP

'Mr. Kazama who returned home'

(2) *hatsudou-shita shinyou-kumiai-kyusai-ginkou-no setsuritsu-mo*...
invoke credit union relief bank GEN establishment

'the establishment of the invoking credit union relief bank'

(3) *shibunsyo-gizou-to gaikoku-jin-touroku-hou-ihan-no utagai-de*
private document falsification and foreigner registration law violation GEN suspicion INS

'on suspicion of the private document falsification and the violation of the foreigner registration law'

Figure 1: Example sentences.

entity recognition using a two-stage machine learning method. Different from previous work, this method regards a compound noun as a labeling unit (we call it *chunk*, hereafter), and estimates the labels of all the chunks in the *bunsetsu* using a machine learning method. In sentence (3), all the chunks in the second *bunsetsu* (i.e., "*gaikoku*", "*gaikoku-jin*", ···, "*gaikoku-jin-touroku-hou-ihan* ", ···, "*ihan*") are labeled, and in the case that the chunk "*gaikoku-jin-touroku-hou*" is labeled, the information about "*hou*" (law) is utilized in a natural manner. Then, in the *bunsetsu*, the best label assignment is determined. For example, among the combination of "*gaikoku-jin-touroku-hou*" (ARTIFACT) and "*ihan*" (OTHER), the combination of "*gaikoku-jin*" (PERSON) and "*touroku-hou-ihan*" (OTHER), etc., the best label assignment, "*gaikoku-jin-touroku-hou*" (ARTIFACT) and "*ihan*" (OTHER), is chosen based on a machine learning method. In this determination of the best label assignment, as the CKY parsing algorithm, the label assignment is determined by bottom-up dynamic programming. We conducted an experimental on CRL NE data, and achieved an F measure of 89.79, which is higher than previous work.

This paper is organized as follows. Section 2 reviews related work of NER, especially focusing on sequential labeling based method. Section 3 describes an overview of our proposed method. Section 4 presents two machine learning models, and Section 5 describes an analysis algorithm. Section 6 gives an experimental result.

2 Related Work

In Japanese Named Entity Recognition, the definition of Named Entity in IREX Workshop (IREX Committee, 1999) is usually used. In this definition, NEs are classified into eight classes: PERSON, LOCATION, ORGANIZATION, ARTIFACT, DATE, TIME, MONEY, and PERCENT. Table 1 shows example instances of each class.

class	example
PERSON	*Kimura Syonosuke*
LOCATION	*Taiheiyou* (Pacific Ocean)
ORGANIZATION	*Jimin-tou* (Liberal Democratic Party)
ARTIFACT	PL-*houan* (PL bill)
DATE	21-*seiki* (21 century)
TIME	*gozen-7-ji* (7 a.m.)
MONEY	500-*oku-en* (50 billions yen)
PERCENT	20 percent

Table 1: NE classes and their examples.

NER methods are divided into two approaches: rule-based approach and machine learning approach. According to previous work, machine learning approach achieved better performance than rule-based approach.

In general, a machine learning method is formalized as a sequential labeling problem. This problem is first assigning each token (character or morpheme) to several labels. In an SE-algorithm (Sekine et al., 1998), *S* is assigned to NE composed of one morpheme, *B, I, E* is assigned to the beginning, middle, end of NE, respectively, and *O* is assigned to the morpheme that is not an NE[2]. The labels S, B, I, and E are prepared for each NE classes, and thus the total number of labels is 33 (= 8 * 4 + 1).

The model for the label estimation is learned based on machine learning. The following features are generally utilized: characters, type of

[2]Besides, there are IOB1, IOB2 algorithm using only I,O,B and IOE1, IOE2 algorithm using only I,O,E (Kim and Veenstra, 1999).

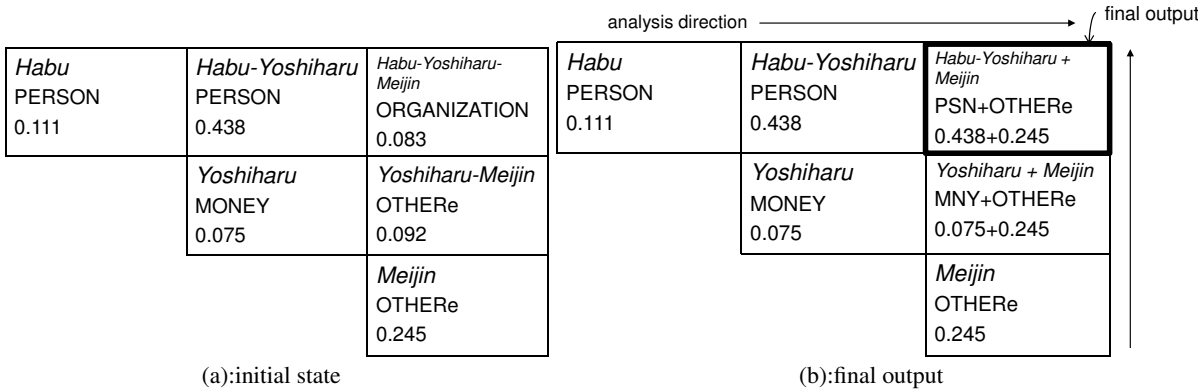


Figure 2: An overview of our proposed method. (the *bunsetsu* "*Habu-Yoshiharu-Meijin*")

character, POS, etc. about the morpheme and the surrounding two morphemes. The methods utilizing SVM or CRF are proposed.

Most of NER methods based on sequential labeling use only local information. Therefore, methods utilizing global information are proposed. Nakano et al. utilized as a feature the word sub class of NE on the analyzing direction in the *bunsetsu*, the noun in the end of the *bunsetsu* adjacent to the analyzing direction, and the head of each *bunsetsu* (Nakano and Hirai, 2004). Sasano et al. utilized cache feature, coreference result, syntactic feature, and caseframe feature as structural features (Sasano and Kurohashi, 2008).

Some work acquired knowledge from unannotated large corpus, and applied it to NER. Kazama et al. utilized a Named Entity dictionary constructed from Wikipedia and a noun clustering result obtained using huge amount of pairs of dependency relations (Kazama and Torisawa, 2008). Fukushima et al. acquired huge amount of category-instance pairs (e.g., "political party - New party DAICHI","company-TOYOTA") by some patterns from a large Web corpus (Fukushima et al., 2008).

In Japanese NER researches, CRL NE data are usually utilized for the evaluation. This data includes approximately 10 thousands sentences in news paper articles, in which approximately 20 thousands NEs are annotated. Previous work achieved an F measure of about 0.89 using this data.

3 Overview of Proposed Method

Our proposed method first estimates the label of all the compound nouns (chunk) in a *bunsetsu*. Then, the best label assignment is determined by bottom-up dynamic programming as the CKY parsing algorithm. Figure 2 illustrates an overview of our proposed method. In this example, the *bunsetsu* "*Habu-Yoshiharu-Meijin*" (Grand Master Yoshiharu Habu) is analyzed. First, the labels of all the chunks ("*Habu*", "*Habu-Yoshiharu*", "*Habu-Yoshiharu-Meijin*", $\cdots$, "*Meijin*", etc.) in the *bunsetsu* are analyzed using a machine learning method as shown in Figure 2 (a).

We call the state in Figure 2 (a) *initial state*, where the labels of all the chunks have been estimated. From this state, the best label assignment in the *bunsetsu* is determined. This procedure is performed from the lower left (corresponds to each morpheme) to the upper right like the CKY parsing algorithm as shown in Figure 2 (b). For example, when the label assignment for "*Habu-Yoshiharu*" is determined, the label assignment "*Habu-Yoshiharu*" (PERSON) and the label assignment "*Habu*" (PERSON) and "*Yoshiharu*" (OTHER) are compared, and the better one is chosen. While grammatical rules are utilized in a general CKY algorithm, this method chooses better label assignment for each cell using a machine learning method.

The learned models are the followings:

- the model that estimates the label of a chunk (*label estimation model*)
- the model that compares two label assignments (*label comparison model*)

The two models are described in detail in the next section.

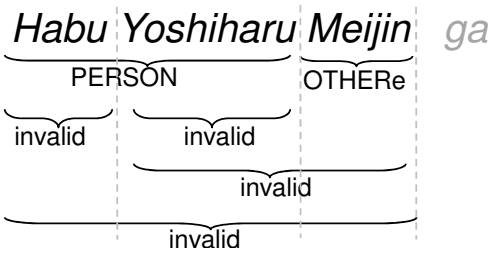


Figure 3: Label assignment for all the chunks in the *bunsetsu* "*Habu-Yoshiharu-Meijin*."

4 Model Learning

4.1 Label Estimation Model

This model estimates the label for each chunk. An analysis unit is basically *bunsetsu*. This is because 93.5% of named entities is located in a *bunsetsu* in CRL NE data. Exceptionally, the following expressions located in multiple *bunsetsus* tend to be an NE:

- expressions enclosed in parentheses (e.g., " '*Himeyuri-no tou*' " (The tower of Himeyuri) (ARTIFACT))
- expressions that have an entry in Wikipedia (e.g., "*Nihon-yatyou-no kai*" (Wild Bird Society of Japan) (ORGANIZATION))

Hereafter, *bunsetsu* is expanded when one of the above conditions meet. By this expansion, 98.6% of named entities is located in a *bunsetsu*[3].

For each *bunsetsu*, the head or tail function words are deleted. For example, in the *bunsetsu* "*Habu-Yoshiharu-Meijin-wa*", the tail function word "*wa*" (TOP) is deleted. In the *bunsetsu* "*yaku-san-bai*" (about three times), the head function word "*yaku*" (about) is deleted.

Next, for learning the label estimation model, all the chunks in a *bunsetsu* are attached to the correct label from a hand-annotated corpus. The label set is 13 classes, which includes eight NE class (as shown in Table 1), and five classes: OTHERs, OTHERb, OTHERi, OTHERe, and invalid.

The chunk that corresponds to a whole *bunsetsu* and does not contain any NEs is labeled as OTHERs, and the head, middle, tail chunk that does not correspond to an NE is labeled as OTHERb, OTHERi, OTHERe, respectively[4].

[3]As an example in which an NE is not included by an expanded *bunsetsu*, there are "*Toru-no Kimi*" (PERSON) and "*Osaka-fu midori-no kankyo-seibi-shitsu*" (ORGANIZATION).

[4]Each OTHER is assigned to the longest chunk that satisfies its condition in a chunk.

1. # of morphemes in the chunk
2. the position of the chunk in its *bunsetsu*
3. character type[5]
4. the combination of the character type of adjoining morphemes
   - For the chunk "Russian Army", this feature is "Katakana,Kanji"
5. word class, word sub class, and several features provided by a morphological analyzer JUMAN
6. several features[6] provided by a parser KNP
7. string of the morpheme in the chunk
8. IPADIC[7] feature
   - If the string of the chunk are registered in the following categories of IPADIC: "person", "location", "organization", and "general", this feature fires.
9. Wikipedia feature
   - If the string of the chunk has an entry in Wikipedia, this feature fires.
   - the hypernym extracted from its definition sentence using some patterns (e.g., The hypernym of "the Liberal Democratic Party" is a political party.)
10. cache feature
    - When the same string of the chunk appears in the preceding context, the label of the preceding chunk is used for the feature.
11. particles that the *bunsetsu* includes
12. the morphemes, particles, and head morpheme in the parent *bunsetsu*
13. the NE/category ratio in a case slot of predicate/noun case frame(Sasano and Kurohashi, 2008)
    - For example, in the case *ga* (NOM) of the predicate case frame "*kaiken*" (interview), the NE ratio "PERSON:0.245" is assigned to the case slot. Hence, in the sentence "*Habu-ga kaiken-shita*" (Mr. Habu interviewed), the feature "PERSON:0.245" is utilized for the chunk "*Habu*."
14. parenthesis feature
    - When the chunk in a parenthesis, this feature fires.

Table 2: Features for the label estimation model.

The chunk that is neither any eight NE class nor the above four OTHER is labeled as invalid.

In an example as shown in Figure 3, "*Habu-Yoshiharu*" is labeled as PERSON, "*Meijin*" is labeled as OTHERe, and the other chunks are labeled as invalid.

Next, the label estimation model is learned from the data in which the above label set is assigned

[5]The following five character types are considered: Kanji, Hiragana, Katakana, Number, and Alphabet.

[6]When a morpheme has an ambiguity, all the corresponding features fire.

[7]http://chasen.aist-nara.ac.jp/chasen/distribution.html.ja

to all the chunks. The features for the label estimation model are shown in Table 2. Among the features, as for feature (3), (5)−(8), three categories according to the position of a morpheme in the chunk are prepared: "head", "tail", and "anywhere." For example, in the chunk "*Habu-Yoshiharu-Meijin*," as for the morpheme "*Habu*", feature (7) is set to be "*Habu*" in "head" and as for the morpheme "*Yoshiharu*", feature (7) is set to be "*Yoshiharu*" in "anywhere."

The label estimation model is learned from pairs of label and feature in each chunk. To classify the multi classes, the one-vs-rest method is adopted (consequently, 13 models are learned). The SVM output is transformed by using the sigmoid function $\frac{1}{1+exp(-\beta x)}$, and the transformed value is normalized so that the sum of the value of 13 labels in a chunk is one.

The purpose for setting up the label "invalid" is as follows. In the chunk "*Habu*" and "*Yoshiharu*" in Figure 3, since the label "invalid" has a relatively higher score, the score of the label PERSON is relatively low. Therefore, when the label comparison described in Section 4.2 is performed, the label assignment "*Habu-Yoshiharu*" (PERSON) is likely to be chosen. In the chunk where the score of the label invalid has the highest score, the label that has the second highest score is adopted.

4.2 Label Comparison Model

This model compares the two label assignments for a certain string. For example, in the string "*Habu-Yoshiharu*", the model compares the following two label assignments:

- "*Habu-Yoshiharu*" is labeled as PERSON
- "*Habu*" is labeled as PERSON and "*Yoshiharu*" is labeled as MONEY

First, as shown in Figure 4, the two compared sets of chunks are lined up by sandwiching "**vs**." (The left one, right one is called the first set, the second set, respectively.) When the first set is correct, this example is positive: otherwise, this example is negative. The max number of chunks for each set is five, and thus examples in which the first or second set has more than five chunks are not utilized for the model learning.

Then, the feature is assigned to each example. The feature (13 dimensions) for each chunk is defined as follows: the first 12 dimensions are used for each label, which is estimated by the label estimation model, and the last 13th dimension is used for the score of an SVM output. Then, for the first and second set, the features for each chunk are arranged from the left, and zero vectors are placed in the remainder part.

positive:
+1 *Habu-Yoshiharu* **vs** *Habu + Yoshiharu*
PSN PSN + MNY
+1 *Habu-Yoshiharu + Meijin* **vs** *Habu + Yoshiharu + Meijin*
PSN + OTHERe PSN + MONEY + OTHERe
⋮
negative:
- 1 *Habu-Yoshiharu-Meijin* **vs** *Habu-Yoshiharu + Meijin*
ORG PSN + OTHERe
⋮

Figure 4: Assignment of positive/negative examples.

Figure 5 illustrates the feature for "*Habu-Yoshiharu*" **vs** "*Habu + Yoshiharu*." The label comparison model is learned from such data using SVM. Note that only the fact that "*Habu-Yoshiharu*" is PERSON can be found from the hand-annotated corpus, and thus in the example "*Habu-Yoshiharu-Meijin*" **vs** "*Habu + Yoshiharu-Meijin*", we cannot determine which one is correct. Therefore, such example cannot be used for the model learning.

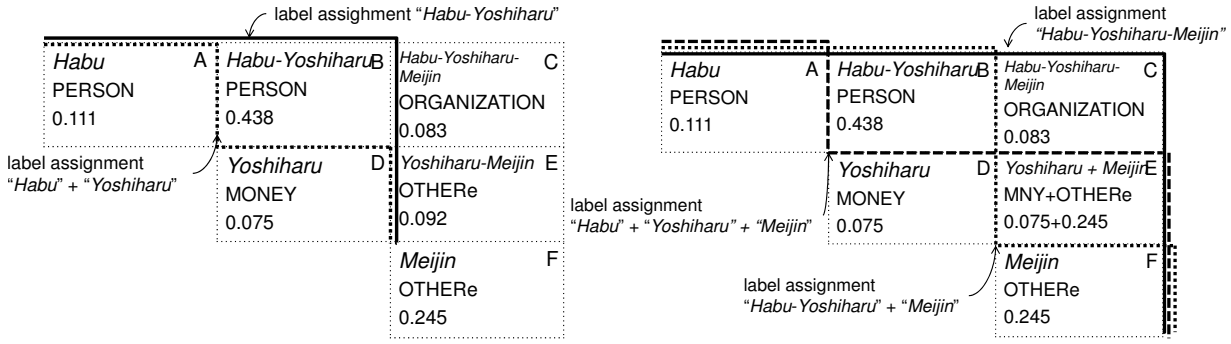
5 Analysis

First, the label of all the chunks in a *bunsetsu* is estimated by using the label estimation model described in Section 4.1. Then, the best label assignment in the *bunsetsu* is determined by applying the label comparison model described in Section 4.2 iteratively as shown in Figure 2 (b). In this step, the better label assignment is determined from bottom up as the CKY parsing algorithm.

For example, the initial state shown in Figure 2(a) is obtained using the label estimation model. Then, the label assignment is determined using the label comparison model from the lower left (corresponds to each morpheme) to the upper right. In determining the label assignment for the cell of "*Habu-Yoshiharu*" as shown in 6(a), the model compares the label assignment "B" with the label assignment "A+D." In this case, the model chooses the label assignment "B", that is, "*Habu - Yoshiharu*" is labeled as PERSON. Similarly, in determining the label assignment for the cell of "*Yoshiharu-Meijin*", the model compares the

chunk	*Habu-Yoshiharu*					*Habu*	*Yoshiharu*			
label	PERSON					PERSON	MONEY			
vector	$\mathbf{V_{11}}$	**0**	**0**	**0**	**0**	$\mathbf{V_{21}}$	$\mathbf{V_{22}}$	**0**	**0**	**0**

Figure 5: An example of the feature for the label comparison model. (The example is "*Habu-Yoshiharu* **vs** *Habu* + *Yoshiharu*", and $\mathbf{V_{11}}$, $\mathbf{V_{21}}$, $\mathbf{V_{22}}$, and **0** is a vector whose dimension is 13.)



(a): label assignment for the cell "*Habu-Yoshiharu*". (b): label assignment for the cell "*Habu-Yoshiharu-Meijin*".

Figure 6: The label comparison model.

label assignment "E" with the label assignment "D+F." In this case, the model chooses the label assignment "D+F", that is, "*Yoshiharu*" is labeled as MONEY and "*Meijin*" is labeled as OTHERe. When the label assignment consists of multiple chunks, the content of the cell is updated. In this case, the cell "E" is changed from "*Yoshiharu-Meijin*" (OTHERe) to "*Yoshiharu + Meijin*" (MONEY + OTHERe).

As shown in Figure 6(b), in determining the best label assignment for the upper right cell, that is, the final output is determined, the model compares the label assignment "A+D+F", "B+F", and "C". When there are more than two candidates of label assignments for a cell, all the label assignments are compared in a pairwise, and the label assignment that obtains the highest score is adopted.

In the label comparing step, the label assignment in which OTHER$_*$ follows OTHER$_*$ (OTHER$_*$ - OTHER$_*$) is not allowed since each OTHER is assigned to the longest chunk as described in Section 4.1. When the first combination of chunks equals to the second combination of chunks, the comparison is not performed.

6 Experiment

To demonstrate the effectiveness of our proposed method, we conducted an experiment on CRL NE data. In this data, 10,718 sentences in 1,174 news articles are annotated with eight NEs. The expression to which it is difficult to annotate manually is labeled as OPTIONAL, and was not used for both

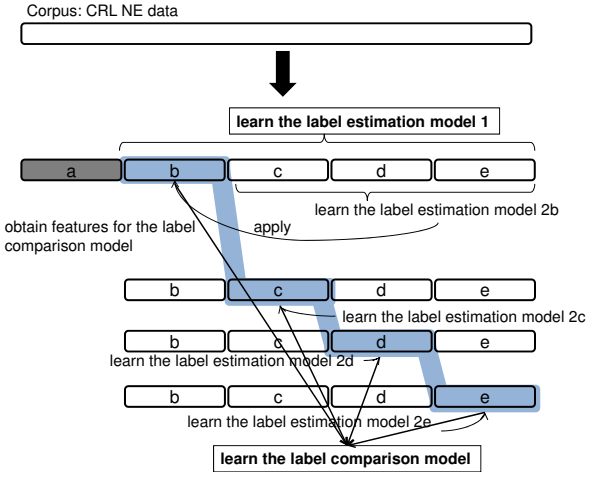


Figure 7: 5-fold cross validation.

the model learning[8] and the evaluation.

We performed 5-fold cross validation following previous work. Different from previous work, our work has to learn the SVM models twice. Therefore, the corpus was divided as shown in Figure 7. Let us consider the analysis in the part (a). First, the label estimation model 1 is learned from the part (b)-(e). Then, the label estimation model 2b is learned from the part (c)-(e), and applying the learned model to the part (b), features for learning the label comparison model are obtained. Similarly, the label estimation model 2c is learned from the part (b),(d),(e), and applying it to the part (c), features are obtained. It is the same with the part

[8]Exceptionally, "OPTIONAL" is used when the label estimation model for OTHER$_*$ and invalid is learned.

	Recall	Precision
ORGANIZATION	81.83 (3008/3676)	88.37 (3008/3404)
PERSON	90.05 (3458/3840)	93.87 (3458/3684)
LOCATION	91.38 (4992/5463)	92.44 (4992/5400)
ARTIFACT	46.72 (349/ 747)	74.89 (349/ 466)
DATE	93.27 (3327/3567)	93.12 (3327/3573)
TIME	88.25 (443/ 502)	90.59 (443/ 489)
MONEY	93.85 (366/ 390)	97.60 (366/ 375)
PERCENT	95.33 (469/ 492)	95.91 (469/ 489)
ALL-SLOT	87.87	91.79
F-measure		89.79

Table 3: Experimental result.

(d) and (e). Then, the label comparison model is learned from the obtained features. After that, the analysis in the part (a) is performed by using both the label estimation model 1 and the label comparison model.

In this experiment, a Japanese morphological analyzer, JUMAN[9], and a Japanese parser, KNP[10] were adopted. The two SVM models were learned with polynomial kernel of degree 2, and β in the sigmoid function was set to be 1.

Table 6 shows an experimental result. An F-measure in all NE classes is 89.79.

7 Discussion

7.1 Comparison with Previous Work

Table 7 presents the comparison with previous work, and our method outperformed previous work. Among previous work, Fukushima et al. acquired huge amount of category-instance pairs (e.g., "political party - New party DAICHI","company-TOYOTA") by some patterns from a large Web corpus, and Sasano et al. utilized the analysis result of corefer resolution as a feature for the model learning. Therefore, in our method, by incorporating these knowledge and/or such analysis result, the performance would be improved.

Compared with Sasano et al., our method achieved the better performance in analyzing a long compound noun. For example, in the *bunsetsu* "*Oushu-tsuujyou-senryoku-sakugen-jyouyaku*" (Treaty on Conventional Armed Forces in Europe), while Sasano et al. labeled "*Oushu*" (Europe) as LOCATION, our method correctly labeled "*Oushu-tsuujyou-senryoku-sakugen-jyouyaku*" as ARTIFACT. Sasano et al. incorrectly labeled "*Oushu*" as LOCATION although they utilized the information about the head of *bunsetsu* "*jyouyaku*" (treaty). In our method, for the cell "*Oushu*", invalid has the highest score, and thus the score of LOCATION relatively drops. Similarly, for the cell "*senryoku-sakugen-jyouyaku*", invalid has the highest score. Consequently, "*Oushu-tsuujyou-senryoku-sakugen-jyouyaku*" is correctly labeled as ARTIFACT.

In the *bunsetsu* "*gaikoku-jin-touroku-hou-ihan*" (the violation of the foreigner registration law), while Sasano et al. labeled "*touroku-hou*" as ARTIFACT, our method correctly labeled "*gaikoku-jin-touroku-hou*" as ARTIFACT. Sasano et al. cannot utilize the information about "*hou*" that is useful for the label estimation since the head of this *bunsetsu* is "*ihan*." In contrast, in estimating the label of the chunk "*gaikoku-jin-touroku-hou*", the information of "*hou*" can be utilized.

7.2 Error Analysis

There were some errors in analyzing a Katakana alphabet word. In the following example, although the correct is that "Batistuta" is labeled as PERSON, the system labeled it as OTHERs.

(4) Italy-*de* *katsuyaku-suru* Batistuta-*wo*
Italy LOC active Batistuta ACC
kuwaeta Argentine
call Argentine
'Argentine called Batistuta who was active in Italy.'

There is not an entry of "Batistuta" in the dictionary of JUMAN nor Wikipedia, and thus only the surrounding information is utilized. However, the case analysis of "*katsuyaku*" (active) is incorrect, which leads to the error of "Batistuta".

There were some errors in applying the label comparison model although the analysis of each chunk is correct. For example, in the *bunsetsu* "HongKong-*seityou*" (Government of HongKong), the correct is that "HongKong-*seityou*" is labeled as ORGANIZATION. As shown in Figure 8 (b), the system incorrectly labeled "HongKong" as LOCATION. As shown in Figure 8(a), although in the initial state, "HongKong-*seityou*" was correctly labeled as ORGANIZATION, the label assignment "HongKong + *seityou*" was incorrectly chosen by the label comparison model. To cope with this problem, we are planning to the adjustment of the value β in the sigmoid function and the refinement of the

[9] http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman-e.html
[10] http://nlp.kuee.kyoto-u.ac.jp/nl-resource/knp-e.html

	F1	analysis unit	distinctive features
(Fukushima et al., 2008)	89.29	character	Web
(Kazama and Torisawa, 2008)	88.93	character	Wikipedia,Web
(Sasano and Kurohashi, 2008)	89.40	morpheme	structural information
(Nakano and Hirai, 2004)	89.03	character	*bunsetsu* feature
(Masayuki and Matsumoto, 2003)	87.21	character	
(Isozaki and Kazawa, 2003)	86.77	morpheme	
proposed method	**89.79**	compound noun	Wikipedia,structural information

Table 4: Comparison with previous work. (All work was evaluated on CRL NE data using cross validation.)

HongKong LOCATION 0.266	HongKong-*seityou* ORGANIZATION 0.205
	seityou OTHERe 0.184

(a):initial state

HongKong LOCATION 0.266	HongKong + *seityou* LOC+OTHERe 0.266+0.184
	seityou OTHERe 0.184

(b):the final output

Figure 8: An example of the error in the label comparison model.

features for the label comparison model.

8 Conclusion

This paper proposed bottom-up Named Entity Recognition using a two-stage machine learning method. This method first estimates the label of all the chunks in a *bunsetsu* using a machine learning, and then the best label assignment is determined by bottom-up dynamic programming. We conducted an experiment on CRL NE data, and achieved an F-measure of 89.79.

We are planning to integrate this method with the syntactic and case analysis method (Kawahara and Kurohashi, 2007), and perform syntactic, case, and Named Entity analysis simultaneously to improve the overall accuracy.

References

Ken'ichi Fukushima, Nobuhiro Kaji, and Masaru Kitsuregawa. 2008. Use of massive amounts of web text in Japanese named entity recognition. In *Proceedings of Data Engineering Workshop (DEWS2008)*. A3-3 (in Japanese).

IREX Committee, editor. 1999. *Proceedings of the IREX Workshop*.

Hideki Isozaki and Hideto Kazawa. 2003. Speeding up support vector machines for named entity recognition. *Transaction of Information Processing Society of Japan*, 44(3):970–979. (in Japanese).

Daisuke Kawahara and Sadao Kurohashi. 2007. Probabilistic coordination disambiguation in a fully-lexicalized Japanese parser. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL2007)*, pages 304–311.

Jun'ichi Kazama and Kentaro Torisawa. 2008. Inducing gazetteers for named entity recognition by large-scale clustering of dependency relations. In *Proceedings of ACL-08: HLT*, pages 407–415.

Erik F. Tjong Kim and Jorn Veenstra. 1999. Representing text chunks. In *Proceedings of EACL '99*, pages 173–179.

Vajay Krishnan and Christopher D.Manning. 2006. An effective two-stage model for exploiting non-local dependencies in named entity recognition. pages 1121–1128.

John Lafferty, Andrew McCallun, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference (ICML'01)*, pages 282–289.

Asahara Masayuki and Yuji Matsumoto. 2003. Japanese named entity extraction with redundant morphological analysis. In *Proceeding of HLT-NAACL 2003*, pages 8–15.

Keigo Nakano and Yuzo Hirai. 2004. Japanese named entity extraction with bunsetsu features. *Transaction of Information Processing Society of Japan*, 45(3):934–941. (in Japanese).

Ryohei Sasano and Sadao Kurohashi. 2008. Japanese named entity recognition using structural natural language processing. In *Proceeding of Third International Joint Conference on Natural Language Processing*, pages 607–612.

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 (WVLC-6)*, pages 171–178.

Vladimir Vapnik. 1995. *The Nature of Statistical Learning Theory*. Springer.