Robust Extraction of Named Entity Including Unfamiliar Word

Masatoshi Tsuchiya[†] **Shinya Hida**[‡] **Seiichi Nakagawa**[‡]
[†]Information and Media Center / [‡]Department of Information and Computer Sciences,
Toyohashi University of Technology
tsuchiya@imc.tut.ac.jp, {hida,nakagawa}@slp.ics.tut.ac.jp

Abstract

This paper proposes a novel method to extract named entities including unfamiliar words which do not occur or occur few times in a training corpus using a large unannotated corpus. The proposed method consists of two steps. The first step is to assign the most similar and familiar word to each unfamiliar word based on their context vectors calculated from a large unannotated corpus. After that, traditional machine learning approaches are employed as the second step. The experiments of extracting Japanese named entities from IREX corpus and NHK corpus show the effectiveness of the proposed method.

1 Introduction

It is widely agreed that extraction of named entity (henceforth, denoted as NE) is an important subtask for various NLP applications. Various machine learning approaches such as maximum entropy(Uchimoto et al., 2000), decision list(Sassano and Utsuro, 2000; Isozaki, 2001), and Support Vector Machine(Yamada et al., 2002; Isozaki and Kazawa, 2002) were investigated for extracting NEs.

All of them require a corpus whose NEs are annotated properly as training data. However, it is difficult to obtain an enough corpus in the real world, because there are increasing the number of NEs like personal names and company names. For example, a large database of organization names(Nichigai Associates, 2007) already contains 171,708 entries and is still increasing. Therefore, a robust method to extract NEs including unfamiliar words which do not occur or occur few times in a training corpus is necessary.

This paper proposes a novel method of extracting NEs which contain unfamiliar morphemes using a large unannotated corpus, in order to resolve the above problem. The proposed method consists of two steps. The first step is to assign the most similar and familiar morpheme to each unfamiliar morpheme based on their context vectors calculated from a large unannotated corpus. The second step is to employ traditional machine learning approaches using both features of original morphemes and features of similar morphemes. The experiments of extracting Japanese NEs from IREX corpus and NHK corpus show the effectiveness of the proposed method.

Table 1: Statistics of NE Types of IREX Corpus

NE Type	Frequency	(%)
ARTIFACT	747	(4.0)
DATE	3567	(19.1)
LOCATION	5463	(29.2)
MONEY	390	(2.1)
ORGANIZATION	3676	(19.7)
PERCENT	492	(2.6)
PERSON	3840	(20.6)
TIME	502	(2.7)
Total	18677	

2 Extraction of Japanese Named Entity

2.1 Task of the IREX Workshop

The task of NE extraction of the IREX workshop (Sekine and Eriguchi, 2000) is to recognize eight NE types in Table 1. The organizer of the IREX workshop provided a training corpus, which consists of 1,174 newspaper articles published from January 1st 1995 to 10th which include 18,677 NEs. In the Japanese language, no other corpus whose NEs are annotated is publicly available as far as we know.[1]

2.2 Chunking of Named Entities

It is quite common that the task of extracting Japanese NEs from a sentence is formalized as a chunking problem against a sequence of mor-

[1] The organizer of the IREX workshop also provides the testing data to its participants, however, we cannot see it because we did not join it.

Proceedings of ACL-08: HLT, Short Papers (Companion Volume), pages 125–128,
Columbus, Ohio, USA, June 2008. ©2008 Association for Computational Linguistics

phemes. For representing proper chunks, we employ IOB2 representation, one of those which have been studied well in various chunking tasks of NLP (Tjong Kim Sang, 1999). This representation uses the following three labels.

- **B** Current token is the beginning of a chunk.
- **I** Current token is a middle or the end of a chunk consisting of more than one token.
- **O** Current token is outside of any chunk.

Actually, we prepare the 16 derived labels from the label **B** and the label **I** for eight NE types, in order to distinguish them.

When the task of extracting Japanese NEs from a sentence is formalized as a chunking problem of a sequence of morphemes, the segmentation boundary problem arises as widely known. For example, the NE definition of IREX tells that a Chinese character "米 (bei)" must be extracted as an NE means *America* from a morpheme "訪米 (hou-bei)" which means *visiting America*. A naive chunker using a morpheme as a chunking unit cannot extract such kind of NEs. In order to cope this problem, (Uchimoto et al., 2000) proposed employing translation rules to modify problematic morphemes, and (Asahara and Matsumoto, 2003; Nakano and Hirai, 2004) formalized the task of extracting NEs as a chunking problem of a sequence of characters instead of a sequence of morphemes. In this paper, we keep the naive formalization, because it is still enough to compare performances of proposed methods and baseline methods.

3 Robust Extraction of Named Entities Including Unfamiliar Words

The proposed method of extracting NEs consists of two steps. Its first step is to assign the most similar and familiar morpheme to each unfamiliar morpheme based on their context vectors calculated from a large unannotated corpus. The second step is to employ traditional machine learning approaches using both features of original morphemes and features of similar morphemes. The following subsections describe these steps respectively.

3.1 Assignment of Similar Morpheme

A context vector V_m of a morpheme m is a vector consisting of frequencies of all possible unigrams and bigrams,

$$V_m = \begin{pmatrix} f(m, m_0), & \cdots & f(m, m_N), \\ f(m, m_0, m_0), & \cdots & f(m, m_N, m_N), \\ f(m_0, m), & \cdots & f(m_N, m), \\ f(m_0, m_0, m), & \cdots & f(m_N, m_N, m) \end{pmatrix},$$

where $M \equiv \{m_0, m_1, \ldots, m_N\}$ is a set of all morphemes of the unannotated corpus, $f(m_i, m_j)$ is a frequency that a sequence of a morpheme m_i and a morpheme m_j occurs in the unannotated corpus, and $f(m_i, m_j, m_k)$ is a frequency that a sequence of morphemes m_i, m_j and m_k occurs in the unannotated corpus.

Suppose an unfamiliar morpheme $m_u \in M \cap \overline{M_F}$, where M_F is a set of familiar morphemes that occur frequently in the annotated corpus. The most similar morpheme $\hat{m_u}$ to the morpheme m_u measured with their context vectors is given by the following equation,

$$\hat{m_u} = \operatorname*{argmax}_{m \in M_F} sim(V_{m_u}, V_m), \qquad (1)$$

where $sim(V_i, V_j)$ is a similarity function between context vectors. In this paper, the cosine function is employed as it.

3.2 Features

The feature set F_i at i-th position is defined as a tuple of the *morpheme feature* $MF(m_i)$ of the i-th morpheme m_i, the *similar morpheme feature* $SF(m_i)$, and the *character type feature* $CF(m_i)$.

$$F_i = \langle\ MF(m_i),\ SF(m_i),\ CF(m_i)\ \rangle$$

The morpheme feature $MF(m_i)$ is a pair of the surface string and the part-of-speech of m_i. The similar morpheme feature $SF(m_i)$ is defined as

$$SF(m_i) = \begin{cases} MF(\hat{m_i}) & \text{if } m_i \in M \cap \overline{M_F} \\ MF(m_i) & \text{otherwise} \end{cases},$$

where $\hat{m_i}$ is the most similar and familiar morpheme to m_i given by Equation (1). The character type feature $CF(m_i)$ is a set of four binary flags to indicate that the surface string of m_i contains a Chinese character, a *hiragana* character, a *katakana* character, and an English alphabet respectively.

When we identify the chunk label c_i for the i-th morpheme m_i, the surrounding five feature sets $F_{i-2}, F_{i-1}, F_i, F_{i+1}, F_{i+2}$ and the preceding two chunk labels c_{i-2}, c_{i-1} are refered.

Morpheme Feature			Similar Morpheme Feature			Character Type Feature	Chunk Label
	(English translation)	POS		*(English translation)*	POS		
今日 (kyou)	(*today*)	Noun–Adverbial	今日 (kyou)	(*today*)	Noun–Adverbial	$\langle 1, 0, 0, 0 \rangle$	`O`
の (no)	gen	Particle	の (no)	gen	Particle	$\langle 0, 1, 0, 0 \rangle$	`O`
石狩 (Ishikari)	(*Ishikari*)	Noun–Proper	**関東 (Kantou)**	(*Kantou*)	Noun–Proper	$\langle 1, 0, 0, 0 \rangle$	`B-LOCATION`
平野 (heiya)	(*plain*)	Noun–Generic	平野 (heiya)	(*plain*)	Noun–Generic	$\langle 1, 0, 0, 0 \rangle$	`I-LOCATION`
の (no)	gen	Particle	の (no)	gen	Particle	$\langle 0, 1, 0, 0 \rangle$	`O`
天気 (tenki)	(*weather*)	Noun–Generic	天気 (tenki)	(*weather*)	Noun–Generic	$\langle 1, 0, 0, 0 \rangle$	`O`
は (ha)	top	Particle	は (ha)	top	Particle	$\langle 0, 1, 0, 0 \rangle$	`O`
晴れ (hare)	(*fine*)	Noun–Generic	晴れ (hare)	(*fine*)	Noun–Generic	$\langle 1, 1, 0, 0 \rangle$	`O`

Figure 1: Example of Training Instance for Proposed Method

	⟶ Parsing Direction ⟶				
Feature set	F_{i-2}	F_{i-1}	F_i	F_{i+1}	F_{i+2}
Chunk label	c_{i-2}	c_{i-1}	$\boxed{c_i}$		

Figure 1 shows an example of training instance of the proposed method for the sentence "今日 (kyou) の (no) 石狩 (Ishikari) 平野 (heiya) の (no) 天気 (tenki) は (ha) 晴れ (hare)" which means "*It is fine at Ishikari-plain, today*". "関東 (Kantou)" is assigned as the most similar and familiar morpheme to "石狩 (Ishikari)" which is unfamiliar in the training corpus.

4 Experimental Evaluation

4.1 Experimental Setup

IREX Corpus is used as the annotated corpus to train statistical NE chunkers, and M_F is defined experimentally as a set of all morphemes which occur five or more times in IREX corpus. Mainichi Newspaper Corpus (1993–1995), which contains 3.5M sentences consisting of 140M words, is used as the unannotated corpus to calculate context vectors. MeCab[2](Kudo et al., 2004) is used as a preprocessing morphological analyzer through experiments.

In this paper, either Conditional Random Fields(CRF)[3](Lafferty et al., 2001) or Support Vector Machine(SVM)[4](Cristianini and Shawe-Taylor, 2000) is employed to train a statistical NE chunker.

4.2 Experiment of IREX Corpus

Table 2 shows the results of extracting NEs of IREX corpus, which are measured with F-measure through 5-fold cross validation. The columns of "Proposed" show the results with SF, and the ones of "Baseline" show the results without SF. The column of "NExT" shows the result of using NExT(Masui et al., 2002), an NE chunker based on hand-crafted rules, without 5-fold cross validation.

Table 2: NE Extraction Performance of IREX Corpus

	Proposed		Baseline		NExT
	CRF	SVM	CRF	SVM	
ARTIFACT	0.487	0.518	0.458	0.457	-
DATE	0.921	0.909	0.916	0.916	0.682
LOCATION	0.866	0.863	0.847	0.846	0.696
MONEY	0.951	0.610	0.937	0.937	0.895
ORGANIZATION	0.774	0.766	0.744	0.742	0.506
PERCENT	0.936	0.863	0.928	0.928	0.821
PERSON	0.825	0.842	0.788	0.787	0.672
TIME	0.901	0.903	0.902	0.901	0.800
Total	0.842	0.834	0.821	0.820	0.732

Table 3: Statistics of NE Types of NHK Corpus

NE Type	Frequency (%)	
DATE	755	(19%)
LOCATION	1465	(36%)
MONEY	124	(3%)
ORGANIZATION	1056	(26%)
PERCENT	55	(1%)
PERSON	516	(13%)
TIME	101	(2%)
Total	4072	

As shown in Table 2, machine learning approaches with SF outperform ones without SF. Please note that the result of SVM without SF and the result of (Yamada et al., 2002) are comparable, because our using feature set without SF is quite similar to their feature set. This fact suggests that SF is effective to achieve better performances than the previous research. CRF with SF achieves better performance than SVM with SF, although CRF and SVM are comparable in the case without SF. NExT achieves poorer performance than CRF and SVM.

4.3 Experiment of NHK Corpus

Nippon Housou Kyoukai (NHK) corpus is a set of transcriptions of 30 broadcast news programs which were broadcasted from June 1st 1996 to 12th. Table 3 shows the statistics of NEs of NHK corpus which were annotated by a graduate student except

[2] http://mecab.sourceforge.net/

[3] http://chasen.org/~taku/software/CRF++/

[4] http://chasen.org/~taku/software/yamcha/

Table 4: NE Extraction Performance of NHK Corpus

	Proposed		Baseline		NExT
	CRF	SVM	CRF	SVM	
DATE	0.630	0.595	0.571	0.569	0.523
LOCATION	0.837	0.825	0.797	0.811	0.741
MONEY	0.988	0.660	0.971	0.623	0.996
ORGANIZATION	0.662	0.636	0.601	0.598	0.612
PERCENT	0.538	0.430	0.539	0.435	0.254
PERSON	0.794	0.813	0.752	0.787	0.622
TIME	0.250	0.224	0.200	0.247	0.260
Total	0.746	0.719	0.702	0.697	0.615

Table 5: Extraction of Familiar/Unfamiliar NEs

	Familiar	Unfamiliar	Other
CRF (Proposed)	0.789	0.654	0.621
CRF (Baseline)	0.757	0.556	0.614

for ARTIFACT in accordance with the NE definition of IREX. Because all articles of IREX corpus had been published earlier than broadcasting programs of NHK corpus, we can suppose that NHK corpus contains unfamiliar NEs like real input texts.

Table 4 shows the results of chunkers trained from whole IREX corpus against NHK corpus. The methods with SF outperform the ones without SF. Furthermore, performance improvements between the ones with SF and the ones without SF are greater than Table 2.

The performance of CRF with SF and one of CRF without SF are compared in Table 5. The column "Familiar" shows the results of extracting NEs which consist of familiar morphemes, as well as the column "Unfamiliar" shows the results of extracting NEs which consist of unfamiliar morphemes. The column "Other" shows the results of extracting NEs which contain both familiar morpheme and unfamiliar one. These results indicate that SF is especially effective to extract NEs consisting of unfamiliar morphemes.

5 Concluding Remarks

This paper proposes a novel method to extract NEs including unfamiliar morphemes which do not occur or occur few times in a training corpus using a large unannotated corpus. The experimental results show that SF is effective for robust extracting NEs which consist of unfamiliar morphemes. There are other effective features of extracting NEs like N-best morpheme sequences described in (Asahara and Matsumoto, 2003) and features of surrounding phrases described in (Nakano and Hirai, 2004). We will investigate incorporating SF and these features in the near future.

References

Masayuki Asahara and Yuji Matsumoto. 2003. Japanese named entity extraction with redundant morphological analysis. In *Proc. of HLT–NAACL '03*, pages 8–15.

Nello Cristianini and John Shawe-Taylor. 2000. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press.

Hideki Isozaki and Hideto Kazawa. 2002. Efficient support vector classifiers for named entity recognition. In *Proc. of the 19th COLING*, pages 1–7.

Hideki Isozaki. 2001. Japanese named entity recognition based on a simple rule generator and decision tree learning. In *Proc. of ACL '01*, pages 314–321.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Appliying conditional random fields to japanese morphological analysis. In *Proc. of EMNLP2004*, pages 230–237.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of ICML*, pages 282–289.

Fumito Masui, Shinya Suzuki, and Junichi Fukumoto. 2002. Development of named entity extraction tool NExT for text processing. In *Proceedings of the 8th Annual Meeting of the Association for Natural Language Processing*, pages 176–179. (in Japanese).

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

Nichigai Associates, editor. 2007. *DCS Kikan-mei Jisho*. Nichigai Associates. (in Japanese).

Manabu Sassano and Takehito Utsuro. 2000. Named entity chunking techniques in supervised learning for japanese named entity recognition. In *Proc. of the 18th COLING*, pages 705–711.

Satoshi Sekine and Yoshio Eriguchi. 2000. Japanese named entity extraction evaluation: analysis of results. In *Proc. of the 18th COLING*, pages 1106–1110.

E. Tjong Kim Sang. 1999. Representing text chunks. In *Proc. of the 9th EACL*, pages 173–179.

Kiyotaka Uchimoto, Ma Qing, Masaki Murata, Hiromi Ozaku, Masao Utiyama, and Hitoshi Isahara. 2000. Named entity extraction based on a maximum entropy model and transformation rules. *Journal of Natural Language Processing*, 7(2):63–90, Apr. (in Japanese).

Hiroyasu Yamada, Taku Kudo, and Yuji Matsumoto. 2002. Japanese named entity extraction using support vector machine. *Transactions of Information Processing Society of Japan*, 43(1):44–53, Jan. (in Japanese).