Analysis and Robust Extraction of Changing Named Entities

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

Abstract

This paper focuses on the change of named entities over time and its influence on the performance of the named entity tagger. First, we analyze Japanese named entities which appear in Mainichi Newspaper articles published in 1995, 1996, 1997, 1998 and 2005. This analysis reveals that the number of named entity types and the number of named entity tokens are almost steady over time and that $70 \sim 80\%$ of named entity types in a certain year occur in the articles published either in its succeeding year or in its preceding year. These facts lead that $20 \sim 30\%$ of named entity types are replaced with new ones every year. The experiment against these texts shows that our proposing semi-supervised method which combines a small annotated corpus and a large unannotated corpus for training works robustly although the traditional supervised method is fragile against the change of name entity distribution.

1 Introduction

It is widely agreed that extraction of named entity (henceforth, denoted as *NE*) is an important subtask for various NLP applications, such as information retrieval, machine translation, information extraction and natural language understanding. Several conferences like Message Understanding Conference(Grishman and Sundheim, 1996) and the IREX workshop (Sekine and Eriguchi, 2000) were conducted to encourage researchers of NE extraction and to provide its common evaluation basis.

In Japanese NE extraction, it is quite common to apply morphological analysis as preprocessing stage which segments a sentence into a sequence of morphemes. After that, either a pattern matcher based on hand-crafted rules or a statistical chunker is employed to extract NEs from a sequence of morphemes. 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. These researches show that machine learning approaches are more promising than approaches based on hand-crafted rules if a large corpus whose NEs are properly annotated is available as training data.

However, it is difficult to obtain an enough corpus in the real world because of the increasing number of NE types and the increasing time gap between the training corpus and the test corpus. There is the increasing number of NE types like personal names and company names in the real world. For example, a large database of organization names(Nichigai Associates, 2007) already contains 171,708 types and is still increasing. Because annotation work is quite expensive, the annotated corpus may become obsolete in a short period of time. Both of two factors expands the difference of NE distribution between the training corpus and the test corpus, and it may decrease the performance of the NE tagger as shown in (Mota and Grishman, 2008). Therefore, a robust method to extract NEs which do not occur or occur few times in a training corpus is necessary.

This paper focuses on the change of NEs over time and its influence on the performance of the NE tagger. First, we annotate NEs in Mainichi Newspaper articles published in 1996, 1997, 1998 and 2005, and analyze NEs which appear in these texts and an existing corpus. It consists of Mainichi Newspaper articles published in 1995, thus, we get an annotated corpus that spans 10 years. This analysis reveals that the number of NE types and the number of NE tokens are almost

Proceedings of the 2009 Named Entities Workshop, ACL-IJCNLP 2009, pages 161–167,
Suntec, Singapore, 7 August 2009.

Table 1: Statistics of NE categories of IREX corpus

NE Categories	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	

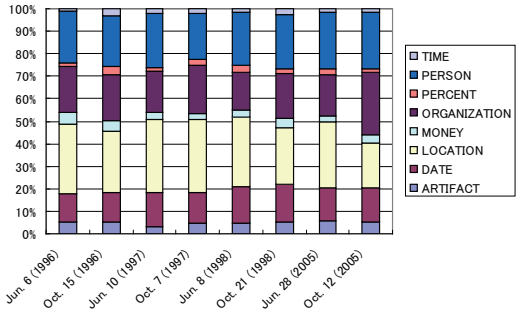

Figure 1: Distribution of NE categories

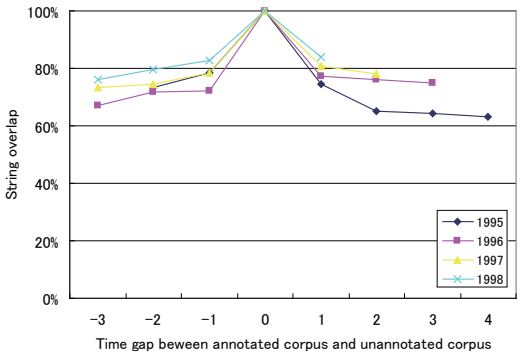

Figure 2: Overlap ratio of NEs over years

steady over time and that that $70 \sim 80\%$ of NE types in a certain year occur in the articles published either in its succeeding year or in its preceding year. These facts lead that $20 \sim 30\%$ of named entity types are replaced with new ones every year. The experiment against these corpora shows that the traditional supervised method is fragile against the change of NE types and that our proposing semi-supervised method which combines a small annotated corpus and a large unannotated corpus for training is robust against the change of NE types.

2 Analysis of Changing Named Entities

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 categories in Table 1. The organizer of the IREX workshop provided a training corpus (henceforth, denoted as *IREX corpus*), which consists of 1,174 Mainichi Newspaper articles published from January 1st 1995 to 10th which include 18,677 NEs. In the Japanese language, no other corpora whose NEs are annotated are publicly available as far as we know.[1] Thus, IREX corpus is referred as a golden sample of NE distribution in this paper.

2.2 Data Description

The most homogeneous texts which are written in different days are desirable, to explore the influence of the text time frame on NE distribution. Because IREX corpus is referred as a golden sample in this paper, Mainichi Newspaper articles written in different years than IREX corpus is suitable. Thus, ordinal days of June and October in 1996, 1997, 1998 and 2005 are randomly selected as sampling days.

Because annotating work is too expensive for us to annotate all articles published in sampling days, thirty percent of them are only annotated. Each article of Mainichi Newspaper belongs into 16 categories like front page articles, international stories, economical stories, political stories, editorial columns, and human interest stories. Because these categories may influence to NE distribution, it is important to keep the proportion of categories in the sampled texts to the proportion in the whole newspaper, in order to investigate NE distribution over the whole newspaper. Therefore, thirty percent articles of each category published at sampling days are randomly selected and annotated in accordance with the IREX regulation.

2.3 Analysis of Annotated Samples

Table 2 shows the statistics of our annotated corpus. The leftmost column of Table 2 (whose pub-

[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.

Table 2: Statistics of sampling texts

Published date	1995	1996		1997		1998		2005	
	Jan. 1∼10	Jun. 5	Oct. 15	Jun. 10	Oct. 7	Jun. 8	Oct. 21	Jun. 23	Oct. 12
# of articles	1174	120	133	106	117	96	126	90	99
# of characters	407881	60790	53625	46653	50362	51006	67744	49038	44344
# of NE types	6979	1446	1656	1276	1350	1190	1226	1230	1113
# of NE tokens	18677	2519	2652	2145	2403	2126	2052	1902	2007
# of NE types / # of characters	0.0171	0.0238	0.0309	0.0274	0.0268	0.0233	0.0181	0.0251	0.0251
# of NE tokens / # of characters	0.0458	0.0414	0.0495	0.0460	0.0477	0.0417	0.0303	0.0388	0.0453

Table 3: Overlap of NE types between texts published in different years

Published date of annotated corpus A	Published year of unannotated corpus U						
	1993	1994	1995	1996	1997	1998	1999
Jan. 1∼10 (1995)	73.2%	78.6%	—	74.4%	65.0%	64.4%	63.3%
Jun. 6, Oct. 15 (1996)	67.2%	71.7%	72.2%	—	77.3%	76.0%	75.1%
Jun. 6, Oct. 7 (1997)	71.2%	73.4%	74.4%	78.6%	—	80.8%	78.6%
Jun. 8, Oct. 21 (1998)	72.5%	74.6%	76.2%	79.7%	82.7%	—	84.0%
Jun. 23, Oct. 12 (2005)	62.3%	64.1%	66.8%	68.7%	71.2%	72.9%	73.8%

lish date is January 1st to 10th in 1995) is corresponding to IREX corpus, and other columns are corresponding to articles annotated by ourselves. Table 2 illustrates that the normalized number of NE types and the normalized number of NE tokens are almost steady over time. Figure 1 shows the distributions of NE categories for sampling texts and that there is no significant difference between them.

We also investigate the relation of the time gap between texts and NE types which appear in these texts. The overlap ratio of NE types between the annotated corpus A published in the year Y_A and the annotated corpus B published in the year Y_B was defined in (Mota and Grishman, 2008) as follows

$$type\_overlap(A, B) = \frac{|T_A \cap T_B|}{|T_A| + |T_B| - |T_A \cap T_B|},$$

where T_A and T_B are lists of NE types which appear in A and B respectively. However, it is impossible to compute reliable $type\_overlap$ in our research because enough annotated texts are unavailable. As an alternative of $type\_overlap$, the overlap ratio of NE types between the annotated corpus A and the unannotated corpus U published in the year Y_U is defined as follows

$$string\_overlap(A, U) = \frac{\sum_{s \in T_A} \delta(s, U)}{|T_A|},$$

where $\delta(s, U)$ is the binary function to indicate whether the string s occurs in the string U or not. Table 3 shows $string\_ratio$ values of annotated texts. It shows that $70 \sim 80\%$ of T_A appear in the preceding year of Y_A, and that $70 \sim 80\%$ of T_A appear in the succeeding year of Y_A.

Figure 2 shows the relation between the time gap $Y_U - Y_A$ and $string\_ratio(A, U)$. Suppose that all NEs are independent and equivalent on their occurrence probability and that $string\_ratio(A, U)$ is equal to 0.8 when the time gap $Y_U - Y_A$ is equal to one. When the time gap $Y_{U'} - Y_A$ is equal to two years, although this assumption leads that $string\_ratio(A, U')$ will be equal to 0.64, $string\_ratio(A, U')$ in Figure 2 is greater than 0.7. This suggests that NEs are not equivalent on their occurrence probability. And more, Table 4 shows that the longer time span of the annotated text increases the number of NE types. These facts lead that some NEs are short-lived and superseded by other new NEs.

3 Robust Extraction of Changing Named Entities

It is infeasible to prepare a large annotated corpus which covers all increasing NEs. A semi-supervised learning approach which combines a small annotated corpus and a large unannotated corpus for training is promising to cope this problem. (Miller et al., 2004) proposed the method using classes which are assigned to words based on the class language model built from a large unannotated corpus. (Ando and Zhang, 2005) pro-

Table 4: Number of NE types and Time Span of Annotated Text

	1995	1995∼1996	1995∼1997	1995∼1998	1995∼2005
ARTIFACT	541 (1.00)	743 (1.37)	862 (1.59)	1025 (1.89)	1169 (2.16)
DATE	950 (1.00)	1147 (1.21)	1326 (1.40)	1461 (1.54)	1583 (1.67)
LOCATION	1403 (1.00)	1914 (1.36)	2214 (1.58)	2495 (1.78)	2692 (1.92)
MONEY	301 (1.00)	492 (1.63)	570 (1.89)	656 (2.18)	749 (2.49)
ORGANIZATION	1487 (1.00)	1890 (1.27)	2280 (1.53)	2566 (1.73)	2893 (1.95)
PERCENT	249 (1.00)	319 (1.28)	353 (1.42)	401 (1.61)	443 (1.78)
PERSON	1842 (1.00)	2540 (1.38)	3175 (1.72)	3683 (2.00)	4243 (2.30)
TIME	206 (1.00)	257 (1.25)	291 (1.41)	314 (1.52)	332 (1.61)
Total	6979 (1.00)	9302 (1.33)	11071 (1.59)	12601 (1.81)	14104 (2.02)

(Values in brackets are rates of increase comparing to 1995.)

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	⟨1, 0, 0, 0, 0, 0⟩	O
の (no)	gen	Particle	の (no)	gen	Particle	⟨0, 1, 0, 0, 0, 0⟩	O
石狩 (Ishikari)	(*Ishikari*)	Noun–Proper	**関東 (Kantou)**	(*Kantou*)	Noun–Proper	⟨1, 0, 0, 0, 0, 0⟩	B-LOCATION
平野 (heiya)	(*plain*)	Noun–Generic	平野 (heiya)	(*plain*)	Noun–Generic	⟨1, 0, 0, 0, 0, 0⟩	I-LOCATION
の (no)	gen	Particle	の (no)	gen	Particle	⟨0, 1, 0, 0, 0, 0⟩	O
天気 (tenki)	(*weather*)	Noun–Generic	天気 (tenki)	(*weather*)	Noun–Generic	⟨1, 0, 0, 0, 0, 0⟩	O
は (ha)	top	Particle	は (ha)	top	Particle	⟨0, 1, 0, 0, 0, 0⟩	O
晴れ (hare)	(*fine*)	Noun–Generic	晴れ (hare)	(*fine*)	Noun–Generic	⟨1, 1, 0, 0, 0, 0⟩	O

Figure 3: Example of Training Instance for Proposed Method

posed the method using thousands of automatically generated auxiliary classification problems on an unannotated corpus. (**?**) proposed the semi-supervised discriminative model whose potential function can treat both an annotated corpus and an unannotated corpus.

In this paper, the method proposed by (Tsuchiya et al., 2008) is employed, because its implementation is quite easy. It 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 Conditional Random Fields(CRF)[2](Lafferty et al., 2001) using both features of original morphemes and features of similar morphemes.

This section gives the detail of this method.

3.1 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 morphemes. For representing proper chunks, we employ IOB2 representation, one of representations 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 categories, 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 a 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 analyze the influence of

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

the text time frame.

3.2 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,

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

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

3.3 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 Eqn. 1. The character type feature $CF(m_i)$ is a set of six binary flags to indicate that the surface string of m_i contains a Chinese character, a *hiragana* character, a *katakana* character, an English alphabet, a number and an other character respectively.

$\longrightarrow$ Chunking Direction $\longrightarrow$

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 4: Chunking Direction

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 referred as shown in Figure 4.

Figure 3 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.

3.4 Experimental Result

Figure 5 compares performances of the proposed method and the baseline method over the test texts which were published in 1996, 1997, 1998 and 2005. The proposed method combines a small annotated corpus and a large unannotated corpus as already described. This experiment refers IREX corpus as a small annotated corpus, and refers Mainichi Newspaper articles published from 1993 to the preceding year of the test text published year as a large unannotated corpus. For example, when the test text was published in 1998, Mainichi Newspaper articles published from 1993 to 1997 are used. The baseline method is trained from IREX corpus with CRF. But, it uses only MF and CF as features, and does not use SF. Figure 5 illustrates two points: (1) the proposed method outperforms the baseline method consistently, (2) the baseline method is fragile to changing of test texts.

Figure 6 shows the relation between the performance of the proposed method and the size of unannotated corpus against the test corpus published in 2005. It reveals that that increasing unannotated corpus size improves the performance of the proposed method.

4 Conclusion

In this paper, we explored the change of NE distribution over time and its influence on the per-

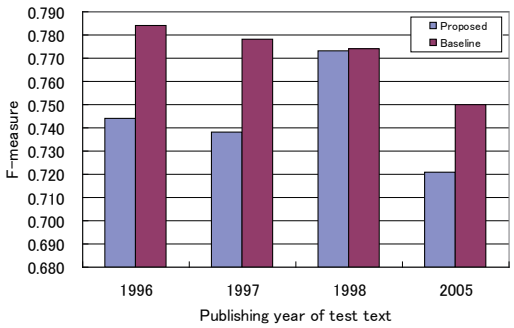

Figure 5: Comparison between proposed method and baseline method

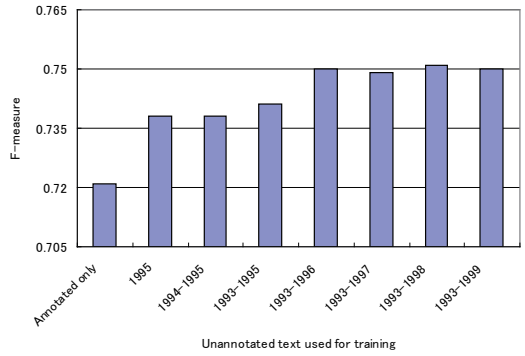

Figure 6: Relation of performance and unannotated corpus size

formance of the NE tagger. First, we annotated Mainichi Newspaper articles published in 1996, 1997, 1998 and 2005, and analyzed NEs which appear in these texts and IREX corpus which consists of Mainichi Newspaper articles published in 1995. This analysis illustrated that the number of NE types and the number of NE tokens are almost steady over time, and that $70 \sim 80\%$ of NE types seen in a certain year occur in the texts published either in its succeeding year or in its preceding year. The experiment against these texts showed that our proposing semi-supervised NE tagger works robustly although the traditional supervised NE tagger is fragile against the change of NE types. Based on the results described in this paper, we will investigate the relation between the performance of NE tagger and the similarity of its training corpus and its test corpus.

References

Rie Kubota Ando and Tong Zhang. 2005. A high-performance semi-supervised learning method for text chunking. In *Proc. of ACL '05*, pages 1–9, June.

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

Ralph Grishman and Beth Sundheim. 1996. Message understanding conference-6: a brief history. In *Proc. of the 16th COLING*, pages 466–471.

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.

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.

Scott Miller, Jethran Guinness, and Alex Zamanian. 2004. Name tagging with word clusters and discriminative training. In *Proc. of HLT-NAACL 2004*, pages 337–342, May.

Cristina Mota and Ralph Grishman. 2008. Is this NE tagger getting old? In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, May.

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.

Masatoshi Tsuchiya, Shinya Hida, and Seiichi Nakagawa. 2008. Robust extraction of named entity including unfamiliar word. In *Proceedings of ACL-08: HLT, Short Papers*, pages 125–128, Columbus, Ohio, June. Association for Computational Linguistics.

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).