Difficulty Indices for the Named Entity Task in Japanese

Chikashi NOBATA
Kansai Advanced Research Center, Communications Research Laboratory
588-2 Iwaoka, Iwaoka-cho, Nishi-ku, Kobe, Hyogo 651-2492, JAPAN
nova@crl.go.jp

Satoshi SEKINE
Computer Science Department, New York University
715 Broadway, 7th floor, New York, NY 10003, USA
sekine@cs.nyu.edu

Jun'ichi TSUJII
Department of Information Science, University of Tokyo
Hongo 7-3-1, Bunkyo-ku, Tokyo 113-0033, JAPAN
tsujii@is.s.u-tokyo.ac.jp

Abstract

We propose indices to measure the difficulty of the named entity (NE) task by looking at test corpora, based on expressions inside and outside the NEs. These indices are intended to estimate the difficulty of each task without actually using an NE system and to be unbiased towards a specific system. The values of the indices are compared with the systems' performance in Japanese documents. We also discuss the difference between NE classes with the indices and show useful clues which will make it easier to recognize NEs.

1 Introduction

The named entity (NE) task is one of the subtasks of information extraction (IE), which were defined in the 6th Message Understanding Conference (MUC-6)(1995) to improve modularity and portability of IE systems. The main objective of the NE task is to identify some specific noun phrases in the target documents. The evaluations of the systems for the NE tasks have been performed in MUCs, there are 7 entity classes defined to identify, such as ORGANIZATION, PERSON, LOCATION, DATE, TIME, MONEY and PERCENT. The type of the articles is restricted for the final goal, information extraction. The target events are executive succession in MUC-6, and space vehicle or rocket launching in MUC-7 (1998), while the classes of NE were the same in both conferences.

IREX (1999) was the workshop for the evaluation of information retrieval (IR) and NE task held in Japan. The documents used in these tasks are Japanese newspaper articles. The NE task in IREX is basically based on that in MUCs, and ARTIFACT class is newly defined, which contains the name of products, laws, prizes.

In the IREX NE tasks, we have a dry run and a formal run evaluation. There are two tasks in the formal run. One task is called in this paper a "general" task. In the general task, the type of documents is not specified, therefore NE systems need to deal with all the types of newspaper articles, similar to that in the dry run. We call the other task a "restricted" task. In the restricted task, articles are restricted to the ones which describe arrest events by the police.

The MUCs and the IREX workshop have evaluated the performances of the IE systems, and the tendency of system performances show partially the difficulty of the IE tasks. However, the evaluation of the systems do not clearly show in detail why the task is difficult or what kind of information should utilized in the IE task. The investigation into the test corpora are necessary to evaluate the

Table 1: Numerical comparison between tasks

	Dry run	General	Restricted
Articles	36	72	20
Words	11173	21321	4892
Characters	20712	39205	8990

performance of the systems more precisely.

There is some research that estimates the difficulty of the IE tasks by looking at the test corpora. Bagga et. al (1997) created a semantic network from test corpora in MUCs, and set the level of the events based on the created network to analyze the performances of the IE systems. In the NE tasks, Palmer et. al (1997) investigated the test corpora at Multi-lingual Entity task (1996) and estimated the lower bound of the performance of NE task in six different languages.

In this paper, we propose some indices to evaluate the difficulty of the named entity task by looking at a test corpus, based on expressions inside and outside the entities. The basic notion to evaluate the difficulty is the variety of expression. To identify named entities properly, a certain amount of knowledge is necessary. We make the hypothesis that the greater the variety of expression is, the more effort to create knowledge would be required. The validity of these indices is evaluated by the correlation with the NE systems which are the participants in the IREX workshop.

First, we introduce the NE tasks performed in the IREX workshop in Section 2, then define indices to estimate the difficulty of the NE tasks. Section 3 describes the index based on the frequency of tokens. Entities, words and characters are checked as candidates of tokens. In Section 4, we have another point of view to describe the method which extracts useful tokens related with each NE class. A set of these tokens show the easiness of recognizing entities in each class, that is, the inverted index for the difficulty. In Section 5, the indices based on expressions around entities are described and evaluated.

Table 2: Average F-measures in the IREX NE tasks

Class	Dry run	General	Restricted
Org.	0.56	0.57	0.55
Person	0.71	0.68	0.69
Loc.	0.66	0.70	0.68
Art.	0.19	0.26	0.58
Date	0.84	0.86	0.89
Time	0.69	0.83	0.90
Money	0.91	0.86	0.91
Percent	1.00	0.86	—
All	0.66	0.70	0.75

2 IREX NE tasks

In Table 1, the basic statistics in the IREX NE tasks are shown, such as the number of articles, words and characters of the test corpora, and the average F-measures[1] in the IREX NE tasks are shown in Table 2. In the general task and restricted task of the formal run, we compared the index with the average F-measures of the 15 NE systems, which are the participants of the IREX workshop. In the dry run, we used the F-measure which is the result of University of Tokyo's NE system because we don't have the results of all the participants in the dry run.

In Table 2, we can see that entities in the first four NE classes (ORGANIZATION, PERSON, LOCATION and ARTIFACT) are more difficult to be identified than those in the second four NE classes (DATE, TIME, MONEY and PERCENT). We therefore discuss the two groups separately if it is necessary. We call the first four NE classes the "ENAMEX" group, and the second four NE classes the "TIMEX-NUMEX" group in the following part of this paper, based on the denotation of the MUCs.

[1] F-measure is defined and used in MUCs, which is a measurement combining *Recall* and *Precision*. *Recall* is the percentage of the correct answers among the answers in the key provided by human. *Precision* is the percentage of the correct answers among the answers proposed by the system. F-measure is defined using *Recall*(R) and *Precision*(P) as $2PR/(P+R)$.

Table 3: Number of different entities in NE classes

Class	Dry run	General	Restricted
Org.	131	187	48
Person	113	217	71
Loc.	89	191	78
Art.	31	39	9
Date	71	126	49
Time	16	32	15
Money	28	13	7
Percent	6	16	-
All	482(485)	818(821)	277

3 Frequency of tokens

3.1 Frequency of entities

At first, we will define an index that shows the difficulty of NE tasks using the frequency of tokens and also the variety of them. The assumption here is that a variety of expressions make it difficult to define entities of NE class or create knowledge base for the class.

Table 3 shows the number of different entities in NE classes for the tasks in the IREX workshop. The number of PERCENT entities in the restricted task remains blank, because the test corpus doesn't have any PERCENT entities. The number of different entities in "All" is a little fewer than the sum of the number of different entities in each class. The reason is that three entities are categorized into more than one class in the dry run and in the general task of the formal run, respectively. Although these entities also make the NE task difficult, we don't use them as an index because the number of entities is small.

The number of different entities alone is influenced by the corpus size and not suitable for the index that should uniformly show the difficulty for different NE tasks, therefore it should be normalized. The first index for the difficulty of NE we introduce is the normalized number of different entities. The index, *Frequency of Entities(FE)*, is defined in Equation 1:

$$FE = \frac{D_E}{N_E} \quad (1)$$

D_E denotes the number of different entities in a class, and N_E is the frequency of entities in a class.

Table 4: Values of *Frequency of Entities(FE)*

Class	Dry run	General	Restricted
Org.	0.61 (=131/214)	0.48 (=187/389)	0.65 (=48/ 74)
Person	0.67 (=113/169)	0.61 (=217/355)	0.73 (=71/ 97)
Loc.	0.46 (=89/192)	0.46 (=190/416)	0.74 (=78/106)
Art.	0.71 (=30/ 42)	0.80 (=39/ 49)	0.69 (=9/ 13)
Date	0.33 (=36/110)	0.18 (=51/277)	0.24 (=17/ 72)
Time	0.46 (=11/ 24)	0.27 (=16/ 59)	0.53 (=10/ 19)
Money	0.09 (= 3/ 33)	0.13 (= 2/ 15)	0.12 (= 1/ 8)
Percent	0.50 (= 3/ 6)	0.29 (= 6/ 21)	— —
All	0.53 (=415/790)	0.45 (=706/1581)	0.60 (=235/389)

When *FE* is calculated, numeral figures are considered to be the same, and replaced by the meta character '#'. In the NE classes, temporal(DATE, TIME) or numerical(MONEY, PERCENT) ones can be identified more easily than the others which will result from the uniformity of numerals. When each distinct numeral is recognized as the same character, the number of different entities is reduced in temporal or numerical NE classes and match our intuition better. The values of *FE* for each class in different tasks are shown in Table 4.

3.2 Frequency of words / characters

We can define another indices, *Frequency of Words(FW)* and *Frequency of Character(FC)*, since entities in the definition of *FE* can be replaced with words or characters to estimate the difficulty of the NE task. The expectation here is that words or characters have higher frequency than the entire entities, therefore the index based on words or characters will be more robust than entities with regard to the corpus size.

Word segmentation for calculating *FW* is performed using a morphological analyzer JUMAN (Matsumoto et al., 1997). When the boundary of an entity is different from the result of segmentation, the word on the edge

Table 5: Values of *Frequency of Characters(FC)*

Class	Dry run	General	Restricted
Org.	0.29 (=258/883)	0.20 (=365/1792)	0.38 (=139/365)
Person	0.39 (=222/575)	0.26 (=319/1228)	0.48 (=148/311)
Loc.	0.30 (=186/618)	0.19 (=284/1491)	0.34 (=155/462)
Art.	0.53 (=131/245)	0.50 (=175/ 347)	0.58 (= 34/ 59)
Date	0.16 (= 44/282)	0.07 (= 54/ 737)	0.07 (= 15/226)
Time	0.18 (= 12/ 66)	0.09 (= 16/ 182)	0.14 (= 10/ 71)
Money	0.06 (= 4 / 72)	0.09 (= 3/ 34)	0.12 (= 2/ 16)
Percent	0.38 (= 5 / 13)	0.10 (= 7/ 58)	— —
All	0.20 (=555/2754)	0.12 (=717/5869)	0.24 (=355/1510)

of the entity is split again. Each distinct numeral is also recognized as the same character in *FW* and *FC*. The values of *FW* have basically the same tendency as those of *FC*, therefore we show only the values of *FC* in Table 5. *FC* shows the difference between classes more clearly than *FE*. The values of *FC* have especially decreased in the NE classes of the TIMEX-NUMEX group.

3.3 Validity of indices

We estimated the correlation coefficients between the defined indices and the F-measure of the NE systems over eight NE classes, as shown in Table 6. The greater value of the defined three indices indicates that identifying entities is more difficult, therefore the negative correlation with F-measures is desirable for these indices.

We can see that all of the indices have quite high correlation with F-measure. While *FW*, *FC* have the lower correlation than *FE* in the dry run, they have higher correlation in the two tasks of the formal run. Considering that F-measure of the dry run is just one system's result, the evaluated performance about the tasks of the formal run is more reliable. Characters are therefore better tokens than entities and words to estimate the difficulty of Japanese NE tasks.

Table 6: Correlation coefficients between indices and F-measure

Task	*FE*	*FW*	*FC*
Dry run	-0.66	-0.63	-0.61
General	-0.91	-0.92	-0.97
Restricted	-0.80	-0.87	-0.89

4 Token index

In this section, we focus on finding useful clues to identify entities. If a class has some specific tokens, identifying entities in each class will be easier than in other classes. In the indices we defined earlier, only the frequency of each token in one class is considered. In the previous assumption, if the tokens often appear in one class, they are regarded as good indicators of the class, and make the identifying the entities of the class more easily. However, if such tokens appear broadly in the whole document, our assumption is not correct. The frequency in the whole corpus should be considered in addition to the frequency in one class.

4.1 Character index

We define *Character Index for each character (CI_c)* to indicate how peculiar the character is to the class. We chose here characters as tokens for the new index, since they are shown to be good indices in the previous result. The definition itself can be applied for the other type of tokens. CI_c for a character c in a class L is defined in Equation 2:

$$CI_c = \frac{n_L(c)}{N_C^L} \frac{n_L(c)}{n(c)} \quad (2)$$

Where, $n_L(c)$ is the frequency of c in the class L, $n_{(c)}$ is the frequency of c in the whole corpus. N_C^L is the number of characters in the class L. The first item in the right side $\frac{n_L(c)}{N_C^L}$ denotes the percentage of the appearance of the character c in the class L, and the second term $\frac{n_L(c)}{n(c)}$ denotes how the appearance of the character c is different between in the class L and in the whole corpus. If c is shown mostly in the class L, $n_L(c)$ is close to $n(c)$.

Table 7: Values of *Character Index(CI)*

Class	Dry run	General	Restricted
Org.	0.34	0.31	0.45
Person	0.51	0.45	0.59
Location	0.38	0.40	0.56
Artifact	0.21	0.15	0.27
Date	0.39	0.48	0.60
Time	0.36	0.40	0.47
Money	0.47	0.51	0.51
Percent	0.33	0.27	—
All	0.57	0.58	0.71

The summation of CI_c of characters in the class can be used for another index to show the difficulty of NE tasks. The index, *Character Index(CI)*, reverses the measure to show the difficulty. The greater the value is, the easier identifying expressions in the class is. Therefore positive correlation with F-measure is desirable.

Table 7 shows the result of CI about every class. When all the characters in the class L are shown only in L, CI becomes the maximum value 1. On the other hand, when the characters in the class L uniformly appear in the entire corpus, CI becomes close to the minimum value $\frac{N_C^L}{N}$, where N is the number of characters in the corpus.

4.2 Validity of CI

The second column of Table 8 shows the correlation coefficient between CI and performances of systems. The results show that CI is not a so suitable index to illustrate the difficulty of tasks as the indices we defined earlier. One possible reason is that CI is the summation of the CI_c values of all the characters in one class. Characters that have lower value of CI_c are not likely to make it easier to identify entities in the NE class, so these characters should be removed when CI is calculated. Some threshold is required to select the characters related with the class. To decide the suitable threshold for CI, we observed the change of the correlation between CI and F-measure in the different IREX tasks when the threshold for CI_c is changed.

Figure 1 shows the relationship between the correlation coefficient and the threshold for CI_c. The axis for the threshold for CI_c has a logarithmic scale. We can see that the correlation coefficient between the indices and F-measures once increased and then decreased in every task when the threshold is gradually lowered. The best correlation coefficients with F-measure from the graph are shown with the thresholds in the column "peak" of Table 8. These values are comparable with the results of the previous indices, although the thresholds are simply obtained by comparisons with the performance of NE systems.

Not all NE classes require such thresholds for CI_c. To show the distribution of the CI more clearly, we investigated the CI_c values for characters in the two groups separately, the ENAMEX group and the TIMEX-NUMEX group. Figure 2 shows that the value of CI_c for each character when characters are listed in the descending order. We can see that some characters are strongly related with NE classes in the TIMEX-NUMEX group, while such characters do not appear in the ENAMEX group.

This result implies that a lot of characters are related with the NE classes in the ENAMEX group, but the contribution of each character is small because of the low frequency. On the other hand, a small number of characters are strongly related with NE classes in the TIMEX-NUMEX group, and the combination of these characters and numerals covers most of the entities in the classes. Consequently, the threshold for CI_c is necessary to the NE classes in the TIMEX-NUMEX group, because it is expected that the number of characters used by the NE systems are restricted.

Figure 3 shows the relationship between the correlation coefficient and the threshold for CI_c about the NE classes in the ENAMEX group, and Figure 4 is about those in the TIMEX-NUMEX group. We can see that the two groups have the completely different tendency about the correlation with F-measures, and that the above observation is true.

Table 8: Correlation coefficients between *CI*s and F-measure

Task	*CI*	peak(threshold)
Dry run	0.62	0.86($CI_c = 0.005$)
General	0.75	0.88($CI_c = 0.004$)
Restricted	0.49	0.96($CI_c = 0.009$)

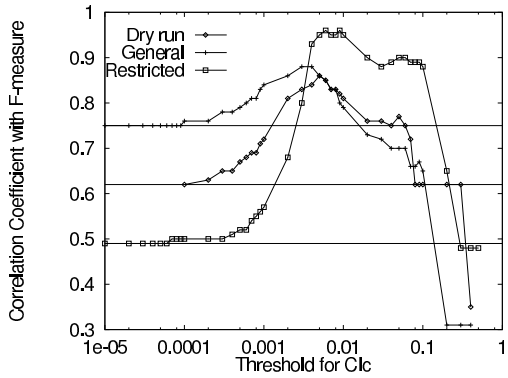

Figure 1: Change of correlation between *CI* and F-measure

4.3 Characters selected by *CI*

Table 9 shows the characters which have high CI_c value in the NE classes of the TIMEX-NUMEX group. The target task is the general one of the formal run. The character "#" indicates the entire numerals.

We can see there are a few characters which have comparatively high CI_c value in the NE classes of the TIMEX-NUMEX group. Especially in the MONEY and PERCENT class, only one character contributes most of the *CI* value of the class. It is natural that the character "円" (the unit of Japanese currency) is dominant in the MONEY class, and that the character "%" corresponds to the PERCENT class. Other characters are also intuitively peculiar to the corresponding class. Though the frequency of numerals in the corpus is quite large, the above clues are necessary to connect numerals into each class, because numerals can often appear in any classes in the TIMEX-NUMEX group.

5 Indices based on context words

In this section, we consider expressions around entities. Only the observation in the entities is not sufficient to describe the difficulty of NE tasks. Even when the class itself has various entities, they can be easily identified if the surrounding words are restricted, i.e. prefixes and suffixes which are not a part of entity. Therefore, it will be reasonable to define the index based on expression around entities.

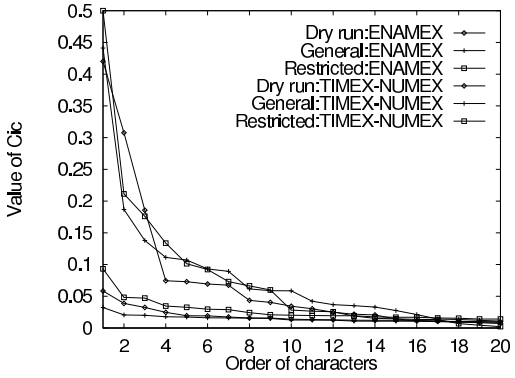

Figure 2: Values of CI_c for each character

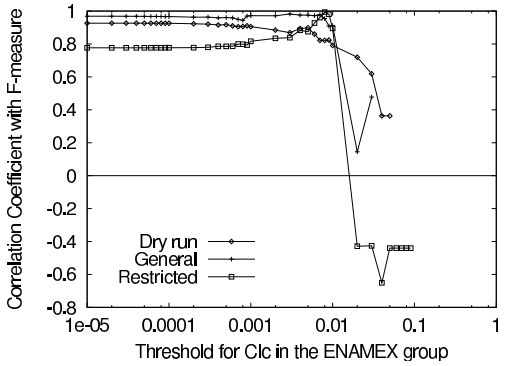

Figure 3: Correlation coefficients between *CI*s and F-measure in the NE classes of the ENAMEX group

5.1 Context word index

The index based on words around entities, *Context Word Index (CWI)*, is defined in Equation 3, which is similar to that of *CI*. The difference is that m is added at the denominator of the second term. The m denotes the range of words that are regarded as the context words.

$$CWI = \sum_w \frac{n_L(w)}{N_{CW}^L} \frac{n_L(w)}{m \cdot n(w)} \qquad (3)$$

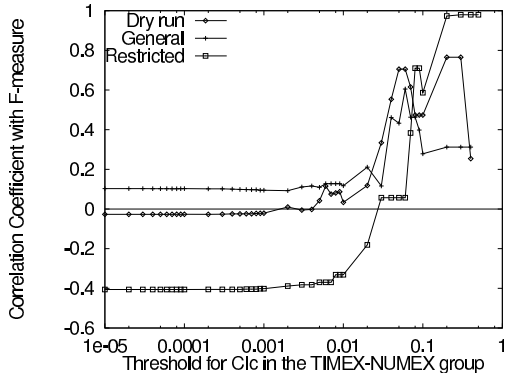

Figure 4: Correlation coefficients between *CI*s and F-measure in the NE classes of the TIMEX-NUMEX group

Table 9: High CI_c characters in the TIMEX-NUMEX group

Class	CI_c	$n_L(c)$	Character
Date	0.1113	277	＃ (numerals)
	0.1071	143	日 (day)
	0.0931	75	月 (month)
	0.0893	98	年 (year)
	0.0421	31	昨 (last)
Time	0.1868	34	午 (noon)
	0.0586	32	時 (time)
	0.0368	23	後 (after)
	0.0352	8	夜 (night)
	0.0330	6	夕 (evening)
Money	0.4412	15	円 (currency unit)
	0.0588	2	銭 (currency unit)
	0.0091	17	＃
Percent	0.1379	8	%
	0.0616	5	倍 (times)
	0.0276	4	半 (half)
	0.0212	4	割 (tenth of)
	0.0134	27	＃

Table 10 shows the values of *CWI*, when the range of words is set to 1 word. The highest value among NE classes for each task is in bold type.

5.2 Words selected by *CWI*

Table 11 shows the values of the correlation coefficient between *CWI*s and F-measures. There is not so strong correlation as shown in the previous indices. We have to say this index does not provide enough information to show how the expressions around entities should be utilized. However, we can see some context words which seem useful for the NE task. We show some examples in Table 12, 13, and Table 14. Figures beside each word in tables means the value of CWI_w and the frequency around the class($n_L(w)$).

Table 11: Correlation coefficients between *CWI*s and F-measure

Task	*CWIpre:* Preceding words			
	$m=1$	$m=2$	$m=3$	$m=4$
Dry run	-0.07	-0.34	-0.53	-0.49
General	0.66	-0.01	-0.01	-0.04
Restricted	0.67	0.39	0.46	0.20
Task	*CWIfol:* Following words			
	$m=1$	$m=2$	$m=3$	$m=4$
Dry run	-0.01	-0.24	-0.07	-0.09
General	0.14	0.29	0.00	0.02
Restricted	0.06	0.46	0.36	0.10

Table 12: High *CWIpre* words of TIME entities in the general task

$CWIpre_w$	$n_L(w)$	Word
0.1805	35	＃日 (the #th day)
0.0920	8	同日 (the same day)
0.0086	1	＃年＃月＃日 (the #th day of the #th month in the #th year)
0.0067	5	同 (the same)
0.0057	1	昨年＃月＃日 (the #th day of the #th month last year)

In Table 10, TIME class has relatively high *CWIpre* values in all tasks than any other class. As shown in Table 12, the reason is that TIME expressions usually follow DATE entities. PERSON class has also relatively high *CWIfol* values than any other class in all tasks, and the example words in Table 13 show that common suffixes and titles often follow person names. These words should be helpful for identifying the name of people in texts. ARTIFACT, MONEY and TIME classes have high *CWI* value in the restricted task of the formal run, and one word contributes most of the part as shown in Table 14. The reason would be that the articles in the test corpus are restricted to the arrest events, and that the usage of words are rather fixed than other types of articles.

6 Conclusion

We defined indices to show the difficulty of the NE tasks and evaluate the validity by compar-

Table 10: Values of *Context Word Index(CWI)*

Class	Dry run		General		Restricted	
	Preceding	Following	Preceding	Following	Preceding	Following
Org.	0.23	0.30	0.16	0.22	0.15	0.20
Person	0.18	**0.47**	0.17	**0.53**	0.16	**0.58**
Loc.	0.22	0.35	0.20	0.21	0.29	0.27
Art.	0.09	0.10	0.05	0.18	0.02	0.56
Date	0.13	0.25	0.15	0.22	0.14	0.33
Time	**0.29**	0.07	**0.29**	0.20	**0.44**	0.40
Money	0.14	0.20	0.25	0.28	0.37	0.45
Percent	0.07	0.04	0.12	0.27	—	—
All	0.32	0.41	0.30	0.36	0.34	0.43

Table 13: High *CWIfol* words of PERSON entities

Task	$CWIfol_w$	$n_L(w)$	Word
Restricted	0.0471	30	容疑者 (suspect)
Restricted	0.0407	33	（
Dry run	0.0406	28	氏 (Mr. or Ms.)
General	0.0370	54	さん (Mr. or Ms.)
Restricted	0.0340	13	さん
Dry run	0.0228	17	さん
General	0.0214	29	氏
General	0.0170	28	」
General	0.0164	25	被告 (defendant)

Table 14: High *CWIfol* words in the restricted task

Class	*CWIfol*	$CWIfol_w$	Word
Artifact	0.5640	0.5470	違反 (violation)
Time	0.4015	0.3876	ごろ (about, around)
Money	0.4460	0.3750	相当 (equal to)

ing them with the performance of the NE systems which participated to the IREX workshop. The correlation coefficients between the defined indices and the system performances are quite high, the best result is 0.97. We also proposed the methods to select useful characters or words to make it easier to recognize named entities, and showed examples.

We would like to improve the index based on the contextual expressions around entities further, and our ultimate goal is to automatically acquire useful information in order to recognize the named entities in the given domain.

Acknowledgements

We would like to thank our colleagues at University of Tokyo. In particular, Dr. Nigel Collier gave us help and suggestions.

References

ARPA. 1996. Message Understanding Evaluation and Conference.

Amit Bagga and Alan W. Biremann. 1997. Analyzing the Complexity of a Domain With Respect To An Information Extraction Task. In *The Tenth International Conference on Research on Computational Linguistics(ROCLING X)*, pages 175–184, August.

DARPA. 1995. *Proceedings of the Sixth Message Understanding Conference(MUC-6)*, Columbia, MD, USA, November. Morgan Kaufmann.

DARPA. 1998. *Proceedings of the Seventh Message Understanding Conference(MUC-7)*, Fairfax, VA, USA, May.

IREX Committee. 1999. *Proceedings of the IREX Workshop*, KKR Hotel, Tokyo, Japan, September.

Yuji Matsumoto, Sadao Kurohashi, Osamu Yamaji, Yuu Taeki, and Makoto Nagao, 1997. *Japanese morphological analyzing System: JUMAN*. Kyoto University, Nara Institute of Science and Technology.

David D. Palmer and David S. Day. 1997. A Statistical Profile of the Named Entity Task. In *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP'97)*, pages 190–193.