Constructing a Japanese Basic Named Entity Corpus of Various Genres

Tomoya Iwakura[1]**, Ryuichi Tachibana** [2]**, and Kanako Komiya** [3]

[1] Fujitsu Laboratories Ltd. [2] Commerce Link Inc. [3] Ibaraki University

Abstract

This paper introduces a Japanese Named Entity (NE) corpus of various genres. We annotated 136 documents in the Balanced Corpus of Contemporary Written Japanese (BCCWJ) with the eight types of NE tags defined by Information Retrieval and Extraction Exercise. The NE corpus consists of six types of genres of documents such as blogs, magazines, white papers, and so on, and the corpus contains 2,464 NE tags in total. The corpus can be reproduced with BCCWJ corpus and the tagging information obtained from https://sites.google.com/site/projectnextnlpne/en/ .

1 Introduction

Named Entity (NE) recognition is a process by which the names of particular classes and numeric expressions are recognized in text. NEs include person names, locations, organizations, dates, times, and so on. NE recognition is one of the basic technologies used in text processing, including Information Extraction (IE), Question Answering (QA), and Information Retrieval (IR).

For the development of NE recognizers in early stage, newspaper articles have been mainly used. For example, the following data sets consist of newspaper articles: eight types of basic Japanese NE recognition data sets for Information Retrieval and Extraction Exercise (IREX) (IREX Committee, 1999), the CoNLL'03 shared task (Tjong Kim Sang and De Meulder, 2003), an English fine-grained NE type that includes 64 classes (Weischedel and Brunstein, 2005), and Sekine's extended NE hierarchy that includes about 200 classes of NEs (Sekine et al., 2002).

As for Sekine's extended NE hierarchy, NE corpus have been created on various genres documents such as blogs, white papers and so on, in BCCWJ (Maekawa et al., 2010).[1] However, compared with the corpus for Sekine's extended NE hierarchy, which covers several genres, corpus for Japanese basic NEs have been created for fewer genres of documents such as newspaper articles of IREX and leading sentences of Web pages (Hangyo et al., 2012).

This paper introduces a Japanese Named Entity (NE) corpus of various genres called BCCWJ Basic NE corpus. BCCWJ Basic NE corpus was created for the sake of expanding genres of documents for Japanese basic NE researches. The corpus includes 136 documents in the Balanced Corpus of Contemporary Written Japanese (BCCWJ) core data annotated with the eight types of NE tags defined by IREX. The corpus contains 2,464 NE tags in total and the genres of the documents are following: Yahoo! Chiebukuro (OC)[2], White Paper (OW), Yahoo! Blog (OY), Books (PB), Magazines (PM) and Newspapers (PN). This corpus includes genres of documents that have not been targeted in existing NE corpus for IREX definition. (IREX Committee, 1999; Hangyo et al., 2012).

2 IREX NE Definition

IREX committee defined the eight NE types: ARTIFACT, LOCATION, ORGANIZATION, PERSON, DATE, MONEY, PERCENT and TIME. Table 1 shows the eight NE types and their examples. In addition to the eight NE types, OPTIONAL, for ambiguous NEs, were defined.

For example, from the following sentence, "Mr. Miyazaki comes from Miyazaki." in English, an

[1] The IREX definition is not a subset of the Sekine's extended NE hierarchy.

[2] Yahoo! Chiebukuro consists documents from a QA site on the Web.

Proceedings of the Sixth Named Entity Workshop, joint with 54th ACL, pages 41–46,
Berlin, Germany, August 12, 2016. ©2016 Association for Computational Linguistics

NE type	Example
ARTIFACT	Nobel Prize
LOCATION	Japan
ORGANIZATION	Foreign Ministry
PERSON	Tom White
DATE	May, 5th
MONEY	100 yen
PERCENT	100%
TIME	10:00 p.m.

Table 1: The eight NE types defined by IREX and the examples.

NE recognizer should extract the first Miyazaki as PERSON and the second one as LOCATION because NE types are decided by context in the IREX definition.

$\langle PER\rangle$ 宮崎 $\langle /PER\rangle$	さん	は
(Miyazaki)	(Mr.)	(postposition)
$\langle LOC\rangle$ 宮崎 $\langle /LOC\rangle$	出身	
(Miyazaki)	(comes from)	

PER and LOC in the above example indicate PERSON and LOCATION.

3 BCCWJ Basic NE corpus

We annotated 136 documents included in BCCWJ core data with IREX-defined NE tags by the following procedure.[3] We choose the same documents of a Japanese morphological analysis corps.[4]

- Initial annotation: Six annotators, the authors and three university students, annotated all the documents with NEs. Each document was annotated by only a member.
- Modification: Four of the annotators checked all the annotated documents again and modified annotation errors. Annotation disagreements are resolved based on discussion of annotators.
- Packaging： We prepared a package only including annotated tags with the positions in each documents. Users having BCCWJ can reproduce the BCCWJ Basic NE corpus with the package.

Table 2 shows the number of documents and NE tags of each genre in BCCWJ Basic NE corpus. For comparing purpose, the statistics of IREX data. The number of documents of BCCWJ Basic NE corpus is more than the sum of the number of the IREX evaluation data: GENERAL data, ARREST DATA. In addition, BCCWJ Basic NE corpus includes documents other than newspapers such as Yahoo! Chiebukuro and White Paper.

Table 3 shows the statistics of BCCWJ Basic NE corpus. Table 4 shows the percentage of each NE in a genre. We see from these statistics that BCCWJ Basic NE corpus has different property compared IREX. For example, we see that Yahoo! Chiebukuro and White Paper include more ARTIFACT than newspapers and Magazine includes more PERSON than the other genres.

4 Example Uses of BCCWJ Basic NE corpus

This section describes some example uses of BCCWJ Basic NE corpus.

4.1 Evaluation of an NE recognizer

We evaluated KNP that extracts the eight types of NEs listed in Table 1 based on the IREX definition. KNP is one of the freely available state of the art NE recognizers. We used Japanese morphological analyzer JUMAN version 7.01 [5] as a morphological analyzer of KNP version 4.12 [6].

Table 5 shows accuracy of KNP on BCCWJ Basic NE corpus. KNP was evaluated with Recall, Precision and F-measure:

- Recall = NUM / (the number of correct NEs)
- Precision = NUM / (the number of words and word chunks recognized as NEs by KNP)

[3] We referred the annotation guideline careted by IREX committee: https://nlp.cs.nyu.edu/irex/NE/df990214.txt . This site is only Japanese.

[4] http://plata.ar.media.kyoto-u.ac.jp/mori/research/NLR/JDC/ClassA-1.list

[5] http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN

[6] http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP

BCCWJ Genre	 the number of documents	 the number of NEs
Yahoo! Chiebukuro (OC)	74	175
White Paper (OW)	8	656
Yahoo! Blog (OY)	34	307
Books (PB)	5	399
Magazines (PM)	2	319
Newspapers (PN)	13	705
Total	136	2,561 (2,464)

IREX Genre	 the number of documents	 the number of NEs
CRL	1174	19,262
DRY	36	832
NET	46	973
AT	23	466
AR	20	397
GE	72	1,667
Total	1,371	23,597 (22,822)

Table 2: The number of NEs in BCCWJ Basic NE corpus and the IREX data set. The documents of NEs for IREX data is the number of news articles. IREX data set consists of the following data created from Mainichi Shimbun news articles: CRL_NE_DATA.idx (CRL), DRYRUN03.idx (DRY), NEtraining981031.idx (NET), ARREST_TRAIN.idx (AT), ARREST01.idx (AR), GENERAL03.idx (GE). The numbers between parentheses in Total columns indicate the number of NEs excluding OPTIONAL.

- F-measure = 2 × Recall × Precision / (Recall + Precision)

where NUM is the number of correct NEs recognized by KNP.

Compared with Newspapers, KNP showed lower performance on Yahoo! Chiebukuro and Yahoo! Blog. One of the reasons seems that KNP was trained with IREX CRL data that consists of news articles. Another reason is Yahoo! Chiebukuro and Yahoo! Blog includes more abbreviations and colloquial expressions than newspapers. Furthermore, KNP also showed lower performance on White Paper even if White Paper documents were written language. One of the reasons seems that White Paper includes more ARTIFACT NEs than Newspapers The accuracy of KNP for ARTIFACT was lower than the other NEs on Newspapers.

From this evaluation, we see that we can evaluate NE recognizers with different perspective by using different genres of documents.

4.2 The Other Expected Use

We also expect that BCCWJ Basic NE corpus contributes to the following research.

- NE recognition for colloquial expressions: Yahoo! Blog contributes to NE recognition researches for colloquial expressions because Yahoo! Blog includes more colloquial expressions than Newspapers and White Paper
- Domain Adaptation: BCCWJ Basic NE corpus includes six genres of documents, therefore, we expect BCCWJ Basic NE corpus is useful for the research of domain adaptation (Daumé III, 2007).
- Revision learning for NE recognition: We also have uploaded not only latest annotation but also older versions of NE annotation results. Therefore, we can use the corpus as an error detection research or revision learning like Japanese morphological analysis (Nakagawa et al., 2002).
- Comparison of annotation performance on different genres of documents: We can use

BCCWJ

Genre	ART	DATE	LOC	MON	OPT	ORG	PERC	PERS	TIME
OC	54	19	57	9	8	19	0	6	3
OW	163	129	140	9	39	128	33	15	0
OY	25	60	52	7	9	61	11	79	3
PB	29	50	87	0	24	26	6	169	8
PM	13	42	32	5	4	17	1	203	2
PN	24	165	188	59	13	118	38	78	22
Total	308	465	557	89	97	369	89	550	37

IREX

Data	ART	DATE	LOC	MON	OPT	ORG	PERC	PERS	TIME
CRL	747	3567	5463	390	585	3676	492	3840	502
DRY	42	110	192	33	42	214	6	169	24
NET	67	137	255	32	47	270	19	138	8
AT	11	69	165	19	7	80	3	94	18
AR	13	72	106	8	8	74	0	97	19
GE	49	277	416	15	86	389	21	355	59
Total	929	4232	6597	497	775	4703	541	4693	630

Table 3: The number of NEs in BCCWJ Basic NE corpus. ART, LOC, MON, OPT, ORG, PERC and PERS indicate ARTIFACT, LOCATION, MONEY, OPTIONAL, ORGANIZATION, PERCENT and PERSON, respectively. The others are same as ones in Table 2.

this corpus for evaluating annotation performance and annotation methods on different genres of documents. One of the examples is described in (Komiya et al., 2016). The paper compared the following two methods to annotate a corpus via non-expert annotators for named entity (NE) recognition task. The first one is an annotation method by revising the results of an existing NE recognizer. The other is an annotation method by hand from the beginning.

5 Conclusion

This paper introduced a Japanese Named Entity (NE) corpus of various genres called BCCWJ Basic NE corpus. We annotated 136 documents in the BCCWJ with the eight types of NE tags defined by IREX. Users having BCCWJ can reproduce use the corpus by using the annotation information of the corpus distributed at a web site. Users having BCCWJ can reproduce use the corpus by using the annotation information of the corpus distributed at the following web site: https://sites.google.com/site/projectnextnlpne/en/ .

Acknowledgments

We appreciate Ai Hirata, Maiko Yamazaki and Masaaki Ichihara for their help of building this corpus.

References

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proc. of ACL'07*.

Masatsugu Hangyo, Daisuke Kawahara, and Sadao Kurohashi. 2012. Building a diverse document leads corpus annotated with semantic relations. pages 535–544.

IREX Committee. 1999. *Proc. of the IREX workshop*.

Kanako Komiya, Masaya Suzuki, Tomoya Iwakura, Minoru Sasaki, and Hiroyuki Shinno. 2016. Comparison of annotating methods for named entity corpora. In *Proc. of LAW-X*.

Kikuo Maekawa, Makoto Yamazaki, Takehiko Maruyama, Masaya Yamaguchi, Hideki Ogura, Wakako Kashino, Toshinobu Ogiso, Hanae Koiso, and Yasuharu Den. 2010. Design, compilation, and preliminary analyses of balanced corpus of contemporary written japanese. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*.

BCCWJ

Genre	ART	DATE	LOC	MON	OPT	ORG	PERC	PERS	TIME
OC	30.86%	10.86%	32.57%	5.14%	4.57%	10.86%	0%	3.43%	1.71%
OW	24.85%	19.66%	21.34%	1.37%	5.95%	19.51%	5.03%	2.29%	0%
OY	8.14%	19.54%	16.94%	2.28%	2.93%	19.87%	3.58%	25.74%	0.98%
PB	7.27%	12.53%	21.80%	0%	6.02%	6.52%	1.50%	42.35%	2.01%
PM	4.08%	13.17%	10.03%	1.57%	1.25%	5.33%	0.31%	63.63%	0.63%
PN	3.40%	23.40%	26.68%	8.37%	1.84%	16.74%	5.39%	11.06%	3.12%

IREX

Genre	ART	DATE	LOC	MON	OPT	ORG	PERC	PERS	TIME
CRL	3.88%	18.52%	28.36%	2.02%	3.04%	19.08%	2.55%	19.94%	2.61%
DRY	5.05%	13.22%	23.08%	3.97%	5.05%	25.72%	0.72%	20.31%	2.88%
NET	6.89%	14.08%	26.21%	3.29%	4.83%	27.75%	1.95%	14.18%	0.82%
AT	2.36%	14.81%	35.41%	4.08%	1.50%	17.17%	0.64%	20.17%	3.86%
AR	3.27%	18.14%	26.69%	2.02%	2.02%	18.64%	0%	24.43%	4.79%
GE	2.94%	16.62%	24.95%	0.90%	5.16%	23.33%	1.26%	21.30%	3.54%

Table 4: The percentage of each NE in a genre. The meanings of items are same as ones in Table 2.

Tetsuji Nakagawa, Taku Kudo, and Yuji Matsumoto. 2002. Revision learning and its application to part-of-speech tagging. In *Proc. of ACL'02*, pages 497–504.

Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata. 2002. Extended named entity hierarchy. In *Proc. of LREC'02*.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proc. of CoNLL'03*, pages 142–147.

R. Weischedel and A. Brunstein. 2005. Bbn pronoun coreference and entity type corpus. linguistic data consortium.

NE / Genre	Yahoo! Chiebukuro	White Paper
ARTIFACT	12.70 (7.41, 44.44)	45.69 (32.52, 76.81)
DATE	68.42 (68.42, 68.42)	77.52 (77.52, 77.52)
LOCATION	82.69 (75.44, 91.49)	86.47 (82.14, 91.27)
MONEY	100.00 (100.00, 100.00)	88.89 (88.89, 88.89)
ORGANIZATION	33.33 (26.32, 45.45)	70.83 (79.69, 63.75)
PERCENT	0 (0, 0)	96.88 (93.94, 100.00)
PERSON	33.33 (50.00, 25.00)	59.57 (93.33, 43.75)
Total	56.20 (46.11, 71.96)	72.12 (68.56, 76.08)

NE / Genre	Yahoo! Blog	Books
ARTIFACT	10.53 (8.00, 15.38)	48.78 (34.48, 83.33)
DATE	71.58 (56.67, 97.14)	51.69 (46.00, 58.97)
LOCATION	68.00 (65.38, 70.83)	57.99 (56.32, 59.76)
ORGANIZATION	50.00 (42.62, 60.47)	39.13 (34.62, 45.00)
PERCENT	95.24 (90.91, 100.00)	60.00 (50.00, 75.00)
PERSON	68.75 (69.62, 67.90)	72.67 (66.86, 79.58)
TIME	50.00 (33.33, 100.00)	80.00 (75.00, 85.71)
Total	63.06 (56.71, 71.01)	62.56 (56.80, 69.61)

NE / Genre	Magazines	Newspapers
ARTIFACT	72.73 (61.54, 88.89)	37.50 (37.50, 37.50)
DATE	86.08 (80.95, 91.89)	86.24 (85.45, 87.04)
LOCATION	28.07 (50.00, 19.51)	86.11 (82.01, 90.64)
MONEY	100.00 (100.00, 100.00)	94.12 (94.92, 93.33)
ORGANIZATION	66.67 (64.71, 68.75)	70.05 (64.41, 76.77)
PERCENT	100.00 (100.00, 100.00)	93.15 (89.47, 97.14)
PERSON	62.82 (53.69, 75.69)	87.34 (88.46, 86.25)
TIME	50.00 (50.00, 50.00)	72.73 (57.14, 100.00)
Total	60.56 (58.73, 62.50)	82.70 (79.77, 85.85)

Table 5: Accuracy of KNP on BCCWJ Basic NE corpus. The value indicates F-measure (Recall, Precision).