Comparison of Annotating Methods for Named Entity Corpora

Kanako Komiya[1] **Masaya Suzuki**[1] **Tomoya Iwakura**[2] **Minoru Sasaki**[1] **Hiroyuki Shinnou**[1]
Ibaraki University[1] Fujitsu Laboratories Ltd.[2]
4-12-1 Nakanarusawa, Hitachi-shi, Ibaraki, 316-8511 JAPAN
1-1, Kamikodanaka 4-chome, Nakahara-ku, Kawasaki, Kanagawa, 211-8588 JAPAN
{kanako.komiya.nlp, 13t4038a}@vc.ibaraki.ac.jp,
iwakura.tomoya@jp.fujitsu.com,
{minoru.sasaki.01, hiroyuki.shinnou.0828}@vc.ibaraki.ac.jp

Abstract

We compared two methods to annotate a corpus via non-expert annotators for named entity (NE) recognition task, which are (1) revising the results of the existing NE recognizer and (2) annotating NEs only by hand. We investigated the annotation time, the degrees of agreement, and the performances based on the gold standard. As we have two annotators for one file of each method, we evaluated the two performances, which are the averaged performances over the two annotators and the performances deeming the annotations correct when either of them is correct. The experiments revealed that the semi-automatic annotation was faster and showed better agreements and higher performances on average. However they also indicated that sometimes fully manual annotation should be used for some texts whose genres are far from its training data. In addition, the experiments using the annotated corpora via semi-automatic and fully manual annotation as training data for machine learning indicated that the F-measures sometimes could be better for some texts when we used manual annotation than when we used semi-automatic annotation.

1 Introduction

The crowdsourcing made annotation of the training data cheaper and faster (Snow et al., 2008). Snow et al. evaluated non-expert annotations but they did not discuss the difference in the annotation qualities depending on how to give them the corpus. Therefore, we compared the two methods to annotate a corpus, which are semi-automatic and fully manual annotations, to examine the method to generate high quality corpora by non-experts. We investigate Japanese named entity (NE) recognition task using a corpus that consists of six genres to examine the annotation qualities depending on the genres.

The annotation of NE task is difficult for non-experts because its definition has many rules, and some of them are complicated. Therefore, the semi-automatic annotation seems a good way to decrease the annotation errors. However, sometimes the existing system also can make mistakes, especially on corpora in other genres but newswires, because it is trained only from the newswire corpus. Therefore, we compare the two methods to annotate a corpus, which are the semi-automatic and fully manual annotations and discuss them, from the point of view of time, agreement, and performance based on the gold standard to generate high quality corpora by non-experts. We also discuss the difference in performances according to the genres of the target corpus as we used the multi-genre corpus for analysis.

2 Related Work

Snow et al. (2008) evaluated non-expert annotations through comparing with expert annotations from the point of view of time, quality, and cost. Alex et al. (2010) proposed agile data annotation, which is iterative, and compared it with the traditional linear annotation method. van der Plas et al. (2010) described the method to annotate semantic roles to the French corpus using English template to investigate the cross-lingual validity. Marcus et al. (1993) compared the semi-automatic and fully manual annotations to develop the Penn Treebank on the POS tagging task and the bracketing task. However, as far as we know, there is no paper which compared the semi-automatic and

Proceedings of LAW X – The 10th Linguistic Annotation Workshop, pages 59–67,
Berlin, Germany, August 11, 2016.

fully manual annotations to develop high quality corpora via non-expert annotators.

We investigate the named entity recognition (NER) task. NER involves seeking to locate and classify elements in text into predefined categories, such as the names of people, organizations, and locations, and has been studied for a long time. Information Retrieval and Extraction Exercise (IREX)[1] defined the nine tags including eight types of NEs, i.e., organization, person, artifact, date, time, money, and percent as well as the option tag for shared task of Japanese NER. However, only newswires were used for this task. For the researches of NER, Hashimoto et al. (2008) generated extended NE corpus based on the Balanced Corpus of Contemporary Japanese (BCCWJ) (Maekawa, 2008)[2] . Tokunaga et al. (2015) analyzed the eye-tracking data of annotators of NER task. Sasada et al. (2015) proposed the NE recognizer which is trainable from partially annotated data.

In 2014, researchers analyzed the errors of Japanese NER using the newly tagged NE corpus of BCCWJ, which consists of six genres as Japanese NLP Project Next [3] (Iwakura, 2015; Hirata and Komachi, 2015; Ichihara et al., 2015). Ichihara et al. (2015) investigated the performance of the existing NE recognizer and showed that the errors increased in the genres far from the training data of the NE recognizer. This paper indicates that the semi-automatic annotation can make some errors on the corpus far from the training data.

We evaluate the semi-automatic and fully manual annotations for Japanese NER task, from the point of view of time, agreement, and performance based on the gold standard to generate high quality corpora by non-experts.

3 Comparison of Annotating Method

This paper compared the following two methods to annotate a corpus.

KNP+M Semi-automatic annotation, which is revising the results of the existing NE recognizer: KNP (Sasano and Kurohashi, 2008) [4]

Manual Fully manual annotation, whichi is annotating NEs only by hand

[1]http://nlp.cs.nyu.edu/irex/index-j.html
[2]http://pj.ninjal.ac.jp/corpus_center/bccwj/
[3]https://sites.google.com/site/projectnextnlp/
[4]http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP

		Method X				
		Tag 1	Tag 2	...	Tag n	Sum
Method Y	Tag 1	a_{11}	a_{21}	...	a_{n1}	a_{01}
	Tag 2	a_{12}	a_{22}	...	a_{n2}	a_{02}
	...	...	...	...	...	...
	Tag n	a_{1n}	a_{2n}	...	a_{nn}	a_{0n}
	Sum	a_{10}	a_{20}	...	a_{n0}	a_{00}

Table 1: The number of tag matching between two annotaters

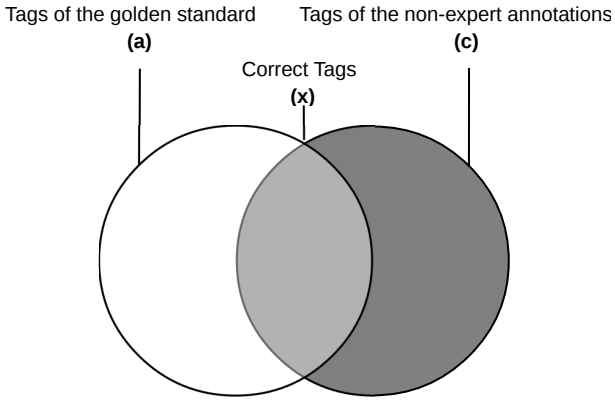

Figure 1: Example of a set of tags

We investigated the annotation time for each text, the observed agreement and Kappa coefficient of annotations, and the precision, the recall, and the F-measure based on the gold standard.

The observed agreement and Kappa coefficient are calculated as equ. (1) and equ. (2) respectively when the numbers of tag matching between two annotaters are as shown in Table 1.

$$d = \frac{\sum_{i=1}^{n} a_{ii}}{a_{00}} \quad (1)$$

$$\kappa = \frac{a_{00}\sum_{i=1}^{n} a_{ii} - \sum_{i=1}^{n} a_{i0}a_{0i}}{(a_{00})^2 - \sum_{i=1}^{n} a_{i0}a_{0i}} \quad (2)$$

The precisions, the recalls, and the F-measures are calculated as equ. (3), equ. (4), and equ. (5) when we have the set of tags as Figure 1.

$$p = \frac{n(x)}{n(c)} \quad (3)$$

$$r = \frac{n(x)}{n(a)} \quad (4)$$

$$f = \frac{2pr}{p+r} \quad (5)$$

4 Experiment

We used 136 texts extracted from BCCWJ, which are available as ClassA[5]. BCCWJ consists of six genres, "Q & A sites" (OC), "white papers" (OW), "blogs" (OY), "books" (PB), "magazines" (PM), and "newswires" (PN). Table 2 shows the summary of the numbers of documents and tags of each genre.

Sixteen non-experts assigned the nine types of NE tag of IREX to the plain texts after reading the definitions [6]. Every annotator annotated 34 texts, which is 17 texts via **KNP+M** and **Manual**, respectively, which makes two sets of corpus for each method. Eight annotators began with **KNP+M**, and the rest began with **Manual** to address the bias of the proficiency. Annotation time is recorded for each text. We calculated the averaged annotation time for one set of corpus, i.e., 136 texts, for each method. Therefore, the documents matched in size when the annotation times were compared. We used the newest corpus of BCCWJ by 2016/2/11 (Iwakura et al., 2016) [7] as the gold standard. We used KNP Ver. 4.11 and JUMAN Ver. 7.0 for windows [8].

The performances were evaluated based on the rules defined for IREX. In other words, the annotations were deemed correct if and only if both the tag and its extent were correct except for the cases of the optional tags. When the optional tag was assigned to some words in the gold standard, the annotations were deemed correct if (1) the words were not annotated by any tags or (2) a word or some words in that extent were annotated by any tags including the optional tag.

As we have two annotators for one file of each method, we evaluated the two performances based on golden standard, which are the averaged performances over the two annotators and the performances deeming the annotations correct when either of them is correct. We investigate the latter performances since we usually integrate the results of two annotators when we generate corpora.

In addition, we used the corpora which are annotated via **Manual** or **KNP+M** as the training data for supervised learning of NER to test the quality of the annotations for the machine learning. The training mode of KNP was used for the experiments. Therefore, the features for training are the same as the original KNP, which are the morpheme itself, character type, POS tag, category if it exists, cache features, syntactic features, and caseframe features (Sasano and Kurohashi, 2008). We used KNP Ver. 4.16 and JUMAN Ver. 7.01 for Linux for training-mode. We used the five-fold cross validation. Since two persons annotated each file for each method, we used two annotations for the training data of each method. Every test set of each validation includes the texts from as many genres as possible.

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

[6]KNP does not extract optional tags.

[7]https://sites.google.com/site/projectnextnlpne/en

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

Method	Observed	Kappa
KNP+M	**0.79**	**0.75**
Manual	0.57	0.50
Both	0.64	0.58

Table 3: Micro-averaged observed agreement and Kappa coefficient of each method (All)

5 Result

Tables 3 and 4 show the micro and macro-averaged observed agreement (Observed) and Kappa coefficients (Kappa) of each method of all the genres. Tables 5 and 6 summarize those of each genre. **KNP+M** and **Manual** in the tables are the agreement values between the two annotators of each method, respectively. **Both** in the tables are averaged values of every combination pairs in the four annotators of the both two methods. Table 7 shows the averaged annotation time for one text according to each method.

Tables 8 and 9 show the averaged precisions (P), recalls (R), and F-measures (F) of each method of all the genres. They are average over the two annotators. Tables 10 and 11 summarize those of each genre. The fully automatic annotation, which is the results of original KNP without revising are also shown in these tables as **KNP**. **Avg.** in the tables indicates the average of **KNP+M** and **Manual**. The higher observed agreements, Kappa coefficients, precisions, recalls, and F-measures among the two methods are written in bold.

Next, we investigated the performances deeming the annotations correct when either of the two annotators is correct. Tables 12 and 13 show the precisions (P), the recalls (R), and the F-measures (F) of each method of all the genres. Tables 14 and 15 summarize those of each genre. The fully

Genre	Doc	Artifact	Date	Location	Money	Organization	Percent	Person	Time	Optional	All
OC	74	44	18	65	9	18	0	6	0	8	168
OW	8	86	143	147	9	136	33	15	0	26	595
OY	34	23	61	59	7	64	10	79	3	17	323
PB	5	32	49	100	0	19	5	174	9	20	408
PM	2	9	24	36	5	18	1	216	3	1	313
PN	13	24	166	192	60	123	37	78	22	20	722
ALL	136	218	461	599	90	378	86	568	37	92	2,529

Table 2: Summary of number of documents and tags

Method	Observed	Kappa
KNP+M	**0.66**	**0.48**
Manual	0.52	0.29
Both	0.52	0.31

Table 4: Macro-averaged observed agreement and Kappa coefficient of each method (All)

Genre	Method	Observed	Kappa
OC	KNP+M	**0.62**	**0.54**
OC	Manual	0.47	0.34
OC	Both	0.52	0.41
OW	KNP+M	**0.78**	**0.73**
OW	Manual	0.41	0.28
OW	Both	0.55	0.46
OY	KNP+M	**0.69**	**0.63**
OY	Manual	0.58	0.50
OY	Both	0.57	0.49
PB	KNP+M	**0.76**	**0.68**
PB	Manual	0.67	0.56
PB	Both	0.71	0.61
PM	KNP+M	**0.87**	**0.84**
PM	Manual	0.61	0.55
PM	Both	0.69	0.64
PN	KNP+M	**0.86**	**0.75**
PN	Manual	0.81	0.65
PN	Both	0.80	0.65

Table 5: Micro-averaged observed agreement and Kappa coefficient of each method

Genre	Method	Observed	Kappa
OC	KNP+M	**0.58**	**0.27**
OC	Manual	0.50	0.15
OC	Both	0.47	0.14
OW	KNP+M	**0.80**	**0.73**
OW	Manual	0.45	0.36
OW	Both	0.59	0.50
OY	KNP+M	**0.63**	**0.47**
OY	Manual	0.50	0.29
OY	Both	0.47	0.30
PB	KNP+M	**0.63**	**0.54**
PB	Manual	0.60	0.43
PB	Both	0.62	0.48
PM	KNP+M	**0.87**	**0.83**
PM	Manual	0.62	0.55
PM	Both	0.69	0.63
PN	KNP+M	**0.88**	**0.74**
PN	Manual	0.74	0.56
PN	Both	0.77	0.59

Table 6: Macro-averaged observed agreement and Kappa coefficient of each method

Method	Averaged time
KNP+M	0:03:19
Manual	0:05:23

Table 7: Tagging time for each method

automatic annotation, which is the results of KNP without revising are also shown in these tables as **KNP** here again.

In addition, we examined the performances of the system trained with the corpora annotated via **KNP+M** and **Manual**. Tables 16 and 17 show the precisions (P), the recalls (R), and the F-measures (F) of each method of all the genres. Tables 18 and 19 summarize those of each genre. The results of original KNP are also shown in these tables as **KNP** here again.

The differences between **KNP** and **KNP+Manual**, **KNP** and **Manual**, and **Manual** and **KNP+Manual** of the precisions and the recalls in Tables 8 and 16 and those of the precisions in Table 14 are statistically significant according to chi-square test. However, the differences between **KNP** and **KNP+Manual** and **KNP** and **Manual** are statistically significant but that between **Manual** and **KNP+Manual** is not significant according to chi-square test when we compared the recalls of Table 12. In addition, the asterisk in the tables of micro-averaged accuracies for each genre, i.e., Tables 10, 14, and 18, means the difference between presicions or recalls of **Manual** and **KNP+Manual** is statistically significant according to a chi-square test. The level of significance in the test was 0.05. When macro-averaged accuracies were compared, the differences were not significant due to the decrease of the samples of the test.

Method	P	R	F
KNP	77.64%	68.09%	72.55%
KNP+M	**84.03%**	**81.41%**	**82.70%**
Manual	75.22%	72.74%	73.96%
Avg.	79.63%	77.07%	78.33%

Table 8: Micro-averaged precision, recall, and F-measure of each method (All)

Method	P	R	F
KNP	47.43%	39.81%	43.29%
KNP+M	**55.30%**	**54.72%**	**55.01%**
Manual	52.54%	51.06%	51.77%
Avg.	53.92%	52.87%	53.39%

Table 9: Macro-averaged precision, recall, and F-measure of each method (All)

6 Discussion

6.1 Agreements and Time

First, Tables 3 and 4 show that the observed agreements and Kappa coefficients of **KNP+M** are higher than those of **Manual** in both micro and macro averages. This is similar in every genre according to Tables 5 and 6. We think this is because that the tags assigned by KNP still remain after the annotators revised the results of KNP. The agreement values of **Both** are usually higher than or similar to those of **Manual** but the macro-averaged Kappa coefficient of **Both** (0.14) is lower than that of **Manual** (0.15) more than one point (0.01) in OC, which indicates the results of annotators greatly vary. These results indicate that there can be some NEs which require more rules to extract in OC because the definition we used was developed for only the newswires. In addition, Table 3 shows that Kappa coefficients indicate good agreement for **KNP+M** and moderate agreement for **Manual** when they are micro-averaged, and Table 4 shows that they indicate moderate agreement for **KNP+M** and poor agreement for **Manual** when they are macro-averaged. Since micro average is an average over NEs, and macro average is that over texts, it means that the agreement values of some texts which include a few NEs were low.

In addition, Table 7 shows that the annotation time for one text of **KNP+M** is approximately two minutes shorter on average than that of **Manual**. These results indicate that **KNP+M** is faster and shows better agreement than **Manual**. The difference in time was significant according to F test. The level of significance is 0.01.

Genre	Method	P	R	F
OC	KNP	72.38%	47.50%	57.36%
OC	KNP+M	***77.74%**	75.31%	**76.51%**
OC	Manual	66.93%	**80.06%**	72.91%
OC	Avg.	71.76%	77.69%	74.61%
OW	KNP	78.87%	78.60%	78.73%
OW	KNP+M	***81.68%**	***84.62%**	**83.12%**
OW	Manual	64.62%	67.22%	65.90%
OW	Avg.	73.11%	75.90%	74.48%
OY	KNP	73.42%	56.86%	64.09%
OY	KNP+M	***85.47%**	***75.00%**	**79.90%**
OY	Manual	79.81%	68.13%	73.51%
OY	Avg.	82.67%	71.56%	76.71%
PB	KNP	75.00%	59.54%	66.38%
PB	KNP+M	**78.54%**	**73.58%**	**75.98%**
PB	Manual	77.85%	72.84%	75.27%
PB	Avg.	78.20%	73.21%	75.62%
PM	KNP	60.61%	57.69%	59.11%
PM	KNP+M	88.51%	**86.38%**	**87.43%**
PM	Manual	**89.68%**	84.94%	87.24%
PM	Avg.	89.08%	85.66%	87.34%
PN	KNP	88.44%	78.49%	83.17%
PN	KNP+M	***87.87%**	***85.11%**	**86.47%**
PN	Manual	77.46%	72.12%	74.70%
PN	Avg.	82.77%	78.61%	80.64%

Table 10: Micro-averaged precision, recall, and F-measure of each method

6.2 Performances Averaged over Annotators

Next, we evaluate the performances of the methods based on the gold standard. First, we evaluate the average over the two annotators.

We can see the precisions, the recalls, and the F-measures of **KNP+M** are higher than those of **Manual** in both micro and macro averages, according to Tables 8 and 9. This is similar in every genre in micro average according to Table 10, except the recall of OC and the precision of PM. When we see these two exceptions, we can see that those of **KNP** are considerably lower than those of other genres. The topic of OC was far from newswires, and a name of person was misrecognized as name of location many times in PM. This fact indicates that the performances of **KNP+M** directly depend on those of **KNP**.

Table 11 shows that the macro-averaged precisions, recalls, and F-measurs of **KNP+M** are better than those of **Manual** in OW, OY, and PN but those of **Manual** are better in OC, PB, and PM, except the recall of PM. We think this is because **KNP** are better than **Manual** in the precisions, the recalls, and the F-measures in OW and PN and the precisions in OY. OW and PN are similar to the training data set of KNP, i.e., newswires, which makes the performances in them better (Ichihara et al., 2015). These results indicate that **KNP+M**

Genre	Method	P	R	F
OC	KNP	30.74%	25.55%	27.91%
OC	KNP+M	38.83%	40.75%	39.77%
OC	Manual	**41.80%**	**43.84%**	**42.79%**
OC	Avg.	40.31%	42.29%	41.28%
OW	KNP	76.84%	80.45%	78.60%
OW	KNP+M	**82.98%**	**85.47%**	**84.21%**
OW	Manual	69.91%	72.65%	71.25%
OW	Avg.	76.45%	79.06%	77.73%
OY	KNP	57.99%	44.37%	50.27%
OY	KNP+M	**68.33%**	**62.94%**	**65.53%**
OY	Manual	55.79%	49.32%	52.36%
OY	Avg.	62.06%	56.13%	58.95%
PB	KNP	66.04%	45.84%	54.12%
PB	KNP+M	71.02%	64.63%	67.67%
PB	Manual	**81.37%**	**67.48%**	**73.77%**
PB	Avg.	76.19%	66.05%	70.76%
PM	KNP	60.31%	66.37%	63.19%
PM	KNP+M	82.34%	**87.00%**	84.61%
PM	Manual	**85.64%**	83.94%	**84.78%**
PM	Avg.	83.99%	85.47%	84.73%
PN	KNP	87.51%	77.70%	82.31%
PN	KNP+M	**87.76%**	**85.06%**	**86.39%**
PN	Manual	78.37%	71.60%	74.83%
PN	Avg.	83.06%	78.33%	80.63%

Table 11: Macro-averaged precision, recall, and F-measure of each method

Method	P	R	F
KNP	77.64%	68.09%	72.55%
KNP+M	**91.34%**	**88.92%**	**90.11%**
Manual	86.76%	88.28%	87.53%

Table 12: Micro-averaged precision, recall, and F-measure of each method (All) deeming the annotations correct when either of two annotators is correct

is better than **Manual** to annotate corpora by non-experts, in particular, the texts in some genres similar to the training data of KNP. However, sometimes **Manual** should be used for some texts, whose genres are far from newswires.

6.3 Sum-Set Performances of Two annotators

Next, we investigate the performances deeming the annotations correct when either of the two annotators is correct. Tables 12 and 13 show that the precision, the recall, and F-measure of **KNP+M** are also better than those of **Manual** even if we deemed the annotations correct when either of the two annotators was correct. However, the difference greatly decreased comparing with Tables 8 and 9, i.e., the performances averaged over the annotators. In particular, the difference between **KNP+M** (62.92%) and **Manual** (62.09%) was less than one point when the macro-averaged F-measures were compared. We think

Method	P	R	F
KNP	47.43%	39.81%	43.29%
KNP+M	**63.48%**	**62.37%**	**62.92%**
Manual	61.96%	62.22%	62.09%

Table 13: Macro-averaged precision, recall, and F-measure of each method (All) deeming the annotations correct when either of two annotators is correct

Genre	Method	P	R	F
OC	KNP	72.38%	47.50%	57.36%
OC	KNP+Manual	**86.79%**	86.25%	86.52%
OC	Manual	85.63%	**90.51%**	**88.00%**
OW	KNP	78.87%	78.60%	78.73%
OW	KNP+Manual	***91.20%**	**91.20%**	**91.20%**
OW	Manual	75.71%	89.07%	81.85%
OY	KNP	73.42%	56.86%	64.09%
OY	KNP+Manual	**93.62%**	**87.13%**	**90.26%**
OY	Manual	92.91%	85.90%	89.27%
PB	KNP	75.00%	59.54%	66.38%
PB	KNP+Manual	87.05%	81.87%	84.38%
PB	Manual	**89.86%**	**86.32%**	**88.05%**
PM	KNP	60.61%	57.69%	59.11%
PM	KNP+Manual	92.65%	**93.55%**	93.10%
PM	Manual	***97.26%**	92.81%	**94.98%**
PN	KNP	88.44%	78.49%	83.17%
PN	KNP+Manual	***93.29%**	**90.33%**	**91.79%**
PN	Manual	89.19%	87.25%	88.21%

Table 14: Micro-averaged precision, recall, and F-measure of each method deeming the annotations correct when either of two annotators is correct

this is because the manual annotations vary and one of the two annotators usually annotates the NEs correctly. As Tables 8 and 9 showed, the non-expert annotators often make mistakes because the definitions of NEs for IREX include so many rules and therefore, the annotators sometimes overlooked some rules when they annotated the texts. However, the experimental results revealed that the performances of the fully manual annotations were almost comparable to those of the semi-automatically annotations when we have two annotators. Moreover, Tables 14 and 15 indicate that the F-measures of **Manual** are better than those of **KNP+M** in OC, PB, and PM. These results are like those in Table 11 but not like those in Table 10, which means that the better method varies depending on the genres even if the performances were micro-averaged when we deemed the results correct when either of two annotator was correct.

Furthermore, we compared Table 8 with Table 12 and Table 9 with Table 13 to compare the performances of annotations by one annotator and

Genre	Method	P	R	F
OC	KNP	30.74%	25.55%	27.91%
OC	KNP+M	46.30%	47.35%	46.82%
OC	Manual	**49.16%**	**50.88%**	**50.01%**
OW	KNP	76.84%	80.45%	78.60%
OW	KNP+M	**91.09%**	90.96%	**91.02%**
OW	Manual	82.55%	**91.39%**	86.74%
OY	KNP	57.99%	44.37%	50.27%
OY	KNP+M	**78.69%**	**73.63%**	**76.07%**
OY	Manual	67.84%	65.47%	66.63%
PB	KNP	66.04%	45.84%	54.12%
PB	KNP+M	83.51%	77.94%	80.63%
PB	Manual	**93.98%**	**85.91%**	**89.76%**
PM	KNP	60.31%	66.37%	63.19%
PM	KNP+M	85.74%	93.17%	89.30%
PM	Manual	**97.58%**	**93.45%**	**95.47%**
PN	KNP	87.51%	77.70%	82.31%
PN	KNP+M	**93.36%**	**90.09%**	**91.70%**
PN	Manual	88.94%	86.39%	87.64%

Table 15: Macro-averaged precision, recall, and F-measure of each method deeming the annotations correct when either of two annotators is correct

Method	P	R	F
KNP	77.64%	68.09%	72.55%
KNP+M	**74.14%**	**38.11%**	**50.34%**
Manual	67.21%	28.52%	40.05%

Table 16: Micro-averaged precision, recall, and F-measure of each method (All) when the annotated data were used for training

those by two annotators. The results in Tables 8 and 9 could be considered as the annotations by one annotator because they are averages over annotators. These four tables show that the results of annotations by two annotators are always better than those by one annotator. In particular, the performances by two annotators of **Manual** are always better than those by one annotator of **KNP+M**. Since the better methods varies depending on the genres in both micro and macro averages when the performances of annotations by two annotators are compared, these results indicate that we should use not only **KNP+M** but also **Manual** in real situation.

6.4 Annotated Corpora as Training Data

Finally, we evaluate the performances of machine learning when we used the annotated corpora via **KNP+M** and **Manual** as the training data. Tables 16 and 17 show that the precision, the recall, and F-measure of **KNP+M** are better than those of **Manual** when we used the annotated corpora as the training data for KNP. However, Tables 18 and 19 show that the micro-averaged precisions

Method	P	R	F
KNP	47.43%	39.81%	43.29%
KNP+M	**40.41%**	**23.55%**	**29.76%**
Manual	31.44%	16.16%	21.34%

Table 17: Macro-averaged precision, recall, and F-measure of each method (All) when the annotated data were used for training

Genre	Method	P	R	F
OC	KNP	72.38%	47.50%	57.36%
OC	KNP+M	**88.46%**	**28.75%**	**43.40%**
OC	Manual	84.21%	20.00%	32.32%
OW	KNP	78.87%	78.60%	78.73%
OW	KNP+M	***74.45%**	***53.16%**	**62.03%**
OW	Manual	54.69%	35.85%	43.31%
OY	KNP	73.42%	56.86%	64.09%
OY	KNP+M	**83.62%**	***31.70%**	**45.97%**
OY	Manual	80.00%	18.30%	29.79%
PB	KNP	75.00%	59.54%	66.38%
PB	KNP+M	70.41%	**30.67%**	**42.73%**
PB	Manual	**73.29%**	27.58%	40.07%
PM	KNP	60.61%	57.69%	59.11%
PM	KNP+M	**55.05%**	**19.23%**	**28.50%**
PM	Manual	51.76%	14.10%	22.17%
PN	KNP	88.44%	78.49%	83.17%
PN	KNP+M	76.00%	***43.30%**	**55.17%**
PN	Manual	**78.26%**	35.90%	49.22%%

Table 18: Micro-averaged precision, recall, and F-measure of each method when the annotated data were used for training

in PB and PN, the macro-averaged precisions in PB and PN, and the macro-averaged F-measure in PB were not the case. The exception of the macro-averaged F-measure shows that sometimes the annotation of **Manual** is better training data than **KNP+M**.

Tables 16 and 17 show the difference in the precisions between the original KNP and other methods are not so large comparing with those of the recalls. In particular, **KNP+M** and **Manual** were better than the original KNP when the micro-averaged precisions in OC and OY were compared according to Table 18. The performances of **KNP+M** and **Manual** were low because the amount of the training data was so small comparing with the original KNP. However, these results show that the precisions will be better than original KNP even if we use a small training data in some genres.

7 Conclusion

We compared the semi-automatic and fully manual annotations to investigate the annotation qualities by non-experts. The methods we investigated

Genre	Method	P	R	F
OC	KNP	30.74%	25.55%	27.91%
OC	KNP+M	**24.32%**	**15.88%**	**19.22%**
OC	Manual	17.34%	12.24%	14.35%
OW	KNP	76.84%	80.45%	78.60%
OW	KNP+M	**71.59%**	**56.71%**	**63.29%**
OW	Manual	62.55%	42.52%	50.63%
OY	KNP	57.99%	44.37%	50.27%
OY	KNP+M	**52.32%**	**24.40%**	**33.28%**
OY	Manual	30.82%	9.184%	14.15%
PB	KNP	66.04%	45.84%	54.12%
PB	KNP+M	51.46%	**23.63%**	32.39%
PB	Manual	**64.93%**	21.65%	**32.47%**
PM	KNP	60.31%	66.37%	63.19%
PM	KNP+M	**54.56%**	**29.20%**	**38.04%**
PM	Manual	53.43%	24.63%	33.72%
PN	KNP	87.51%	77.70%	82.31%
PN	KNP+M	75.21%	**43.71%**	**55.28%**
PN	Manual	**77.88**%	37.01%	50.17%

Table 19: Macro-averaged precision, recall, and F-measure of each method when the annotated data were used for training

were **KNP+M**, which was revising the results of the existing NE recognizer, and **Manual**, which was annotating NEs only by hand. We investigated Japanese NER task. We evaluated the annotation time, the observed agreement, Kappa coefficients, and the precisions, the recalls, and the F-measures based on the gold standard. As two annotators annotated each text for each method, we evaluated the precisions, the recalls, and the F-measures averaged over annotators and those deeming the results correct when either of them was correct. The experiments revealed that **KNP+M** was faster and showed better agreements and higher performances than **Manual** on average but sometimes **Manual** should have been used for some texts whose genres were far from newswires. Finally the experiments using the annotated corpora via **KNP+M** or **Manual** indicated that the F-measures sometimes could be better for some texts when we used **Manual** than when we used **KNP+M**.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 15K16046 and contribution from Fujitsu Laboratories Ltd.

References

Bea Alex, Claire Grover, Rongzhou Shen, and Mijail Kabadjov. 2010. Agile corpus annotation in practice: An overview of manual and automatic annotation of cvs. In *Proceedings of Fourth Linguistic Annotation Workshop, ACL 2010*, pages 29–37.

Taiichi Hashimoto, Takashi Inui, and Koji Murakami. 2008. Constructing extended named entity annotated corpora (in japanese). *IPSJ SIG Technical Reports（NLP）*, 2008-NL-188:113–120.

Ai Hirata and Mamoru Komachi. 2015. Analysis of named entity recognition for texts of various genres (in japanese). *NLP2015 Error Analysis Workshop*. https://docs.google.com/viewer?a=v&pid=sites&srcid=ZGVmYXVsdGRvbWFpbnxwcm9qZWN0bmV4dG5scHxneDo1ZGYxOTg3YWE1MDIzOTRi.

Masaaki Ichihara, Kanako Komiya, Tomoya Iwakura, and Maiko Yamazaki. 2015. Error analysis of named entity recognition in bccwj. *NLP2015 Error Analysis Workshop*. https://docs.google.com/viewer?a=v&pid=sites&srcid=ZGVmYXVsdGRvbWFpbnxwcm9qZWN0bmV4dG5scHxneDoxZTY1MWY4YTBjNmNjNzIx.

Tomoya Iwakura, Ryuichi Tachibana, and Kanako Komiya. 2016. Constructing a japanese basic named entity corpus of various genres. *Proceedings of NEWS 2016*.

Tomoya Iwakura. 2015. Error analysis of named entity extraction (in japanese). *NLP2015 Error Analysis Workshop*. https://docs.google.com/viewer?a=v&pid=sites&srcid=ZGVmYXVsdGRvbWFpbnxwcm9qZWN0bmV4dG5scHxneDo1ZTg0ZmJmYmRjNThmN2I1.

Kikuo Maekawa. 2008. Balanced corpus of contemporary written japanese. In *Proceedings of the 6th Workshop on Asian Language Resources (ALR)*, pages 101–102.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: the penn treebank. *Computational Linguistics - Special issue on using large corpora: II*, 19:313–330.

Tetsuro Sasada, Shinsuke Mori, Tatsuya Kawahara, and Yoko Yamakata. 2015. Named entity recognizer trainable from partially annotated data. In *Proceedings of the PACLING 2015*, pages 10–17.

Ryohei Sasano and Sadao Kurohashi. 2008. Japanese named entity recognition using structural natural language processing. In *Proceedings of IJCNLP 2008*, pages 607–612.

Rion Snow, Brendan O'Conner, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast – but is it good? evaluation non-expert annotation for natural lanugage tasks. In *Proceedings of the 2008 Conference on Emprical Methods in Natural Language Processing (EMNLP)*, pages 254–263.

Takenobu Tokunaga, Jin Nishikara, Tomoya Iwakura, and Nobuhiro Yugami. 2015. Analysis of eye tracking data of annotators for named entity recognition task (in japanese). *IPSJ SIG Technical Reports（NLP）*, 2015-NL-223:1 – 8.

Lonneke van der Plas, Tanja Samardžić, and Paola Merlo. 2010. Cross-lingual validity of propbank in the manual annotation of french. In *Proceedings of Fourth Linguistic Annotation Workshop, ACL 2010*, pages 113–117.