Error Analysis of Named Entity Recognition in BCCWJ

Masaaki Ichihara[1]
Tomoya Iwakura[2]
Kanako Komiya[1]
Maiko Yamazaki[3]
[1]Ibaraki University, [3]Fujitsu Laboratories Ltd., [2]Tokyo Institute of Technology
{11t4004s@hcs, kkomiya@mx}.ibaraki.ac.jp, iwakura.tomoya@jp.fujitsu.com, yamazaki@lr.pi.titech.ac.jp

1 Introduction

Named Entity Recognition is a process by which named entities (NEs) such as the names of persons, locations, and artifacts are extracted. Most named entity recognition techniques have been studied on news articles, however, their performances on different domain texts such as blogs, books and magazines are still not evaluated well. This paper reports an error analysis of KNP on six domains for revealing causes of errors for further improvement of NE recognition[1].

2 Error Analysis of KNP on BCCWJ

Japanese dependency and case structure analyzer KNP[2] ([2] and [3]) was used as the named entity recognizer. The versions we used were KNP Ver.4.11 and JUMAN Ver.7.0.

The six genres, "Q & A sites", "white papers", "blogs", "books", "magazines", and "newspaper articles", in Balanced Corpus of Contemporary Written Japanese (BCCWJ) were used as the target corpora.

One hundred thirty six texts extracted from BCCWJ, they are available as ClassA[3], were used for the experiments.

They were manually annotated with nine kinds of NE that were defined by Information Retrieval and Extraction Exercise (IREX)[4]. These NE types are the names of persons, locations, artifacts, dates, times, moneys, percents, and optional[5]. The annotation was done by five members of NE team of the Project Next NLP, and checked by four members of it.

We compared KNP outputs with the manually annotated texts and analyzed errors.

Table 1 shows the performances of KNP. The equations of recall, precision, accuracy, and F-measure are as follows. "Correct", the numerator of recall, precision, and accuracy, is the number of the correct answers of KNP. "Annotated", the denominator of recall, denotes the number of the NEs that were manually annotated. "KNP outputs", the denominator of precision, denotes the number of the NEs that KNP output. The denominator of accuracy is the logical sum (OR) of "Annotated" and "KNP outputs". The denominators of recall, precision, and accuracy vary because KNP sometimes cannot extract some NEs and sometimes extracts wrong information. Also, an NE that the system output sometimes consists of multiple annotated NEs as illustrated by an example in Figure 1 and vice versa. Table 1 shows the recall is lower than the precision.

KNP: ⟨PERSON⟩ 韓露 ⟨/PERSON⟩

Annotation :
⟨LOCATION⟩ 韓 ⟨/LOCATION⟩
⟨LOCATION⟩ 露 ⟨/LOCATION⟩

Figure 1: An example of an NE KNP output includes multiple annotated NEs

$$Recall = \frac{Correct}{Annotated} \quad (1)$$

$$Precision = \frac{Correct}{KNPoutputs} \quad (2)$$

$$Accuracy = \frac{Correct}{Annotated \bigcup KNPoutputs} \quad (3)$$

$$F - measure = \frac{2Recall \cdot Precision}{Recall + Precision} \quad (4)$$

[1]This paper is an English version of (Ichihara et al., 2015) [1] with additional information and some corrections.
[2]http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?KNP
[3]http://plata.ar.media.kyoto-u.ac.jp/mori/research/NLR/JDC/ClassA-1.list
[4]http://nlp.cs.nyu.edu/irex/index-e.html
[5]KNP does not extract optional tags.

Table 1: Performances of KNP

Performance	Rate	Correct	Denominator
Recall	61.79%	1632	2641
Precision	74.79%		2182
Accuracy	57.95%		2816
F-measure	67.68		

The errors were classified into the following five types. Examples were shown with description.

No extraction The error where KNP did not extract tokens as an NE though they were annotated.

KNP: サクラ大戦

Annotation :
⟨ARTIFACT⟩ サクラ大戦 ⟨/ARTIFACT⟩

No annotation The error where KNP extracted tokens as an NE though they were not annotated.

KNP: ⟨PERSON⟩ 日勤 ⟨/PERSON⟩

Annotation : 日勤

Wrong range The error where KNP extracted tokens as an NE and only the range was wrong. (The extracted tokens were partially annotated or they were the part of the annotated tokens.)

KNP1: ⟨PERSON⟩ ワシントン佐藤千矢子 ⟨/PERSON⟩

Annotation1 :
⟨PERSON⟩ 佐藤千矢子 ⟨/PERSON⟩

KNP2: 北大西洋条約機構（ＮＡＴＯ）⟨ORGANIZATION⟩ 欧州連合軍 ⟨/ORGANIZATION⟩

Annotation2 : ⟨ORGANIZATION⟩ 北大西洋条約機構（ＮＡＴＯ）欧州連合軍 ⟨/ORGANIZATION⟩

Wrong tag The error where KNP extracted tokens as an NE and only the tag type was wrong.

KNP: ⟨PERSON⟩ ウベア ⟨/PERSON⟩

Annotation :
⟨LOCATION⟩ ウベア ⟨/LOCATION⟩

Wrong range and tag The error where KNP extracted tokens as an NE and both the range and the tag type were wrong.

KNP: ⟨PERSON⟩ ワシントン佐藤千矢子 ⟨/PERSON⟩

Annotation :
⟨LOCATION⟩ ワシントン ⟨/LOCATION⟩

Table 2: Summary of errors

Error type	Num	Rate
No extraction	619	52.28%
No annotation	159	13.43%
Wrong range	162	13.68%
Wrong tag	127	10.73%
Wrong range and tag	117	9.88%
All errors	1184	100.00%

Table 2 shows a summary of errors. These errors were counted by the logical sum (OR) of annotated NEs and KNP outputs. The most frequent error was “No extraction” and it accounted for more than half of the total errors. The second most frequent error was “Wrong range” and most of them were the errors where extracted tokens were the part of the annotated tokens.

Table 3 shows a summary of errors by types of NEs. These errors were also counted by the logical sum (OR) of annotated NEs and KNP outputs. “Correct” and “Error” are the numbers of the correct answers and the errors of KNP. “Total” is the sum of “Correct” and “Error”. “No extraction” and “ Errors with extraction” in the table mean the numbers of “No extraction” and the errors other than “No extraction”, respectively. “No extraction rate” is the ratio of “No extraction” in “Error”.

Table 3 shows that no extraction rates of “ARTIFACT”, “PERCENT”, “TIME”, and “OPTIONAL” are especially high. At the same time, there are small number of NEs of “PERCENT” and “TIME ”in the corpora. Therefore, we can see “ARTIFACT” is the big reason why the no extraction rate of all tags is high. No extraction rate of “OPTIONAL” is 100% because KNP does not extract OPTIONALs and this is another reason.

Table 3 also shows that most of “TIME”, “MONEY”, and “PRECENT” were correctly tagged by KNP if they were tagged. Most of the errors when they were extracted are those of “ORGANIZATION”, “PERSON”, and “LOCATION”. The sum of errors of “ARTIFACT” and “DATE” are less than 30% of all errors when they were extracted.

Table 4 shows the accuracies and the rates of no extraction in “Total” according to the tag type. “Accuracy” is the ratio of the correct answers in “Total”, the sum of correct answers and errors of KNP, and “No extraction/Total” is the ratio of no extraction in it. These errors were also counted by the logical sum (OR) of annotated NEs and KNP outputs.

Table 4 shows that the accuracy of “ARTIFACT” is particularly low comparing with the other tags. The same table shows the ratio of no extraction in “Total” is also high. Therefore, we could see that “No extraction” of “ARTIFACT” is the biggest cause

Table 3: Summary of errors by types of NEs

Tag	Correct	Error	Total	No extraction	Errors with extraction	No extraction rate
ARTIFACT	90	259	349	192	67	74.13%
DATE	343	145	488	62	83	42.76%
LOCATION	409	226	635	72	154	31.86%
MONEY	88	4	92	2	2	50.00%
ORGANIZATION	236	200	436	77	123	38.50%
PERCENT	79	12	91	10	2	83.33%
PERSON	364	222	586	88	134	39.64%
TIME	23	9	32	9	0	100.00%
OPTIONAL	0	107	107	107	0	100.00%
All Tags	1632	1184	2816	619	565	52.28%

Table 4: Accuracies and rates of no extraction in "Total" according to the tag type

Tag	Accuracy	No extraction/Total
ARTIFACT	25.79%	55.01%
DATE	70.29%	12.70%
LOCATION	64.41%	11.34%
MONEY	95.65%	2.17%
ORGANIZATION	54.13%	17.66%
PERCENT	86.81%	10.99%
PERSON	62.12%	15.02%
TIME	71.88%	28.13%
OPTIONAL	0.00%	100.00%
All Tags	57.95%	21.98%

of the errors of KNP and the main reason of low recall.

3 Error Analysis of "No Extraction"

The target corpora we used consisted of six genres, "Q & A sites", "white papers", "blogs", "books", "magazines", and "newspaper articles", in BCCWJ.

Table 5 shows a summary of errors by genres of texts. These errors except "No extraction" are those that KNP output. "Correct" and "Error" are the number of the correct answers and the errors of KNP. "Total" is the sum of "Correct" and "Error". " No extraction" and " Errors with extraction" in the table mean the numbers of "No extraction" and the errors other than "No extraction", respectively. "No extraction rate" is the ratio of "No extraction" in "Error". "Docs" is the number of documents of the genre.

The total number of errors (1169) and total number of errors with extraction (550) are different from those in Tables 2 and 3 (1184 and 565). This is because some NEs that KNP output include multiple annotated NEs.

Table 6: Accuracies and rates of no extraction in "Total" according to the genre

Genre	Accuracy	No extraction/Total
Q & A	40.00%	44.21%
White paper	58.73%	20.63%
Blog	50.74%	27.89%
Book	50.35%	28.07%
Magazine	53.45%	14.66%
Newspaper	72.27%	15.49%
All	58.26%	22.10%

In addition, the number of words varies according to the genre. We think this is a reason why the total number of the NEs was not proportional to the number of the documents.

Table 5 shows that the genre whose no extraction rate was the highest was "Q & A sites" and the genre with the lowest rate was "magazines".

Table 6 shows the accuracies and the rates of no extraction in "Total" according to the genre. "Accuracy" is the ratio of the correct answers in "Total", the sum of correct answers and errors of KNP, and "No extraction/Total" is the ratio of no extraction in it. These errors except "No extraction" are those that KNP output. "Accuracy" of "All" (58.26%) is different from "Recall" in Table 1 (61.79%) because the number of the NEs KNP output was different from the number of the NEs that were annotated by humans.

Table 6 shows that "newspaper articles" is the genre whose accuracy is the highest. We think this is because KNP was trained with newspaper articles of MAINICHI SHIMBUN. Table 6 also shows the genre with the lowest accuracy was "Q & A sites". We think this is because the writing style of Q & A sites was the most different from that of newspaper articles. The same table shows that the genre whose no extraction rate was the highest was "Q & A sites"

Table 5: Summary of errors by genres of texts

Genre	Correct	Error	Total	No extraction	Errors with extraction	No extraction rate	Docs
Q & A	76	114	190	84	30	73.68%	74
White paper	427	300	727	150	150	50.00%	8
Blog	171	166	337	94	72	56.63%	34
Book	217	214	431	121	93	56.54%	5
Magazine	186	162	348	51	111	31.48%	2
Newspaper	555	213	768	119	94	55.87%	13
All Genres	1632	1169	2801	619	550	52.95%	136

and the genre with the lowest rate was “magazines”.

3.1 No Extraction of Q & A Sites

“Q & A sites” was the genre whose accuracy was the lowest. The examples of no extraction errors in “Q & A sites” are shown as follows.

i Many names of products, characters, and medicines were not extracted.
・サクラ大戦 (Sakura Wars)・スーパーファミコン (Super Nintendo Entertainment System) ・アクトレイザー (ActRaiser)・バイオハザード 4 (Resident Evil 4)・仮面ライダー (Kamen Rider)・ウルトラマン (Ultraman)・ガンダム (Gundam)・ミノスタシン (Minostacin) ・アスピリン (Aspirin)

ii Abbreviations were not extracted.
Formal names are noted in brackets.
・マリオワールド (Mario World)（スーパーマリオワールド (Super Mario World)）
・GC(ニンテンドーゲームキューブ (Nintendo GameCube))・JNB(ジャパンネット銀行 (Japan Net Bank)) ・LA(ロサンゼルス (Los Angeles))

iii The unusual date expressions were not extracted.
・九十／十一／二十一 (90/11/21)

iv Hiragana expressions were sometimes wrongly parsed.
・“さとし (Satoshi)” in “知恵ぶくらー・さとし (CHIEBUKURER Satoshi)” should be the name of person but it is wrongly parsed as “悟る (Satoru)”: a verb.

v NEs written in alphabets and numbers were not extracted.
・ＰＳ２ ・ＩＳＤＮ ・ＪＲ（“ＪＲ西 (JR East)” were extracted.)・Ｏｕｔｌｏｏｋ Ｅｘｐｒｅｓｓ

3.2 No Extraction of Newspaper Articles

“Newspaper articles” was the genre whose accuracy was the highest. The examples of no extraction errors in “newspaper articles” are shown as follows.

i Some NEs with specific prefixes and suffixes were not extracted.
・半～(half **, ex. half time) ・～圏 (** region, ex. 首都圏 (capital region), 三大都市圏 (three major metropolitan areas)) ・～地域 (** area)・～ポイント (** point) ・同～ (same **, ex. 同～年 (same ** year), 同日 (same day), 同年秋 (same year autumn))

ii OPTIONALs were not extracted because KNP does not extract optional tags.

iii The unusual English expressions in Japanese sentences were not extracted.
・KOERA ・JAPAN

iv Brackets sometimes cause the errors.
・【フェニックス（米アリゾナ州） (【Phoenix (Arizona, US))

v NEs that consist of general nouns were not extracted. This could be the reason why the names of products and characters were not extracted.
・昼寝 (Hirune, a nap)・ザウルス (Zaurus)・ファミリーマート (Family Mart) ・シャープ (Sharp)・ルネサンス (The Renaissance)
・“Softbank” sometimes could be extracted and sometimes could not. They were parsed as nominative case when they were extracted and as “in clause” when they were not.

4 Discussion

According to the examples described, we think that the lack of knowledge in the dictionary and the errors of the parser are the big reasons of the errors of the named entity recognition. In particular, the names of artifacts including the names of products or characters are often new words that were coined. These NEs are not in the dictionary KNP uses and therefore, they should be judged if they were the NEs or not depends on the features of the surrounding patterns and the syntactic features. As a result, the correct parsing would be important for the NEs that cannot use dictionary information. However, the casual writing style like Q & A sites causes the errors in

morphological analysis and parsing. We think that if the sentences of these informal writing styles could be correctly analyzed and parsed, the errors would be decreased. The training of texts with informal writing styles could be the solution of this problem. In addition, most of the NEs that were not extracted by KNP were found in Wikipedia or other Web sites. This information also could help the recall improve.

5 Conclusion

This paper reports an error analysis of the named entity recognizer KNP on six domains for revealing causes of errors. The texts of BCCWJ were manually annotated and compared with the automatically tagged texts. The analysis revealed that the most frequent error was "No extraction": the case where the tokens were not extracted by KNP though they were annotated. It also revealed that "No extraction" of "ARTIFACT" is the biggest cause of low recall and "Q & A site" is the genre whose accuracy is the lowest. We focused on the no extraction errors and found out that the lack of dictionary information and the various writing styles cause these errors.

Acknowledgements

This work was partially supported by JSPS KAKENHI Grant Number 24700138. We would like to thank Dr. Ryohei Sasano who provides us the helpful information about KNP and team members of NE team of Project Next NLP.

References

[1] Masaaki Ichihara, Maiko Yamazaki, and Kanako Komiya. Error analysis of named entity extraction in bccwj (bccwj における固有表現抽出のエラー分析). 第 7 回 コーパス日本語学ワークショップ 発表論文集, p. to appear, 2015.

[2] Ryohei Sasano and Sadao Kurohashi. Japanese named entity recognition using non-local information (in japanese). *IPSJ Journal*, Vol. 49, No. 11, pp. 3765–3776, 2008.

[3] 笹野遼平, 河原大輔, 黒橋禎夫, 奥村学. 構文・述語項構造解析システム knp の解析の流れと特徴. 言語処理学会, 第 19 回年次大会 発表論文集, pp. 110–113, 2013.