Vol.2015-NL-221 No.13
Vol.2015-SLP-106 No.13
2015/5/26

複数ジャンルを対象とした基本固有表現タグ付きコーパスの作成

岩倉 友哉[1,a)]　平田 亜衣[2,b)]　立花 竜一[2,c)]　山崎 舞子[3,d)]　市原 正陽[4,e)]　古宮 嘉那子[4,f)]

概要：本稿では，BCCWJ の複数ジャンルの文書を対象とした固有表現タグ付きコーパスを紹介する．本コーパスは，BCCWJ のコアデータに含まれる Yahoo!知恵袋，白書，Yahoo!ブログ，書籍，雑誌，新聞の 6 分野，136 文書から構成されており，IREX で定義された 8 種類の固有表現タグが合計 2,464 表現に付与されている．

Constructing a Japanese Basic Named Entity Corpus of Various Genres

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 with the eight types of NE tags defined by IREX. The NE corpus consists of six types of genres of documents such as blogs, magazines, white papers, and so on. The corpus contains 2,464 NE tags in total.

1. はじめに

固有表現 (Named Entity = NE) 抽出は，テキストに出現する人名や地名などの固有名詞や，日付や時間などの数値表現を抽出する技術である．たとえば，日本語の固有表現の種類としては，IREX [1] で定義された 8 種類および，約 200 種類が定義された拡張固有表現 [2] が提案されている．

IREX の定義においては，IREX の NE タスク向けに，毎日新聞を対象に作成したデータセット [*1] や，Web 文書の冒頭 3 文に固有表現タグを含む各種言語情報を人手で付与した京都大学 Web 文書リードコーパス [3] が作成されている．拡張固有表現においては，新聞記事に加え，複数ジャンルの文書を含む BCCWJ[4] に対してタグ付けされたコーパスが作成されている [5]. [*2]

本稿では，IREX の定義に基づく基本的な固有表現抽出のための複数ジャンルの文書に対象とした「BCCWJ 基本固有表現コーパス」を紹介する．本コーパスは，BCCWJ のコアデータに含まれる「Yahoo!知恵袋」,「白書」,「Yahoo!ブログ」,「書籍」,「雑誌」,「新聞」の 6 分野，136 文書に対し，IREX で定義された 8 種類の固有表現タグを付与したものである．これらには，従来の IREX の定義に基づき作成されたコーパスには含まれていない「書籍」や「雑誌」といったジャンルの文書が含まれており，IREX の定義に基づく固有表現抽出の評価の幅を広げることが期待される．このコーパスは，Project Next NLP [6] における固有表現抽出グループの活動で作成したものであり，Project Next NLP の固有表現抽出のページで公開する．入手および利用方法については，「付録：BCCWJ 基本固有表現コーパス利用方法」の節にて紹介する．

2. IREX の定義

今回のタグ付けに用いた IREX の固有表現抽出の定義の概要を述べる．詳しくは，IREX の固有表現抽出定義のページ [*3] を参照願いたい．

[1] 株式会社富士通研究所
[2] 首都大学東京
[3] 東京工業大学
[4] 茨城大学
[a)] iwakura.tomoya@jp.fujitsu.com
[b)] hirata-ai@ed.tmu.ac.jp
[c)] tachibana-ryuichi@ed.tmu.ac.jp
[d)] yamazaki@lr.pi.titech.ac.jp
[e)] 11t4004s@hcs.ibaraki.ac.jp
[f)] kkomiya@mx.ibaraki.ac.jp

[*1] 次から入手可能．http://nlp.cs.nyu.edu/irex/Package/IREXfinalB.tar.gz (2015 年 4 月 23 日確認)

[*2] 2015 年 4 月の時点では，言語資源協会から配布されている．http://www.gsk.or.jp/catalog/gsk2014-a/ (2015 年 4 月 23 日確認)

[*3] http://nlp.cs.nyu.edu/irex/NE/df990214.txt (2015 年 4 月 23

表 1 IREX で定義された固有表現クラスと例

固有表現クラス	例
ARTIFACT	ノーベル化学賞
LOCATION	日本
ORGANIZATION	外務省
PERSON	村山富市
DATE	5 月 5 日
MONEY	100 円
PERCENT	100 %
TIME	5 月 5 日

表 2 BCCWJ および IREX データの固有表現タグ付与数の内訳．データ数の項目は，BCCWJ は文書数，IREX データは記事数．IREX データは，毎日新聞を対象とした CRL_NE_DATA.idx (CRL), DRYRUN03.idx (DRY), NE-training981031.idx (NET)，ARREST_TRAIN.idx (AT)，ARREST01.idx (AR)，GENERAL03.idx (GE) の 5 種類のデータセット．Total の丸括弧の中の数値は OPTIONAL タグを除いた数．

BCCWJ

データ	データ数	固有表現タグ総数
Yahoo!知恵袋	74	175
白書	8	656
Yahoo!ブログ	34	307
書籍	5	399
雑誌	2	319
新聞	13	705
Total	136	2,561 (2,464)

IREX

データ	データ数	固有表現タグ総数
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)

IREX では，ARTIFACT （製品名，法律名などの固有物名），LOCATION（場所表現），ORGANIZATION（組織名），PERSON （人名），DATE （日付表現），MONEY（金額表現），PERCENT （ 割合表現 ），TIME （時間表現）の合計 8 種類の固有表現クラスが定義されている．表 1 に IREX の固有表現の例を示す。IREX の定義では，表記によらず，文脈によって決まる意味に基づき，抽出を行なう必要がある．そこで，以下の「宮崎」が 2 回出現する文では，次のようなタグ付けを行なう．

⟨*LOCATION*⟩ 宮崎 ⟨/*LOCATION*⟩ 出身の ⟨*PERSON*⟩ 宮崎 ⟨/*PERSON*⟩ さん．

さらに，IREX の定義では，タグ付けが困難と判断された場合には，「OPTIONAL」というタグを用いることも可能である．

3. BCCWJ 基本固有表現コーパスの作成

BCCWJ のコアデータに含まれる 136 文書 [*4] を対象とし，以下の手順でコーパスの作成を行なった．

- タグ付け： 各メンバーの担当文書を決めてタグ付け．この時点では，各文書には 1 名の担当者だけを割当てて実施．
- タグ修正： 最初にタグ付けした結果を集約後に，メンバーが 1 名ずつ順に，全体を確認し修正するという流れで実施．
- 配布準備：タグ情報だけを抜き出し，BCCWJ を用意すれば利用できる形にパッケージング．

表 2 に，2015 年 4 月 14 日版における，各ジャンルのデータセットに含まれるデータ数および固有表現の総数を載せる．また，比較のために，IREX データセットの内訳を載せる．今回作成したコーパスには，「書籍」や「雑誌」 といった IREX が対象とする新聞記事や京都大学 Web 文書リードコーパスが対象とする Web ページとは異なるジャンルのデータが含まれている．また，固有表現の総数としては，IREX の本試験の評価データである「総合課題 (GE)」および 「限定課題 (AR)」より多い．

表 3 に各データの固有表現の内訳を，表 4 に各データにおける各固有表現クラスの全体に占める割合を示す．これらの表から，BCCWJ の「新聞」や毎日新聞に対してタグ付けされた IREX データと比較し，「Yahoo!知恵袋」 や「白書」は ARTIFACT の割合が多いといった特徴や，「雑誌」は人名の占める割合が高いといった，異なる特徴を持つデータセットであることがわかる．

4. BCCWJ 基本固有表現コーパスの利用例

本節では，本固有表現タグ付きコーパスの利用例を紹介する．

4.1 IREX の定義に基づく固有表現抽出器の評価

利用例の一つとして，IREX の定義に基づき固有表現を抽出する KNP の評価結果を紹介する．今回の評価では，形態素解析器 JUMAN [*5] のバージョン 7.01 および KNP の [*6] バージョン 4.12 を利用した．

表 5 は，2015 年 4 月 14 日版における KNP による固有表現抽出の精度である．評価には，以下の Recall，Precision，F-measure を用いた．

[*4] 対象の文書一覧は次のページにて参照できる．(2015 年 4 月 23 日確認) http://plata.ar.media.kyoto-u.ac.jp/mori/research/NLR/JDC/ClassA-1.list

[*5] JUMAN のページ (2015 年 4 月 23 日確認) http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN

[*6] KNP のページ (2015 年 4 月 23 日確認) http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP

表 3 BCCWJ および IREX データの固有表現タグ付与数の内訳．ART は ARTIFACT，LOC は LOCATION，OPT は OPTIONAL，ORG は ORGANIZATIO の略．その他の各項目は，表 2 と同じ意味．

BCCWJ

データ	ART	DATE	LOC	MONEY	OPT	ORG	PERCENT	PERSON	TIME
Yahoo!知恵袋	54	19	57	9	8	19	0	6	3
白書	163	129	140	9	39	128	33	15	0
Yahoo!ブログ	25	60	52	7	9	61	11	79	3
書籍	29	50	87	0	24	26	6	169	8
雑誌	13	42	32	5	4	17	1	203	2
新聞	24	165	188	59	13	118	38	78	22
Total	308	465	557	89	97	369	89	550	37

IREX

データ	ART	DATE	LOC	MONEY	OPT	ORG	PERCENT	PERSON	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

表 4 BCCWJ および IREX データの固有表現タグの割合の内訳．その他の項目は，表 2 と表 3 と同じ意味．

BCCWJ

データ	ART	DATE	LOC	MONEY	OPT	ORG	PERCENT	PERSON	TIME
Yahoo!知恵袋	30.86%	10.86%	32.57%	5.14%	4.57%	10.86%	0%	3.43%	1.71%
白書	24.85%	19.66%	21.34%	1.37%	5.95%	19.51%	5.03%	2.29%	0%
Yahoo!ブログ	8.14%	19.54%	16.94%	2.28%	2.93%	19.87%	3.58%	25.74%	0.98%
書籍	7.27%	12.53%	21.80%	0%	6.02%	6.52%	1.50%	42.35%	2.01%
雑誌	4.08%	13.17%	10.03%	1.57%	1.25%	5.33%	0.31%	63.63%	0.63%
新聞	3.40%	23.40%	26.68%	8.37%	1.84%	16.74%	5.39%	11.06%	3.12%

IREX

データ	ART	DATE	LOC	MONEY	OPT	ORG	PERCENT	PERSON	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%

- Recall = NUM / (抽出すべき正しい固有表現の数)
- Precision = NUM / (KNP が抽出した固有表現の数)
- F-measure = 2 × Recall × Precision / (Recall + Precision)

NUM は KNP により正しく抽出された固有表現の数である．

「Yahoo!知恵袋」や「Yahoo!ブログ」といったデータと比較し，「新聞」では，高い精度が得られている．この理由の一つとしては，KNP が使用している主な学習データは新聞記事にタグ付けされた IREX の CRL データであるのに対し，「Yahoo!知恵袋」や「Yahoo!ブログ」には，新聞記事にはあまり含まれない口語表現や省略表記が含まれるためだと考えられる．また，「白書」は，新聞記事と同様に，正しい書き言葉で記載されていると考えられるが，「新聞」と比較し，低い精度となった．この理由の一つとしては，「白書」が「新聞」においてもあまり高い精度が得られない ARTIFACT を多く含むことが理由の一つと考えられる．このように，複数ドメインの文書を用いることで，異なる視点での評価が行なえると期待される．

4.2 Project Next NLP での利用例

Project Next NLP の活動では，BCCWJ 基本固有表現

表 5 BCCWJ 基本固有表現コーパス (2015 年 4 月 14 日版) の各ジャンルのデータにおける KNP の評価．各項目の数値は，”F-measure (Recall, Precision)”．

NE / データ	Yahoo!知恵袋	白書
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 / データ	Yahoo!ブログ	書籍
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 / データ	雑誌	新聞
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)

コーパスを次の分析に利用した．

- KNP の誤りパタンの分析 [7]: 固有表現の範囲認識の誤り，固有表現のタイプ判別誤り，抽出漏れといった誤りパタンの観点からの KNP の分析．
- ドメイン別データの学習による精度評価および KNP との比較 [8]: 異なるドメインのデータでの精度調査および学習の効果調査．
- 固有表現抽出における辞書利用に関する調査 [9]: BCCWJ 基本固有表現コーパスに出現した固有表現を形態素解析器の辞書に登録した後の精度変化の調査および辞書登録後に残る誤りの分析．

4.3 その他の利用方法案

本コーパスには，複数ジャンルの文書が含まれていることから次のような研究にも利用できると期待される．

- 口語を含む文書における固有表現抽出の研究：たとえば，「Yahoo!ブログ」は，「新聞」や「白書」と異なり，口語表現が多く含まれるため，書き言葉だけでなく，口語表現を対象とした固有表現抽出の研究にも利用できると期待される．
- 転移学習 [10] の研究：複数ジャンルが含まれるため，新聞記事から白書といった異なるジャンルへの適用といった場面を想定した研究を行なうためのデータとして利用できると期待される．

5. まとめ

本稿では，Project Next NLP における固有表現抽出グループの活動において作成した BCCWJ 基本固有表現コーパスを紹介した．本コーパスは BCCWJ を準備すれば利用できる形にて公開している．

付録：BCCWJ 基本固有表現コーパス利用方法

本コーパスは，BCCWJ とタグ情報が含まれたパッケージを用意した後，以下の手順で復元できる．その他の情報は，公開ページや，パッケージに含まれる README.txt を参照願いたい．

(1) BCCWJ を用意．
(2) 次のページからコーパス復元のためのパッケージを入手．2015 年 4 月時点では，2015 年 4 月 14 日版が最新．
https://sites.google.com/site/projectnextnlpne/

(3) ダウンロードしたパッケージを展開し，作成されたディレクトリに移動．その後，復元のための perl のスクリプトを実行．

- Unix 系 OS であれば，次のように実行．-d で指定するのは，BCCWJ のコアデータのディレクトリ．
  - perl tools/gendata.prl -d core_M-XML
- Windows の場合 -w を指定し，次のように実行．
  - perl tools\gendata.prl -d core_M-XML -w

(4) 実行後に，そのディレクトリ以下に，nedata というディレクトリが作成される．「ne」で終わるファイルが固有表現タグ付き文書である．

謝辞

Project Next NLP での固有表現抽出班の活動にあたり，京都大学 Web 文書リードコーパスを，京都大学の河原准教授からご提供いただきました．また，東工大の笹野助教からは，KNP の実装の詳細につきまして，お教えいただきました．ここに感謝の意を表します．

参考文献

[1] IREX Committee: *Proc. of the IREX workshop* (1999).
[2] Sekine, S., Sudo, K. and Nobata, C.: Extended Named Entity Hierarchy, *Proc. of LREC'02* (2002).
[3] 萩行正嗣，河原大輔，黒橋禎夫：多様な文書の書き始めに対する意味関係タグ付きコーパスの構築とその分析，自然言語処理，Vol. 21, No. 2, pp. 213–247 (2014).
[4] 前川喜久雄：KOTONOHA 『現代日本語書き言葉均衡コーパス』の開発 (〈特集〉資料研究の現在)，日本語の研究，Vol. 4, No. 1, pp. 82–95 (2008).
[5] 橋本泰一，乾　孝司，村上浩司：拡張固有表現タグ付きコーパスの構築，情報処理学会研究報告. 自然言語処理研究会報告，Vol. 2008, No. 113, pp. 113–120 (2008).
[6] 関根　聡：Project Next NLP 概要 (2014/3　2015/2) (2015).
[7] Ichihara, M., Komiya, K., Iwakura, T. and Yamazaki, M.: Error Analysis of Named Entity Recognition in BCCWJ, エラー分析ワークショップ (言語処理学会年次大会 2015) (2015).
[8] 平田亜衣，小町　守：様々なジャンルのテキストに対する固有表現認識の分析，エラー分析ワークショップ (言語処理学会年次大会 2015) (2015).
[9] 岩倉友哉：固有表現抽出におけるエラー分析，エラー分析ワークショップ (言語処理学会年次大会 2015) (2015).
[10] 神嶌敏弘：転移学習，人工知能学会誌，Vol. 25, No. 4, pp. 572–580 (2010).