BCCWJ における固有表現抽出のエラー分析

市原　正陽（茨城大学工学部　情報工学科）
山崎　舞子（東京工業大学　大学院総合理工学研究科）
古宮　嘉那子（茨城大学工学部　情報工学科）

Error Analysis of Named Entity Extraction in BCCWJ

Masaaki Ichihara(Department of Computer and Information Sciences, Ibaraki University)
Maiko Yamazaki(InterdisciplinaryGraduate School of Science and Engineering,
Tokyo Institute of Technology)
Kanako Komiya(Department of Computer and Information Sciences, Ibaraki University)

要旨

テキスト中に含まれる固有表現を正しく認識することは，自然言語で書かれたテキストに含まれる情報を誤りなく取得するうえで必要である．よって，本研究では「現代日本語書き言葉均衡コーパス」よりランダムサンプリングをしたテキストを京都大学の「日本語構文・格・照応解析システム KNP」にかけ，その結果に含まれるエラーの分析を行った．分析結果から，KNP の固有表現抽出機能が固有表現の抽出を誤るのは，形態素解析や構文解析の誤り，辞書の知識不足が大きな要因と考えられることが分かった．

１．はじめに

固有表現抽出とは，テキストの中から人名や地名，商品名などの固有表現を自動的に抽出する処理である．しかし，誤った情報を抽出することや，本来抽出したい固有表現が抽出できないことがままある．そのため，本稿では，現在の固有表現抽出システムを使用して得られたエラーに対してエラー分析を行う．

２．使用システムおよび使用コーパス

日本語のコーパスとして「現代日本語書き言葉均衡コーパス」（BCCWJ）（Maekawa (2008)）を用いる．システムは固有表現を抽出するために「日本語構文・格・照応解析システム KNP[1]」（KNP）を使用する．KNP では CRF を用いた系列ラベリングに基づいて固有表現の解析を行っている．また KNP では，固有表現抽出を行う際の素性として形態素情報のほかに「キャッシュ素性」や「係り先素性」などを使用している（笹野ら（2008））．

また，本研究では固有表現を分類するために Information Retrieval and Extraction Exercise[2]（IREX）で定義された組織名，人名，地名，固有物名，日付表現，時間表現，金額表現，割合表現，オプショナルの 9 つの固有表現を使用した．

[1] http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP
[2] http://nlp.cs.nyu.edu/irex/index-j.html

３．BCCWJ における固有表現抽出のエラー分析手法

３．１　BCCWJ における KNP のエラー分析

今回エラーの分析をするにあたって BCCWJ のうち「YAHOO!知恵袋」「白書」「YAHOO!ブログ」「書籍」「雑誌」「新聞」の 6 つからランダムサンプリングした計 136 個のテキストに対して人手によって IREX で定義された 9 つの固有表現タグを付けた．これを正解として比較を行っていく．また， KNP の固有表現の解析を行うオプションである-ne を使うことで，それらのテキストの平文から固有表現タグの付いた平文を出力した．その後，それらの人手と KNP のタグが付けられたテキストのペアを比較することでエラーに対して分析を行った．

３．２　BCCWJ コーパスへの IREX のタグ付け

IREX の固有表現タグの人手による付与は，テキストを 5 分割したものに対して Project Next NLP の NE のタスクのメンバー5 人がそれぞれタグ付けを行った．5 分割したテキスト群のうちの一つを対象とする時にはそれぞれ「hi」「ichi」「iwa」「ko」「ta」とする．

３．３　BCCWJ コーパスにおけるエラー抽出

人の手によってタグの付けられたテキストと KNP によってタグの付けられたテキストの比較を行い，エラーの種類によって分類して分析を行った．

４．BCCWJ における固有表現抽出のエラー結果

４．１　KNP が付与したタグの正解率

表 1 に KNP の付けたタグ全体の正解していた数，不正解していた数と正解の割合を示す．

表 1 固有表現の正解不正解の内訳

	正解	不正解	総数	正解率
hi	297	194	491	60.49%
ichi	195	99	294	66.33%
iwa	303	187	490	61.84%
ko	385	385	770	50.00%
ta	452	319	771	58.63%
総数	1632	1184	2816	57.95%

KNP の付けた固有表現タグは半分以上が人手で付けたものと一致した．

４．２　タグの範囲に対する分析

タグの範囲に対する分類として，以下の 5 種類に分類を行った．

KNP なし：KNP は固有表現として抽出しなかったが，正解は固有表現だったもの
人手なし：KNP は固有表現として抽出したが，正解は固有表現ではなかったもの
範囲別：KNP は固有表現として抽出したが，正解と固有表現の範囲だけが異なっていたもの
タグ別：KNP は固有表現として抽出したが，正解と固有表現の種類だけが異なっていたもの
両方別：KNP は固有表現として抽出したが，正解と固有表現の範囲，種類がともに異なっていたもの

比較方法としては文字位置が人手で付けたタグの範囲よりも KNP が狭い範囲でタグをつけていたもの，人手で付けたタグの範囲よりも KNP が広い範囲でタグをつけていたもの，人手で付けたタグの範囲と KNP が付けたタグの範囲が一部分だけ被っているものは，それぞれ別々のエラーとしてカウントした．

そのため一方では一つの固有表現としてタグが付けられたものが，もう一方では分割されて固有表現としてタグが付けられていた場合，分割されている方の数だけエラーとしてカウントされている．その例を図 1 として以下に示す．

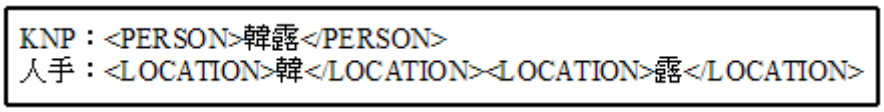

図 1 人手で付けた固有表現が KNP の出力した固有表現の中に 2 つ入っている例

図 1 と同様に KNP の出力した固有表現が人手で付けた固有表現の内側に入っていても，同じように分割されている方をカウントする．

KNP の付けたタグと人手で付けたタグの比較を行った結果を表 2 に示す．

表 2 KNP のエラーの内訳

	KNP なし	人手なし	範囲別	タグ別	両方別	エラー総数
hi	98	33	34	15	14	194
ichi	48	21	16	6	8	99
iwa	133	30	14	3	7	187
ko	212	34	38	72	29	385
ta	128	41	60	31	59	319
総数	619	159	162	127	117	1184

結果から，5 分割したすべてにおいて，KNP がタグをつけられていないエラーの数が最も多く，全体の半分以上のエラーがこれに含まれていた．次に多かったのは，タグは同様のものが付けられているが，付けられている範囲が異なっているものだった．このうち，一部分だけが被っているエラーはごく少数で，その内のほとんどは人手で付けたタグの範囲の方が広かった．

４．３　KNP が誤って付けたタグに対する分析

表 3 には KNP がタグを付けた中で，人手で付けたものと違っていたものの内訳を示す．表 3 にある 8 つの固有表現タグは，KNP によって付けられていた固有表現タグである．

ORG：ORGANIZATION，組織名，政府組織名を表す
PERS：PERSON，人名を表す
LOC：LOCATION，地名を表す
ART：ARTIFACT，固有物名を表す
DATE：DATE，日付表現を表す
TIME：TIME，時間表現を表す
MONEY：MONEY，金額表現を表す
PERC：PERCENT，割合表現を表す

表 3 タグごとの内訳

	ORG	PERS	LOC	ART	DATE	TIME	MONEY	PERC	総数
hi	27	6	19	14	30	0	0	0	96
ichi	8	34	3	3	3	0	0	0	51
iwa	22	5	16	6	1	0	2	2	54
ko	31	37	76	9	20	0	0	0	173
ta	35	52	40	35	29	0	0	0	191
総数	123	134	154	67	83	0	2	2	565

この結果から，「TIME」「MONEY」「PERCENT」に関しては，KNP は間違って固有表現タグを付けることが少ないことがわかる．また，「ARTIFACT」や「DATE」に関しても誤っているものがあるが，合わせて KNP が誤って固有表現タグを付けたもののうち 3 割に満たなかった．そして，KNP が固有表現タグを付けた誤りのうち「ORGANIZATION」「PERSON」「LOCATION」の 3 つが，誤りの大部分を占めていることが分かった．

５．KNP が固有表現タグを付与できなかったエラーに対する分析

表 2 から分かるように KNP が固有表現のタグを付ける際に出るエラーの中で最も数が多いのは，KNP が固有表現のタグを付けられないエラーだったため，それに関して分析を行った．

５．１　各タスクのエラーの割合

今回エラーを取得するために使用したテキストは BCCWJ のコアデータである「OC」「OW」「OY」「PB」「PM」「PN」の 6 つで，それぞれ「YAHOO!知恵袋」「白書」「YAHOO!ブログ」「書籍」「雑誌」「新聞」の 6 つのタスクから取得されたものである．それらのタスクごとのエラーの割合を表 4 に示す．

タグ無：KNP がタグを付けなかったエラーの数
タグ有：KNP がタグを付けたエラーの数（範囲の間違い，タグの間違いも含む）
タグ無割合：不正解の合計数に対する KNP がタグを付けなかったエラーの割合

表 4 タスクごとのエラーの割合[3]

all	正解	タグ無	タグ有	合計	不正解の合計	タグ無割合	文書数
YAHOO!知恵袋	76	84	30	190	114	73.68%	74
白書	427	150	150	727	300	50.00%	8
YAHOO!ブログ	171	94	72	337	166	56.63%	34
書籍	217	121	93	431	214	56.54%	5
雑誌	186	51	111	348	162	31.48%	2
新聞	555	119	94	768	213	55.87%	13
合計	1632	619	550	2801	1169	52.95%	136

[3] 表 3 ではタグの付けられたエラーの総数が 565 個だったものが表 4 では 550 個になっているのは，表 1 では人手と KNP 両方からみたエラーの数を表おり，表 4 では KNP のエラーに関してのみ注目しているため．

表 4 で文書数と合計数に比例関係がないのは，一つの文書内にある文字数がジャンルによって大きく異なるためである．また，それぞれのジャンルの内「YAHOO!知恵袋」が最も不正解の中でタグを付けられないエラーの割合が多く，逆に「雑誌」が一番タグを付けられないエラーの割合が低かった．

５．２　各タスクの正解率

「YAHOO!知恵袋」「書籍」「YAHOO!ブログ」「書籍」「雑誌」「新聞」それぞれの正解率と全体の合計に対するタグ無の割合を表 5 に示す．

タグ無割合：正解，不正解両方の合計数に対する KNP がタグを付けなかったエラーの割合

表 5 タスクごとの正解率とタグ無の割合

all	正解率	タグ無割合	精度	再現率	F 値
YAHOO!知恵袋	40.00%	44.21%	71.70%	43.93%	54.48%
白書	58.73%	20.63%	74.00%	63.35%	68.27%
YAHOO!ブログ	50.74%	27.89%	70.37%	55.70%	62.18%
書籍	50.35%	28.07%	70.00%	52.54%	60.03%
雑誌	53.45%	14.66%	62.63%	57.76%	60.10%
新聞	72.27%	15.49%	85.52%	73.80%	79.23%
合計	58.26%	22.10%	74.79%	61.79%	67.68%

表 5 から分かるように「新聞」の正解率が一番高かった．また「YAHOO!知恵袋」の正解率が一番低く，そのほかのタスクの正解率はその 2 つと比べると，正解率の差は少なかった．「新聞」の正解率が一番高かったのは，KNP は毎日新聞データを訓練事例としているためだと考えられる．また，「YAHOO!知恵袋」のタスクが 6 つのタスクの中で最も正解率が低いのは，新聞と文体が遠いからではないかと考えられる．また，正解，不正解の内のタグ無の割合は「雑誌」の割合が最も低く，「YAHOO!知恵袋」の割合が最も高かった．

５．３　固有表現タグの付けられなかった形態素の分析

表 5 の正解率から，最も割合の低かった「YAHOO!知恵袋」と最も割合の高かった「新聞」に含まれる形態素に対して分析を行った．

５．３．１　「YAHOO!知恵袋」内の固有表現タグの付けられなかった形態素の分析

ⅰ.商品名やキャラクター名が取れない事が多い．

実際に取れなかった商品名やキャラクター名，薬品名の一部

・サクラ大戦　・スーパーファミコン　・アクトレイザー　・バイオハザード 4

・仮面ライダー　・ウルトラマン　・ガンダム　・ミノスタシン　・アスピリン

ⅱ.略されたものが取れない．

ⅰの影響が強いのかもしれないが，略された商品名も取れていない．

・スーパーマリオワールドは取れてマリオワールドは取れない

・GC(ニンテンドーゲームキューブ)　・JNB(ジャパンネット銀行)　・LA(ロサンゼルス)

ⅲ.特殊な日付の表現が取れない．

・九十／十一／二十一

ⅳ.ひらがなで表記されていると誤って解析してしまう

”知恵ぶくらー・さとし”と記述されたファイルがあり，本来”さとし”は PERSON と取って欲しいのだが，動詞の”悟る”として解析されていた.

ⅴ.略称でなくてもアルファベットやアラビア数字と組み合わさったものが取れない

・ＰＳ２　・ＩＳＤＮ　・ＪＲ（ＪＲ西となった部分は正しく取れていた）

・Ｏｕｔｌｏｏｋ　Ｅｘｐｒｅｓｓ

５．３．２　「新聞」内の固有表現タグの付けられなかった形態素の分析

Ⅰ.基本的に取れないものがある

・半～(時間表現など様々)　・～圏(首都圏，三大都市圏)　・～地域　・～ポイント

・同～(同～年，同日，同年秋)

半日や首都圏，ユーロ地域などが誤りとして確認でき，正解には含まれていなかった.

ただし，半分は PERCENT として取得できていた.

Ⅱ.英語や日本語などを OPTIONAL として取れなかった.

本来「<OPTIONAL>英</OPTIONAL>語」「<OPTIONAL>日本</OPTIONAL>語」のように取れてほしい．しかしそもそも KNP の機能として OPTIONAL と付ける機能はない.

Ⅲ.英語表記で書かれることが少ないものが取れなかった

・KOERA　・JAPAN

Ⅳ.付近にその形態素に関する情報があっても（があると取れなかった.

・【フェニックス（<LOCATION>米アリゾナ州</LOCATION>）

Ⅴ.一般名詞やそれが組み合わさったようなものは取れないことが多かった.

ｉ（商品名やキャラクター名が取れないことが多い）の原因も同様である可能性がある

・昼寝　・ザウルス　・ファミリーマート　・シャープ　・ルネサンス

(ソフトバンクが取れている所と取れていないところがあった．取れているものはガ格に，取れていないものは文節内と解析されていた．)

６．考察

分析から，KNP の固有表現抽出機能が固有表現の抽出を誤るのは，形態素解析や構文解析の誤り，辞書の知識不足が大きな要因と考えられる．特に固有物名(ARTIFACT)は商品名などが対象となるため，他の固有表現より造語が分類されやすく，その場合一般名詞の組み合わされたパターンが分類される可能性が高いと考えられる．そのため KNP の場合先行文脈やその単語に対する係り受けの関係などからその単語が固有表現なのか推察しなければならず，正しい構文解析は重要である.

また，構文解析するにあたっても新聞などより口語的なものを扱う可能性も十分あり，そういった場合，助詞が抜けている事などが構文解析の妨げとなる事は多いと推察できる.

そのため，新聞とは書かれ方の大きく異なる文書からも学習することで，特定ジャンルでない文書から固有表現を抽出しようとする場合効果的である可能性が高い．また，取ることのできなかった固有表現の大半が wikipedia などネット上に情報があることが確認できたため，それらを辞書に取りこむことでより正確な固有表現抽出の実現が期待できる.

謝辞

本研究は，文部科学省科学研究費補助金[若手 B(No:24700138)]の助成により行われました．ここに，謹んで御礼申し上げます．

また，KNP についての質問に快く答えてくださった，東京工業大学の笹野遼平先生に謹んで御礼申し上げます．

また，Project Next NLP の NE 班の班長である岩倉友哉先生をはじめ，班員の皆様方には多くのご協力をいただきました．謹んで御礼申し上げます．

参考文献

[1]笹野遼平，黒橋禎夫(2008)「大域的情報を用いた日本語固有表現認識」情報処理学会論文誌,Vol.49No.11,pp.3765-3776

[2]笹野遼平，河原大輔，黒橋禎夫，奥村学(2013)「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

[3] Kikuo Maekawa(2008). Balanced corpus of contempo-rary written japanese. In ALR 2008, pp. 101-102

付録

今回対象としたBCCWJのコアデータ内の136ファイル

YAHOO! 知恵袋	OC01_00001	OC01_00002	OC01_00003	OC01_00004	OC01_00005
	OC01_00006	OC01_00007	OC02_00001	OC02_00002	OC02_00003
	OC02_00004	OC02_00006	OC02_00007	OC02_00008	OC03_00001
	OC03_00005	OC04_00001	OC04_00002	OC04_00003	OC05_00001
	OC05_00003	OC05_00004	OC05_00006	OC06_00001	OC06_00008
	OC08_00001	OC08_00002	OC08_00004	OC08_00006	OC09_00001
	OC09_00002	OC09_00003	OC09_00004	OC09_00006	OC09_00008
	OC10_00001	OC10_00003	OC10_00005	OC10_00006	OC10_00007
	OC11_00001	OC11_00002	OC11_00004	OC11_00005	OC11_00006
	OC11_00007	OC12_00002	OC12_00003	OC12_00004	OC12_00005
	OC12_00006	OC12_00007	OC12_00008	OC13_00001	OC13_00002
	OC13_00003	OC13_00004	OC13_00005	OC13_00006	OC13_00007
	OC13_00008	OC14_00001	OC14_00003	OC14_00004	OC14_00005
	OC14_00006	OC14_00007	OC14_00008	OC15_00001	OC15_00002
	OC15_00004	OC15_00006	OC15_00007	OC15_00008	
白書	OW6X_00000	OW6X_00002	OW6X_00003	OW6X_00007	OW6X_00008
	OW6X_00009	OW6X_00011	OW6X_00013		
YAHOO! ブログ	OY01_00082	OY01_00137	OY01_00148	OY01_00185	OY02_00095
	OY04_00001	OY04_00027	OY04_00173	OY06_00060	OY06_00146
	OY06_00168	OY07_00097	OY07_00135	OY07_00164	OY08_00115
	OY08_00137	OY08_00156			
書籍	PB11_00006	PB12_00001	PB22_00002	PB43_00001	PB59_00001
雑誌	PM11_00002	PM24_00003			
新聞	PN1a_00002	PN1d_00001	PN1d_00002	PN1f_00002	PN1g_00002
	PN2c_00002	PN2g_00002	PN3b_00001	PN3c_00002	PN4b_00001
	PN4c_00001	PN4c_00002	PN4f_00001		