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

平田 亜衣　　小町 守
首都大学東京 システムデザイン研究科
{hirata-ai@ed, komachi@}tmu.ac.jp

1 はじめに

固有表現認識（Named Entity Recognition）とは，情報抽出の技術の一つであり，テキストから人名，地名，商品名などといった固有表現と呼ばれる表現を自動的に認識する処理のことである．固有表現認識は自然言語処理の重要な技術であり，生の文章からの文書要約や，質問応答などで必要不可欠である．

固有表現抽出はタグ付きコーパスとして新聞記事が十分な量あるので，テストデータが新聞記事であれば教師ありの学習が適応できる．しかし，Yahoo!知恵袋などで出現するような表現は新聞記事に登場しないものがあるため，うまく対応できない表現があるといったの問題点がある．

実際に日本語解析システムの KNP （バージョン 4.12）の固有表現モジュールを用いて，現代日本語書き言葉均衡コーパス（BCCWJ）[1] の Yahoo!知恵袋の文章の一部に固有表現抽出を行った例が表 1 である．入力の文章に対して，“バンプレスト” にしか PERSON の固有表現しか付かないが，“仮面ライダー” や “ウルトラマン” には ARTIFACT のタグが付くべきである．また，“バンプレスト” も企業名であり，PERSON よりも ORGANIZATION のタグが付くべきである．このように新聞記事で学習された固有表現認識器は，新聞記事には出現しない商品名や企業名に対しては取りこぼしが発生したり，企業名に “さん” をつけるなどの口語的表現に対しては人名と誤認識してしまうという問題がある．また他の誤りとして日付表現などが抽出できないなどの問題が確認された．

トレーニングとテストのデータのジャンルが違うことから起こる，固有表現認識がうまくいかない原因として考えられるものとして，形態素解析などの下のレイヤーの解析誤り，テストデータのジャンルの固有表現タグ付きコーパスの不足，テストデータのジャンルの固有表現辞書の不足，などが考えられる．

この論文では固有表現タグ付きコーパスの不足に焦点を当て，テストデータと同じジャンルのトレーニングデータを用いることで既存の手法での抽出では抽出できなかった誤りや，間違った誤りを分析することを目的とする．

表 1: BCCWJ を KNP で固有表現認識した例

スーファミに詳しい方！ <PERSON> バンプレスト </PERSON> さんのソフトで、仮面ライダーやウルトラマン、ガンダムが２頭身で一緒になって戦うソフトの名前なんでしたっけ??

2 関連研究

固有表現抽出の手法は Conditional Random Fields（CRF）[2] や Support Vector Machine（SVM）[3] を用いた機械学習に基づく手法がよく使用されている．

KNP の固有表現モジュールを作成した 笹野ら [4] は，SVM を用いて固有表現抽出を行っている．笹野らの研究では Web 文書をテストデータにして実験を行っているが，トレーニングデータは新聞記事のものになっているために，固有表現認識における Web テキストをトレーニングデータとした場合の性能は分からない．[1]

また Kazama ら [5] は Wikipedia や Web 文書から EM ベースでのクラスタリングによって大規模なクラスタ情報を抽出し，大規模な固有名詞に関する辞書を構築し，それを CRF での固有表現認識に適応させた実験を行っている．しかしこの研究も固有表現を Web データから抽出しているものの，新聞記事をトレーニングデータ・テストデータとしているために，Web データをトレーニングデータに用いた場合の影響は分からない．

[1] KNP の固有表現抽出のモジュールは公開されており，このモジュールは CRF を用いている．そしてトレーニングデータとして新聞記事を対象としたものになっている．

Web 文書からの固有表現認識を行ったものとして，新里ら [6] がレストラン属性をつけたタグ付きコーパスを作成し，SVM を用いて固有表現認識を行っている．しかし，この研究は Web データを用いて固有表現認識器を学習しているものの，後述する IREX で定義された固有表現ではなく，レストラン分野に特化したものになっている．

3 実験

3.1 固有表現

日本語の固有表現の種類は，IREX [7] で定義された 8 種類のものと，さらに細かく分類した 200 種類の拡張固有表現 [8] が提案されている．今回は IREX で定義された固有表現を用いて実験を行う．

IREX で定義された固有表現は “ORGANIZATION”, “PERSON”, “LOCATION”, “ARTIFACT”, “DATE”, “TIME”, “MONEY”, “PERCENT” の 8 種類あり，重複や入れ子表現のない唯一のタグを割り当てる．表 9 は IREX で定義された固有表現とその例である．そして，例えば “日本銀行” のように，地名としての “日本” と組織名としての “日本銀行” とで表現が重なっている場合は，原則的に長い単位の表現，つまり “日本銀行” 全体を組織名として抽出する．[2]

3.2 BCCWJ・京都大学 Web リードコーパス・CRL 固有表現データを用いた固有表現認識

今回使用するデータとして現代日本語書き言葉均衡コーパス (BCCWJ)，京都大学 Web 文書リードコーパス Version 0.9 [9]，CRL 固有表現データの 3 つを用いる．

BCCWJ の今回使うデータとして，コアデータの Class-A のファイルに対して Peject Next NLP [10] の固有表現タスクで 5 名によって固有表現が付与されたデータを使用する．表 2 は Class-A のデータに含まれる 6 種類のデータごとの文数と固有表現の数である．表 10 では実際にタグ付けした一部の例文を挙げる．

また，表 3 は固有表現の種類による内訳である．BCCWJ には様々なジャンルのデータが含まれており，ブログや知恵袋などといった，新聞とは違った砕けた表現で書かれているものや，新聞や白書などといった固い表現で書かれたものなど多様なジャンルの文書が含まれている．表 3 から分かるように，新聞（PN）では LOCATION と DATE が多いが，白書（OW）と知恵袋（OC）で LOCATION の次に多いのは ARTIFACT である．また，書籍（PB）・雑誌（PM）・ブログ（OY）の 3 ジャンルにおいて最も多い固有表現は PERSON であり，いずれのジャンルにおいても新聞とは異なった固有表現の分布をしており，新聞記事をトレーニングデータとして学習された固有表現認識器が別ジャンルでは誤認識しやすいことが予想される．

表 2: BCCWJ の Class-A ファイルに対してのジャンル・文数とタグ付けされた固有表現の数

	ジャンル	文数	固有表現の数
OW	白書	504	625
PB	書籍	511	375
PN	新聞	505	685
PM	雑誌	492	314
OC	Yahoo!知恵袋	504	166
OY	Yahoo!ブログ	515	293
計		3,031	2,458

京都大学 Web 文書リードコーパスには様々な Web 文書が含まれており，そのリード（冒頭）3 文，計 7,500 文に対して形態素情報や IREX の基準に基づいて固有表現の情報が付与されたものを使用する．

CRL 固有表現データには毎日新聞の記事約 1 万文に対して固有表現のタグ付けをしたものである．KNP はこのデータを用いてトレーニングしている．

3.3 素性とモデル

本研究では CRF によって固有表現認識器を作成する．系列ラベリングのベースラインの素性として，表 4 のように対象の表層と品詞，その前後 2 つの表層とその品詞，またその組み合わせの素性を用いる．

固有表現認識では Inside/Outside 法のバリエーションの 1 つである IOB2 が用いられることがあり，“B” が固有表現の系列の始めを表し，“I” は B に続く固有表現を表し，それ以外を “O” で表す．しかし，KNP などの先行研究では Start/End 法を用いて高い精度を達成している．これは IOB2 の B・I・O の他に “S” と “E” が追加されたものであり，“S” は 1 形態素のみで固有表現であるものを表し，“E” は 2 形態素以上の固有

[2] コーパスや KNP の出力では “OPTIONAL” という固有表現が出現することがあるが，これはコーパス作成時にタグ付けがタグ付け判定者にも困難な場合に付けられる表現である．実験をする際は OPTIONAL は無視することにする．

表 3: BCCWJ のジャンルに対しての固有表現の種類の内訳

	ORGANIZATION	PERSON	LOCATION	ARTIFACT	DATE	TIME	MONEY	PERCENT
OW	129（20.6%）	33（5.3%）	144（23%）	165（26.4%）	129（20.6%）	0（0%）	10（1.6%）	33（5.3%）
PB	25（6%）	169（45.1%）	89（23.7%）	30（8%）	50（13.3%）	8（2.1%）	0（0%）	4（1.1%）
PN	118（17.2%）	78（11.4%）	185（27.0%）	24（3.5%）	161（23.5%）	21（3.1%）	60（8.8%）	38（5.5%）
PM	17（5.4%）	202（64.3%）	32（10.2%）	13（4.1%）	0（0%）	2（0.6%）	5（1.6%）	1（0.3%）
OC	19（11.4%）	6（3.6%）	57（34.3%）	54（32.5%）	18（10.8%）	3（1.8%）	9（5.4%）	6（3.6%）
OY	59（20.1%）	79（26.9%）	47（16.0%）	29（0.99%）	58（19.8%）	3（1.0%）	7（2.4%）	11（3.8%）

表 4: 固有表現認識に用いた素性とラベルの例

位置	表層	品詞	固有表現ラベル
i−2	同時	名詞-普通名詞-一般	O
i−1	に	助詞-格助詞	O
i	M S N	名詞-普通名詞-一般	B-ARTIFACT
i+1	メッセンジャー	名詞-普通名詞-一般	E-ARTIFACT
i+2	も	助詞-係助詞	O

表現の最後に用いられる．今回の実験では Start/End 法を用いて実験を行う．

3.4 ツール

BCCWJ と京都大学 Web 文書リードコーパスには形態素情報が付与されているため，BCCWJ では短単位，京都大学 Web 文書リードコーパスでは元から付与されている形態素情報を用いる．CRL 固有表現データには形態素情報が付与されていないので，MeCab（バージョン 0.996，IPA 辞書）を用いて形態素解析を行う．

CRF を行うツールとして CRF++（バージョン 0.58）[3] を用い，比較対象として KNP の固有表現モジュール（バージョン 4.12）を用いる．

3.5 評価

BCCWJ･京都大学 Web 文書リードコーパス･CRL 固有表現データをそれぞれ文単位で 10 分割し，CRF++ を用いて 10 分割交差検定を行う．また，KNP の固有表現モジュールで，10 分割交差検定を行った時のテストデータを用いてテストをしたものと比較する．

今回評価の指標として precision と recall と F1 スコアを用いて比較を行う．

[3]http://crfpp.googlecode.com/svn/trunk/doc/index.html

3.6 結果

まず BCCWJ のジャンルを考慮しない全てのデータ・京都大学 Web 文書リードコーパス・CRL 固有表現データでの CRF++，KNP を用いた実験結果が表 5 である．

次に，表 6 は BCCWJ 全体に対して固有表現認識を行った場合の固有表現タグごとの実験結果である．

また表 7 は BCCWJ のジャンルごとで実験した結果である．

4 考察

表 5 から，Web などから収集したコーパスである BCCWJ と京都大学 Web リードコーパスがテストデータのときは知恵袋やブログなどで学習した CRF++ の方が若干高い精度となった．CRL 固有表現データに KNP を適用した結果が非常に高くなっているのは，KNP のトレーニングデータに CRL 固有表現データが含まれているためであると思われる．

表 6 から，“DATE”，“TIME”，“MONEY”，“PERCENT” の固有表現で F 値が CRF++が KNP での結果に比べて大幅に低くなっている．これは JUMAN から得た情報にはカテゴリ名が付与されていることが原因の一つではないかと思われる．今回用いた CRF++ の素性テンプレートでは，単語の表層と品詞の情報しか用いていないため，データスパースネスの問題がある．KNP は CRF++ を用いた BCCWJ での実験よりもカテゴリ名素性が多い分，精度が上がったのではないかと思われる．

また，表 7 より，新聞記事で学習された KNP は白書や新聞のような堅い文章で高い再現率となった．一方，CRF++ はいずれのジャンルにおいても KNP より高い適合率となった．この原因の一つとして表 2 にあるように，OC（Yahoo!知恵袋）と OY（Yahoo!ブログ）での固有表現の数が少ないことが原因ではないかと思われる．全体の固有表現が少ないために，トレー

表 5: BCCWJ，京都大学 Web リードコーパス，CRL 固有表現データに CRF++，KNP を用いたときの固有表現認識の結果

	KNP			CRF++		
	precision	recall	F1	precision	recall	F1
BCCWJ	68.06	62.82	65.34	85.39	61.34	71.39
京都大学 Web リードコーパス	61.43	81.81	80.17	94.86	90.22	92.48
CRL 固有表現データ	97.02	97.63	97.32	86.02	85.66	85.84

表 6: BCCWJ 全体に KNP，CRF++ を用いて固有表現認識を行った固有表現タグごとの実験の結果

	KNP			CRF++		
	precision	recall	F1	precision	recall	F1
ORGANIZATION	60.10	75.57	66.96	75.12	51.11	60.84
PERSON	46.64	55.09	50.52	78.79	76.36	77.56
LOCATION	48.49	55.09	51.58	70.39	76.36	73.25
ARTIFACT	51.89	28.08	36.45	80.91	58.63	67.99
DATE	63.06	82.37	71.43	71.22	61.06	65.75
TIME	56.67	65.67	60.84	61.67	33.67	43.56
MONEY	66.52	95.00	78.25	66.32	61.49	63.82
PERCENT	69.91	93.83	80.13	75.61	46.50	57.58

表 7: BCCWJ のジャンルごとの KNP，CRF++，CRF++（CRL）

	KNP			CRF++			CRL 固有表現データを追加		
	precision	recall	F1	precision	recall	F1	precision	recall	F1
OW（白書）	53.84	95.65	68.76	90.51	66.34	76.56			
PB（書籍）	60.53	49.10	54.22	79.63	42.43	55.37			
PN（新聞）	72.35	75.71	73.99	85.23	50.36	63.33			
PM（雑誌）	44.29	46.87	45.56	83.52	60.36	70.07			
OC（Yahoo!知恵袋）	68.29	45.80	54.83	90.40	33.74	49.14	77.95	49.32	60.42
OY（Yahoo!ブログ）	64.14	60.67	62.36	68.72	39.11	49.85	69.49	53.11	60.21

ニングデータから文脈のパターンや，固有表現を学習することができず，適合率は高いが再現率は低い結果にとどまったのではないかと思われる．

さらに，CRF++ で F1 スコアが最も低かったウェブテキスト 2 ジャンルについて，トレーニングデータに CRL 固有表現データを追加して実験を行った．その結果より，OC（Yahoo!知恵袋）での結果が KNP での結果よりも precision，recall，F1 のいずれにおいてもスコアが上回った．また，OY（Yahoo!ブログ）では KNP と同程度の精度となった．この結果から OC においてはトレーニングデータを単純に増やすことで KNP での結果よりも高い結果が得られるということが分かった．これは表 6 から分かるように CRF++ は PERSON，LOCATION，ARTIFACT で KNP より高い性能を示しており，OC は LOCATION と ARTIFACT を合わせると全体の 6 割を占めるためであると考えられる．一方 OY では同程度の精度となったが，これは OY で全体の 4 割を占める ORGANIZATION と DATE において，CRF++ は KNP に劣るため，PERSON や LOCATION における性能向上と相殺しているためだと思われる．

表 8 では固有表現認識ができた例と誤った例を挙げている．下線が引いてある箇所が正しいタグである．

KNP でうまく固有表現認識ができたが，CRF++ では認識できなかったものとして，“ガンバ大阪” や “APEC” のようなものがあり，これは CRF++ではトレーニングデータに “テレ朝” や “APEC” がないが，KNP のトレーニングデータである CRL 固有表現データに存在し，前後の文脈パターンや固有表現を認識できていたのではないかと思われる．

また，KNP では “クマ” や “嵐山” などの人名を認識できなかったり，地名と誤認識してしまうことがあった．“嵐山” は地名と人名として両方あり得る固有名詞だが，CRF++ での結果ではうまく前後の文脈から正しく認識することができた．“ポケモンカード” という単語は KNP ではうまく認識ができないが，これは CRF++でのトレーニングデータに “ポケモンカード” という単語が出現したのでうまく認識することができたと考えられる．[4]

[4]今回，“ポケモンカード” という単語は OC（Yahoo!知恵袋）に 7 回出現している．

表 8: 固有表現認識の例

KNP	CRF++
KNP でうまく固有表現認識ができた例	
<ORGANIZATION> ガンバ大阪 </ORGANIZATION>に兄がいることから （Yahoo!ブログ）	ガンバ大阪に兄がいることから
さらに，<ORGANIZATION>APEC</ORGANIZATION>以外の国々との間でも （白書）	さらに，APEC 以外の国々との間でも
<PERSON> ケリー </PERSON>上院議員は （新聞）	ケリー上院議員は
◆ <LOCATION> 米 </LOCATION>が新規原発を建設の方針（新聞）	◆米が新規原発を建設の方針
「<ARTIFACT> 高性能林業機械化促進基本方針 </ARTIFACT>」に基づき （白書）	「高性能林業機械化促進基本方針」に基づき
人生の <PERCENT> 三分の一 </PERCENT>くらいは （書籍）	人生の三分の一くらいは
CRF++でうまく固有表現認識ができた例	
PS2 でも発売されますよ。(Yahoo!知恵袋)	<ARTIFACT>PS2</ARTIFACT> でも発売されますよ。
クマに相談したら「水羊羹持って謝りに行け」って。(雑誌)	<PERSON> クマ </PERSON> に相談したら「水羊羹持って謝りに行け」って。
<LOCATION> 嵐山 </LOCATION> はボコボコにしてた。(雑誌)	<PERSON> 嵐山 </PERSON> はボコボコにしてた。
ポケモンカードをタダで配るというではないか !! （書籍）	<ARTIFACT> ポケモンカード </ARTIFACT> をタダで配るというではないか !!
KNP でも CRF++でもうまく固有表現認識ができなかった例	
Yahoo 掲示板なんかに本音があったり （Yahoo!知恵袋） （本来なら “Yahoo 掲示板” に artifact のタグが付与される）	Yahoo 掲示板なんかに本音があったり
近くには <ARTIFACT> 千光寺 </ARTIFACT> などもあり坂の多い街で映画のロケ地にもなっています.(Yahoo!ブログ) （本来なら “千光寺” に LOCATION のタグが付与される）	近くには千光寺などもあり坂の多い街で映画のロケ地にもなっています.
文化多様性条約策定の動き（白書） （本来なら “文化多様性条約策定” に ARTIFACT のタグが付与される）	文化多様性条約策定の動き
ゴールデン街へ行けば必ず誰かに絡まれたからね、素人に。(雑誌) （本来なら “ゴールデン街” に LOCATION のタグが付与される）	ゴールデン街へ行けば必ず誰かに絡まれたからね、素人に。

また，KNP でも CRF++ でもどちらでも誤認識や認識しなかったものに “Yahoo!掲示板” や “ゴールデン街” といった単語があった．これは CRL 固有表現データには出現しない単語であり，CRF++ でもうまく認識が出来なかったので文脈からのパターンでもうまく認識ができなかったのではないかと思われる．また，本研究で用いた CRF++ の素性に Kazama ら [5] のようなクラスタリング素性を使用することで，データスパースネスの解消が期待できる．

5 おわりに

この論文では BCCWJ，京都大学 Web 文書リードコーパス，CRL 固有表現データに対して CRF，KNP を用いて固有表現認識を行い，精度を比較した．新聞記事でトレーニングを行った KNP に対して，BCCWJ や京都大学 Web 文書リードコーパスでトレーニングしたものの方が精度がわずかに高い結果となった．

今後の課題としては，KNP で行っているような係り受けなどの情報を入れることで口語表現などに対しても適応するできるのかどうかなどの分析ができると考えられる．

謝辞

京都大学 Web リードコーパスを公開前に使用させていただいた京都大学の河原先生に感謝いたします．

A 付録

表 9 は IREX で定義された固有表現とその例である．表 10 は BCCWJ 内のジャンルの例文である．

表 9: IREX で定義された固有表現とその例

	種類		例
固有名詞的表現	組織名，政府組織名	ORGANIZATION	通産省，自民党，全日空ホテル
	人名	PERSON	アリス，寅さん，若乃花
	地名	LOCATION	日本，太平洋，豊田駅
	固有物名	ARTIFACT	魚沼産コシヒカリ，サンフランシスコ条約，ギリシャ神話
時間表現	日付表現	DATE	前日，４月３日，２１世紀
	時間表現	TIME	午後７時，明け方，昨夜
数値表現	金額表現	MONEY	１１４円，数十兆円
	割合表現	PERCENT	１５％，数十パーセント

表 10: BCCWJ 内のジャンルの例文

ジャンル	例
Yahoo!ブログ	いろいろとブログめぐりとかしてたら、どうやら <PERSON> まさやん </PERSON> の曲が <ORGANIZATION> ＮＨＫ </ORGANIZATION> ドラマ「<ARTIFACT> 監査法人 </ARTIFACT>」というドラマの主題歌になるようで。 携帯サイトにもしっかり書いてありますね。 タイトル書いてないので、新曲かどうかは不明ですけど。 <PERSON> くまざき </PERSON> さんの <ARTIFACT> 撮りバカ日報 </ARTIFACT> では何か撮影してますね〜。
新聞	<ORGANIZATION> 財務省 </ORGANIZATION> が <DATE> ９日 </DATE> 発表した <DATE> ４月末 </DATE> の外貨準備高は <MONEY> 三千六百二十六億千百万ドル </MONEY> となり、<DATE> 前月末 </DATE> に比べ <MONEY> 十一億三千九百万ドル </MONEY> 増と、３カ月ぶりにプラスに転じた。 外国為替市場でユーロが対ドルで上昇したことに伴い、ユーロ建て資産が増えたことが主因。

参考文献

[1] 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.

[2] 福岡健太. Semi-Markov conditional random fields を用いた固有表現抽出に関する研究. Master's thesis, 奈良先端科学技術大学院大学情報科学研究科, 2006.

[3] 山田寛康, 工藤拓, 松本裕治. Support vector machine を用いた日本語固有表現抽出. 情報処理学会論文誌, Vol. 43, No. 1, pp. 44–53, 2002.

[4] 笹野遼平, 黒橋禎夫. 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. 49, No. 11, pp. 3765–3776, 2008.

[5] Jun'ichi Kazama and Kentaro Torisawa. Exploiting wikipedia as external knowledge for named entity recognition. In *EMNLP-CoNLL*, pp. 698–707, 2007.

[6] 新里圭司, 関根聡, 吉永直樹, 鳥澤健太郎. 固有表現抽出手法を用いたレストラン属性情報の自動認識. 言語処理学会第 12 回年次大会講演論文集 (2006), D1-2, pp. 98–96, 2006.

[7] IREX ワークショップ予稿集, 1999.

[8] 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 (NL-188-17), pp. 113–120, 2008.

[9] 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.

[10] Project next nlp. `https://sites.google.com/site/projectnextnlp/`.