拡張固有表現タグ付きコーパスの構築

- 白書，書籍，Yahoo!知恵袋コアデータ -

橋本 泰一 † 　中村 俊一 ‡

† 東京工業大学 統合研究院

‡ 日本システムアプリケーション

hashimoto@iri.titech.ac.jp, s-nakamura@jsa.co.jp

1 はじめに

「机」「椅子」「空」「愛」といった一般的な概念を表す表現ではなく,「夏目漱石」「東京オリンピック」「日本」などの物，イベントや考え方を表す言語表現を固有表現と呼ぶ．固有表現は，質問応答，情報抽出，機械翻訳，テキストマイニングなどの応用技術に用いられる自然言語処理におけて重要な基礎知識である．日本語においては，評価型ワークショップ IREX において，新聞記事をベースに固有表現タグ付きコーパス（CRL 固有表現データ[1]）が構築され，そのコーパスをもとに日本語における固有表現認識に関する研究が進んだ．そして，様々な固有表現認識手法 [6–9, 11, 12, 14, 16, 17] が提案されてきた．

IREX で定義された固有表現の種類は，組織名，人名，地名，日付表現，時間表現，金額表現，割合表現，固有物名の 8 種類である．この 8 種類のタグを毎日新聞記事（1,174 記事数）に付与したコーパスが CRL 固有表現データである．しかし，このコーパスを利用して開発された固有表現認識器を，質問応答システム，情報抽出システムやテキストマイニングに利用しようとしても実際に認識できる固有表現の種類が少なく，新聞以外の分野の文書に対する精度も十分満足のいくレベルではない．さらなる高度な言語処理システムの発展を促すためには，より詳細に定義された固有表現の定義のもと様々な分野の文書に対してタグ付けされた言語資源の作成が必要である．

本研究では，200 種類の固有表現を定義した「関根の拡張固有表現階層」をもとに，様々な分野の文書に固有表現タグが付与された大規模なコーパスの構築を目指している．複数の分野の文書のコーパスを作成するにあたり，大規模日本語コーパスである BCCWJ を利用した．本論文では，白書，書籍，Yahoo!知恵袋の 3 分野の文書に対して固有表現タグ付けを行ったコーパスについて報告する．さらに，機械学習アルゴリズムの一つである CRF をベースに固有表現認識に関する評価実験を行い，精度が約 80%，再現率が約 46%，F 値が約 60%という結果を得た．

[1] http://nlp.cs.nyu.edu/irex/index-j.html

2 関根の拡張固有表現階層

大規模な固有表現タグ付きコーパス構築に向けて，固有表現の定義として,「関根の拡張固有表現階層」（以下，拡張固有表現）[2]を採用した.「関根の拡張固有表現階層」は，MUC(Message Understanding Conference) プロジェクトで策定された固有表現の定義 [1]，それを基に策定された日本における IREX プロジェクトの定義 [3]，ACE(Automatic Content Extraction) プロジェクト[3]の定義をもとに，関根が拡張を行った固有表現の定義 [2,4,5] である．関根は，質問応答システム，情報抽出，機械翻訳，情報検索，要約などの自然言語処理技術への応用を目的として，この定義の策定を行っている．

拡張固有表現の大きな特徴は，固有表現の種類の豊富さである．MUC では，組織名，人名，地名，日付表現，時間表現，金額表現，割合表現の 7 種類，IREX では，MUC の 7 種類に固有物名を加えた 8 種類を固有表現として定義している．一方，拡張固有表現（バージョン 7.1.0）では，物やイベントなどの「名前」「時間表現」「数値表現」を最初の階層とし，最大 4 階層で構成され，200 種類の固有表現の定義が定められている．これは様々な自然言語処理技術への応用を考慮し，新聞記事や百科事典などに見られる概念や単語を考慮していることに起因する．

3 拡張固有表現タグ付きコーパス

拡張固有表現のタグ付きコーパスについては，橋本らの報告がある [15]．橋本らは，拡張固有表現 (Version

[2] http://nlp.cs.nyu.edu/ene/

[3] http://www.itl.nist.gov/iad/mig/tests/ace/

表 1: 拡張固有表現タグ付きコーパスの概要

	文書数	総文字数	平均文字数	形態素数	平均形態素数	表現数		平均表現数
						異なり	のべ	
白書コア	62	351,649	5671.8	228,651	3687.9	5,276	11,819	190.6
書籍コア	83	369,491	4451.7	234,540	2825.7	4,884	14,206	171.2
知恵袋コア	938	179,345	191.2	110,649	118.0	3,270	5,609	6.0
毎日新聞	8,584	3,643,361	424.4	-	-	63,545	252,763	29.4
白書	400	2,340,364	5850.9	-	-	23,857	74,203	185.5
CRL	1,174	593,763	505.8	-	-	7,153	19,254	16.4

7.1.0) の定義に則って，毎日新聞および BCCWJ に含まれる白書に対しタグ付けを行った．このコーパスは，毎日新聞は 8,584 記事に対し，のべ 252,763 個，異なり 79,632 個のタグを付与し，白書は 400 文書に対し，のべ 74,203 個，異なり 23,857 個のタグを付与した．これまで固有表現認識に関する研究に用いられていた CRL 固有表現データは毎日新聞 (1,174 記事，のべタグ数 19,254 個，異なりタグ数 7,153 個) にタグ付けされたものであった．そのため，橋本らのコーパスは，従来のコーパスに比べ十分に大規模なコーパスであった．

しかし，橋本らが構築したコーパスは新聞記事と白書と 2 種類のジャンルのコーパスのみであり，研究対象をさらに広げるためにもジャンルを増やす必要がある．また，従来のコーパスは形態素情報が付与されていないため，固有表現認識手法の比較において問題があった．従来の固有表現認識タスクにおける従来手法のほとんどは，機械学習をベースで形態素情報を素性としてもちいるものが多い．しかし，コーパスには形態素情報が付与されていないため，JUMAN，ChaSen，Mecab などの形態素解析器を用いる必要があった．しかし，形態素解析器の種類やバージョンによって形態素解析結果が変化し，その結果固有表現認識結果も大きく変動するため，手法の比較検討を行う際に論文に記載された性能を再現することが困難であった．固有表現認識手法の比較実験および検討を正確に行うことができるように，共通の形態素情報付きのコーパスの作成が必要である．

本論文では，特定領域研究「日本語コーパス」[10] で構築されている BCCWJ の白書，書籍，Yahoo!知恵袋各コアデータに対してタグ付けを行った．これまでの新聞記事と白書の 2 種類に加え，新たに書籍と Web の 2 種類の分野のコーパスを構築した．さらに，コアデータには短単位の形態素情報が人手により付与されているため，共通の形態素情報が利用できるようになる．各コーパスの概要を表 1 に示す．

白書コアデータ（62 文書，総形態素数 228,651）に対して，のべ 11,819 個，異なり 5,276 個の固有表現が，書籍コアデータ（83 文書，総形態素数 234,540) に対して，のべ 14,206 個，異なり 4,884 個の固有表現が，Yahoo!知恵袋コアデータ (938 文書，総形態素数 110,649) に対して，のべ 5,609 個，異なり 3,270 個の固有表現が付与された．白書および書籍コアデータは比較的 1 文書のサイズが大きく，1 文書に対して 100 以上の固有表現が付与されている．一方，Yahoo!知恵袋は 1 文書のサイズが小さく，付与された固有表現の数は非常に少ない．また，Yahoo!知恵袋は文書の内容が多種多様であるため，同一の固有表現が出現するケースが少なかった．

総文字数を比較してみると，3 コアデータ全体では，CRL 固有表現データを上回っている．さらに，これまで構築された毎日新聞，白書を加えると，拡張固有表現タグ付きコーパス全体では，CRL 固有表現データの 10 倍以上の大きさのコーパスが構築できた．今後の固有表現に関する自然言語処理の発展に十分活用できると思われる．

4 評価実験と考察

4.1 評価実験

白書，書籍，Yahoo!知恵袋の 3 コアデータを用いて，拡張固有表現認識の評価実験を行った．固有表現認識手法として，機械学習アルゴリズムの一つである Conditional Random Fields（CRF）を用いた手法を，同定するチャンクタグの方式は IOB2 を採用した．入力はコアデータに人手付与された形態素および品詞とし，学習および認識で用いた素性は，該当形態素および単語，前後 2 形態素および単語，前 2 拡張固有表現タグ，を利用する図 1．また，認識方向は文頭から文末に向かって行った．認識するタグの種類が多く，同時にす

位置	形態素	品詞	固有表現
i - 2	文部	名詞-普通名詞-一般	B-Goverment
i - 1	科学	名詞-普通名詞-サ変可能	I-Goverment
i	省	接尾辞-名詞的-一般	I-Goverment
i + 1	に	助詞-格助詞	O
i + 2	おい	動詞-一般	O
i + 3	て	助詞-接続助詞	O

図 1: 学習および認識に使用する素性

表 2: 評価実験結果

	精度	再現率	F 値
白書コア	86.5	49.7	63.1
書籍コア	84.2	39.7	53.4
知恵袋コア	81.8	24.2	37.3
3 コア	85.6	46.1	59.8
ALL	86.7	74.7	80.2

表 3: 3 コアの実験結果の内訳

	精度	再現率	F 値
3 コア	85.6	46.1	59.8
白書コア	86.2	54.3	66.6
書籍コア	85.0	44.0	57.4
知恵袋コア	82.8	33.7	47.9

表 4: ALL の実験結果の内訳

	精度	再現率	F 値
ALL	86.7	74.7	80.2
白書コア	83.0	63.0	71.6
書籍コア	84.0	50.1	62.2
知恵袋コア	74.1	39.9	51.9
白書	81.9	73.5	77.4
毎日新聞	88.6	77.9	82.9

べての拡張固有表現タグに認識することが困難であったため，各拡張固有表現タグごとに学習と認識を行った．そのため，認識された拡張固有表現タグが互いに入れ子や交差する場合も存在する．

評価実験に用いたコーパスは，白書コアデータのみ（白書コア），書籍コアデータのみ（書籍コア），Yahoo!知恵袋コアデータのみ（知恵袋コア），白書・書籍・Yahoo!知恵袋の 3 コアデータすべて（3 コア），すべてのコーパス（3 コアデータ，コア以外の白書，毎日新聞）（ALL）のの 5 種類を用いた．それぞれ，10 分割交差検定により実験を行った．コア以外の白書および毎日新聞は，Unidic（バージョン 1.3.12）[13] と Mecab（バージョン 0.98）を用いて形態素解析を行った．精度，再現率，F 値で評価を行い，その結果を表 2 に示す．3 コアおよび ALL における実験結果の各データごとの評価の内訳をそれぞれ表 3，表 4 に示す．

4.2 考察

表 2 から，白書，書籍，Yahoo!知恵袋ともに精度が 80%を超えており，タグ数が増加したにも関わらず比較的高い精度を示している．しかし，再現率においては，最も結果がよかった白書においても約 50%しかなく，Yahoo!知恵袋においては約 24%と非常に低い結果であった．どのコアデータにおいても再現率が比較的低く，再現率の向上が今後の課題の一つと言える．

3 コアデータをすべて用いた場合，全体の評価結果は，精度約 85%，再現率約 46%，F 値約 60%であった．表 3 から，どのコアデータにおいても，他のコアデータが学習データ増えたことによる精度の向上はあまり見られない．しかし，再現率は，白書と書籍においては約 5%，Yahoo!知恵袋においては約 9%と大きな向上が確認できた．

すべての拡張固有表現コーパスを用いた場合，全体の評価結果は，精度約 87%，再現率約 75%，F 値約 80%であった．表 4 から，新聞記事および白書の大幅なデータの影響で，各コアデータの再現率の大幅な向上が確認できた．しかし，どのコアデータに対しても精度が大きく低下している．これは，白書および毎日新聞には，正しい形態素解析情報が付与されていないため，形態素解析結果の誤りが学習結果に影響を与えていると考えられる．

また，ALL における毎日新聞の成績は他のコーパスに比べ非常に高い．この原因は，コーパスの規模が大きいこともその一因であると思われるが，それ以外に新聞記事が比較的固有表現を認識しやすい文章であることに起因しているのではないかと考えている．従来，CRL 固有表現データによる固有表現認識においては，

笹野ら [17] の F 値 89.43 が最も良い成績である．笹野らの手法が新聞記事以外の文書への適用や固有表現タグ数の増加により，どのような結果が得られるのかについて興味深い課題である．

5 まとめ

本論文では，新たな固有表現タグ付きコーパスの構築に向けて，拡張固有表現タグを付与したコーパスについて報告した．

「関根の拡張固有表現階層」の定義 (Version 7.1.0) に則って，200 種類の固有表現を，白書，書籍，Yahoo! 知恵袋に対してタグ付けを行った．その結果，特定領域研究「日本語コーパス」で構築されている白書コアデータに対して，のべ 11,819 個，異なり 5,276 個の固有表現が，書籍コアデータに対して，のべ 14,206 個，異なり 4,884 個の固有表現が，Yahoo!知恵袋コアデータに対して，のべ 5,609 個，異なり 3,270 個の固有表現が付与された．コアデータを利用することにより，人手により形態素情報が利用することが可能になっている．また，今回構築したコーパスはおおよそ CRL 固有表現データと同規模である．

構築したコーパスを用いた固有表現認識に関する評価実験においては，CRF をベースとした手法で，精度が約 80%，再現率が約 46%，F 値が約 60%という結果を得た．

特定領域研究「日本語コーパス」では，白書，書籍，Yahoo!知恵袋以外にも，新聞記事や雑誌といった新たな分野のコアデータを構築している．このプロジェクトの成果を利用して，新たな分野の文書に対して，コーパスを構築していく予定である．また，構築したコーパスを用いて固有表現認識の発展に向けて，新たな認識手法の研究に取り組んで行くことも今後の課題である．

謝辞

本実験を実施するにあたり，ニューヨーク大学の関根聡氏には，拡張固有表現タグ定義のご提供，およびタグ修正作業に対する多大なる助言をいただきました．ここに，心より感謝の意を表します．

参考文献

[1] Ralph Grishman and Beth Sundheim. Message Understanding Conference - 6: A Brief History. In *COLING-96*, 1996.

[2] Satoshi Sekine. Extended named entity ontology with attribute information. In *In Proceedings of the 5th International Conference on Language Resources and Evaluation*, 2008.

[3] Satoshi Sekine and Hitoshi Isahar. IREX: IR and IE Evaluation project in Japanese. In *LREC2000*, pp. 1977–1980, 2000.

[4] Satoshi Sekine and Chikashi Nobata. Definition, dictionary and tagger for extended named entities. In *In Proceedings of the Forth International Conference on Language Resources and Evaluation*, 2004.

[5] Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata. Extended Named Entity Hierarchy. In *LREC2002*, 2002.

[6] 渡辺一郎, 桝井文人, 福本淳一. 固有表現抽出ツール NExT の精緻化とユーザビリティの向上. 言語処理学会第 10 回年次大会, 2004.

[7] 土屋雅稔, 肥田新也, 中川聖一. 非頻出語に対して頑健な日本語固有表現の抽出. 情報処理学会自然言語処理研究会, pp. 1–6, 2008.

[8] 山田寛康. Shift-reduce 法に基づく日本語固有表現抽出. 情報処理学会自然言語処理研究会 (NL-179-3), pp. 13–18, 2007.

[9] 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, 2004.

[10] 前川喜久雄. 代表性を有する大規模日本語書き言葉コーパスの構築. 人工知能学会誌, Vol. 24, No. 5, pp. 616–622, 2009.

[11] 中野桂吾, 平井有三. 日本語固有表現抽出における文節情報の利用. 情報処理学会論文誌, Vol. 45, No. 3, 2004.

[12] 乾孝司, 村上浩司, 橋本泰一, 内海和夫, 石川正道. 接尾辞情報を利用した文書からの組織名抽出. 人工知能学会論文誌, Vol. 24, No. 6, pp. 469–478, 2009.

[13] 伝康晴. 多様な目的に適した形態素解析システム用電子化辞書. 人工知能学会誌, Vol. 24, No. 5, pp. 640–646, 2009.

[14] 浅原正幸, 松本裕治. 日本語固有表現抽出におけるわかち書き問題の解決. 情報処理学会論文誌, Vol. 45, No. 5, 2004.

[15] 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. 情報処理学会自然言語処理研究会 (2008-NL-188), 2008.

[16] 岩倉友哉. Stacking の効率的な学習方法と日本語固有表現抽出での評価. 情報処理学会自然言語処理研究会 (2005-NL-167), pp. 21–28, 2005.

[17] 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.