人物・組織エンティティに特化した固有表現抽出器の開発

仲野 友規　　乾 孝司
筑波大学大学院 システム情報工学研究科

y.nakano@mibel.cs.tsukuba.ac.jp inui@cs.tsukuba.ac.jp

1 はじめに

自然言語処理の基礎技術のひとつに固有表現抽出(Named Entity Recognition，NER）がある．固有表現の中でも人名や組織名といった人物・組織エンティティをあらわす表現の自動抽出は以下のような応用において必須の技術である．

- 文書マスキング [1]：プライバシー保護やビジネス上の機密保持の観点から，テキストデータ中の個人や企業を特定しうる情報を除去する．
- 社会分析 [2]：テキストデータを解析し，社会的影響の大きい事故や事件，イベント等を分析する．この際の重要な分析観点のひとつとして当事者（5W1H の Who）が据えられる．

これらの背景から，本研究では人物・組織エンティティに特化した固有表現抽出器の開発に取り組んでいる．また，このように特定の固有表現に特化することで，より詳細な分析ができ，それに応じた対策を効率よく実施することが可能となる．昨年，我々は人物・組織エンティティを指す固有表現に対して「関根の拡張固有表現階層[*1]」の情報が付与された，橋本らの拡張固有表現タグ付きコーパス [3, 4] を使った難易度評価を行った [5]．本稿では，難易度評価の対象となった固有表現を一般的な固有表現抽出器を用いて抽出した結果の誤り分析を行い，得られた知見をもとに抽出器を改良し抽出性能の向上を図ったので，その結果を報告する．

2 人物・組織向け固有表現クラス体系

汎用性の高い固有表現の分類として，関根によって定義された全 200 種類の固有表現クラスからなる「関根の拡張固有表現階層」がある．本研究では，この中から人物および組織エンティティをあらわす固有表現が属する可能性がある固有表現クラスを選び出すことで，新たに人物および組織に特化した固有表現クラス体系「人物・組織クラスサブセット」を定義し用いている．

人物・組織クラスサブセットは，関根の元定義と同様に階層構造を持っており，第 1 階層は PERSON，ORGANIZATION，GOE，GPE，P_VOCATION [*2]の 5 種類，第 2 階層は 40 種類の固有表現クラスから構成されている．

3 難易度レベル

我々は固有表現の難易度評価を行うにあたり，NE_Rate，Majority_Rate という 2 つの指標を導入した [5]．固有表現 x の NE_Rate(以下，NR) は，x が固有表現であるか否かの曖昧性を示し，この値が高い表現 x は値の低い表現と比べると抽出が容易であると考えられる．固有表現 x の Majority_Rate（以下，MR）は，x の固有表現クラス間の曖昧性を示し，この値が高い表現 x はサブセットにおける固有表現クラスの曖昧性が少ないので，この値の低い表現と比べると抽出が容易であると言える．そして，両指標をそれぞれ軸とする 2 次元の領域を考え，各指標それぞれに分割点を定めて領域を 4 つに分割することを通して難易度レベルを定義した．レベル 1，$2_{\mathrm{NR>MR}}$，$2_{\mathrm{NR<MR}}$，3 とあり，数字が大きいほど抽出が難しいことを表している．

4 誤り分析

4.1 分析方法

一般的な素性に基づく固有表現抽出器の結果に対して誤り事例を分析することで，抽出器の問題点を把握し，抽出器の性能改善に役立てる．誤り分析の対象となる固有表現抽出器は，CRF を用いた手法で一般的な素性を用いており，我々と同様に拡張固有表現タグ付きコーパスを用いているという理由から，南ら [6] の論文を参考にして構築した．単語認定は UniDic 辞書[*3] を採用した MeCab でおこなった．考慮する素性は前後窓枠 = 2 の範囲の文字，文字種，品詞である．文字単位で処理するため，品詞には SE 形式でラベル付けすることにより，文字単位の素性とする．CRF の実装には CRFsuite [*4] を用いた．クラスラベルの全ての遷移素性を考慮するため，`feature.possible_transitions=1` とした．正則化等の細部設定のチューニングは行っていない．

以上の実験条件のもと，拡張固有表現タグ付きコーパスに対して，サブセットの第 2 階層の粒度で 5 分割交差検定

[*1] https://sites.google.com/site/extendednamedentityhierarchy/

[*2] 元定義では，POSITION _VOCATION だが，本稿では紙面の都合上，略称を使う．他の名称も曖昧性がない場合は適宜，略称を使う．

[*3] https://ja.osdn.net/projects/unidic/

[*4] http://www.chokkan.org/software/crfsuite/

表 1　固有表現クラス毎の誤り数

クラス	部分一致誤り	クラス誤り	未抽出誤り	過抽出誤り	合計 1[*5]	合計 2[*6]	トークン
PER	287	113	1,041	296	1,441(9.2)	1,737(11.1)	15,642
ORG	956	719	2,184	808	3,859(16.0)	4,667(19.4)	24,090
GPE	875	1,498	2,101	1,623	4,474(11.9)	6,097(16.3)	37,493
GOE	121	323	302	68	746(28.5)	814(31.1)	2,618
P_VOC	1,754	6	4,222	1,356	5,982(19.7)	7,338(24.1)	30,422
MIX	50	332	268	27	650(36.4)	677(37.9)	1,785
合計	4,043	2,991	10,118	4,178	17,152	21,330	112,050

をおこない抽出実験を実施した．交差検定の分割データにおいて，人物・組織クラスサブセットに該当する全トークンを学習データに使用した．そのうち，難易度評価対象の固有表現のトークンを分析の対象とした．正解判定はチャンクの厳密一致とし，部分一致は不正解とする．

4.2　分析結果

表 1 に固有表現クラス毎の誤り数を示す．ここで，括弧内の値は固有表現クラス内での誤り率 [%] を表す．表中の「MIX」は，同一のトークンに複数の固有表現クラスが割当てられている特殊な事例に対する特別クラスである．これは抽出実験の便宜上の設定であり，以降の議論では注目しない．

■部分一致　正解固有表現のチャンクと抽出器の出力が完全一致ではなく部分的に重なっている状態．

■クラス誤り　チャンク範囲は完全一致しているものの，クラスを誤ってしまう状態．この誤りは，抽出器がどれだけ文脈の情報を拾えているかに依存しているといえる．

■未抽出誤り　正解固有表現のチャンクを抽出できなかった（全て O タグとしてしまう）誤り．

■過抽出誤り　上記 3 つの誤りは正解の固有表現を基準とした誤りであるのに対し，過抽出は固有表現でない部分を固有表現として抽出してしまう誤りである．したがって，適合率低下の原因となりやすい．また，サブセットでない固有表現クラスに属する固有表現の一部となっている部分をサブセットの固有表現として抽出した場合にも，誤りとなってしまう．

クラス誤り以外の誤りは，文脈の情報の他に複合語や長い固有表現を上手く捉えられていないということが原因として考えられる．これらを踏まえ，改良方針を以下のように立てた．

1. 全体の抽出性能を底上げするような素性の提案
2. 複合語等長い固有表現や文脈の情報を上手く拾えるような素性の提案

[*5] 部分一致から未抽出誤りまでの，固有表現として正しく抽出されなかった誤り数

[*6] 過抽出も含めた全誤り数

5　素性の提案

5.1　参考にした関連研究

Toh ら [7] は CRF の素性に，英語の固有表現抽出で一般的な素性に加え，知識ベース内の目標単語が属しているリストの種類を素性とする Gazetteer 素性や，外部コーパスを用いて単語をクラスタリングし，単語が所属しているクラスタ番号を素性とする単語クラスタ素性を用いた．同様のデータを用いて固有表現抽出性能を争うワークショップ [8] では，CRF を用いた手法の中で 1 位の性能となった．

中野ら [9] は，文書を文節単位に区切ることにより，目標トークンが所属する文節や隣接する文節の情報を取り入れた．アルゴリズムは SVM を用いている．CRL（通信総合研究所）固有表現データを用いた実験では，当時報告されていた SVM を用いた固有表現抽出手法の中で最高の精度が得られたと述べている．

笹野ら [10] は，目標形態素の窓枠内の形態素の情報だけでなく，先行文の同一形態素や共参照関係にある表現，所属文節の係り先文節といった大域的な情報を取り入れた．データには，IREX の定義に基づき作成された CRL 固有表現データ，IREX の公式のテストデータ，Web テキストに IREX の定義に基づき固有表現タグを付与したデータの 3 種類を使用しており，どのデータに対しても固有表現抽出の精度が向上することを確認している．

5.2　Gazetteer 素性

前節でも触れたが，Toh ら [7] は知識ベースに目標単語のエントリが存在した場合，そのエントリが属するリストの種類を素性として用いており，抽出性能を向上させることを示した．我々の知る限り，CRF を用いた日本語固有表現抽出でこの素性を用いている研究は存在しない．そこで本研究では，Gazetteer 素性を日本語固有表現抽出に適用し，改良方針 1 である抽出性能の底上げを図る．

Toh らが用いている知識ベースは，我々と対象言語が異なるため用いることが出来ない．そこで，知識ベースの「エントリ」を「Wikipedia のページタイトル」，「エントリが属するリスト」を「Wikipedia のページが属するカテゴリ」とした．そして，2017 年 8 月 1 日時点の Wikipedia 日本語版のダンプ[*7] から，ページタイトルとページが属するカテゴリの ID が対応している知識ベースを作成し用いた．まず，

[*7] https://dumps.wikimedia.org/jawiki/

目標形態素の基本形をクエリとして知識ベースを引き，カテゴリ ID を得る（クエリをタイトルとするページがない場合，ページが属するカテゴリがない場合は 0 が返される）．そして，文字単位の素性とするため品詞素性と同様にカテゴリ ID の前に SE 形式でラベル付けする．カテゴリ ID が複数ある場合も，すべてにラベルを付け素性とする．

5.3 文節情報に基づく素性

以下は改良方針 2 に対する素性である．なお，係り受け解析は CaboCha を用い，文単位で行う．

■所属文節主辞素性 中野ら [9] は目標トークンの所属文節内の主辞の情報を素性に取り入れている．本研究でも，複合語は同文節になりやすく，複合語の主辞は文節の主辞になりやすいということから，所属文節の主辞を素性として用いる．具体的には，所属文節の主辞の基本形をクエリとして知識ベースを引き，カテゴリ ID が 0 でない場合はカテゴリ ID，0 の場合は主辞の基本形をそのまま素性とする．

■係り先文節主辞素性 クラス誤りの原因として考えられるのは，前後 2 文字の範囲では文脈をほぼ考慮できないためということである．前節でも述べたように，笹野ら [10] は係り先文節の主辞を素性として用い，抽出性能が向上することを示した．そこで，本研究でも，所属文節の係り先文節の情報を用いる．具体的には，係り先文節の主辞の基本形をクエリとして知識ベースを引き，カテゴリ ID が 0 でない場合はカテゴリ ID，0 の場合は主辞の基本形をそのまま素性とする．係り先がない場合は特別な素性（本研究では “N/A”）とする．

■所属文節格助詞素性 笹野ら [10] は係り先文節の主辞と共に，目標形態素の所属文節の格も素性として用いた．本研究でも，所属文節の格助詞を素性として用いる．格助詞が存在しない場合は，特別な素性（本研究では “N/A”）とする．

6 評価実験

6.1 実験方法

提案素性の有効性を確かめるため，評価実験を行った．使用するデータや評価対象は 4.1 節で述べたものと同様である．抽出性能は，出力結果に対する適合率，再現率，F 値で測定した．

6.2 実験結果

抽出性能一覧を表 2 に示す．また，難易度レベルと F 値の関係一覧を表 3 に示す．ベースラインは 4.1 節で述べた素性セットであり，これに対して提案素性を追加する．

Gazetteer 素性が抽出性能の底上げ向上に貢献していることがわかる．そして，Gazetteer 素性に加え，文節情報に基づく素性を組み合わせたいずれの場合においても，ベースラインを上回る抽出性能であることがわかる．最高性能は，「+ 所属文節主辞素性」であった．しかし，Gazetteer 素性のみ追加した場合より，それに文節情報に基づく素性を追加した場合の性能が劣っている場合も多い．特に，係り先文節主辞素性が含まれている場合の性能が低い．しかし，難易度レベル別に見てみると，それぞれの難易度レベルで最も性能が良かったのは，文節情報に基づく素性を含む場合であり，文節情報に基づく素性の有効性を確認できたといえる．興味深いのは，全体の F 値が最も高かった「+ 所属文節主辞素性」は，トークン数が多いレベル 1 のみ F 値が最高で，レベル 2，3 に関しては，他の組み合わせに劣っている場合があるということである．実際に，レベル $2_{\mathrm{NR>MR}}$ に関しては「全素性」や「− 所属文節格助詞素性」，レベル $2_{\mathrm{NR<MR}}$，3 に関しては「− 係り先文節主辞素性」の方が性能がいいことがわかる．レベル 2，3 の固有表現が多い状況下では，全体の抽出性能の優劣が入れ替わる可能性が考えられる．

表 2 素性と抽出性能一覧

追加素性	適合率	再現率	F 値
（ベースライン）	0.894	0.847	0.868
Gazetteer 素性	0.891	0.868	0.879
+[*8]所属文節主辞素性	0.896	0.868	**0.881**
+ 係り先文節主辞素性	0.888	0.858	0.871
+ 所属文節格助詞素性	0.891	0.870	0.879
−[*9]所属文節主辞素性	0.888	0.857	0.871
− 係り先文節主辞素性	0.895	0.868	0.880
− 所属文節格助詞素性	0.894	0.863	0.877
全素性	0.894	0.861	0.876

表 3 難易度レベルと F 値の関係一覧

追加素性	1	$2_{\mathrm{NR>MR}}$	$2_{\mathrm{NR<MR}}$	3
（ベースライン）	0.918	0.654	0.739	0.626
Gazetteer 素性	0.926	0.659	0.771	0.651
+ 所属文節主辞素性	**0.927**	0.668	0.776	0.667
+ 係り先文節主辞素性	0.919	0.656	0.760	0.641
+ 所属文節格助詞素性	0.926	0.659	0.773	0.654
− 所属文節主辞素性	0.919	0.656	0.760	0.639
− 係り先文節主辞素性	0.926	0.668	**0.777**	**0.670**
− 所属文節格助詞素性	0.923	**0.671**	0.768	0.660
全素性	0.922	**0.671**	0.767	0.661

6.3 分析

適合率に強く影響するのはクラス誤り，過抽出誤りであるため，適合率が最も高かった「+ 所属文節主辞素性」はそれらの誤りが最も少なかった．同様に，再現率に強く影響するのは未抽出誤りであるため，再現率が最も高かった「+ 所属文節格助詞素性」は未抽出誤りが最も少なかった．

「Gazetteer 素性」については，未抽出誤りを大きく減らすことが出来ていた．Gazetteer 素性を追加したことで，固有表現になりやすいという情報をベースラインと比べて上手く得ることができたからだと考えられる．以下の正解事例は，いずれもベースラインでは抽出できていなかった．

[*8] + 素性名：Gazetteer 素性とその素性のみを追加した場合
[*9] − 素性名：Gazetteer 素性とその素性以外の文節情報に基づく素性を追加した場合

市場に参戦してきたのが “サッポロビール⟨Company⟩”。

サッチャー政権及び “保守党⟨Political_Party⟩” の支持率

これらの抽出固有表現はいずれも Wikipedia にページがあり，いくつかのカテゴリに属している．一方で，クラス誤り，過抽出誤りが増えており，Gazetteer 素性は同じ表層形（基本形）であれば同じ素性が追加され，文脈を考慮できないためであると考えられる．クラス誤りについては以下のような誤り事例が見られた．

“川崎⟨City⟩” のようにパスをつないで攻めるチーム

正解は Pro_Sports_Organization である．過抽出誤りについては以下のような誤り事例が見られた．

我が “国鉄⟨Corporation_Other⟩” 鋼輸出の７．２％を占め

この固有表現も，Wikipedia にページがあり，幾つかのカテゴリに属している．

最も性能の高かった「+ 所属文節主辞素性」については，「Gazetteer 素性」と比べ適合率が向上しているが，部分一致誤り以外の誤りを減らすことができていた．これについては，導入理由にもあるように複合語や長い固有表現を捉えられているからだと考えられる．固有表現の長さ別に性能を見てみると，窓枠外でトークン数が多い長さ（文字数）6，7 の固有表現の F 値が向上していることがわかった．以下のような正解事例が見られた．

三木 “内閣総理大臣⟨Position_Vocation⟩” は，

「Gazetteer 素性」では，“三木内閣⟨Cabinet⟩” と抽出されていた．

性能があまり良くなかった係り先文節主辞素性を含む組み合わせについては，未抽出誤りが増加してしまっている点で共通していた．文分割の誤りや Yahoo!知恵袋などの口語体のデータを含んでいることによる，係り受け解析の解析誤りが原因として考えられる．

7 おわりに

本稿では，人物・組織エンティティに特化した固有表現抽出器の開発に取り組み，誤り分析の結果から CRF の素性を新たに提案することで一般的な固有表現抽出器の性能の向上を図った．抽出性能全体の底上げのため導入した Gazetteer 素性については，ベースラインから性能を大きく向上させることが出来た．また，文節情報に基づく素性については，Gazetteer 素性と所属文節主辞素性を追加した場合で全体の F 値が 0.881 と最大となり，長い固有表現の抽出性能が上がっていることが要因として挙げられた．また，それぞれの難易度レベルに対して有効な組み合わせを確認することができた．

今後の課題としては，より詳細な分析や，係り先文節主辞素性が奮わなかった原因と考えられる係り受け解析の解析誤りを減らすために，より高精度な文分割を行うことが挙げられる．提案素性が他の固有表現クラスのサブセットでも有効かどうかを確かめるのも興味深い課題である．

謝辞

本研究の一部は科研費（15K20884）の助成を受けて実施されました．

参考文献

[1] 伊川洋平, 宅間大介, 金山博. 安全語のアンマスキングによる機密情報マスキングシステム (情報抽出). 電子情報通信学会技術研究報告. DE, データ工学, Vol. 106, No. 150, pp. 79–84, jul 2006.

[2] 橋本泰一, 村上浩司, 乾孝司, 内海和夫, 石川正道. 文書クラスタリングによるトピック抽出および課題発見. 社会技術研究論文集, Vol. 5, pp. 216–226, 2008.

[3] 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.

[4] 橋本泰一, 中村俊一. 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-. 言語処理学会 第 16 回年次大会 発表論文集, pp. 916–919, 2010.

[5] 仲野友規, 乾孝司. 人物・組織エンティティに対する固有表現抽出課題の難易度評価. 言語処理学会 第 23 回年次大会 発表論文集, pp. 246–249, 2017.

[6] 南和江, 藤井康寿, 土屋雅稔, 中川聖一. 大規模コーパスを用いた固有表現抽出手法の検討. 言語処理学会 第 17 回年次大会 発表論文集, pp. 328–331, 2011.

[7] Zhiqiang Toh, Bin Chen, and Jian Su. Improving Twitter Named Entity Recognition using Word Representations. In *Proceedings of the ACL 2015 Workshop on Noisy User-generated Text*, pp. 141–145, July 2015.

[8] Timothy Baldwin, Marie-Catherine de Marneffe, Bo Han, Young-Bum Kim, Alan Ritter, and Wei Xu. Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. In *Proceedings of the Workshop on Noisy User-generated Text*, pp. 126–135, Beijing, China, July 2015. Association for Computational Linguistics.

[9] 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, pp. 934–941, mar 2004.

[10] 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.