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

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 [9]：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 [7]：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．

我々は現在，抽出対象として人物・組織エンティティに関する固有表現クラスに限定した固有表現抽出器の開発を進めている．この中で，抽出対象となる固有表現の難易度評価を実施したので，本稿ではその結果を報告する．

2 難易度評価とは

固有表現抽出器の研究開発をおこなう過程の中で誤り分析が行われる [3]．誤り分析は現在の抽出器の問題点を把握するために不可欠な過程と言える．また，誤り分析とともに難易度評価も重要である [6]．難易度評価とは，抽出器に依存せず，固有表現そのものが持つ特性に基づいて固有表現の自動抽出の難しさを評価することである．誤り分析と難易度評価を実施することは，抽出器の改良計画を検討する際に役立つと期待でき，例えば，抽出を誤った事例群を難易度の高い事例群と低い事例群に分割し，個別の改良計画を立てることができる．

固有表現抽出の難易度評価の先行研究として，野畑ら [6] の研究がある．野畑らは「表現の多様性が抽出を難しくする」という考えに基づいて，固有表現クラスに対する抽出難易度を示す指標を提案した．本研究では，野畑らの指標と併用できる指標として，固有表現クラスではなく，個々の固有表現に対する抽出難易度を示す２つの指標（NE_Rate と Majority_Rate），およびそれらの要約となる難易度レベルの定義を検討した．提案指標は，固有表現クラスの集合およびその部分集合を考慮するすることで，任意の部分集合に該当する固有表現の抽出難易度を求めることができる．本研究では，この指標を利用して，後述する「人物・組織クラスサブセット」の固有表現に対する難易度評価を実施した．また，ある固有表現抽出器を題材にして抽出器の結果と難易度を照らし合わせることで，抽出器の抽出性能と難易度レベルの相関を調査した．

3 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[*1]」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，以降の難易度評価に用いる．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，以下の通り，第１階層（大文字）では５種類，第２階層（小文字）では 40 種類の固有表現クラスから構成される．

1. 人名（PERSON）
   - Person
2. 組織名（ORGNIZATION）
   - Organization_Other，International_Organization，Show_Organization，Family，Ethnic_Group_Other，Nationality，Sports_Organization_Other，Pro_Sports_Organization，Sports_League，Corporation_Other，Company，Company_Group，Political_Organization_Other，Government，Political_Party，Cabinet，Military
3. 政府を持つ地域名（GPE）
   - GPE_Other，City，County，Province，Country
4. 組織名の属性を持つ施設（GOE）
   - GOE_Other，Public_Institution，School，Research_Institute，Market，Park，Sports_Facility，Museum，Zoo，Amusement_Park，Theater，Worship_Place，Car_Stop，Station，Airport，Port
5. 地位職業名（P_VOCATION[*2]）
   - P_Vocation

[*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．

4 難易度の定義

4.1 2つの難易度指標

固有表現の抽出難易度を定義するにあたり，以下に示す2つの指標（NE_Rate と Majority_Rate） を導入する．

まず，以下で使う記号を説明する．固有表現タグ付きコーパスを D とする．ある文字列表現を x であらわす時，D 中に現れる x のトークンを x_i，そのトークン数を N_x とする．T を「関根の拡張固有表現階層」に含まれる固有表現クラス（200 種類）の集合，S を人物・組織クラスサブセット第2階層（40 種類）のクラスの集合 ($S \subset T$) とする．また，$I_1(x_i, c)$ は，トークン x_i が固有表現クラス c に属するエンティティを表している場合に 1，それ以外は 0 を返す指示関数，$I_2(x_i, C)$ は，x_i が固有表現クラス集合 C の要素となるいずれかのクラスに属するエンティティを表している場合に 1，それ以外は 0 を返す指示関数であるとする．

このとき，各指標はそれぞれ次の式で与えられる．

$$\mathrm{NE\_Rate}(x, T, D) = \frac{\sum_{x_i \in D} I_2(x_i, T)}{N_x - \sum_{\langle axb \rangle_j \in D} I_2(\langle axb \rangle_j, T)} \quad (1)$$

$$\mathrm{Majority\_Rate}(x, S, T, D) = \frac{\max_{s \in S}\{\sum_{x_i \in D} I_1(x_i, s)\}}{\sum_{x_i \in D} I_2(x_i, T)} \quad (2)$$

式 (1) の $\mathrm{NE\_Rate}(x, T, D)$(以下，NR) は，コーパス D に出現する表現 x のすべてのトークンのうち，T に含まれるいずれかの固有表現を表しているトークンの割合を示す．$\mathrm{NR} = 1$ となる表現は単純な辞書ベースの手法で過不足なく抽出可能であるので，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．なお，分母の $\langle axb \rangle_j$ は x を含むトークン（$(\mathrm{length}(x) < \mathrm{length}(axb))$）である．タグ付きコーパスの作成時，固有表現が入れ子構造（例えば，“筑波大学” という 組織名とその中の “筑波” という地名）になっている場合は入れ子の外側のみを認定するので，N_x から $\langle axb \rangle_j$ に関する数を引くことで値を調整している．

式 (2) の $\mathrm{Majority\_Rate}(x, S, T, D)$(以下, MR)は, 式 (1) の分子を構成するトークンのうち，人物・組織クラスサブセット S に含まれる固有表現を表しているトークンの中で最多数を占めるクラスに属するトークンの割合を示す．この値が高い表現 x は S 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．

4.2 難易度指標に基づく難易度レベルの定義

上記の2つの指標を使って，人物・組織クラスサブセットに関する固有表現の抽出難易度を定義する．両指標をそれぞれ軸とする2次元の領域（以降，難易度領域と呼ぶ）を考えると，各表現はこの領域中の1点で示される．後述するように，難易度領域にプロットされた様子を観察することである程度の難易度の概要を把握することが可能である．ここではさらに，各指標それぞれに分割点を定めて難易度領域を4つに分割することを通して難易度レベルを定義する．難易度領域とその分割結果の概念図を図 1 に示す．今回は，分割後の部分領域に対して図に示した難易度レベルを割りあて，各表現の難易度を求めることにする．両指標の定義から，レベル1に該当する表現は抽出が比較的容易であり，レベル3に該当する表現は抽出が困難であることを意味する．レベル2は抽出を難しくする原因に基づき2種類に区別している．すなわち，$2_{\mathrm{NR>MR}}$ に該当する表現は，コーパス中に出現すれば固有表現を表すことが多いが，固有表現クラス間の曖昧性が大きい．そのため，この部分領域に該当する表現は固有表現クラスの曖昧性解消が抽出精度向上のボトルネックとなりやすい．一方，$2_{\mathrm{NR<MR}}$ に該当する表現は，固有表現クラス間の曖昧性は少ないが，固有表現とならないケースが多くある．そのため，この部分領域に該当する表現は固有表現であるか否かの判別がボトルネックになると考えられる．

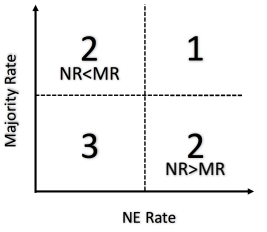

図 1 難易度領域と難易度レベル

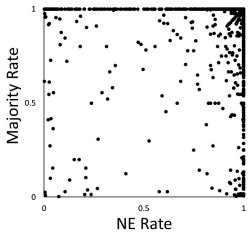

図 2 難易度領域へのプロット結果

5 難易度評価と相関調査

5.1 検討項目

前節で述べた難易度レベルに基づいて，固有表現タグ付きコーパスに出現する固有表現の抽出難易度を評価する．次に，固有表現抽出器の抽出結果と難易度を照らし合わせることで，既存の固有表現抽出器の抽出性能と難易度レベルの相関を調査する．

5.2 固有表現の難易度評価

「関根の拡張固有表現階層」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 [4, 5] に対して固有表現の抽出難易度を評価する．このコーパス中で何らかの固有表現タグが付与されている表現 x のうち，$I_2(x_i, T) \geq 10$ かつ $I_2(x_i, S) \geq 1$ を満たす 2,493 件の固有表現を評価対象にする[*3]．

評価対象のそれぞれに対して，先述の2つの指標 NR および MR の値を計算し，難易度領域へプロットした結果を図 2 に示す．本コーパスの場合，多くの固有表現が NR と MR のいずれかの指標値が1であることがわかる．

次に，それぞれの難易度指標の平均値（NR：0.934，MR：

[*3] トークンの出現数をカウントする際，固有表現タグが部分重複している事例はアノテーション誤りの可能性があるため取り除いた．

表 2　難易度レベル毎の固有表現の例

レベル	固有表現の例 (NR, MR)	左列の第 1 番目の例における出現文脈の例
1	日本 (0.958, 0.998)，米国 (0.994, 0.998)， 文部省 (1, 1), 村山 (0.995, 1), 田中 (1, 1)	今の 日本$_{\langle \mathrm{Country} \rangle}$ には、 トラックで 日本$_{\langle \mathrm{Country} \rangle}$ 記録を出すつもり つまり、日本$_{\langle \mathrm{Country} \rangle}$ における新しい
$2_{\mathrm{NR}>\mathrm{MR}}$	東京 (0.990, 0.548)，新潟 (0.982, 0.660), ダイエー (1, 0.702)，毎日新聞 (1, 0.409), 高校生 (0.973, 0.017)	高速バスで 東京$_{\langle \mathrm{Province} \rangle}$ から現地入り 過度に 東京$_{\langle \mathrm{City} \rangle}$ に集中していたものを 打診を受けたのは、東京$_{\langle \mathrm{Company} \rangle}$、日本興業、住友
$2_{\mathrm{NR}<\mathrm{MR}}$	委員 (0.201, 1)，記者 (0.668, 1), 兵 (0.179, 1)，学生（0.681, 1), メンバー (0.354, 1)	を批判する 委員$_{\langle \mathrm{P_Vocation} \rangle}$ を抑え、 について 委員$_{\langle \mathrm{P_Vocation} \rangle}$ がただしたところ から調査委員として派遣された
3	タイ (0.213, 0.903)，王 (0.343, 0.581), 日 (0.106, 0.846)，金 (0.274, 0.612), 京 (0.330, 0.667)	タイ$_{\langle \mathrm{Country} \rangle}$ では水不足が深刻だ 味付けされた タイ$_{\langle \mathrm{Fish} \rangle,\langle \mathrm{Food_Other} \rangle}$ と 本塁打の大会タイ記録

表 1　難易度レベルの割り当て分布

レベル	割当件数	ave.NR	ave.MR
1	1,834	0.997	0.998
$2_{\mathrm{NR}>\mathrm{MR}}$	272	0.994	0.609
$2_{\mathrm{NR}<\mathrm{MR}}$	268	0.604	0.998
3	119	0.563	0.565
合計	2,493		

0.935）を難易度領域の領域分割点とし，この分割点に従って，各固有表現へ難易度レベルを割り当てた．難易度レベルの分布を表 1 に示す．また，各レベルに属する固有表現の例を表 2 に示す．

表 1 から，本コーパスにおける今回の設定ではレベル 1 に該当する固有表現が多いことがわかる．また，レベル 2 は 2 種類に区別したが，どちらも一定数の固有表現が該当していることが確認できる．また，他のレベルより数は少ないものの，レベル 3 の固有表現も存在することがわかった．また，表 2 に例示しているが，各レベルの固有表現を観察することで，レベルごとに幾つか特色があることがわかった．例えば，レベル $2_{\mathrm{NR}>\mathrm{MR}}$ には GPE（Geo-Political Entity）の第 2 階層となるクラスの表現，レベル $2_{\mathrm{NR}<\mathrm{MR}}$ には P_VOCATION の表現が集まりやすかった．さらに，レベル 3 は固有表現クラスに偏りはないが，短い文字列で構成される固有表現が多く集まっていた．レベル $2_{\mathrm{NR}>\mathrm{MR}}$ に集まる GPE となる固有表現を観察すると，正式名称の省略形で書かれる事例が少なからず含まれていた．これは，企業やスポーツチームのように地域性をもつエンティティは “東京◯◯銀行” や “浦和□□” のように名称に地名がはいることがあり，文脈からエンティティが明らかであれば地名（“東京” や “浦和”）のみの省略形として記述されることによる．レベル 3 にも省略形で書かれる事例があるが，こちらは省略形の文字列が他の固有表現と重ならないためレベル $2_{\mathrm{NR}>\mathrm{MR}}$ ではなくレベル 3 となっている．各難易度レベルのトークン数の詳細については表 5 も参照されたい．

5.3　固有表現抽出器の抽出性能と難易度レベルの相関調査

5.3.1　調査に使用する固有表現抽出器

固有表現抽出器の抽出性能と難易度レベルの相関を調査する．まず，調査に使用する固有表現抽出器について説明する．固有表現抽出器は，南ら [8] の論文を参考にして構築した．南らは拡張固有表現タグ付きコーパスを使って CRF[1] に基づく抽出器を構築しており，本稿では彼らの実験条件にあわせて抽出器を構築した．ただし，系列ラベリングの処理単位については条件を変更した．南らは文字単位の系列を処理しているが，本稿では実験負荷を抑制する目的で単語単位の処理に変更した．単語認定は UniDic 辞書[*4] を採用した MeCab[*5] でおこなった．その他の実験条件は以下の通りである．チャンク表現は IOB2 表現である．考慮する素性は前後窓枠＝2 の範囲の単語および品詞である．CRF の実装には CRFsuite[2] を用いた．クラスラベル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，人物・組織クラスサブセットの第 1 階層の粒度，すなわち 5 種類の固有表現（PERSON, ORGANIZATION, GPE, GOE, P_VOCATION）に対して，5 分割交差検定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った．また，トークンのうち，前節と同じ条件となる $I_2(x_i, T) \geq 10$ かつ $I_2(x_i, S) \geq 1$ を満たす固有表現のトークンを評価データに使用した．学習および評価ともに，コーパス中の固有表現タグと単語区切りが不整合を起こすトークンおよびその周辺に出現するトークンは抽出対象から除外した．その結果，抽

[*4] `https://ja.osdn.net/projects/unidic/`
[*5] `http://taku910.github.io/mecab/`

表 3　調査に使用する固有表現抽出器の抽出性能

固有表現クラス	適合率	再現率	F 値	トークン数
PERSON	0.931	0.909	0.920	15,074
ORGANIZATION	0.872	0.864	0.868	23,103
GPE	0.863	0.915	0.888	33,562
GOE	0.829	0.736	0.779	2,534
P_VOCATION	0.863	0.851	0.857	28,942
Mix	0.877	0.690	0.772	1,293
平均 / 合計	0.874	0.878	0.876	104,508

表 4　難易度レベルと F 値の関係

レベル	PER	ORG	GPE	GOE	P_VOC
1	0.945	0.898	0.900	0.867	0.895
$2_{NR>MR}$	0.786	0.741	0.885	0.450	0.666
$2_{NR<MR}$	0.846	0.705	0.760	0.585	0.768
3	0.667	0.602	0.780	0.601	0.464

出対象は 2,483 件の固有表現（104,508 トークン）となった．

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．この際，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して評価している．結果を表 3 に示す．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．

5.3.2　相関調査

表 3 で示した抽出器の抽出性能（F 値）と抽出対象となっている固有表現の難易度レベルに相関があるか調査する．人物・組織クラスサブセット第 1 階層の 5 クラスについて，難易度レベルごとに F 値を算出した結果を表 4 に示す．この表から，すべての固有表現クラスで難易度レベルが上がるほど F 値が低下する傾向があることがわかる．一部のクラスでレベル 2 とレベル 3 で結果が反転することもあるが，レベル 1 は常にもっとも高い F 値となっている．つまり，人物・組織クラスサブセットに対する抽出器の抽出性能は難易度レベルと相関があり，レベル 1 の固有表現に対しては比較的良好な抽出性能を示すが，レベルが上がるに従って性能低下を引き起こしていることがわかる．

6　おわりに

本稿では，固有表現抽出の難易度評価のための指標を検討し，人物・組織エンティティを表す固有表現を対象にして，難易度評価を実施した．その結果として，難易度レベルごとの固有表現の分布状況を確認した．また，難易度レベルごとに固有表現に幾つか特色があることがわかった．また，求めた難易度レベルと固有表現抽出器の抽出性能の間の相関を調査し，人物・組織クラスサブセットに対する抽出器の抽出性能はレベル 1 の固有表現に対しては比較的良好な抽出性能を示すが，レベルが上がるに従って性能低下を引き起こしていることを確認した．

表 5　固有表現クラス毎のトークン数

レベル	PER	ORG	GPE	GOE	P_VOC
1	12,507	19,339	25,217	1,844	20,884
$2_{NR>MR}$	1,061	1,847	5,903	453	135
$2_{NR<MR}$	1,054	1,094	1,064	151	7,569
3	452	823	1,378	86	354
合計	15,074	23,103	33,562	2,534	28,942

今後の予定としては，誤り分析および野畑らの指標と併用した難易度評価をおこない，そこで得られる知見を人物・組織エンティティに関する固有表現クラスに限定した固有表現抽出器の開発へ反映させる予定である．

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] John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the eighteenth international conference on machine learning, ICML*, Vol. 1, pp. 282–289, 2001.

[2] Naoaki Okazaki. CRFsuite: a fast implementation of Conditional Random Fields (CRFs), 2007.

[3] 岩倉友哉. 固有表現抽出におけるエラー分析. 言語処理学会 第 21 回年次大会 ワークショップ, 2015.

[4] 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.

[5] 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.

[6] 野畑周, 関根聡, 辻井潤一. 日本語固有表現抽出の難易度を示す指標の提案と評価. 自然言語処理, Vol. 10, No. 1, pp. 3–26, 2003.

[7] 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.

[8] 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.

[9] 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, 2006.