入れ子構造を持つ固有表現の階層的抽出

岩崎慧悟[†1]　阪本浩太郎[†1]　渋木英潔[†2]　森辰則[†1]

[†1]横浜国立大学大学院環境情報学府 [†2]国立情報学研究所
E-mail: {i-keigo,sakamoto,mori}@forest.eis.ynu.ac.jp, shib@nii.ac.jp

1　はじめに

固有表現抽出は質問応答システムや文書要約といった自然言語処理のシステムで重要となる．しかし，ほとんどの固有表現抽出器は固有表現を文の中での最長の単位による抽出しか対象としていない．一方で固有表現において入れ子構造となる固有表現が存在する．入れ子構造を持つ固有表現は世界史の教科書や医学論文といったテキストで出現する事が多い．図 1 に入れ子構造を持つ固有表現の例を示す．入れ子構造を持つ固有表現が内包する固有表現も重要な情報を持っている事もあり，これらを抽出することにより，既存の自然言語処理において役立てる事が期待できる．こうした入れ子構造を持つ固有表現抽出に関する研究が近年行われはじめた。

本稿では入れ子構造を持つ固有表現を入れ子の深さに対応する層に分け階層的に抽出する手法について述べる．入れ子構造を持つ固有表現において低層 (初めの層) を短い表現 (一番内側の表現) とするか否かといった抽出器の構造の違いで抽出精度が変わる可能性がある．また，それらの抽出器を組み合わせる手法にもいくつかの方法が考えられる．世界史の教科書を対象とした評価実験によりそれらの手法について比較を行い，その結果について述べる．

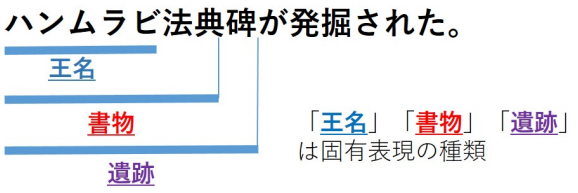

図 1: 入れ子構造を持つ固有表現の例

2　関連研究

入れ子構造を持つ固有表現の抽出に関する研究は近年行われはじめている． Marinho ら [1] は医学の分野において特定の症状の表現を抽出するの際に固有表現の持つ入れ子構造に関する研究の規則を人手で定めて，その規則を適用する文脈を学習させる．Yaseen ら [2] は生物学，医学の文献から高々深さが 2 である入れ子構造を持つ固有表現を抽出するために一番長い表現 (一番外側の表現) を抽出した後，その固有表現を入力としてこれを分割し種類を同定する．しかし，これらの先行研究は予め人手で定めた入れ子構造のみを扱えたりと制限があったり，入れ子構造の深さが高々2 に限定されていた．本研究では任意の深さと構造を持つ入れ子を扱える固有表現抽出手法を提案する．

笹野ら [3] は固有表現抽出の際に先行文における同一形態素の解析結果，共参照関係にある表現の解析結果，係り先から得られる情報，固有表現情報を付与した格フレームを用いた格解析から得られる情報という 4 つの大域的な情報を用いる．本研究では固有表現抽出の際に笹野らの提案した情報を用いないが，抽出対象となる未知テキスト全文に渡って複数回現れる固有表現という大域的な情報を用いる．

3　階層的な固有表現抽出

3.1　階層構造

固有表現の入れ子の深さに対応する層を仮定し，層を追って順番に固有表現を抽出する事を考える．層の高低について，より先に処理される層を「低い層」，より後に処理される層を「高い層」と呼ぶ．この時，長い表現ほど低い層として扱うものをピラミッド型構造と呼ぶ．短い表現ほど低い層として扱うものを逆ピラミッド型構造と呼ぶ．図 2，図 3 にピラミッド型構造と逆ピラミッド型構造の例を示す．下線付きのテキストは固有表現の種類を表す．ピラミッド型構造では 1 番長い表現は必ず 1 層目に位置するが，それ以外の表現は例え同じ表現であっても入れ子の深さによってどの階層に位置するのかが変わる．一方，逆ピラミッド型構造では 1 番短い表現は必ず 1 層目に位置すると共に同じ固有表現は必ず同じ層に現れる．このように層の処理の順番の違いにおいて抽出される固有表現が変わりうる．

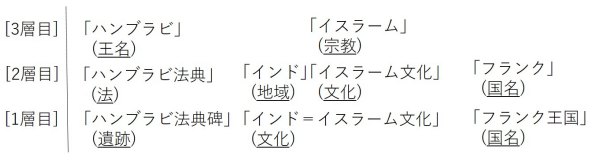

図 2: ピラミッド構造の例

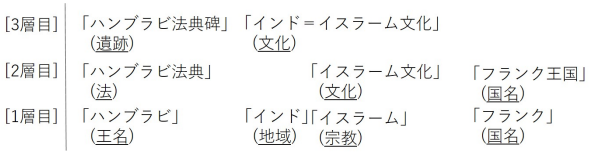

図 3: 逆ピラミッド構造の例

3.2 処理の流れ

固有表現抽出器は N 文字から構成される文 $X = (x_1, x_2, ...x_N)$ を入力として，文字それぞれの要素が持つ素性を用いて固有表現のラベル列 $Y = (y_1, y_2, ..., y_N)$ を出力する系列ラベリングにより定義出来る．

階層的な固有表現抽出では各層において現れる固有表現のみを対象とし，これを抽出するように固有表現抽出器を学習させる．与えられた文とそこから導出出来る基本素性から各層に現れる固有表現のラベル列を推定するように抽出器を学習する．未知テキストにおける抽出時には対応する層に対して学習された抽出器を用いてそれぞれの層において与えられた文と素性を入力として，固有表現のラベル列を層毎に出力する．

ここで，基本素性に加えて，先に抽出されるより低層の抽出結果も素性に加えることも考えられる．図 4 にピラミッド型構造において低層の抽出結果を利用した階層的固有表現抽出の処理の流れを示す．逆ピラミッド型においても同様の処理の流れとなる．図矢印の左の青枠が入力となる素性となっている．矢印右の赤枠が入力の文字列に対応した固有表現ラベル列となっている．1 層目の抽出には基本素性のみを使用して抽出を行う。2 層目は基本素性と 1 層目の抽出結果を素性として用いる。3 層目は基本素性と 1,2 層目の抽出結果も素性として用いる．層の数は学習コーパスに現れる固有表現の内最大の深さの表現の深さに合わせる．前の層で抽出された固有表現の情報を用いることで抽出結果の精度の改善を期待できる．

4 重複表現の利用

抽出対象となる未知テキスト全文に渡って複数回現れる固有表現を重複表現と呼ぶ．抽出対象とするテキスト全文に対して固有表現抽出を行い，2 つの階層構造の抽出結果から「コンスタンティノープル」に関する抽出結果を全てまとめた例を図 5 に示す．このように固有表現の抽出結果が全て同一の固有表現ラベルになるとは限らない．世界史の教科書にように同一表現に対して，同一のラベルが振られるべきであることを前提とするのであれば，投票により最多のラベルに統一することが考えられる．そこで，重複表現について 2 つの構造における抽出結果を集計し一番多い固有表現のラベルを統一することにより，固有表現の訂正を行う．それぞれの層の抽出終える毎に，その層までに抽出された最多のラベルに仮に統一する．最後に全ての層に渡って最多なラベルを選ぶ．

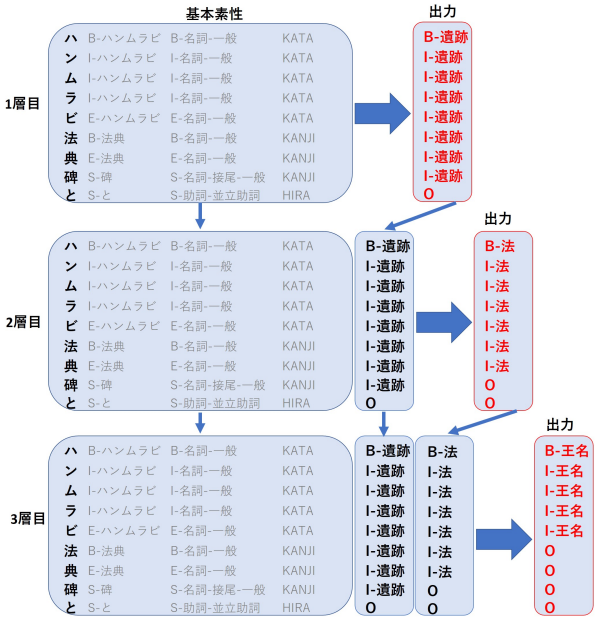

図 4: ピラミッド型構造において低層の抽出結果を利用した階層的な固有表現抽出

	ピラミッド型	逆ピラミッド型
コンスタンティノープル	身分	都市
コンスタンティノープル	王族	都市
コンスタンティノープル	都市	都市
コンスタンティノープル	国名	O(固有表現ではない)

図 5: 重複表現の抽出例

最多のラベルが複数である場合の処理について考える．例えば,「コンスタンティノープル」の抽出結果を集めた時,「都市」と「国名」の両方が最多のラベルとなる際には同点の解消が必要となる。この場合いずれのラベルの内どれが正しいのか判断が難しいため，ラベルの訂正を行わない．

5 評価実験

本章では，評価実験の実験目的，実験設定，評価手法について説明する．

5.1 実験目的

提案手法が以下の条件においてどのような抽出精度となるか比較検討する．1) 階層構造の違い,2) 低層の抽出結果の利用の有無，3) 重複表現の利用の有無

5.2 実験設定

Asahara[4] らの研究では日本語の固有表現抽出において，単語単位で行うより文字単位で行った方が抽出

精度が高くなることを示している．本研究でも文字単位で固有表現抽出を行う．

ベースとなる固有表現抽出器として SVM(Suport Vector Machine) に基づく系列ラベリング手法を用いる．山田ら [5] の研究と同様に，多値分類化には one-versus-rest 方式を用いる．また，多項式カーネル関数を用い次数は 2 とする．

山田ら [5] の研究と同様に文末から文頭に向けて解析する左向き解析を用いる．

基本となる素性について，i 番目の文字の固有表現のラベルを推定する時，i-2 から i+2 までの文字自身，その文字が所属する形態素，その形態素の品詞，および文字種を素性として用いる．既に解析した i-2 から i-1 の固有表現のラベルも素性として用いる．

評価用のコーパスとしては山川書籍の世界史の教科書 [6] の内, 先史時代から中世ヨーロッパまでに現れる 1624 文に対して入れ子構造を持つ固有表現の注釈付けをしたものを用いた．このコーパスでは 3 層目までの階層を構築することで全ての固有表現を網羅することが出来た．ピラミッド型構造では，1 層目で 5368 個，2 層目で 2090 個，3 層目で 252 個，また逆ピラミッド型では 1 層目で 5464 個，2 層目で 2003 個，3 層目で 243 個の固有表現が確認出来た．

5.3 評価手法

各条件に対し，五分割交差検証を行い，それらの平均の精度 (Precision), 再現率 (Recall)，F 値で比較を行う．

5.4 実験結果

実験結果を表 1 に示す．表の「i 層目」は i 層目の抽出結果のみに注目した場合に対応する.「層全体」はすべての層に渡って抽出結果を集計した場合に対応する．

6 考察

6.1 階層構造の違い

ピラミッド型構造に比べて逆ピラミッド型構造の方が層全体の精度，再現率，F 値が良い．ピラミッド型構造では同一の固有表現が別々の層に存在しうるのに対し，逆ピラミッド型では同一の固有表現が同じ層に集中し，抽出器が学習が促進されるからであると思われる．ピラミッド型の固有表現抽出は固有表現で 1 番長い表現を 1 層目に入れるので,「ハンムラビ法典碑」といった長い表現が集まりやすいが,「ハンムラビ」といった短い表現が単独で出てきた場合も 1 層目に入る．逆ピラミッド型では一番短い表現を 1 層目に入れるので，そのため本来，2 層目以降で抽出すべき短い固有表現を 1 層目で誤って抽出する例がみられた．

6.2 低層の抽出結果の利用の有無

低層の抽出結果を利用した方が，利用しない場合に比べて 2 層目以降，及び層全体の F 値は上がっている．「ハンムラビ法典」といった固有表現を抽出する際に逆ピラミッド型であれば,「ハンムラビ」といったその固有表現に内包される固有表現であったり，ピラミッド型であれば「ハンムラビ法典碑」といったその固有表現を内包する固有表現を用いることは有効であると示された．

6.3 重複表現の利用の有無

重複表現を利用すると，利用しない場合と比べて，2 層目以降の精度が減少し，再現率，及び F 値が改善されている．これには抽出自体に失敗していた固有表現を正しく抽出出来るようになった例も要因の一つである.「ローマ法大全」や「宋代文化」といった入れ子構造を持つ固有表現は全ての固有表現を抽出できるとは限らない．重複表現を用いない手法における実際の抽出例を見るとピラミッド型では「宋代文化」「宋」だけが取れて逆ピラミッド型では「宋」「宋代」だけが取れていた．しかし，最多のラベルで固有表現を統一することにより，それぞれの構造で固有表現の情報を用いることで抽出誤りを訂正することに成功している．

精度が下がる原因としては，少数の固有表現のラベルが正解であった場合，その表現も誤りのラベルで上書きされてしまうことにある．1 つの場所においてどちらの構造でも同じ誤りのラベルを付与してしまうとその誤りラベルが多数派で選ばれてしまう可能性が強くなる．例えば,「エジプト」という固有表現をラベル「国名」で抽出したかったのだが，逆ピラミッド型で「エジプト」をラベル「地域」で多く抽出していた，結果「エジプト」をラベル「国名」で抽出していた場所もラベル「地域」で訂正してしまっていた．そのため，ラベルの総数で 2 番目に多いラベルである場合は訂正を行わないといった手法も試す必要がある．

7 おわりに

本稿では入れ子構造を持つ固有表現を抽出する手段として階層的固有表現抽出手法を提案した．低層の抽出結果を利用した場合，抽出精度が向上した．ピラミッド型構造より逆ピラミッド型構造の方が層全体の抽出精度はよい．層全体の精度を上げるため重複表現の利用について検討した．実験の結果，重複表現の利用が 2 層目以降の再現率向上に役立つことが分かった．

今後の課題としてはコーパスの拡充が挙げられる．また，今回は固有表現抽出器のベースとして SVM を使用したが，LSTM などの深層学習を用いた手法も検討する必要がある．

表 1: 実験結果

		低層の抽出結果の利用:無し			低層の抽出結果の利用:有り					
		重複表現の利用:無し						重複表現の利用:有り		
		P	R	F	P	R	F	P	R	F
ピラミッド	1 層目	**83.7**	75.7	79.5	**83.7**	75.7	79.5	83.2	**77.0**	**80.0**
	2 層目	83.6	74.4	78.7	**85.7**	76.2	80.7	82.1	**82.6**	**82.3**
	3 層目	79.6	47.8	59.3	**85.3**	62.8	72.2	76.8	**79.1**	**77.9**
	層全体	83.9	75.3	79.4	**84.2**	75.4	79.6	82.6	**78.6**	**80.6**
逆ピラミッド	1 層目	**86.6**	77.9	**82.0**	**86.6**	77.9	**82.0**	84.7	**78.7**	81.6
	2 層目	**88.9**	76.8	82.4	88.3	80.2	84.1	85.0	**84.9**	**84.9**
	3 層目	**84.7**	56.1	67.4	82.0	62.0	70.6	81.6	**73.2**	**77.1**
	層全体	**87.2**	77.2	81.9	87.0	78.0	82.2	84.7	**80.1**	**82.3**

8 謝辞

本研究の前身となる基礎的検討を行った松永詠介氏と神貴久氏に深く感謝いたします.

参考文献

[1] Zita Marinho, Afonso Mendes, Sebãstiao Miranda and David Nogueira. Hierarchical Nested Named Entity Recognition. Proceedings of the 2nd Clinical Natural Language Processing Workshop,pp.28-34,2019

[2] Usama Yaseen, Pankaj Gupta and Hinrich Schütze. Linguistically Informed Relation Extraction and Neural Architectures for Nested Named Entity Recognition in BioNLP-OST 2019. Proceedings of The 5th Workshop on BioNLP Open Shared Tasks,pp.132-142,2019

[3] 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.

[4] Masayuki Asahara and Yuji Matsumoto. Japanese Named Entity Extraction with Redundant Morphological Analysis. Proc. HLT-NAACL2003, pp.8-15 ,2003.

[5] 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.445-53,2002.

[6] 株式会社山川出版社. 世界史 B 詳説世界史改訂版 (世 B-016)