XML 文書における構造の素性を用いた照応による人物検索

Person Retrieval on XML Documents by Coreference that Uses Structural Features

米井 由美 ♥ 岩井原 瑞穂 ♦
吉川 正俊 ♠

Yumi YONEI Mizuho IWAIHARA
Masatoshi YOSHIKAWA

複数キーワードによる検索は，キーワード間の意味的な関係が重要である．人物の事柄に対する検索では，「人物名」と「人物の属性」によるキーワード設定が可能である．しかし，これらのキーワード間の関係を考慮しない場合，検索結果として返される文書には，「人物の属性」が検索キーワードに設定した「人物名」と異なる人物の属性である内容の文書が含まれることが起こりうる．係受け解析や照応解析を用いて，キーワードが意味的な関係を持つ文書を抽出することで，検索精度を向上することが考えられるが，XML などの構造文書においては，照応の起き方が文書構造に大きく影響を受けると考えられる．本論文では，そのことを XML 文書における構造の素性を用いた照応により確かめ，この照応解析を利用した人物検索について述べる．

As for retrieval by two or more keywords, semantic relation between keywords is important. For retrieving information about a person, it is common to search by a pair of keywords consisting of person's name and his/her attribute. However, if semantic relation between keywords is not considered, the documents that describe different person's attribute may be retrieved. By using dependency analysis and coreference analysis, it is possible to retrieve the contents in which query keywords have semantic dependencies and improve search precision. However, as for structural documents such as the XML, coreference is often influenced by the document structure. In this paper, we confirm it by the coreference that uses structural features of XML documents, and we describe our person retrieval that uses the structural coreference.

♥ 学生会員 京都大学大学院情報学研究科修士課程 yyonei@db.soc.i.kyoto-u.ac.jp
♦ 正会員 京都大学大学院情報学研究科 iwaihara@i.kyoto-u.ac.jp
♠ 正会員 京都大学大学院情報学研究科 yoshikawa@i.kyoto-u.ac.jp

```
<article>
  <name> 亀田興毅 </name>
  <body>
    <section></section>
    <section>
       <title> 来歴 </title>
       <item> 大阪府大阪市西成区天下茶屋出身。
       </item>
       <item> 11 歳の時、父・史郎からボクシングを
       教わるようになる。 </item>
    </section>
     ...
  </body>
</article>
```

図 1 XML 文書と照応
Fig. 1 XML document and coreference

1. はじめに

現在のキーワード検索は，キーワードの出現頻度や出現位置をもとに行われている．複数キーワードによる検索では，キーワード間に意味的な関係をもって設定されることが多い．そのため，検索システムはキーワード間の意味的な関係を考慮する必要がある．人物の事柄に対する検索では,「人物名」と「人物の属性」によるキーワード設定が可能である．しかし，これらのキーワード間の意味的な関係を考慮しない場合，検索結果として返される文書には,「人物の属性」が検索キーワードに設定した「人物名」と異なる人物の属性である内容の文書が含まれることが起こりうる．そこで,「人物名」と「人物の属性」におけるキーワード検索で，キーワード同士の関係を考慮することで，人物検索における検索精度の向上を図ることを本研究の目的とする．

自然言語処理で研究されている係受け解析や照応解析を用いて，キーワードの「人物名」と「人物の属性」が，その人物に関する記述として出現している文章を抽出することにより，検索精度を向上することが考えられる．文書中の同一指示対象を同定する照応解析は，文字列一致や意味などの手がかりをもとに機械学習により行われる．この手がかりとなる属性のことを素性という．しかし，係受け解析や照応解析を検索に適用することは，大規模辞書等の言語リソースの確保が必要であることや Web ページに対して精度がまだ十分でないことなどの課題がある．

一方，XML や HTML などの構造文書においては，照応が文書構造上の位置関係に大きく影響すると考えられる．たとえば，図 1 のように，人物名が記事名を示す name ノードに出現する場合，その記事全体に，その人物に関する内容が書かれている可能性が高い．そのため，item ノードに出現する「大阪府大阪市西

成区天下茶屋出身。」の文には，主語が省略されているが，記事の人物の出身情報が記述されているのが最も確からしいと判断できる．2 つ目の item ノードに「父・史郎」という別の人物が出現しているが，これは「大阪府大阪市西成区天下茶屋出身。」の文とテキスト上では一文しか離れていない．しかし，異なる item では，話題が変わっている場合が多く，照応関係を持つ確率が低いといえる．このように，2 つの語の間に照応関係が存在するか判定する際に，文書の論理的構造の位置関係から求められる照応の起きる確率を利用することが考えられる．そこで，XML 文書における構造の素性が照応に有効であることを実験により確かめ，この照応解析を利用した人物検索について述べる．

本論文では，2 章で関連研究について述べる．3 章では，文書構造の素性を用いた照応と，その人物検索への適用手法について述べる．4 章では Wikipedia の XML 文書を題材に，構造の素性を用いた照応解析の実験を行った結果とその考察について述べ，5 章でまとめる．

2. 関連研究

自然言語処理解析を Web 検索に用いた研究に，インターネットの評判抽出が挙げられる [1][2]．文献 [1] では，照応解析の手法を意見抽出に適用することで，依存関係にない属性と評価値の対を抽出することを行っている．文献 [2] では，多商品分野に渡る大量の自由記述文から,「対象」と「評価」の抽出を行い，抽出された「対象」と「評価」を，同一文中で対応付け，対の構造を抽出している．

Web の文書構造を考慮して，オブジェクト同士の関連を求める関連研究として，表解析や画像検索などが挙げられる．表は，属性と属性値の関係，caption と表全体の関係などの構造をもとに，情報抽出に利用されている [3][4]．文献 [3] では，単一の表形式を入力とし，表形式中のセル間の類似度を用いて属性-属性値の関係を認識する手法を提案している．文献 [4] では，類似するデータの配置によって表構造を解釈し，さらにデータの内容に基づいて似た情報が記述されている表を統合している．画像検索は，周辺テキストから画像と関連するセンテンスやキーワードを抽出する手法が研究されている [5][6]．文献 [5] では，画像が出現するノードから親ノードへかけて 3 ノード間のノードを取得し，その範囲を画像を説明する文章の候補としている．文献 [6] では，Web 上の画像に対し，画像の前後に出現するテキスト，文書構造，リンク構造に基づいて，画像の使用状況を表す 3 種類の Web 文脈を定義している．

本研究においては，オブジェクトを画像とその関連キーワード，表の属性，属性値に留まらず，HTML や XML の構造文書に出現する 2 つのキーワード「人物名」,「人物の属性」の間の関連を文書構造から解析することを行う．また，自然言語的な素性を用いることによって，2 つのオブジェクトが長文のテキストに出現する場合も対応できると考えられる．

3. 提案手法

本研究の XML 文書における照応解析による人物検索についての提案手法を述べる．

3.1 照応解析

照応解析のひとつの手法として，素性をもとにした機械学習がある．本論文では，XML 文書を入力として，文書に出現する先行詞候補と照応詞候補の組合せに対して自然言語的な素性と文書構造の素性を抽出し，それらをもとに機械学習により照応関係を求める方法について検討する．

3.1.1 素性

自然言語的な素性 (linguistic features)

自然言語処理に用いられる素性は文献 [7] を参考にする．大きく分類すると，以下の 4 種類の素性がある．

- 語彙的な情報を用いた素性 (文字列一致)
  照応詞候補と先行詞候補の二つの文字列の文字列一致の情報を素性として導入する．文字列一致には，”完全一致”，”前方一致”，”後方一致”，”主辞 (最右の内容語) の一致”，”部分一致”，”構成文字列の一致”などが挙げられる．一般に，文字列が一致するほど，照応詞候補と先行詞候補が同一指示関係であると考えられる．
- 形態・統語的な情報を用いた素性 (文法)
  照応詞候補と先行詞候補それぞれの品詞，指示詞，助詞，対象とする名詞句の連体修飾要素の時制などの文法的な情報を素性として導入する．たとえば，指示連体詞「その」が名詞句に係る場合は，その名詞句は定名詞句である可能性が高いといったことが挙げられる．
- 意味的な情報を用いた素性 (意味)
  辞書や，係受け解析システムが出力した固有表現のタグを用いて，意味的な情報を素性として導入する．固有表現のタグには，人物名，地名などの固有表現に対する”PERSON”,”LOCATION”などのタグがあり，Cabocha [8] では自動的に付与される．照応詞候補と先行詞候補の意味属性が同じであれば，それらが同一指示関係である可能性が高い．
- 名詞句間の距離情報を用いた素性 (名詞句間の距離)
  照応詞候補と先行詞候補の距離が離れるほど同一指示関係とならない可能性が高い．そこで，照応詞候補と先行詞候補の文間の距離を素性として導入する．

文書構造の素性 (structural features)

XML や HTML などの構造文書においては，照応の起き方が文書構造に大きく影響を受けると考えられる．文書内の出現位置が離れていても，XML の木構造上で親子関係にあることで照応の可能性が強くなる．また，照応詞候補と先行詞候補がテキスト上では近隣に出現していたとしても，木構造上の距離が離れていれば，照応が起きない場合がある．そこで，XML の構造を考慮した素性を照応解析に用いることが効果的であると考えた．用いる構造情報は照応詞候補と先行詞候補が出現するテキストノードの葉から共通の祖先ノードまでのパスを抽出した部分木とする．

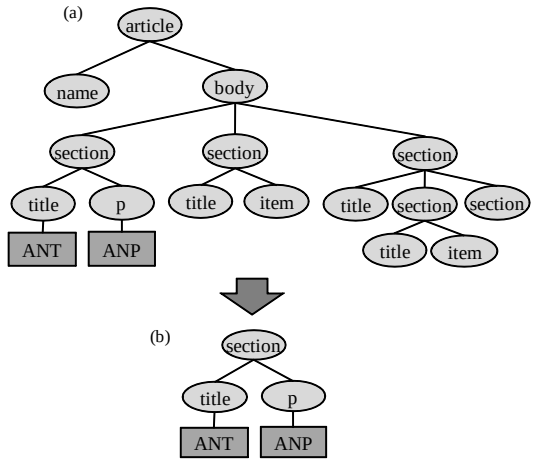

図 2　照応組合せ部分木の生成
Fig. 2　generating the coreference combination subtree

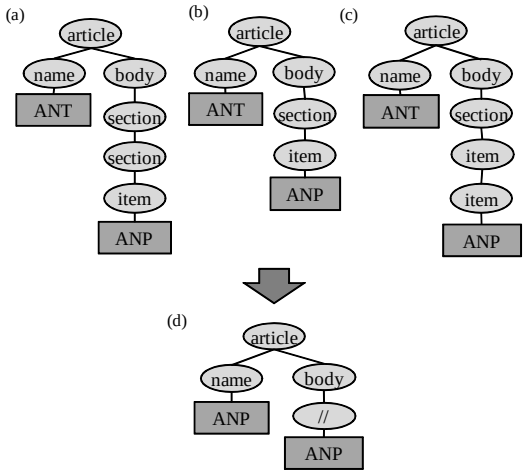

図 3　k-照応組合せ部分木の生成 (k=2)
Fig. 3　generating the k-coreference combination subtree(k=2)

本論文では，この先行詞候補と照応詞候補からなる部分木を，**照応組合せ部分木**と呼ぶ．

図 2 は照応組合せ部分木の生成を説明したものである．図 2(a) が XML 文書木の全体であり，先行詞候補がある section ノードの title ノードに出現し，照応詞候補がその section ノードの段落に出現しているとする．この時，XML 文書木を入力として，照応組合せ部分木が，図 2(b) のように自動的に生成される．これは，先行詞候補と照応詞候補の出現位置関係を表現している．このように，照応組合せ部分木は，先行詞候補と照応詞候補の出現位置関係の情報を含んでおり，構造の素性として有効であると考えられる．

しかし，このように XML 文書中に出現する全ての照応詞候補と先行詞候補の組合せの木構造パターンを抽出すると，文書が複雑になるにつれ，構造の素性の種類が非常に多くなってしまう．また，照応組合せ部分木が類似しており，構造的特徴が同じであっても，異なる照応組合せ部分木であれば，素性は異なるため，汎用性を持たない可能性がある．そのため，類似した部分木の構造をある程度一般化してまとめ，素性の個数を削減することを試みる．そこで，照応組合せ部分木の根から深さ k までの部分木を抽出することを行う．本論文ではこれを，**k-照応組合せ部分木**と呼ぶ．k より深い場合は任意の木構造とした部分木とし，木構造の上位に共通部分を持つ照応組合せ部分木をまとめることにより，素性の個数を削減する．

図 3 は k=2 のときの k-照応組合せ部分木の生成を説明したものである．図 3(a)(b)(c) の 3 つの照応組合せ部分木は，全て異なるがどれも先行詞が name ノードに出現し，照応詞が body に出現することが重要な特徴となる．図 3(d) の照応組合せ部分木の根から深さ 2 まで抽出した k-照応組合せ部分木 (k=2) は，このような特徴が表現可能である．そこで，この k-照応組合せ部分木を構造の素性とする．以降，下記のように k-照応組合せ部分木の根，左部分木，右部分木からなる木構造を括弧を用いて再帰的に表現したものを構造の素性のフォーマットとする．

(article (name 先行詞候補)(body 照応詞候補))

これは，図 3(d) の k-照応組合せ部分木を表現している．

3.1.2　学習機械

学習機械として，自然言語処理に応用した研究がいくつかなされている最大エントロピーモデル [9][10] が挙げられる．最大エントロピーモデルとは，事象 x, y が同時に起こる頻度 $C(x, y)$ を訓練データとして，条件付き確率 $p(x, y)$ で表わされる確率モデルを推定するアルゴリズムである．

まず，素性関数と呼ばれる事象の組 (x, y) に対して 1,0 を返す任意の関数を生成し，作った素性関数から制約を定め，この 2 つを使ってモデルを作る．たとえば，事象 x が先行詞候補と照応詞候補の位置関係が，

(article (name 先行詞候補)(body 照応詞候補))

の k-照応組合せ部分木で表現でき，事象 y が先行詞候補と照応詞候補が照応関係であるとき，次のような素性関数が生成される．

$$f_i(x, y) = \begin{cases} 1 & \text{if } x \wedge y \\ 0 & otherwise \end{cases}$$

これは，先行詞候補が記事名に，照応詞候補が記事内に出現しており，それらが照応関係にあるとき 1 を返す関数である．確率モデルは式 (1) で求められる．

$$p(x, y) = \frac{1}{z(x)} e^{\sum_i \lambda_i f_i(x,y)} \tag{1}$$

$$z(x) = \sum_y e^{\sum_i \lambda_i f_i(x,y)} \tag{2}$$

λ はそれぞれの素性を重み付けるパラメータである．また，$z(x)$

は，正規化を行っている．ここで，次のような制約がある．

$$\tilde{P}(f_i) = \sum_{x,y} \tilde{p}(x,y) f_i(x,y) \quad (3)$$

$$P(f_i) = \sum_{x,y} \tilde{p}(x) p(x,y) f_i(x,y) \quad (4)$$

$$P(f_i) = \tilde{P}(f_i) \quad (5)$$

$\tilde{P}(f_i)$ は，学習データによる素性の期待値，$P(f_i)$ はモデルにより得られる素性の期待値である．それぞれの期待値が同じになるようなモデルを作らなければならないという制約がある．モデルのエントロピーは式 (6) で計算される．

$$H(P) = -\sum_{x,y} p(x,y) log p(x,y) \quad (6)$$

これを最大にするようにパラメータ λ を計算することで，最大エントロピーモデルを推定することができる．

訓練データから学習機械により，モデルが生成される．本論文では，素性をもとに生成されたモデルを**照応モデル**と呼ぶことにする．

3.2 照応解析の検索への適用

「人物名」を先行詞とした照応詞が出現している部分には，その人物の記述がなされていると考えられる．たとえば，学習機械により生成された照応モデルから，構造の素性

(article (name 先行詞候補)(body 照応詞候補))

の重みが大きく，照応関係との相関が高いと言える場合，body には name ノードに出現する人物の記述がなされていると考えられる．そのため,「人物の属性」が body に出現する場合は，記事名の「人物名」に関する属性である可能性があるといえる．

このように，XML 文書における「人物名」と「人物の属性」のキーワードによる人物検索において,「人物名」と,「人物の属性」が出現する文の主体となる人物表現との間に，先行詞と照応詞の照応関係があると認められる場合に検索の解とみなすことができる．そこで，先行詞と照応詞をそれぞれに対応させ，学習機械が生成した照応モデルを検索に適用する．このとき，照応関係を持つ確率が高いほど上位にランキングすることが良いと考えられる．

4. 実験

文書構造の素性が，XML 文書における照応に有効であることを確認するため，文書構造の素性を用いた照応解析の実験を行った．

4.1 実験環境

4.1.1 実験データ

Wikipedia [1]を XML 文書に変換した文書を題材とした．記事は以下の 4 つである．

- 例題 1:「亀田興毅」
- 例題 2:「福田康夫」
- 例題 3:「ガソリン国会」
- 例題 4:「安全保障会議」

[1] http://ja.wikipedia.org/

表 1 実験データ
Table 1 Experimental data

	例題 1	例題 2	例題 3	例題 4
照応詞候補数	333	521	6	72
先行詞候補数	75	96	4	4
総組合せ数	12597	22758	18	33
正解数	240	390	7	14
不正解数	12357	22269	11	19

例題 1，例題 2 は記事名が人物名，例題 3，例題 4 は記事名が人物名以外の例である．実験データは表 1 の通りである．人物検索への適用を考えているため，人物に関する照応を調査する．そこで，照応詞候補を文章に出現する人物を表す名詞句やゼロ代名詞，先行詞候補を人物の氏名とした．たとえば，照応詞候補として,「亀田」,「プロボクサー」など，先行詞候補として「亀田興毅」,「亀田大毅」などが挙げられる．また，ある照応詞の先行詞候補は，先行文脈，すなわち，文書の最初から照応詞候補の出現する文までに出現する人物名とする．

今回，先行文脈を照応詞候補が出現する段落内や文の数などで制限しなかった理由は，構造文書では照応詞候補と先行詞候補がテキスト上で離れているが，構造上において照応関係を持つ場合が存在するためである．また，不正解の構造パターンも学習する必要があるためである．表 1 から分かるように，照応詞候補と先行詞候補が照応関係を持つ場合以外はすべて不正解となるため，データ数が多くなるほど，不正解数は正解数に比べて極めて多くなっている．また，訓練データと分類データは，全てのデータから半数ずつランダムに選択した．

4.1.2 素性

自然言語的な素性は，文献 [7] を参考にした．実験では，茶筅 [11] と Cabocha [8] を用いた形態素解析，固有表現タグ付与，係受け解析を行い，すべての素性を自動的に抽出した．辞書を用いた意味の素性においては，EDR 電子化辞書 [12] を用いた．

Wikipedia の構造の素性は，図 2，図 3 の提案手法のアルゴリズムをもとに，自動的に生成した．全ての例題から生成される k=2 とした場合の k-照応組合せ部分木の構造の素性とその分布は，表 2 の通りである．

表 2 の 1) と 2) は同じ item ノード，同じ段落に出現する場合を示す．3) は，先行詞候補が item ノードに，照応詞候補がその子要素に出現しており，親子関係を持つことを表している．4) は，先行詞候補が記事名である name ノードに出現する場合を表しており，5) は先行詞候補が section の title に，照応詞候補がその section 内に出現する場合を示す．また，6) は先行詞候補と照応詞候補が異なる section に出現する場合を示す．6)〜10) は先行詞候補と照応詞候補が木構造の上で兄弟関係を持つことを表

表 2　全ての例題の正解照応の構造パターン分布
Table 2　Distribution of structural pattern about all examples

k-照応組合せ部分木 (k=2)	正解数
1) (item 先行詞 照応詞)	91
2) (p 先行詞 照応詞)	153
3) (item 先行詞 (normalist 照応詞))	4
4) (article (name 先行詞)(body 照応詞))	357
5) (section (title 先行詞)(section 照応詞))	4
6) (body (p 先行詞)(section 照応詞))	42
7) (body (section 先行詞)(section 照応詞))	0
8) (normalist (item 先行詞)(item 照応詞))	0
9) (section (normalist 先行詞)(normalist 照応詞))	0
10) (section(p 先行詞) (p 照応詞))	0
11) (section (section 先行詞)(section 照応詞))	0
合計	651

している．このように，提案手法によって，構造的な特徴をとらえた素性を生成することができた．また，XML の文書構造で，同じノードに出現する場合に照応関係を持つ頻度が高く，3) の親子関係を持つものや，4)，5) の name や title のノードの性質から，照応関係を持つ傾向がある構造パターンが存在することが言える．一方，先行詞と照応詞が，異なる section や item などの兄弟関係を持つ場合は，照応しないことが言える．(body (p 先行詞候補)(section 照応詞候補) の素性に，照応関係を多く持つのは,「亀田大毅」「亀田史郎」が，最初の段落で記述されて以降,「大毅」,「史郎」などと，照応表現で記述されていたためである．

4.2　評価手法

評価は学習機械により分類されたデータに対して行う．照応関係を正しく同定できた場合を正解とし，精度，再現率を以下の式を用いて求める．

$$\text{精度} = \frac{\text{照応関係を正しく同定できた数}}{\text{実際に照応があると判定された照応の総数}}$$

$$\text{再現率} = \frac{\text{照応関係を正しく同定できた数}}{\text{照応関係があると判定すべき照応の総数}}$$

精度を上げれば再現率が下がり，再現率を上げれば精度が下がる傾向にあるため，精度と再現率の調和平均である F 値を評価尺度として用いる．F 値は以下の式で求められる．

$$F\ 値 = \frac{2 * \text{精度} * \text{再現率}}{\text{精度} + \text{再現率}}$$

4.3　実験結果と考察

本手法の文書構造の素性が，ＸＭＬなどの構造文書における照応に有効であることを確認するため，文書構造の素性を用いた場合と用いなかった場合の照応解析の比較実験を行った．また，文書構造の素性としてどの k-照応組合せ部分木が望ましいか，k-照応組合せ部分木のｋの値を様々に変えて比較実験を行った．

表 3　素性の種類における実験結果
Table 3　Experiment results about performance of features

		例題 1	例題 2	例題 3	例題 4
自然言語	精度	74.3 %	76.0%	51.3 %	31.2%
	再現率	40.8 %	66.8%	57.5 %	48.7%
	F 値	52.7 %	71.1%	54.2 %	38.0%
文書構造 ＋ 自然言語	精度	77.0%	78.9%	69.3%	75.0%
	再現率	48.1%	69.2%	91.7%	54.8%
	F 値	59.2%	73.7%	74.9%	63.3%

表 4　k-照応組合せ部分木における実験結果
Table 4　Experiment results about depth of k-coreference combination subtrees

		例題 1	例題 2	例題 3	例題 4
k=2	精度	77.0%	78.9%	69.3%	75.0%
	再現率	48.1%	69.2%	91.7%	54.8%
	F 値	59.2%	73.7%	74.9%	63.3%
k=3	精度	75.4%	77.6%	69.3%	75.0%
	再現率	46.2%	68.6%	91.7%	54.8%
	F 値	57.3%	72.8%	74.9%	63.3%
k=∞	精度	72.1%	78.0%	69.3%	75.0%
	再現率	49.3%	72.0%	91.7%	54.8%
	F 値	58.6%	74.9%	74.9%	63.3%

4.3.1　素性の種類における比較実験

文書構造の素性を用いた照応解析が，XML 文書における照応に有効かどうかを確認する．そこで，(I) 自然言語的な素性のみを用いた照応解析，(II) 文書構造の素性と自然言語的な素性を用いた照応解析を実装し，比較実験を行った．構造の素性は，k=2 のときの k-照応組合せ部分木を用いた．実験結果は表 3 の通りである．

表 3 から，(I) 自然言語的な素性のみを用いた場合と，(II) 文書構造と自然言語的な素性を用いた場合を比較すると，全ての例題において，文書構造と自然言語的な素性を用いた場合の照応解析のほうが，F 値の値が向上していることが分かる．このことから，文書構造の素性が XML 文書における照応に有効であること言える．

一方，再現率が全体的に低かった．これは，正解例に比べて不正解例が極めて多く，学習が不正解に偏ってしまい，正解とすべき例に対しても不正解と判定する場合が多かったためである．

4.3.2　k-照応組合せ部分木における比較実験

照応組合せ部分木の根からの深さ 2 までを抽出した部分木 (k=2 のときのｋ-照応組合せ部分木) を構造の素性とした場合，照応組合せ部分木の根からの深さ 3 までを抽出した部分木 (k=3) を構造の素性とした場合，照応組合せ部分木そのまま (k=∞) を構造の素性とした場合による照応解析の比較実験を行った．また，素性は，文書構造と自然言語的な素性を用いる．表 4 はその実験結果である．

例題 3，例題 4 においては，深さ 3 以上の照応組合せ部分木が存在していなかったため，全て同じ評価を得た．例題 1，例題 2 において，k=2 と k=3 のときの k-照応組合せ部分木による照応を比較すると，k=2 のときの方が精度，再現率ともに良い結果を得ている．素性の削減を行わず，そのままの照応組合せ部分木を構造の素性とした場合の照応 ($k=\infty$) は，精度は低いが再現率が 3 つの中で最も高いという結果を得た．F 値は，例題 1 では k=2，例題 2 では $k=\infty$ が高くなっており，どちらがより良いか判断できない．しかし，照応組合せ部分木そのままを構造の素性として用いると，文書が複雑になるにつれ素性の個数が莫大になってしまう．また，汎用性を失い，他の文書に適用できないという問題が生じる．そのため，素性の個数を削減できるという点で，今回は k=2 の k-照応組合せ部分木を採用する．

5. おわりに

本論文では，XML 文書における構造の素性を用いた照応により，人物検索の精度を向上することを行った．構造の素性として，k-照応組合せ部分木を提案し，XML などの構造文書において，文書構造の素性を用いた照応が有効であることを確認するため，Wikipedia の XML 文書を題材として 4 つの例題を作成し，実験を行った．

実験の結果，自然言語的な素性のみによる照応解析と文書構造と自然言語的な素性による照応解析の F 値を比較すると，全ての例題において，文書構造と自然言語的な素性による照応が良い結果を得ていた．そのため，文書構造の素性が構造文書の照応に有効であることが言える．また，文書構造の素性としてどの k-照応組合せ部分木が望ましいか，k-照応組合せ部分木の k の値を様々に変えて比較実験を行った．評価結果に大きな差は見られなかったが，素性の個数が削減できるという点から，今回は $k=2$ のときの k-照応組合せ部分木を採用した．

今後の課題として，正負例の偏りの問題を解消し，再現率の向上を行うことが挙げられる．また，例題を増やして，他の Wikipedia の記事やその他の HTML 文書においても実験を行い，本手法の有効性を確認したい．実際に本手法を適用した検索システムを実装することも重要な課題である．さらなる発展として，人物だけでなくその他の固有表現の照応解析も行い，人物検索だけでなく，エンティティ検索に本手法を適用することが考えられる．

[謝辞]

本研究の一部は，平成１９年度科研費基盤研究 (B)(課題番号 18300031), および科学技術振興機構 戦略的国際科学技術協力推進事業 「アイデンティティ連 携におけるリスクを考慮した個人情報共有方式」による．

[文献]

[1] 小林のぞみ，飯田龍，乾健太郎，松本祐治，照応解析手法を利用した属性ー評価値対および意見性情報の抽出，言語処理学会第 11 回年次大会発表論文集, 2005.

[2] 二本木智洋，住田一男，文の構造化による口コミ評価の分析・検索，インタラクション 2002 論文集，pp175-176, 2002.

[3] Hsin-Hsi Chen, Shih-Chung Tsai, Jin-He Tsai, Mining Tables from Large Scale HTML Texts, 18th International Conference, Computational Linguistics, pp.166-172，2000.

[4] Minoru Yoshida, Kentaro Torisawa, Junichi Tsujii, Extracting ontologies from World Wide Web via HTML tables, Pacific Association for Computational Linguistics, pp.332-341,2001.

[5] 出原博，藤本典幸，竹野浩，萩原兼一， WWW 画像検索における画像周辺の HTML 構文構造を考慮した画像説明文の抽出手法, 信学技報，DE2005-136，2005.

[6] 是津耕司，田中克己，Web からの画像の文脈情報の抽出と提示，DEWS，6-p-05，2003.

[7] 飯田龍, 乾健太郎, 松本裕治, 関根聡， 最尤先行詞候補を用いた日本語名詞句同一指示解析, 情報処理学会論文誌, Vol 46, No. 3，2005.

[8] 工藤拓，松本裕治，Support Vector Machine を用いた Chunk 同定, 自然言語処理, Vol. 9, No. 5, pp.3-21，2002.

[9] Andrew Kehler, Probabilistic Coreference in Information Extraction,CoRR, cmp-lg/9706012,1997.

[10] Adam L. Berger, Stephen A. Della Pietra, Vincent J. Della Pietra， A Maximum Entropy Approach to Natural Language Processing, Computational Linguistics, 22 ，1996.

[11] 松本裕治, 北内啓, 平野善隆, 松田寛, 高岡一馬, 浅原正幸, 形態素解析システム『茶筌』version 2.3.3 使用説明書, 奈良先端科学技術大学院大学 ，2003.

[12] 日本電子化辞書研究所，EDR 電子化辞書仕様説明書第 2 版.，Technical Report TR － 045, 1995.

米井 由美 Yumi YONEI
京都大学大学院情報学研究科修士課程在学中．2007 京都大学工学部情報学科卒業．日本データベース学会学生会員．

岩井原 瑞穂 Mizuho IWAIHARA
京都大学大学院情報学研究科准教授．1993 九州大学工学研究科博士後期課程修了，工学博士．情報処理学会，電子情報通信学会，ACM, IEEE 各会員．日本データベース学会正会員．

吉川 正俊 Masatoshi YOSHIKAWA
京都大学大学院情報学研究科教授．1985 京都大学大学院工学研究科博士後期課程了，工学博士．XML データベース，多次元空間索引等の研究に従事．ACM，IEEE Computer Society 各会員．日本データベース学会理事．