# 二段階洗練化手法による新聞記事からの人物説明記述の抽出

西田 成臣† 森 辰則‡
†横浜国立大学 大学院 環境情報学府 ‡横浜国立大学 大学院 環境情報研究院
E-mail: {aki,mori}@forest.eis.ynu.ac.jp

## 1 はじめに

電子化文書の増加に伴い，必要な情報を効果的に抽出する技術が求められている．このような情報の一つに人物情報が存在する．人物情報を抽出することはそれ自体に価値があるだけでなく，得られた情報を利用することで，名寄せや人物の評判情報といった，更なる活用が期待できる．このような背景の下，評価型ワークショップ NTCIR-7[1] では人物情報の質問応答を含むタスクが設定されている．

本稿では，質問応答の一環となる新聞記事から人物の説明記述 (以下，人物説明記述) のまとまりを抽出することを目標として，次の二段階に分けて，抽出する文書部分を洗練化する手法を提案する．まず，解候補となる文書部分に対して，人物説明記述を含むか否かを機械学習により判定を行う．続いて，人物説明記述を含むと判定された文書部分を起点として人物説明記述がどこまで連続していて，終点がどこであるかの判定をする．また，ベースラインとして系列ラベリングに基づく手法を検討し，これとの比較により提案手法の有効性を確認する．

## 2 研究背景と関連研究

質問応答とは，ユーザからの質問に対する答えを返す技術のことであり，そのタスクは主に二種類に大別される．一つは，人名や地名，組織名といった比較的短い言い回しで回答が得られる factoid 型の質問応答であり，もう一つは，定義や理由，方法といった比較的長い言い回しを回答に必要とする non-factoid 型の質問応答である．本研究では与えられた人名に対して，対象となる人物の情報を情報源から抽出することを目的としており，その情報は比較的長い言い回しを必要とすると予想される．ゆえに，人物情報の抽出は non-factoid 型の，特に事物の内容を明らかにする定義型質問応答の一環と考えることができる．

### 2.1 non-factoid 型質問応答

前述のとおり， non-factoid 型質問応答では，定義や理由，方法のように，回答に複数の文を要求する問題に答えることを目的としており，これに対するアプローチには二種類が考えられる．定義や理由といった回答に要求されるものを基にして質問の型を限定し，それぞれの型ごとに処理を行う方法と，質問の型によらず統一的に行う方法である．

本研究では，質問の型を定義型，その中でも人物情報に限定して処理を行う．統一的に行う方法に比べて，型ごとに処理方法を用意するためにシステム構築のコストが高いものの，回答に特化した部分システムを用意することができるため，より精度の高い回答を得られる可能性が高いという利点がある．

### 2.2 定義型質問応答

定義型の質問とは，物事の性質や素性などその定義を問う質問である．

英語の定義型質問に対し， Han ら [2] はコーパスから推定した確率モデルに基づいて，質問に対する回答の内容の関連性と，型に応じた記述スタイルを満たすかという計算を行っている．

直接定義型の質問に関係しているわけではないが，藤井ら [3] は， Web より収集した文書群より事物の用語を説明している部分を抽出，組織化することで事典検索サイト Cyclone を構築しており，用語の説明部分を抽出するという点において定義型質問応答および，本研究と関連している．

## 3 本研究におけるタスク定義

本研究の目的は知りたい人物の説明記述である文書部分の集合を，情報源となる文書集合から抽出することにある．そこで，説明記述に対する抽出処理を考えるため，本研究において抽出の対象とした新聞記事における人物説明記述の調査を行った．調査対象は 98 年の 1 月の記事中から無作為に抽出した記事 400 件とした．なお，本研究で構築する，人物の説明記述を抽出するシステムでは，毎日新聞の 98 年から 01 年までの 4 年分を抽出対象として使用する．次に新聞記事中の人物説明記述の例を示す．

> 田中勇氏（たなか・いさむ）１９６１年早大政経卒、日本アスベスト（現ニチアス）入社。常務を経て９７年６月から専務。神奈川県出身、６３歳。（６月２９日就任。音馬峻社長は会長に）
> ◇エリア・カザン映画監督。１９０９年、トルコ生まれ。１３年、アメリカに移住。４０年代から舞台演出家、映画監督として活躍。『紳士協定』（４７年）、『波止場』（５４年）でアカデミー監督賞受賞。９９年には同特別名誉賞を受賞した。

新聞記事においては，その記事に登場する人物について，補足的な説明を与える意図で，記事の本文とは別に人物の略歴等の説明記述が置かれることが多い．一方で，記事の本文においては，ある人物が関わった出来事等の説明が現れる．人物に対する質問応答においては，上記の両者を回答候補とする立場もあるが，本研究では，人物自身を特徴付ける解説を人物の説明記述であると定義し，前者のみを抽出対象とする．

このような考えのもとに新聞記事中を調査し，発見された人物説明記述において確認された特徴を次に記す．

- 人物の説明記述は複数文に渡り連続して書かれる
- 人物の説明記述の開始文には説明対象の人名が含まれる

これらの特徴に関しては，少なくとも調査した新聞記事においては例外はなかった．

以上の目的および抽出対象の新聞記事の調査結果を考慮すると本研究のタスクは次のようになる．

1. 説明記述を抽出する対象の人名をユーザから受け付け，その人名が登場する文が人物の説明記述の先頭になっているかを判定
2. 1で説明記述の先頭であると判定された文を先頭として後続する文群のどこまでが一連の説明であるかを判定

## 4 提案手法

新聞記事における人物説明の記述スタイルには「〜生まれ」や「〜卒」といった特徴的なものが多く含まれているため，これらの特徴をより効果的に利用するということを考える．このような文の記述スタイルに着目した場合，文の係り受け構造を利用して機械学習を行う手法が効果を上げている [4]．これを踏まえて，本研究においても，文の記述スタイルを活かして，係り受け構造を利用し学習・判定を行うこととした．

### 4.1 系列ラベリングに基づく手法

人物の説明記述が複数の文の列によって構成されていると考えれば，人物の説明記述の抽出は，固有名詞の抽出などの際に行われるチャンク同定の問題へ帰着できると考えられる．チャンクの同定を考える場合にはチャンクをどのように表現するかが重要な問題となる．この問題に対しては，チャンクを構成する各要素に対して，チャンクに関する状態を表すラベルを付与するという手法が広く用いられている．本研究においては，Tjong らにより提案されている [5]，IOB1 法を基にした IOB2 法を文の構成要素として利用する．すなわち，各文に対して，以下の三種のラベルを割り振ることにより，チャンクに関する状態を表現する．チャンク自身は，ラベルの系列により表現される．この手法をベースライン手法とする．

I ： 現在位置の文はチャンクの先頭以外の一部である
O ： 現在位置の文はチャンクに含まれない
B ： 現在位置の文はチャンクの先頭である

本研究での抽出対象で考えれば，人物の説明記述の開始文に B，開始文以外の説明文に I，それ以外の無関係の文に O のラベルを与えることとなる．これにより，文単位に対して状態のラベリングを行ったコーパスを準備し，各文のラベルをクラスとした分類器を機械学習により構成することで，複数文から構成される人物説明記述を抽出することが可能となる．本研究ではラベリングの学習と判定に BACT-0.13[4] を使用することとした．BACT は，ラベル付き順序木の分類器を構成する，Boosting を用いた機械学習システムである．本研究では文に対して，CaboCha を使用して文節単位の係り受け構造を導出した後，各文に対して上記のラベルを分類クラスとして与える．このときの文節単位での係り受け構造において，表層表現のみ利用した．

また，BACT での分類は正負の二値により行われ，この判定のみでは IOB の三値の分類は行うことはできないため，One-vs-Rest 法を用いて三値分類へ拡張した．なお，本研究では各文のクラスの学習・判定に対して，分類の対象となる文とその前後の文の計三文により学習・判定を行うこととした．具体的には次のように行う．学習・判定の対象となる文を係り受け解析し，表層表現の係り受け関係をラベル付き順序木を表現する $S$ 式に変換したものを $S_i$ とし，同様に，対象の前後の文から得た $S$ 式をそれぞれそれぞれ $S_{i-1}, S_{i+1}$ とする．これらの式が一つにまとまっている木を得るために，これらがかかっている先のノードを仮想的に考え，それを文字列"$P$" で表現して，$(P\ S_{i-1}\ S_i\ S_{i+1})$ という構造で一つの $S$ 式にまとめたものにクラスを付与することとした．

### 4.2 二段階洗練化手法

Kaynak らによれば，分類問題において，カスケーディングと呼ばれる，複数の分類器により段階的に分類処理を行う手法を用いることで，複雑さとコストを増加させることなく，精度を上昇させることが出来るという報告がなされている [6]．本研究では，これを踏まえて，人物の説明記述の抽出処理を二段階に分けることにより，抽出結果を洗練化する手法を提案する．

#### 4.2.1 一段階目の抽出法

一段階目の抽出処理では，解候補となる文書部分が人物に関する説明記述を含むかどうかを大まかに判定することを目的とする．ここで，解候補は，人名を開始文に含む文書部分であり，あらかじめ決められた数 $n$ の文からなるものとする．第一段階目においては，人物の説明記述の出現箇所を見つける大まかな判定を行うだけであるので，固定長のパッセージを分類対象としている．具体的には 4.1節で三文を窓幅として一つの事例を構成した方法と同様にして行う．例えば，$n$=5 とした場合には，解候補の先頭の文から五文それぞれを表層表現に基づくラベル付き順序木に対応する $S$ 式に変換する．変換された $S$ 式を $S_1, S_2, S_3, S_4, S_5$ とすれば，これらの式を一つの木にまとめるための仮想的なノードである文字列"$P$" を用いて，$(P\ S_1\ S_2\ S_3\ S_4\ S_5)$ という式を構成し，これを事例として BACT による学習・分類を行う．

#### 4.2.2 二段階目の抽出法

二段階目の抽出処理の目的は，一段階目の判定において人物説明記述を含むと判定された文書部分が人物説明文書の先頭の位置を与えるとした上で，説明記述の範囲を限定することにある．すなわち，一段階目で得られた文書部分の開始文を起点として，後続する各文が人物の説明記述を含むか否かの判定を行い，人物説明記述の末尾となる文を見つけ出す．なお，二段階目では判定対象が人物説明を含むと判断された文書部分であるため，一段階目とは異なり，文書中において，人物の説明記述が現れる箇所のみから学習事例を構成すればよい．

これらを踏まえた上で，二段階目の抽出では，連続する二文を単位として，これを，a) 二文とも人物説明文であるか，b) 一文目が人物説明文であり，二文目が無関係の文であるか，の判定を行う．すなわち，人物の説明記述とその周囲に現れる二文が人物説明文の連続であるなら正のクラスとし，切れ目であるならば負のクラスとして訓練・評価事例を作成する．

## 5 評価実験と考察

評価実験を行うために，訓練用ならびに評価用の注釈付コーパスを，それぞれ，毎日新聞の記事 98 年 1 月一か月分ならびに同 2 月の前半半月分から作成した．3章で述べたように，質問応答では対象となる人名がユーザにより与えられる．そのため，テキスト中に現れる人名を網羅的に抽出することは，ここで対象とする人物説明記述抽出タスクの一部ではないと考える．よって，コーパス作成においても，人名の網羅性については追求せず，コーパスに対して ChaSen2.3.3 を用いて形態素解析を行い，形態素の品詞細分類が「人名」となった形態素を含む文を選択した．そして，その文を含む周辺の文脈について，そこに人物説明記述が現れているのかという判断を行い，現れている場合には，その範囲も判断し，注釈付けを行った．人物説明記述が現れている場合には，人物説明記述が現れている範囲を<Bio judge="yes">...</Bio> というタグで囲んだ．また，人物説明記述が現れていない場合には，負例を作成する目的で，「人名」となった形態素を含む一文を<Bio judge="no">...</Bio> のタグで囲むこととした．なお，負例はこの文を開始文とし，4.2.1節で述べた窓幅である $n$ 文からなる文書部分である．

ここで、系列ラベリングに基づく手法、ならびに、提案手法のいずれにおいても、利用者が与えた人名が含まれる文書部分を対象として抽出処理を行うことが前提となっていることに注意されたい。そのため、各手法における訓練時においてはもちろんのこと、評価時においても文書集合全体を処理対象とするのではなく、上記の正例ならびに負例である文書部分 (と手法に応じた前後数文) のみを処理対象にする。表 1に，コーパスに現れる人物説明記述の数，ならびに，第一段階の評価に用いる負例の数を示す．本研究では，注釈付けした新聞記事コーパスの 98 年 1 月一か月分を訓練事例として，同 2 月前半半月分を評価事例として使用する．なお，人物説明記述の数は，注釈付した人物説明記述の文章のひとまとまりを 1 としている．

表 1: 作成した学習・評価用のデータ

| | 人物説明記述の数 | 負例の数 |
|---|---|---|
| 訓練事例 | 576 | 9019 |
| 評価事例 | 298 | 3796 |

### 5.1 系列ラベリングに基づく手法におけるラベルの分類精度

4.1節で述べた手法に基づいて，注釈付コーパスを IOB2 方式によりラベル付けしたときのラベルの数を表 2に示す．この学習用事例から分類器を構成したときに，評価用事例を分類した結果を表 3に示し，各ラベルごとの Precision， Recall， F 値を表 4に示す．

表 2: IOB2 法により付与したラベルの数

| | I | B | O |
|---|---|---|---|
| 訓練事例 | 2090 | 576 | 37932 |
| 評価事例 | 871 | 278 | 16165 |

各ラベルの分類精度の結果である表 4を見れば，人物説明記述の先頭である B の Precision が極端に悪いものとなっていた．表 3で示したラベルの判定結果によれば，正解ラベルが B の人物説明記述の先頭である文に対しては， I と判定してしまうことは極めて少なかったのに対して，正解ラベルが I の先頭以外の人物説明記述である文に対しては， B となり人物説明記述の先頭であると判定してしまう割合が多いことで Precision を低いものとしてしまっている．この I を B と誤って多く判定していまう影響は I に対するラベルの判定結果にも表れており， I の Recall を低くする要因となっている．

表 3: 系列ラベリングに基づくラベルの判定結果

| 正解ラベル | 判定ラベル | 判定数 |
|---|---|---|
| I | I | 702 |
| | O | 51 |
| | B | 118 |
| O | I | 60 |
| | O | 16040 |
| | B | 65 |
| B | I | 7 |
| | O | 28 |
| | B | 243 |

表 4: 系列ラベリングに基づく手法における各ラベルの分類精度

| | Precision | Recall | F1 |
|---|---|---|---|
| I | 91.3 | 80.1 | 85.3 |
| O | 99.5 | 99.2 | 99.4 |
| B | 66.0 | 87.4 | 75.2 |

これは他のラベルに対して人物説明記述の先頭のデータ数が比較的少なかったことや，人物説明記述の先頭の文と，それ以外の人物説明記述の文が比較的近い特徴表現を持っている場合に，誤った判定をしてしまうことが考えられる．

### 5.2 二段階洗練化手法

系列ラベリングによる実験と同様，作成したコーパスを基に，評価実験を行うこととなるが，提案手法では抽出が二段階に分かれているため，以下では，まずは二段階を独立させてそれぞれで評価実験を行う．

#### 5.2.1 一段階目の文書部分の分類精度

4.2.1節で述べた手法に基づき，一段階目における文書部分の分類実験を行った．このときの評価指標は，Accuracy， Precision， Recall， F 値とした．表 5に実験結果を示す．

表 5: 一段階目の文書部分の分類精度

| Accuracy(%) | Precision(%) | Recall(%) | F1 |
|---|---|---|---|
| 98.9 | 92.5 | 91.6 | 92.1 |

結果を見れば， Precision， Recall ともに 9 割を超える精度となっており，一段階目の時点で比較的高い精度で人物説明記述を含むかどうかという文書部分の分類が可能であることが確認できた．これにより，二段階目の説明境界の抽出時には，人物説明記述文のみを扱い局所的な事例に特化した訓練事例を作成することで，十分な分類精度を実現できるのではないかと考えられる．

#### 5.2.2 二段階目における説明境界の分類精度

4.2.2節で述べた手法に基づき，二段階目における説明記述の境界を発見する問題についての分類実験を行っ

た．作成した事例における人物説明記述の末尾の境界の数及び，隣接する二文がいずれも人物説明記述である場合の数は表6の通りである．この訓練事例を用いて構成された分類器について，評価事例を用いて評価した結果は表7となった．

表 6: 人物説明記述の境界及び隣接二文が人物説明記述の一部である場合の数

| | 末尾の境界数 (負) | 人物説明記述の連続 (正) |
|---|---|---|
| 訓練事例 | 532 | 2046 |
| 評価事例 | 278 | 871 |

表 7: 二段階目における説明境界の分類精度

| Accuracy(%) | Precision(%) | Recall(%) | F1 |
|---|---|---|---|
| 95.9 | 96.9 | 97.7 | 97.3 |

表7によれば，Precision，Recall ともに高い精度で人物説明記述の境界の判定が行えていることが確認できる．これより，やはり 5.2.2節で述べたように，局所的な分類問題とすることにより，精度を高くすることができたと考えられる．

### 5.3 各手法における説明記述の同定精度の評価

前節までにおいて，各手法について，その部品となる分類器の精度を検証したが，本節では，各手法について，最終的な説明記述の抽出精度を比較検討する．比較検討には，前節までと同様，注釈付けした正例と負例の評価事例を使用する．比較検討は具体的には，以下の方法により行う．系列ラベリングに基づく手法では 5.1節における分類結果を用いて，文の列において，B,I,I,...,I,O という系列を見つける．このとき，B から O の直前までの文書部分を説明記述として抽出する．一方で，二段階洗練化手法では，まず 4.2.1節の分類手法により説明記述が含まれていると判断される文書部分を得る．その後に 4.2.2節の分類手法により説明記述の末尾の境界を認定し，説明記述として抽出する．

以上の抽出手法の精度の比較を行うために次のような判定の基準を設ける．開始文から判定を行い，開始文とそれ以外の説明文が過不足なく判定できていれば完全抽出とする．完全抽出に加えて，途中で判定を誤る場合と，反対に正解となる範囲を超えて過剰に抽出してしまう場合とを合わせたものを部分抽出とする．ただし，系列ラベリングに基づく手法における同定において，人物説明の開始が正しく判定されず，I,I,I,I といった系列が見つかった場合でも，その系列が正解の説明記述と重なりがある場合には部分抽出としている．また，無関係な文の連続を誤って抽出した場合には，誤抽出とする．

以上の判定基準を基に，各抽出手法で評価実験を行った．評価指標としては部分抽出と完全抽出のそれぞれに対して Recall，Precision，F 値を用いた．その結果を表8に示す．

結果を F 値により比較すると，系列ラベリングに基づく手法よりも抽出処理を二段階に分けて文書部分を洗練化することで，部分・完全抽出ともに精度の向上が見られた．部分的な抽出を許す場合では F 値で 4 ポイント程度の差となっていたが，過不足の無い抽出のみを解とする場合には，部分的な抽出以上となる 8 ポイントを超える精度の向上となっていた．一段階目の処理では

表 8: 各抽出法の同定精度の比較

| | | Precision(%) | Recall(%) | F1 |
|---|---|---|---|---|
| 系列ラベリング | 部分 | 91.0 | 88.3 | 89.6 |
| | 完全 | 88.4 | 66.8 | 76.1 |
| 二段階洗練手法 | 部分 | 94.2 | 93.3 | 93.8 |
| | 完全 | 93.2 | 77.2 | 84.4 |

文書部分を単位として人物説明記述の判定を行っているが，この判定自体で説明記述の特徴が学習しやすく，これにより一段階目の時点でかなりの精度で判定が出来ている．さらに，二段階目では切れ目を判定するという，限定された状況において，より詳細な分類を行うタスクに限定できたことが，系列ラベリングに基づく手法よりも精度の上昇が見込めたのではないかと考えられる．

## 6 おわりに

本研究では，質問応答の一環となる人物説明記述が含まれる文章の抽出を行う手法として次の二つの手法を提案した．一つは系列ラベリングに基づく手法であり，本研究ではこれをベースラインと設定した．もう一つは，抽出方法を次の二段階に分ける手法である．一段階目の処理では，解候補が人物の説明記述であるかどうかの判定を行い，続いて二段目の処理では，人物説明記述の境界がどこであるかを判定することにより抽出する文書部分の洗練化を行う．

二段階洗練化手法を用いることで，部分的な抽出を許す抽出では 4 ポイント，判定の厳しい過不足のない抽出では 8 ポイントの F 値の向上がみられ，本研究の提案する手法の有効性を確認することができた．

今後の予定としては，現状の情報源である新聞記事から，より情報量が膨大かつ多様性のある Web 記事からの抽出へと拡張してゆくことを考えている．また，本研究では，ある程度まとめられた略歴相当を抽出の対象としているが，利用者が必要とする人物情報はそれ以外にも，関係情報や評判情報といった多様なものが想定される．そこで，利用者が必要としている情報にそった情報を提供できるよう，予め人物情報を分類して，人物情報の明確化を行うことを検討している．加えて，現状では同姓同名といった人名の曖昧性に対処できていないため，この点を考えることも今後の課題となる．

## 参考文献


[1] NTCIR-7 Task Definition. http://aclia.lti.cs.cmu.edu/wiki/TaskDefinition

[2] Kyoung-Soo Han, Young-In Song, Hae-Chang Rim. Probabilistic model for definitional question answering. In Proceedings of *SIGIR* 2006, pp. 212–219, 2006.

[3] Atsushi Fujii, Katunobu Itou, Tetsuya Ishikawa. Cyclone: An Encyclopedic Web Search Site. In Proceedings of The 14th International World Wide Web Conference (WWW2005),pp.1184-1185,2005.

[4] Taku Kudo, Yuji Matsumoto. A Boosting Algorithm for Classification of Semi-Structured Text. In Proceedings of EMNLP 2004,pp301-308,2004.

[5] Erik F. Tjong, Kim Sang, Jorn Veenstra. Representing text chunks. In Proceedings of EACL ' 99,pp.173-179,1999.

[6] Cent Kaynak, Ethem Alpaydin. MultiStage Cascading of Multiple Classifiers: One Man's Noise is Another Man's Data. In Proceedings of the Seventeenth International Conference on Machine Learning,pp.455-462,2000.