Webからの人物の属性情報抽出

People Attribute Extraction from the Web

渡部 啓吾[*1] Danushka Bollegala[*1] 松尾 豊[*2] 石塚 満[*1]
Keigo Watanabe Yutaka Matsuo Mitsuru Ishizuka

[*1]東京大学大学院 情報理工学系研究科
Graduate School of Information Science and Technology, The University of Tokyo

[*2]東京大学大学院 工学系研究科
School of Engineering, The University of Tokyo

Personal names are among one of the most frequently searched items in web search engines. Extracting information in the form of attributes and values for a particular person enables us to uniquely identify that person on the web. For example, although namesakes share the same name they usually have different date of births or affiliations. Given a set of documents retrieved for a particular person, we propose two stage approach to extract values for a set of attributes for that person. In the first stage we mark all potential attribute strings in a given text. The second stage then attempts to select the attribute values relevant to a person name. We use a named entity recognition tool to mark all occurrences of named entities in a given document. We then use a rule-based tagger to identify the variants of the given person name. Next, we employ a combination of rules and pre-compiled attribute value candidate lists to extract values for a given set of attributes. The candidate value lists are manually created using resources available on the web such as Wikipedia. The proposed method is evaluated on the test data collection created for the attribute extraction subtask at the second Web People Search Task (WePS). According to the results in the official evaluation, the proposed method is ranked 5-th among the 15 participating systems.

1. はじめに

Web 上で人は生年月日や職業といった多くの属性情報と結び付けられている．それらの情報を正確に抽出することは，Web 上に出現する人々を明確に識別する上で非常に重要である．例えば同姓同名問題の場合を考えてみると，その人々は同じ名前を共有しているが，所属や国籍，生年月日，出身地などその他の属性は恐らく異なっている．そのため，人物に関する属性情報を抽出することは，同姓同名問題を解決するのに役立つと言える．例えば “Jim Clark” という人名について考えてみると，ある一人はレーシングドライバーであり，またある一人はネットスケープ社の創始者であり，大学教授でもある．この二人を区別したいと思えば，職業という属性に着目することで，ドライバーあるいは教授であると判断することが出来る．

人物の属性情報の抽出タスクは，これまで主に同姓同名問題解決のためのサブタスクとして扱われてきた．Mann らは（名前－生年月日）などのシードを用意し，そこからパターンを生成することで属性の抽出を行い，人名の曖昧性解消を試みた [Mann 03]．また上田ら [上田 08]，木村ら [木村 06] は人物に関連する職業や地方をリストを元に抽出し，同姓同名問題の解決に役立てた．これらの研究は同姓同名問題解決のために，上手く属性情報を利用する一方で，抽出した属性情報自体の精度はあまり重要視されてこなかった．本論文では，従来研究と異なり，属性情報抽出自体の精度を高めることを目的とする．属性情報抽出を上手く行うことが出来れば，同姓同名問題の解決だけでなく，データベースの構築や，検索時のクエリ拡張にも応用が可能である．

本論文では，与えられた人名の属性情報を Web 上の文書から抽出する手法を提案する．対象とする属性は表 4 の 16 種類とする．この属性は第 2 回 Web People Search Task(WePS)[*1] のサブタスク（Attribution Extraction Task）で評価に利用したものを選んだ．本論文では，2 章で提案手法について詳しく説明し，3 章で提案手法の評価を行う．最後に 4 章で本研究についてまとめる．

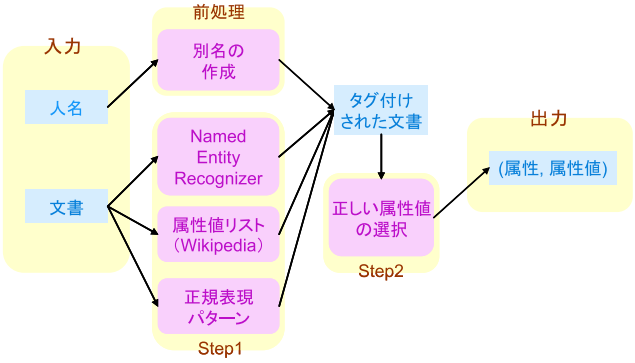

図 1: 全体の流れ

2. 手法

Web は非常に有用なリソースであるが，その文書には新聞や雑誌などに比べ整形されていない文章が多いため，普通に情報を抽出しようとすると，意図しない無関係な語も多く含まれてしまう．そのため，得られた属性の中から適切な属性を選ぶ操作が必要となる．そこで，本論文では以下のように二段階の手法を提案する．

連絡先: 渡部啓吾, 東京大学大学院 情報理工学系研究科, 東京都文京区本郷 7-3-1,watanabe@mi.ci.i.u-tokyo.ac.jp

*1 http://nlp.uned.es/weps/

表 1: 人名の表記ゆれの生成ルールと具体例

ルール	例: George Bush
（名前）＋（苗字）	George Bush
（苗字）＋（名前）	Bush George
間にカンマが入るもの	Bush, George
間に一単語入るもの	George Walker Bush
（名前のイニシャル）＋（苗字）	G Bush
（苗字）＋（カンマ）＋（名前のイニシャル）	Bush, G
（名前のイニシャル）＋（一単語）＋（苗字）	G. W. Bush

表 2: 組み合わせる呼び名

Mr., Mrs. Miss., Ms, Rev., Prof., President, Minister, Prime Minister, General, Madame, Lady, Dr., King., Queen, Vice President, Senator, Lawyer, Major, Maj., General, Gen., Maj. Gen., Major General, Jr.

Step1: 属性値の候補となる語を見つける
Step2: 候補の中からふさわしい属性値を選択する

全体の流れを 1 に示す．以下，各 STEP について説明する．

2.1 前処理

本論文では，Web 上の文書を属性抽出の対象としているが，HTML タグがついている文書では Named Entity Recognizer（以下，NER と呼ぶ）を上手く利用することが難しい．そのため，まず初めに HTML タグを取り除く必要がある[*2]．

また，実際に属性値の候補を抽出したとき，その属性値がどの人名に結び付けられているのかを判断するために，入力となる人物の名前がどこに出現するのかをマークする必要がある．そのため，入力として与えられた人名の別の表現方法を，人手で作ったルールに基づいて生成する．そのルールと具体例を表 1 に示す．このとき，表 2 の呼び名を組み合わせたものも人名の表現とする．

2.2 Step1: 属性値候補のマーク

Step1 では前処理の終ったテキストに対して，属性値候補の出現位置をマークする．属性値の候補となりうるものは，それぞれの属性によって大きく異なる．例えば，生年月日を抽出しようと思えばその候補は数字などの日付表現となるが，所属を抽出しようと思えばその候補は組織の名前などになる．そのため，属性値候補の取得には，それぞれの属性値ごとに以下のような三つの手法を用いる．

リストマッチング 属性値の対象となるエンティティのリストを Wikipedia を元に作り，テキストとのマッチングを行い候補とする．Affiliation，Award，Birthplace，Degree，Major，Nationality，Occupation，School に利用する．

NER によるタグ付け NER[*3]によって人名，地名，組織名がタグ付けされるので，それを候補とする．人名と判断されたものは，Mentor，Relatives に，地名は Birthplace に，組織名は Affiliation に利用する．

正規表現によるマッチ 電話番号のように表現方法が限られているものに対して，正規表現によるマッチングを行う．Date of birth，E-mail，Fax，Phone，Web site に用いる．

属性ごとの詳細を表 3 に示す．

2.3 Step2: 正しい属性値の選別

Step2 では Step1 で得られた属性値の候補全てに対して，信頼度のスコアを計算し，その信頼度がある閾値を超えたものを正しい属性値として取得する．スコアは以下のような基準を元として，ヒューリスティックに決定した．

- 人名の近くに出現するほどスコアを高くする．ただし，他の人名をまたぐ範囲には適用しない．
- 属性ごとに設定された手がかり表現が近くに出現した場合はスコアを高くする．このとき，手がかり表現の前に候補が出現するか，後に出現するかによってもスコアは異なる．手がかり表現としては，正解となる属性値が出現した文脈において，頻出する表現を用いた．手がかり表現の例を表 3 に示す
- 属性値の候補が人名の一部を含んでいる場合はスコアを高くする．ただし，e-mail など一部の属性のみを対象とする．

用いた手がかり表現を表 3 に示す．ただし選別の際に e-mail のみ，例外リストを用いて一部頻出の表現を除いた．

3. 評価

提案手法を WePS2 のデータセットを用いて評価を行った．このデータセットは 30 の人名がそれぞれ出現するウェブ上の文書 3468 ページを対象とし，一人当たりの平均ページ数は 115.6 である．そのページから評価に用いるのに適切で無いと判断された 585 ページが除かれ，評価には 2883 ページが用いられた [Sekine 09]．属性ごとのマッチ数（Match），超過して生成した数（OVG），取れなかった数（Miss），適合率（Precision），再現率（Recall），F 値（F-measure）を表 4 に示す．トータルの F 値は 0.083 と低い値であったが，評価に参加した 15 システムの中で 5 番目の成績であった．F 値が最

*2 本研究では http://www.oluyede.org/files/htmlstripper.py を用いた．
*3 本研究では Stanford Named Entity Recognizer を用いた．（http://nlp.stanford.edu/software/CRF-NER.shtml）

表 3: 属性ごとの抽出条件

Attribute	候補のマーク	手がかり表現
Affiliation	会社名リスト（43040），大学名リスト（1726），NER-組織名	enter, member of, work for...
Award	賞リスト（454）	None
Birthplace	NER-地名	birthplace, born in...
Date of birth	$(\backslash d+/\backslash d\{1,2\}/\backslash d\{1,2\}$...	born, birth...
Degree	学位名リスト（175）	recieve, get, degree...
E-mail	$[\backslash w\backslash -\backslash .]+@(\backslash w[\backslash w\backslash -]+\backslash w\backslash .)+[\backslash w\backslash -]+$...	例外リスト（webmaster@domain...）
Fax	$+\backslash d\{1,3\}[\backslash -\ ]?\backslash(?\backslash d\backslash)?[\backslash -\ ]?\backslash d\{1,5\}$...	fax
Major	専攻名リスト（318）	degree, major in
Mentor	NER-人名	work with, coach, trainer...
Nationality	国籍名リスト（442）	None
Occupation	職業名リスト（666）	position, serve, title...
Other name	前処理で生成した別名	a.k.a., other name...
Phone	$+\backslash d\{1,3\}[\backslash -\ ]?\backslash(?\backslash d\backslash)?[\backslash -\ ]?\backslash d\{1,5\}$...	tel, phone, mobile
Relatives	NER-人名	spouse, brother, sister, wife, father...
School	学校名リスト（25271）	school, graduate...
Web site	$https?://[\backslash w\backslash -\backslash .]+(:\backslash d+)?(/([\backslash w/\_\backslash .]*(\backslash ?\backslash S+)?)?)?$	None

も高いシステム [Chen 09] で 0.122，適合率が最も高いシステム [Chen 09] で 0.304，再現率が最も高いシステム [Balog 09] で 27.4 であるため，属性抽出というタスクは非常に難しいタスクであると言える．

属性ごとに詳細を見ていくと，おおよそ以下の三パターンに分けられることが判る．Birthplace，Date of birth，E-mail，Nationality，Phone など比較的 F 値が高い属性，Award，Degree，Fax，Major，Mentor，Relatives，などそもそもマッチ数が無いかまたはほとんど無い属性，その他残りの属性である．比較的高い F 値を出した属性は，Step1 の段階である程度候補が絞り込めているからだと考えられる．例えば，E-mail や Phone は正規表現で簡単に表現可能であり，Nationality などは国籍についての表現が限られており，その表現が出た際は本人がその国籍である可能性が非常に高いからだと推測できる．そもそもマッチ数が少ない属性の原因は Step1 の候補抽出の段階で，ほとんど正解となる候補が取れていないためである．これは，属性ごとに理由は異なると考えられるが，例えば Fax であれば，電話番号と区別することが難しいこと，Mentor，Relatives であれば NER の性能に左右されてしまうことが原因として挙げられる．また，提案手法は相対的に再現率が高く，適合率が低くなっているが，この精度を全体的に上げるためには Step2 において適切な属性値の選択を行う必要がある．

4. まとめ

本論文では，ウェブ上の文書から人物の属性を抽出する手法を提案した．提案手法は二つの Step から成り，Step1 ではリストや NER，正規表現などを用いて属性値となりうる候補に印をつけ，Step2 ではその候補の中から人名との距離，手がかり表現の有無などを元に適切な属性値を選別する．提案手法は WePS2 のデータセットで評価し，全体の F 値が 0.083 であり，参加 15 システムの中で 5 番目の成績であった．属性の抽出は非常に難しいタスクであるが，人名の曖昧性解消だけでなく，データベースの構築や，検索時のクエリ拡張などにも利用でき，挑戦しがいのあるタスクと言える．提案手法はヒューリスティックな部分が大きいので，今後は機械学習などを用いた抽出手法を考えていきたい．

参考文献

[Mann 03] G. Mann and D. Yarowsky. Unsupervised Personal Name Disambiguation. In *Proceedings of Conference on Natural Language Learning (CoNLL-03)*, pages 33-40, 2003.

[上田 08] 上田洋, 村上晴美, 辰巳昭治. Web 上の同姓同名人物識別のための職業関連情報の抽出. 第 22 回人工知能学会全国大会, 2D2-3, 2007.

[Balog 09] K. Balog, J. He, K. Hofmann, V. Jijkoun, C. Monz, M. Tsagkias, W. Weerkamp, and M. Rijke. The University of Amsterdam at WePS2. In *Proceedings of the 2nd Web People Search Evaluation Workshop (WePS-09) at 18th International World Wide Web Conference*, 2009.

[木村 06] 木村塁，戸田浩之，田中克己. 検索結果スニペットのクラスタリングによる同姓同名人物の特定, DEWS-06, 2Ci11，2006.

[Sekine 09] S. Sekine and J. Artiles. Weps 2 evaluation campaign: overview of the web people search attribute extraction task. In *Proceedings of the 2nd Web People Search Evaluation Workshop (WePS-09) at 18th International World Wide Web Conference*, 2009.

[Chen 09] Y. Chen, S. Lee and C. Huang. PolyUHK: A Robust Information Extraction System for Web Personal Names. In *Proceedings of the 2nd Web People Search Evaluation Workshop (WePS-09) at 18th International World Wide Web Conference*, 2009.

表 4: WePS のデータセットにおける評価

Attribute	Match	OVG	Miss	Precision	Recall	F-measure
Affiliation	709	14151	2378	0.048	0.230	0.079
Award	0	371	264	0.000	0.000	0.000
Birthplace	145	374	154	0.279	0.485	0.355
Date of birth	118	239	251	0.331	0.320	0.325
Degree	18	445	317	0.039	0.054	0.045
E-mail	132	66	77	0.667	0.632	0.649
Fax	0	1	65	0.000	0.000	0.000
Major	8	113	165	0.066	0.046	0.054
Mentor	1	31	342	0.031	0.003	0.005
Nationality	86	660	164	0.115	0.344	0.173
Occupation	260	7009	3032	0.036	0.079	0.049
Other name	37	2647	752	0.014	0.047	0.021
Phone	128	933	91	0.121	0.584	0.200
Relatives	7	989	906	0.007	0.008	0.007
School	74	312	418	0.192	0.150	0.169
Web site	20	723	134	0.027	0.130	0.045
Total	1743	29064	9510	0.057	0.155	0.083