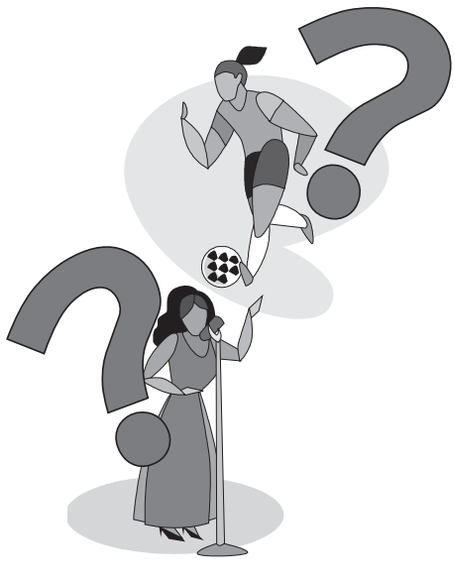
Web検索における人名の曖昧性解消技術の動向

― 同姓同名のクラスタリング ―



関根 聡（ニューヨーク大学）

なぜ人名の曖昧性解消が重要か

「山口百恵」と「佐々木元」

たとえば，Google で，「山口百恵」を検索してみる．当然ながら，元歌手の山口百恵さんという人物の検索結果が大量に表示される（**図-1**）．しかし，Wikipedia には載っているフットサル選手の山口百恵さんの検索結果は上位 10 位内には表示されない．今度は，情報処理学会の会長である「佐々木元」という人名で検索してみる．上位 10 件は，NEC の会長，東北大学の教授，プロの BMX ライダー，翻訳者の 4 名の人物で占められている（**図-2**）[☆1]．もしも，探している人物が，フットサル選手の山口百恵さんだった場合は 12 位まで，また，1960 年代の映画監督の佐々木元さんだった場合には 82 位までリストを見続けなければいけない．さらに，特に有名ではない山口百恵さんや佐々木元さんを捜している場合には 100 位以内にも入っていないかもしれない．このように，Web 上の人物検索における問題の 1 つに，字面上同一の人名が，現実世界にいる複数の人物を指すことがあるという同姓同名の問題が挙げられる．この問題は，根本的には全世界の情報が，インターネットという一枚岩の上に無差別に並べられていることによって生じた新しい種類の問題と捉えることができる．

この同姓同名の問題は言語に依らず存在する．“Michael Jordan” という人名で検索した場合，伝説のバスケットボール選手以外にも，機械学習や人工知能の専門家で UC Berkeley の教授の Michael Jordan 氏は 8 位に出現するし，Tufts University 薬学科の Michael Jordan 氏は 85 位に初めて出現する．また，Wikipedia を見ると，英語で「よくある人名」として引用される “John Smith” の一覧ページには 85 名もの John Smith が並んでいる．

この同姓同名の検索結果が，ページごとに単に順番に並んでいるという状況は非常に不便である．各場面でユーザが探す対象と検索結果の順番に常に相関関係があるわけではない．どこまで見たら，自分の探している人物が見つかるのか分からないし，そもそもその人物がリスト中にあるのかどうかすら分からない．しかし，もしも検索している対象が人物である場合には，同姓同名の検索結果が人物ごとにクラスタリングされて表示されたら非常に便利である．ユーザは検索した人名にマッチする人物の一覧をざっと見て，自分のお目当ての人物を見つけられるであろう．この問題を解くためには，同姓同名を人物ごとに振り分ける技術が必要になってくる．この技術が「人名の曖昧性解消」の技術である．

「女優」と「練習試合」

たとえば，「山口百恵」の検索結果の要約（スニペット）に，「女優」「元歌手」「コンサート」「歌姫」または「三浦友和」「秋桜」などの単語が見つかれば，検索ユーザはその人物は歌手の山口百恵さんだとすぐに気付く．逆にスニペットに「フットサル」「練習試合」という単語が現れたら，「おや？」と感じるに違いない．これは，人物に関する特徴が頭の中にあって，その特徴と関連する単語を見つければ，その人物を特定したと安心できるが，逆に，その特徴とは相容れない単語を見つけたときには違和感

☆1 NEC の会長は，もちろん情報処理学会の会長でもあるのだが，情報処理学会の会長としての言明は残念ながら 135 位までない（2008 年 2 月時点）．また，上位 10 件のうちには形態素解析誤りで，佐々木元兵，佐々木元昭の 2 名がいる．

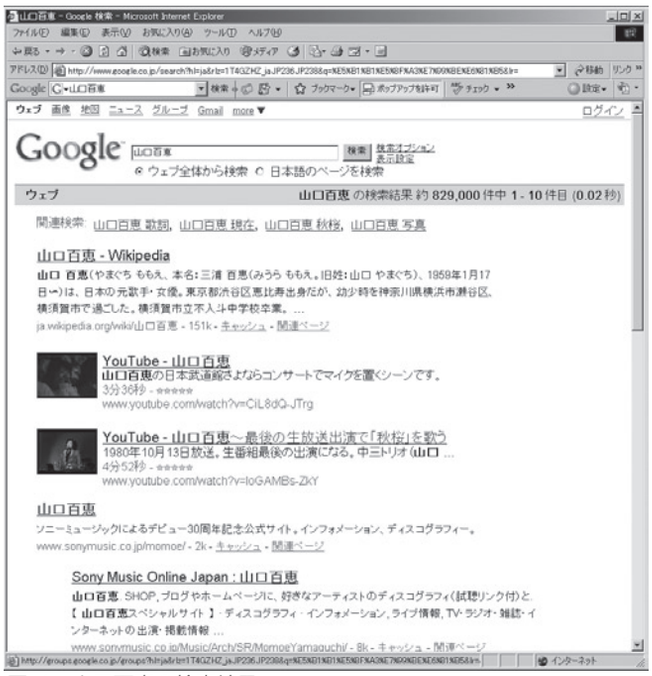


図 -1　山口百恵の検索結果

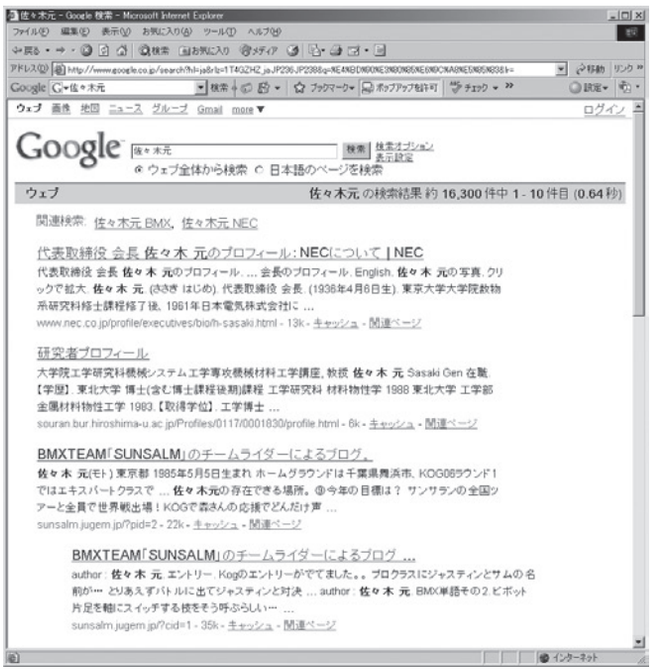


図 -2　佐々木元の検索結果

を覚えてしまう人間の反射条件的な認識手法と言えるものかもしれない．もし，同姓同名の分類をシステムで実現したいのであれば，上記のようなキーワードをテキストの中で認識し，各テキストの特徴を得るという方法が考えられる．つまり，この問題を解くためのシステムを作るためには，各テキストの内容を読解し，きちんと理解をする必要はなさそうである．たとえば，ある人物に頻出するキーワード群の認識や，キーワードの類似度を測定するといったお手軽な手法を使って，人名の曖昧性解消はある程度，実現できそうである．つまり，人物の曖昧性解消を実現するために必要なのは，キーワードを使ったドキュメントの類似度の尺度を設計することと，どのくらい似ていれば同一人物であると判断するかの閾値くらいのようである．

「LSI 開発」と「材料工学」

実際，上記のようなお手軽な方法である程度この問題は解決できる．しかしながら，「ある程度」以上はそう簡単ではない．たとえば，NEC 会長の佐々木元氏は若い頃は，超 LSI 開発の研究者であり，Robert N. Noyce Medal を受賞されるなど研究分野でも活躍されている．それに対して，東北大学の教授の佐々木元氏は機械材料工学講座の教授であり，「金属基複合材料の製造」などの研究をされている．SAMPE Japan, Excellent Paper 等を始めとして輝かしい受賞歴もお持ちである．確かに，現在の NEC の会長のページにあるキーワードは，「就任」「経済」「施策」「環境問題」等の経済的な用語が多く，東北大学の佐々木元氏とは似ていないかもしれない．しかし，単純にキーワードだけでクラスタリングをすると，NEC の佐々木元氏の過去の研究業績に関連するページは東北大学の佐々木元氏の方にクラスタリングされてしまうかもしれない．

人間はなぜこの 2 人を区別できるのであろうか？ 2 人の名前の読みが「はじめ」と「げん」で異なっていることはあまり関係なさそうである．おそらく，NEC の会長をしながら東北大学の教授を兼務することはあり得ないだろうという常識や，それぞれの方の受賞歴や職歴などの経歴を理解するなどの詳細な特徴量を利用して曖昧性を解消しているのだろう．また，推定年齢の違いや NEC という大企業の会長という特異性にもヒントがあるかもしれない．これはその対象や分野の知識を我々がすでに持っているということと深い関連がありそうである．逆に考えると，仮に BMX[☆2] の世界に佐々木元氏が 2 人いたとしても，BMX のことをほとんど知らない筆者には，実は区別がついていないかもしれない．

このように，人名の曖昧性解消の問題は，一見簡単そうであるが，言語理解や世界知識の認識にも通じる深い問題が含まれている．後述するが，これまでも自然言語処理の研究者による研究を始め多くの研究成果が発表されている．人名の照応の研究や，人名だけに限らない言語の曖昧性解消の研究，人物の簡単な属性を認識する研究などと深い関連がある．研究対象として取り組んでい

☆2　土のコースで競争したり，空中技などを披露することのできる自転車．

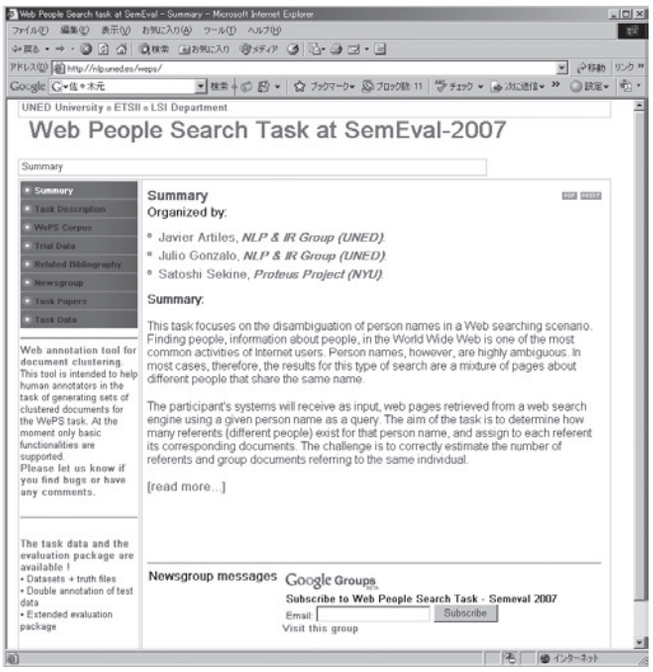


図 -3　WePS の Web サイト（http://nlp.uned.es/weps）

くのに十分面白い課題であり，実際，次に挙げるようなプロジェクトが立ち上がり，盛況に運営されており，技術的にもそして応用的にもホットなトピックになっている．

評価プロジェクト：WePS

筆者がオーガナイザの 1 人となった人名の曖昧性解消の評価プロジェクト「Web People Search（WePS）」が 2006 年末から 2007 年にかけて行われた（**図 -3**）．ここでは，そのプロジェクトについて紹介する．より詳細を知りたい方は文献 1），2）を参照されたい．

タスク

この評価プロジェクトでは，参加者はある人名で検索された Web 検索結果の 100 ドキュメントまでのセットを渡され，その中に含まれるページを人物ごとにクラスタリングするというタスクにより，人名の曖昧性解消の技術を競った．人物へのクラスタリングは，ドキュメント単位で行う．つまり，人物 A に言及されているドキュメントの集合，人物 B に言及されているドキュメントの集合といったものを作ることが要求されている．細かいが，もし，あるドキュメントにおいて，人物 A と人物 B の両方への言及があった場合には両方のクラスタに属するようなかたちにするのが正解となる．

プロジェクトでは，トライアルデータ，トレーニングデータ，テストデータが作成され配布された．人名の曖昧性は，その人名がどのような人名かによって様態が異なることが予想される．つまり，同姓同名の中に非常に有名な人物がいる場合と，特に有名な人物がいない場合では，曖昧さの分布が違い，それらを解消する技術も異なってくる可能性がある．この点に留意し，データを作る際の人名は 3 つの違った種類のリソースからサンプリングすることとした．1 つは国勢調査に挙がった人名など無作為にとれる人名，1 つは Wikipedia の項目に挙がった人名，そして最後の 1 つは学会発表者に挙がっている人名である．これらのリソースからランダムにサンプリングして人名を集め，その人名で Web 検索をすることによりトレーニングデータ，テストデータを作成した．Web ページは，Yahoo! の API を利用してそれぞれの人名に対して 100 ページを上限として収集した．ただし，トレーニングデータにおける無作為の人名データ（表 -1 の WEB03）はすでに研究発表されているもの [4] を流用している．正解データは人手で作成しており，トレーニングデータについては 1 カ所，テストデータについては 2 カ所で正解を作成し，最終的に 1 つのデータにまとめた．データの概要を**表 -1** に挙げる．表には各リソースごとに，1 人名あたりの人物数の平均と，1 人名あたりの検索されたドキュメント数を示している．

トレーニングデータ		
リソース	人物数	ドキュメント数
Wikipedia	23.14	99.00
ECDL06	15.30	99.20
WEB03	5.90	47.20
平均	10.76	71.20
テストデータ		
Wikipedia	56.50	99.30
ACL06	31.00	98.40
国勢調査	50.30	99.10
平均	45.93	98.93

表 -1　データの概要

評価と問題点

評価プロジェクトには世界中から 16 チームが参加した（日本からは 1 チーム）．結果は，Purity（各クラスタにおいて最も多い正解のラベルの割合の加重平均）と Inverse purity（各正解のクラスタにおいて，最も多く含まれたクラスタに属する正解クラスタの要素数の割合の加重平均）に基づいた F 値で評価された．評価結果の F 値が最高であったチームは 0.75 の値であったが，人間の評価はそれぞれ 0.98，0.91 であり，まだ人間のパフォーマンスには及ばない結果であることが分かった．ベ

ースラインとしては，1 ドキュメントにつき 1 クラスタ（Scattered），または全ドキュメントが 1 つのクラスタ（Joined）という単純なベースラインが作れるが，F 値の評価では，全チームが Joined の精度を超え，半数の 7 チームが Scattered の精度を超えた．また，1 ドキュメントが複数のクラスタに含まれることを許しているため，Joined と Scatter の両方を単純に足し合わせたベースラインシステム（Combined）を作成することが可能であった．このベースラインを越えたのは 1 チームしかなかった．これは 1 ドキュメントが複数のクラスタに含まれることは滅多にないのにその特徴を利用した姑息なベースラインであり，評価の形態の問題であると考えている．この問題の解決方法として，新たな評価尺度を検討しているが，現在はまだ確定したものにはなっていない．

今回の評価には 1 つ，データの作成とシステムの構成の双方に関連した問題があった．表 -1 を見ると，トレーニングデータとテストデータの平均人物数の間に大きな乖離があることが分かる．次のセクションで述べるように，多くのシステムは，クラスタの生成のための閾値を，トレーニングデータを利用して決定している．たとえば，クラスタ数がトレーニングデータの平均の 11 になったらクラスタリングを終了するという単純なアイディアを採用しているシステムにおいては，この乖離は非常に大きな問題になってしまう．この乖離の原因は，それぞれのデータをダウンロードした時期が異なっており，それぞれの時点の Yahoo! のシステムの違いに起因している可能性がある．この問題は評価結果の信頼性にもかかわる問題であり，次回の評価では注意してデータ収集をする必要があるという教訓になった．

参加システムの技術

参加システムの技術の概要を紹介する．参加したほとんどのシステムは細かい点で違いはあるが，全体的な流れは非常に似たものであった．まず，各システムは HTML のドキュメントからテキスト部分を抽出する．これは既存のツールを使ったり，自作の簡単なフィルタを使ったりしている．そして，抽出したテキストに対して自然言語処理や Web 解析ツールを使った前処理を行う．ここには，単語の品詞を同定する品詞タガー，名詞句や動詞句を同定するチャンカー，人名，組織名，地名などを同定する固有表現タガー，代名詞などの照応関係を同定する照応解析，テキストに書かれているイベントなどを抽出する情報抽出，電子メールアドレスや URL の抽出などの自然言語解析技術を利用したり，リンク解析や，複数のテストデータをリンクしている Web ページの存在などを抽出しているシステムもあった．次に，これらの解析結果を素性として，ドキュメント間の類似度を計算する．ここでは，前述の素性をベクトルの要素として列挙し，頻度や単語の相対的重要度を測る TF-IDF などで重み付けしたベクトル間のコサイン距離で類似性を求める方法が主流であったが，他にもいくつかの方法が存在した．そして，その類似度を元にクラスタリングを行う．クラスタリングは主に凝集型（Agglomerative）クラスタリングが主流であった．そして，クラスタリングをどこで止めるかという閾値の設定には，トレーニングデータを何らかの形で使う方法が主流であった．それは単純にクラスタ数の平均を利用する方法や，ベクトル類似度の値を使う方法などが利用されていた．ただし，人名によってクラスタの数が大きく変わるため，より客観的な形を求める方法もいくつかあった．たとえば，2 つのドキュメントの中にある固有表現の重複の数を閾値にするなどの方法を試している参加者もあった[☆3]．しかし，閾値の設定にはトレーニングデータはあまり当てにならない点，人名によってクラスタ数などが大きく異なる点を考慮すると，閾値の決定方法は技術的に大きな課題である．決定的な解法は見られていないが，今後の研究に期待したい．また，多くの参加者の議論によると，固有表現の同定が 1 つの重要な要素技術であることが分かる．たとえば，参加チームの 1 つは，多種多様な素性を試してみたが，最終的に，ドキュメントの中に現れる該当人名以外の固有表現のリストのみを使った場合が最も精度が高かったという報告をしている．

このように，第 1 回目の評価プロジェクトは世界中の多くの研究者の参加を得，さまざまな技術が使われたこと，多くの参加者が大枠では同じような技術を使い，標準的な方法が確認されたことなど，さまざまな成果があったと考えている．

人物の属性

評価プロジェクトが終了し，詳細にシステムやデータの分析をしてみると，実際にどのような部分に大きな問題があったのか，技術の進歩はどのような部分の解決によって得られるのかが分かる．今回の評価型プロジェクトの参加者の分析結果，および正解作成者の報告を見ていると，曖昧性の解消のために，各人物の属性を認識することが非常に重要だということが分かった．たとえば，

☆3 実際はもう少し複雑である．詳細は参加者の論文を参照していただきたい．

職業名は一般的に同姓同名の曖昧性解消を行うために非常に重要である．また，何代かにわたり同姓同名の国王がいるという特殊な場合では，后の名前や就任時期が曖昧性解消の決定的な手がかりになる．したがって，人物の属性というものを定義し，それをきちんと抽出することは，人名の曖昧性解消の精度を人間の精度並みに上げるために非常に重要であると考えられる．また，この属性抽出の技術は人名の曖昧性解消だけではなく，いわゆる情報抽出の技術に深い関連があり，この技術を極めることにより幅広い応用が考えられる．我々は 2008 年の秋に向けて 2 回目の評価プロジェクトを行うことを計画しているが，その中で，人物の属性抽出のサブタスクを行うことを考えている．

このサブタスクを計画する上で，一番大きな問題となるのは「人物の属性」とはどのようなものであるか，ということである．評価プロジェクトとして利用することが目的であるため，その属性は，評価として意味をなす程度に幅広くきちんと定義できる範囲のものであると同時に，将来のシステム展開に有用な技術を構築できる属性としてきちんと設定しなければいけない．そのために我々はまず，データを分析し，属性と考えられるものにはどのようなものがあるかを調査した．今回の WePS で使用されたドキュメントのうちの 156 ドキュメントを使用し，作業者に人物の属性と考えられるものを可能な限り多く抽出してもらい，その後で，抽出された属性を分析した．属性は，「人物の属性は属性値です」と言えるもの（下線の部分には「関根聡の国籍は日本です」というように具体例が入る），そして，属性名，属性値が名詞句またはそれ相当で表現できるものに限った．対象文章中では「関根は准教授です」というように，（「職業名」という）属性名が文章中に明示的に書かれていなくてもよい．作業の結果，156 ドキュメントから 123 種類の属性が抽出できた．最も頻度の高い 6 つの属性は，職業名（116），作品（70），所属組織名（70），フルネーム（55），同僚・師匠（41），出身校（41）であった（括弧内の数字はその属性があったドキュメント数．ドキュメント内に複数回その属性が現れても 1 回と数えている）．これらの 123 の属性のうちには評価の対象とするには適さないと思われるような属性がいくつかあった．それは，バスケットボール選手の生涯得点のような分野依存の属性や，配偶者の生年月日のような属性の属性といったものである．また，父，母，兄弟のような属性は親類という属性でまとめることができる．このように，抽出された属性を分析し，それを元に，評価に適し，一般的にも受け入れられると考えられる属性を 16 種類に特定した．その

	属性名	頻度	属性値例
1	生年月日	21	January 1, 1900
2	出生地	24	Tokyo, Japan
3	別名	56	Mister S
4	職業名	141	Associate Professor
5	所属組織名	83	New York University
6	作品	70	Apple Pie Parser
7	賞	26	Best Picture Award
8	学歴	79	PhD.
9	同僚・師匠	48	Ralph Grishman
10	場所	63	New York, USA
11	国籍	5	Japan
12	親類	45	Taro Sekine
13	電話番号	27	+1-212-998-0000
14	FAX 番号	11	+1-212-995-0000
15	電子メール	25	xxx@yy.nyu.edu
16	Web サイト	13	nlp.cs.nyu.edu/sekine

表 -2　16 種類の属性

16 種類の属性を**表 -2** に示す．表 -2 で表されている頻度は，その属性に含まれる調査時の頻度を合計した数字を参考のために載せている．

本サブタスクは，各 Web ページから属性をどのくらい抽出できるかで評価するものであり，各人物についてまとめるという評価までは行わない考えである．人名の曖昧性解消の問題と絡めると，それぞれの技術の評価が純粋には行われなくなってしまうためである．評価は，適合率，再現率，そして F 値で行う予定である．

このタスクに対しては，固有表現抽出，テキストマイニング，パターンマッチング，知識獲得，関係抽出，情報抽出などさまざまな技術が展開され利用されることが期待できる．また，この評価プロジェクトを通じて，人名の曖昧性解消のためだけではなく，さまざまな応用に向けた技術やリソースが開発されることも期待できる．その中には，職業名などの新しいタイプの固有表現の辞書や，アノテーションのためのツールやテキストマイニングのツールなども含まれるであろう．また，このプロジェクトの結果として，技術の進展だけではなく，近未来における新しいアプリケーションの実現も期待している．

関連技術

人名，地名，組織名などの名前に関する曖昧性解消の問題を明示的に定義し解決しようとした研究は Bagga ら [3] による複数文書における照応解析に端を発していると一般的に考えられている．照応解析という技術は，同じ文書内に複数表現されている同一指示対象の同定をい

うが，これを複数文書において試みた研究である．彼らの手法は，まず，人名を表す表現における「要約」をそれぞれの文書において作成し，その要約の特徴を元に類似性を計算し，クラスタリングをすることによって同一指示対象の同定を行う．この方法は WePS でもほぼすべてのシステムのベースとなっている方法である．また，Mann[4] はこのタスクを対象にジョン・ホプキンス大学で博士号を取った．彼の方法では，特徴量は単にコンテキストにある単語ではなく，誕生年，誕生日，職業名，出生地，死亡年などの属性を認識する方法をとっている．これは次の WePS でタスクとしているのと同じ方向である．その他，関連研究の幅広い論文リストは WePS の Web ページから入手できる．

日本においても，本タスクの重要性は以前から認識されており，機械翻訳の研究開発が盛んに行われていた 1990 年前後にも，照応解析の研究が行われてきた．人名などの固有表現が自然言語処理の研究対象として広く認知された 1990 年代後半以降，その曖昧性の解消も重要課題の 1 つとなり，2005 年には日本語での同姓同名の曖昧性解消の研究[5]も行われている．ここでは，各ドキュメントにある人物に特徴を持たせ，人名の共起関係を特徴量としたドキュメントのクラスタリングとして問題を解いている．また，最近は WePS に参加した東京工業大学以外の大学の研究室や企業でも，人名の曖昧性解消の研究が行われており関連論文の数は 20 をくだらない．文献 6)にその一部が紹介されている．

今後の展開

第 1 には，人間のパフォーマンスとシステムの成績の大きな隔たりを解消していかなければいけない．現在の単語の類似性を基本としたナイーブな方法だけでは，人間のパフォーマンスに到達することはほとんど不可能だと思われる．データの解析をしてみると，そこには，人間が当たり前に持っている常識が必要となる場合が少なくない．たとえば，50 歳を超した女性がプロのフットサルの 1 級選手であることはほぼ考えられないとか，BMX の優勝者と NEC の社長が同一人物になることは滅多にないとか，逆に，NEC の社長が過去に 1 級の研究者であることは十分に考えられるといった常識である．このような常識をどのように獲得し形式化し応用するかといった問題が非常に面白い課題として存在する．

また，指示対象の同定の問題には，実は曖昧性解消の問題とは別に，同じ指示対象が別の名前で呼ばれたときにそれを同定する問題が存在する．たとえば，「チャーリー・チャップリン」が「喜劇王」と呼ばれるようなあだ名や，名字だけで呼ばれる場合，外国人の名前のカタカナ表記が統一されないという問題などさまざまな問題が含まれている．もちろん，文章理解のためには，人物が代名詞（彼，彼女）や普通名詞（前社長，容疑者）で呼ばれる場合などの認識も非常に重要である．

本技術の適用対象は Web 検索だけにあるとは限らない．本技術の適用対象がソーシャル・ネットワーキング・サービス（SNS）や科学研究分野であると，ある程度構造化された特徴量が使える場合がある．それは，SNS の友達のリンクや論文の共著，引用などのリンク情報である．このような情報を有効に使える場においては，人名の曖昧性解消の技術は違った種類のものになるかもしれない．

また，本技術はすでに人物検索や人物情報のポータルサイトを目指した検索サイトなどで実用化され始めている．米国では，Spock，Zoominfo，Wink 等が有名であるし，Yahoo! や Google の検索エンジンにも人名の曖昧性解消技術が使われている様子が伺える．

本稿では，Web 上での人名の曖昧性解消を対象にした評価型プロジェクトの WePS を中心に，人名の曖昧性解消の問題について解説した．まだまだやるべき問題が山積しており，今後の研究の動向が注目される．

参考文献

1) WePS Web ページ : http://nlp.uned.es/weps
2) Artiles, J., Gonzalo, J. and Sekine, S. : The SemEval-2007 WePS Evaluation : Establishing a Benchmark for the Web People Search Task, In Proceedings of SemEval 2007, Association for Computational Linguistics (2007).
3) Bagga, A. and Baldwin, B. : Entity-based Cross-document Coreferencing using the Vector Space Model, In Proc. 36th Annual Meeting of the Association for Computational Linguistics (ACL) and 17th Conf. on Computational Linguistics (COLING), pp.79-85 (1998).
4) Mann, G. S. : Multi-Document Statistical Fact Extraction and Fusion, Ph.D thesis, Johns Hopkins University (2006).
5) 佐藤進也，風間一洋，福田健介，村上健一郎：実世界指向 Web マイニングによる同姓同名人物の分離，情報処理学会論文誌データベース，Vol.46, SIG8, pp.26-36 (June 2005).
6) 杉山一成，奥村　学：Web 検索結果における人名の曖昧性解消への半教師有りクラスタリングの適用，情報処理学会研究報告，Vol.2007, No.94, 2007-NL-181 (3), pp.15-20.

（平成 20 年 3 月 31 日受付）

関根　聡（正会員） sekine@cs.nyu.edu

1987 年東京工業大学応用物理学科卒業．松下電器東京研究所入社．1992 年英国 UMIST 大学計算言語学科 MSc．1998 年ニューヨーク大学 Ph.D．同年より研究助教授．2007 年より研究准教授．研究対象は自然言語処理．特に情報抽出，知識獲得に興味を持つ．自然言語処理技術のコンサルタント・技術移転を行う（株）ランゲージ・クラフト研究所を設立し，運営している．