人名の曖昧性解消 評価型プロジェクト WePS

関根聡
ニューヨーク大学

Javier Artiles
UNED, Spain

Julio Gonzalo
UNED, Spain

1. はじめに

インターネット上で頻繁に行なわれる Web 検索の目的のひとつに、人物に関する検索がある。しかしながら、世の中には言語にかかわらず同姓同名の人物が数多くいるため、人名は非常に曖昧性が高い。したがって、人名で Web 検索をすると、数多くの同姓同名の人名が提示されることがよくある。その結果、ユーザーは自分が求めている人物の情報を得るために、新たなキーワードを加えて再検索するか、長いリストの中からその人物の情報を探さなければいけないことになる。

この問題に対する理想的な解として、ユーザーは、ユーザーの求める人名のみを入力し、検索を行ない、その結果は人物毎にクラスタリングされ、ユーザーの求める人物がすぐに分かるような形式で表示されることが考えられる。この解の実現を最終目的として、我々は Web People Search(WePS)のプロジェクトを開催し、人名の曖昧性解消の技術向上を目指している。この評価プロジェクトでは、参加者はある人名の Web 検索結果のセットを受け取り、その中に含まれるページを人物毎にクラスタリングすることが期待されている。このタスクは、例えば、SenseEval で行なわれてきた単語の語義の特定（WSD）を行なう語義の曖昧性のタスクと深い関連性がある。この二つのタスクとも自然言語が内包する曖昧性解消という問題を解決しようとするものであるが、いくつかの違いが存在する。WSD では一般的に、オープンクラスの内容語（普通名詞、形容詞、動詞、副詞）を対象にしている。そのため、語義の境界はあまり明確ではない。それに対して、WePS では人物の境界はほとんどの場合に明確である。また、WSD では通常、辞書の語義を使うため、語義の数は予め知りうるある程度の数に限られているが、WePS では予め人物の数はわからない。したがって、このタスクは実際には「語義の識別(Word Sense Discrimination)」のタスクである。また、曖昧性の数も、少ない物から非常に多い物まで様々である。曖昧性の調査に米国の世論調査のデータがあるが、1 億人が 9 万の名前を使っているという結果を示している。また、WePS のタスクは複数のドキュメントに含まれる同一の言及を認定するドキュメント横断型の照応解析とも関連がある。

人名の曖昧性解消に関する過去の研究には、Bag of words による Vector Space モデルを利用した方法(Bagga and Baldwin 98)や、より深く豊かな素性を使った(Mann and Yarowsky 03)の研究、2 つの名前が同一人物を指しているかどうかを Maximum Entropy Model を使って取り組んだ(Fleischman and Hovy 04)などがある。さらなる研究リストは、本プロジェクトのホームページ(WePS HP)を参照されたい。

本論文では、1 回目の WePS のタスク定義、データ、評価結果、参加システムの技術を概観し、2 回目として計画している新しいサブタスクの意義、調査等についても述べる。

2. タスク定義とデータ

参加チームはある人名の Web 検索結果として 100 ドキュメントを受け取り、そのドキュメント群に含まれる対象人名の曖昧性解消を行なう。人物へのクラスタリングは、ドキュメント単位で行なう。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うかたちである。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。

我々は、トライアルデータ、トレーニングデータ、テストデータを作成し配布した。トライアルデータは（Artiles et al. 06）で作成されたものであり、本論文では説明は割愛する。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、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。一つは世論調査に挙がった人名など無作為に取れる人名、一つはWikipediaの項目に挙がった人名、そして最後の一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た人名を使って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は(Gideon Mann 06)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カ所で作成し最終的に一つのデータにまとめた。データの概要を表1に挙げる。

表1　データの概要

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
世論調査	50.30	99.10
平均	45.93	98.93

3. 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入る要素数の割合の加重平均）に基づいたF値で評価された。1ドキュメントにつき1クラスタという非常に簡単なベースラインを越えたのは7チームであった。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わかった。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、単純だが姑息なベースラインシステム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。

今回の評価にはひとつ、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わかる。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。例えば、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いることから生じるYahoo!のシステムの違いに起因している可能性があるが、実際の所はYahoo!の開発者に聞かない限り不明である。ただ、この問題は評価結果の信頼性にも関わる問題であり、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。

4. 参加システムの技術

参加システムの大部分は以下のような仕組みでシステムを構成していた。まず、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、自作の簡単なファイルターを使ったりして行なっている。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なう。ここには、POSタギング、チャンカー、固有表現タガー、照応解析、事実抽出、電子メイルアドレスやURLの抽出などの自然言語解析技術を利用したり、また、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に置き換え、頻度やTF-IDFなどの重み付きの計算結果を値としたベクトル間のコサイン距離で求める方法が主流であったが、他の方法もいくつか存在した。そして、その類似度を元にクラスタリングを行なう。クラスタリングはボトムアップの凝集型クラスタリングが主流であるが、他の方法も存在した。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。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。例えば、二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（実際はもうちょっと複雑である。詳細は参加者の論文を参照していただきたい）。しかし、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、この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を見ると、固有表現が一つの重要な技術であることがわかる。例えば、参加チームの一つは、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。

5. 人物の属性抽出サブタスク

今回の評価型プロジェクトの参加者のコメント、および正解作成者のコメントを見ていると、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わかった。例えば、何代かにわたり同姓同名の国王がいた場合などには、后の名前や就任時期が曖昧性解消の決定的手がかりになる。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技術は、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は非常に重要な技術になってくると考えられる。また、こ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な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な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。その属性は、評価プロジェクト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開発につながるような属性が存在することを確認し、それを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内の 156 ドキュメントを使用し、タグ付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するという方法をとった。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。「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。もっとも頻度の高い 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）。これらの 123 の属性の内には評価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られる。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。表２に示す。表２で表されている頻度は、先の調査における頻度のうち、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。

表２　１６種類の属性

	属性名	頻度	属性値例
1	生年月日	21	March 5, 1965
2	誕生地	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 Univ.
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	Shigeru Sekine
13	電話番号	27	+1-212-998-3175
14	FAX 番号	11	+1-212-995-4123
15	電子メイル	25	sekine@cs.nyu.edu
16	Web サイト	13	nlp.cs.nyu.edu/sekine

本サブタスクは、各 Web ページから可能な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なわない考えである。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な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 値で行なう予定である。

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、情報抽出など様々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、人名曖昧性解消のためだけではなく、様々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。

6. まとめ

本論文では、2006 年末から 2007 年始めにかけて行なわれた人名の曖昧性解消の評価型プロジェクトである WePS を紹介した。世界各地から１６チームの参加があり、その多様な技術の適応と、その問題点に関する活発な議論、今後の展開に対する積極的で協力的な参加者で盛り上がるなど成功裏に終了したと考えている。その内容をふまえ、2008 年には、サブタスクに人物の属性抽出を加えて、２回目の評価プロジェクトを開催する計画である。興味を持っていただける方々には是非参加をお願いしたいと考えている。

本論文のより詳細な内容は、本プロジェクトのホームページ(WePS HP) または、(Artiles et al. 07)を参照されたい。本プロジェクトで作成されたデータ、ツールもホームページより入手できる。

参考文献

WePS homepage: http://nlp.uned.es/weps/

Javier Artiles, Julio Gonzalo, Satoshi Sekine 2007. The SemEval-2007 WePS Evaluation: Establishing a benchmark for the Web People Search Task.. In Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007).

Javier Artiles, Julio Gonzalo, and Felisa Verdejo. 2005. A Testbed for People Searching Strategies in the WWW In Proceedings of the 28th annual InternationalACM SIGIR conference on Research and Development in Information Retrieval (SIGIR'05), pages 569-570.

Amit Bagga and Breck Baldwin. 1998. Entity-Based Cross-Document Coreferencing Using the Vector Space Model In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics (COLING-ACL'98), pages 79-85.

Michael B. Fleischman and Eduard Hovy 2004. Multidocument person name resolution. In Proceedings of ACL-42, Reference Resolution Workshop.

Gideon S. Mann. 2006. Multi-Document Statistical Fact Extraction and Fusion Ph.D. Thesis.

Gideon S. Mann and David Yarowsky 2003. Unsupervised Personal Name Disambiguation In Proceedings of the seventh conference on Natural language learning at HLT-NAACL, pages 33-40.