2D2-4

語の共起ネットワークを用いたエンティティの別名抽出

Alias Extraction Using Word Co-occurrence Network

本間大輝[*1] Taiki Honma　Danushka Bollegala[*1]　松尾豊[*2] Yutaka Matsuo　石塚満[*1] 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

An entity may have multiple name aliases on the Web. Identifying aliases of a name is important to extract precise information of entities. In this paper, We try to extract an alias ranking of a given name using word co-occurrence network. In the network, nodes are words which frequently co-occur with a given name, and an edge represents existence of a co-occurrence relation. We rank each node based on structural similarity to a given name. Experimental results on a dataset of Japanese celebrities show that the proposed method outperforms all baselines.

1. はじめに

Web の発展に伴い，Web の情報源としての価値が高まっている．また，検索エンジンを用いれば，だれでも Web に効率的にアクセスできるようになった．そのため，検索エンジンを利用し，エンティティに関する知識を Web から自動的に抽出する研究に注目が集まっている．たとえば，商品の評判 [3] を抽出する研究，事実を抽出する研究 [4] などが行われている．

しかし，Web からエンティティに関する知識を抽出する研究において，同姓同名の問題と別名の問題が性能向上の妨げとなっている．なお，本研究では，以下の 4 条件を満たすものをあるエンティティE の別名と定義する．

1. E を参照するために用いられる表現である
2. E への参照として用いる人が 2 人以上いる
3. E の上位概念ではない
4. 条件 1，2，3 を満たすもののうち最も短いもの

たとえば,「松井秀喜」に対し,「ゴジラ」は別名だが,「野球選手」,「4 番」,「ヤンキースのゴジラ」,「ゴジラ松井」などは別名ではないとする．

同姓同名の問題は，1 つの語に複数のエンティティが対応する問題である．この問題については，すでに多くの研究がなされている [1]．しかし，もう一方の，1 つのエンティティに複数の語が対応するという別名の問題はほとんど研究が行われていない．

そこで，本研究では，あるエンティティの名前が与えられたときに，その別名を検索エンジンを利用して Web から自動的に抽出する手法を提案する．提案手法は，ある語の別名らしさを，語の共起のネットワークの構造に着目して測るため，あらゆる種類のエンティティに対し適用可能である．

連絡先: 本間大輝，東京大学大学院 情報理工学系研究科，〒 113-8656 東京都文京区本郷 7-3-1，Tel 03-5841-6774, honma@mi.ci.i.u-tokyo.ac.jp

2. 関連研究

Web で別名を抽出した研究としては外間・北川の研究がある [2]．しかし，外間・北川の手法は「こと」というパターンに頼った別名抽出であるために，人物にしか適用可能でない．提案手法は，人物以外にも適用可能である．Web ではないが，ただのテキストを集めたコーパス内で，別名抽出を行った研究としては，Holzer らの研究 [5] と，Hsiung らの研究がある [6]．これらの研究はどちらも，あらゆる種類の別名に適用可能となっているが，別名の候補がいくつか与えられている，という状況でしか利用できない．提案手法は，別名候補が与えられていなくても適用可能である．

3. 方法

直感的には，テキストの上で，ある名前が果たしている役割と，ほぼ同じ役割を果たしている語が別名，と考えられる．この直感を実現する 1 つの方法は，語の果たしている役割を構文解析などを通して，直接的にテキストから判定するというものである．しかし，このような方法だと，あらゆるエンティティに適用可能にするのは難しい．そこで，本研究では，テキストをまず，語をノードとするネットワークに変換し，次に，そのネットワーク上で，語の役割を捉える．以下，この 2 段階の手順を説明する．

3.1 ネットワークの生成

この手順で生成したいネットワークはたとえば図 1 のようなものである．ネットワークのノードは語であり，エッジは共起関係があるかどうかを表す．ノードは，与えられた名前で検索し，得られたスニペット上位 100 件に含まれる名詞のうち，名前との共起の強い上位 30 語とする．語 w_1 と語 w_2 の共起の強さは次の Jaccard 係数で測る．

$$Jaccard(w_1, w_2) = \frac{Hits(w_1 AND w_2)}{Hits(w_1) + Hits(w_2)} \quad (1)$$

ここで，$Hits(w)$ は語 w を検索したときのヒット件数を表す．ノード間にエッジを張るか張らないかは，Jaccard 係数が閾値を超えるかどうかで判定する．なお，閾値は 10^{-4} とした．

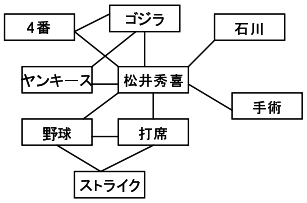

図 1: 松井秀喜に関する語の共起ネットワーク

3.2 語の別名らしさの評価

生成したネットワーク上で，与えられた名前 $Name$ に対する，語 w の別名らしさを次の式 $Score$ で測る．

$$Score(w) = |1hop(w) \cap 1hop(Name)| \quad (2)$$

ここで，$1hop(w)$ は語 w と 1 ホップの関係でつながっている語の集合を表す．$Score$ は，語 w が名前 $Name$ とどれだけ多くの語を共有するかを調べることで，語 w と名前 $Name$ のネットワーク上での役割の類似性を測っている．図 1 で言えば,「ゴジラ」と 1 ホップの関係にある「4 番」と「ヤンキース」は，どちらも「松井秀喜」とも 1 ホップの関係にあるため,「ゴジラ」は別名らしい．一方,「打席」と 1 ホップの関係にあるのは「野球」と「ストライク」だが，これらのうち「松井秀喜」と 1 ホップの関係にあるのは「野球」だけであるため,「打席」は別名らしくない．

4. 実験・考察

提案手法の有効性を確かめるためには，提案手法によって生成できる，別名らしさで並んだ語のランキングの正しさを検証する必要がある．そこで，我々は有名人 20 人に関する名前とその別名の正解データを用意した．さらに，提案手法との比較のために，2 つのベースライン手法と 1 つの理想的手法を用意した．

Low1 ノードをランダムにランキングする

Low2 ノードを与えられた名前との共起が強い順に並べる

Up ノードを最適にランキングする

正解データに対し，提案手法と比較手法の出力するランキングの Mean-Average-Precision(以下 MAP) を調べた結果，表 1 のようになった．また，提案手法に対し,「中田英寿」,「堀江貴文」を入力したときの，出力はそれぞれ表 2，表 3 のようになった．

表 1: 人物の別名に対する各手法の抽出精度の比較

	提案手法	Low1	Low2	Up
MAP	0.380	0.093	0.104	0.700

表 1 から，提案手法が 2 つのベースライン Low1，Low2 を上回ることが分かる．式 2 によるノードの評価が有効に働いているといえる．しかし，理想的手法 Up と比較すると，提案手法の精度は 1/2 程度である．また，理想的手法でも MAP が 1 となっていない．ノードの評価，選択ともに改善の余地があるといえる．提案手法が失敗に終わるのは，主に，別名があいまいな場合であった．表 2 の結果はその典型的な例である．

表 2: 中田英寿に対する別名抽出結果

順位	別名	正否
1	サッカー	×
2	誇り	×
3	ヒデ	○
4	引退	×
5	古巣	×

表 3: 堀江貴文に対する別名抽出結果

順位	別名	正否
1	ホリエモン	○
2	保釈	×
3	判決	×
4	ライブ	×
5	地裁	×

5. おわりに

与えられた名前の別名を Web から自動的に抽出する手法を提案した．評価実験の結果，ベースラインを大きく上回る精度が確認できた．人物以外の正解データで提案手法の性能を検証すること，あいまいな別名も精度よく抽出できるようにすることが今後の課題である．

参考文献

[1] R. Guha and A. Garg. Disambiguating people in search. In Stanford University, 2004.

[2] T. Hokama and H. Kitagawa. Extracting mnemonic names of people from the web. In Proc. of the 9th ICADL, pp. 121-130, 2006.

[3] P. Turney. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In Proc. of the 40th ACL, pp. 417-424, 2002.

[4] M. Pasca, D. Lin, J. Bigham, A. Lifchits, and A. Jain. Organizing and searching the world wide web of facts - step one: the one-million fact extraction challenge. In Proc. of the 21st AAAI, 2006.

[5] R. Holzer, B. Malin, and L. Sweeny. Email alias detection using network analysis. In Proc. of SIGKDD Workshop on Link Discovery: Issues, Approaches, and Applications, 2005.

[6] P. Hsiung, A. Moore, D. Neill, and Jeff Schneider. Alias Detection in Link Data Sets. In Proc. of the 2005 ICIA, 2005.