関係名詞らしさを用いた固有表現間の関係同定

平野 徹　松尾 義博　菊井 玄一郎
日本電信電話株式会社 NTT サイバースペース研究所
{hirano.tohru,matsuo.yoshihiro,kikui.genichiro}@lab.ntt.co.jp

1 はじめに

近年，Web に存在する膨大な量のテキストを知識源として活用するための研究が注目されている [11]．計算機でテキストを知識源として活用するには，テキストから知識を構造化された形で抽出する必要がある．そこで我々は，テキスト中で言及されている，固有表現 X, Y とその関係を示す表現（関係表現）R を［X，Y，R］の構造化された形で抽出することを目指している．

本稿で抽出する［X, Y, R］は，テキスト中で明記・含意されている,「X が Y を[1]R する」もしくは「X の R は Y だ」で表現しうる関係とする．ここで，関係表現 R がテキスト中に出現する［X, Y, R］を**明示的関係**，テキスト中にないものを**暗黙的関係**と呼ぶ（明示的関係には，X，Y，R が異なる文に出現するものも含む）．例えば,「浅田真央の姉の浅田舞は〜」から抽出する［浅田真央，浅田舞，姉］は，関係表現 R=“姉” がテキスト中に出現しているので明示的関係となる．一方,「民主党の小沢一郎氏は〜」から抽出する［民主党，小沢一郎，党員］は，関係表現 R=“党員” がテキスト中にないので暗黙的関係となる．

なお本稿では，抽出する［X, Y, R］と実世界との整合性を考慮しないことに注意されたい．例えば，実世界で X と Y が R の関係を有するとしても，入力テキスト中で言及されていなければ［X, Y, R］は抽出しない．逆もまたしかりである．

テキストから［X, Y, R］を抽出する研究は，抽出したい関係表現 R が与えられる研究 [1, 4, 7, 10, 13] と，関係表現 R も発見する研究 [2, 3] に大別できる．前者の研究では，明示的関係と暗黙的関係の両方が抽出可能だが，抽出したい関係表現の数だけ，統語パタンやシード，学習データの構築などの人的コストがかかる．

一方，後者の研究では，抽出したい関係表現ごとの人的コストはないが，4.1 で後述するように，実データ中の［X, Y, R］のうち，暗黙的関係は約 36%を占めるにも関わらず，抽出することができない．

そこで我々は，抽出したい関係表現ごとの人的コストがなく，また暗黙的関係も抽出できるように，次の 2 ステップで,［X, Y, R］の抽出を行う．

Step 1: 何らかの関係を有する組 X, Y を抽出
Step 2: 抽出された X と Y の関係表現 R を特定

Step1 について，我々は，任意の組 X, Y に対して，何らかの関係（明示的関係と暗黙的関係の両方）を有するか否かの分類問題として，文脈的素性を用いた手法を提案した [5]．Step2 については，Step1 で抽出された組 X, Y ごとに，まず (a) テキスト中に出現する関係表現 R の同定を試み（明示的関係），次に (b) 関係表現 R が未特定の組 X, Y の関係表現 R を推定する（暗黙的関係）．これらのうち，本稿ではまず (a) テキスト中に出現する関係表現の同定に取り組む．

なお暗黙的関係はテキストの文脈に頼ることなく X, Y の組から容易に推定できるデフォルト的なものであることから，(b) については，入力の X, Y に対して既定の関係表現を用意することで対処する．この既定の関係表現は，組織名-人名などの X, Y のクラスの組ごとに人手で定める方法や，大規模テキストコーパスから処理 (a) で抽出した大量の［X, Y, R］を用いて定める方法が考えられる．

本稿では，(a) テキスト中に出現する関係表現 R を同定する手法として，関連研究で注目されていた文の構造情報を用いた手法に加え，新たに，関係を示しやすい語の特徴（関係名詞らしさ）を導入した手法を提案し，その有効性について議論する．

以下，2 節でテキスト中に出現する関係表現の同定について述べ，3 節で関係名詞らしさを導入した手法を提案する．4 節で評価実験の結果を報告し，最後に 5 節でまとめる．

2 テキスト中に出現する関係表現の同定

テキスト中に出現する関係表現の同定とは，入力の組 X, Y の関係表現 R をテキストから発見するタスク

[1]“を” だけでなく，“に”，“で” など任意の格助詞でも良い

図 1: 係り受けの構造情報の例

である．例えば,「浅田真央の姉の浅田舞は銅メダルを獲得した。」と X=“浅田真央”，Y=“浅田舞” を入力として，関係表現 R=“姉” を出力するタスクである．

本タスクにおいて，Banko らは，入力の組 X, Y の間に出現する単語列を対象に，系列ラベリング問題として，CRF を用いた R の同定手法を提案し [2]，長谷川らは，入力の組 X, Y が 10 単語以内にある場合に限り，X と Y の間に出現する単語列を R と同定する手法を提案した [3]．これらの研究は英語で行なわれている．英語ではテキスト中に出現する関係表現 R の約 86% が X と Y の間に出現するため，提案された手法で同定できる．しかしながら，日本語においては，関係表現 R が X と Y の間に出現する割合は約 26%と英語とは大きく異なる．そのため，従来研究を日本語に適用することは難しい．

そこで我々は，テキスト中の全文節を候補として，その中から適切な候補を 1 つ選択し，選択された文節の先頭から主辞までの単語列を関係表現 R と同定することにした．ここで，入力には暗黙的関係の組 X, Y も含まれるため，どの候補も関係表現 R でないと棄却する必要があることに注意されたい．

3 提案手法

本稿では，関連研究 [7, 13] で，その有効性が報告されている文の構造情報を用いた手法に加え，新たに，関係を示しやすい語の特徴（関係名詞らしさ）を導入した手法を提案する．

我々は，テキスト中に出現する関係表現 R を同定するための手がかりとして，各候補文節の単語が，どの程度関係を示しやすいかに注目した．例えば，先の例では，“姉”，“銅メダル”，“獲得した” の候補の中で，“姉” が人と人の関係を最も示しやすいため関係表現 R と判断できる．しかしながら，関係を示す語のリストとして利用できる資源もなく，また関係を示しやすい語を推定する手法も自明ではない．本稿では，動詞は全ての組 X, Y の関係を示し得る語と考え，関係を示しやすい名詞の推定を試みた．

ここでは，提案手法のベースとなる文の構造を用いた手法と，提案手法で導入する関係名詞らしさとその推定手法について説明する．

3.1 文の構造を用いた手法

文の構造情報を用いた手法は，係り受けの構造で，X, Y に対して，関係表現 R が出現しやすい構造があるという考えに基づく．例えば，先の例では，図 1 のような係り受け構造を持ち，X=“浅田真央”，Y=“浅田舞” の間に出現する “姉” が関係表現 R であると判断できる．本稿では，関係表現 R が出現しやすい係り受け構造を見つけるために，入力の X, Y と当該候補を含む最小の係り受け木を，各文節の係りタイプ・表層格と，人名や組織名などの X, Y のクラス，候補の表記・品詞を用いて，図 2 のような木で表現する．この木の部分木を素性とするブースティングに基づく分類器 [9] を用いて，one vs rest 法により，候補の中から適切な候補を 1 つ選択する．

3.2 関係名詞らしさの導入

固有表現 X と Y の関係表現 R が名詞の場合,「X の R は Y だ」で表現でき，名詞 R は項 X を取ると考えられる．この種の名詞は，一項名詞あるいは相対名詞，関係名詞と呼ばれる（以降，代表して関係名詞と呼ぶ）．関係名詞は，基準 X が定まらないと指示対象 Y が定まらない名詞と定義され [8],「X の “関係名詞”」の形で，指示対象の名称 Y を用いずに指示対象を表現できる．例えば，指示対象 Y=“浅田舞” を表現するのに，基準 X=“浅田真央” と関係名詞=“姉” を用いて「浅田真央の姉」と表現できる．

我々は，関係を示しやすい名詞を推定するために，次の 2 つの関係名詞の特徴に基づく素性を導入し，各素性を図 2 で示すように係り受けから作成した素性木内の候補ノードの直下に配置する．

3.2.1 基準 X に基づく関係名詞らしさ（R1）

関係名詞と基準 X の特徴に基づいた関係名詞の推定手法が提案されている [12]．この手法は,「A の B」において，B が関係名詞ならば A は特定の意味範疇に集中し，B が関係名詞以外ならば A は分散するという考えに基づく．

具体的には，大規模テキストコーパスから「A の B」にマッチする名詞 A と名詞 B を収集し，A を分類語彙表のカテゴリ C= $\{ c_1, c_2, \cdots, c_m \}$ に割り当てる．次に，B を名詞 n に固定した時の A のカテゴリの散らばりを次式で算出する．

$$\mathcal{H}c(n) = -\sum_{c \in C} P(c|n) log_m P(c|n) \quad (1)$$

ただし，

$$P(c|n) = \frac{freq(c,n)}{freq(n)} \quad (2)$$

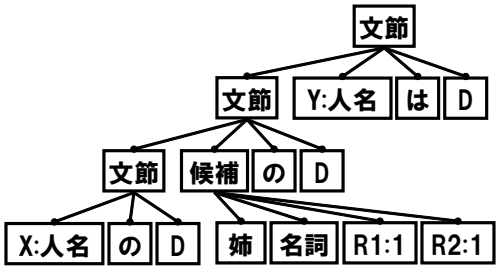

図 2: X=“浅田真央”, Y=“浅田舞”, 候補=“姉” の素性木

最後に，算出された $\mathcal{H}c(n)$ を名詞 n の特徴量として，周辺の正例と負例の密集率を比較する k-NN 推定法によって，n が関係名詞か否かを判定する．

本稿では，上記手法を日本語語彙大系 [6] のカテゴリを用いて実装し，ニュース 1,698,798 記事，ブログ 10,449,468 記事を対象に行なった．抽出された「A の B」は 55,492,891 用例であった．また，k-NN 推定法で用いる学習データとして，4.1 の学習データ中の名詞のうち，関係表現とタグ付けされた名詞を関係名詞，そうでない名詞を関係名詞でないとみなし利用した．

このように関係名詞かを推定した結果を素性として導入する．本稿では，本素性を “R1” と呼び，素性の値は，関係名詞ならば 1，関係名詞以外ならば 0 となる．

3.2.2 指示対象 Y に基づく関係名詞らしさ（R2）

「X の R は Y だ」において，R を名詞 n に固定して，その指示対象 $y \in Y$ を並べてみると，n は固有表現の集合 Y のカテゴリとなっている．例えば，n を “知事” に固定して，その指示対象 $y \in Y$ を並べると,「石原慎太郎」「橋本徹」「東国原英夫」など，人の集合が得られ，“知事” はこの集合のカテゴリである．言い換えると,「n が関係名詞ならば，n は固有表現のカテゴリである」が成り立つ．

そこで，先のテキスト集合から，単純な統語パタン「A は B だ」（但し，A は固有表現に限る）で固有表現のカテゴリリストを獲得した（293,820 用例:頻度 10 回以上）．このリストに n があるかを素性として導入する．本稿では，本素性を “R2” と呼び，素性の値は，カテゴリリストにあれば 1，なければ 0 となる．

ここで,「n が固有表現のカテゴリならば，n は関係名詞である」（逆）は真とは限らないが,「n が固有表現のカテゴリでないならば，n は関係名詞でない」（対偶）は真であるため，本素性は，関係名詞でない名詞の候補を除くフィルターとして効果が期待できる．

4 評価実験

テキスト中に出現する関係表現 R の同定において，提案手法の有効性を調査するために，次の手法を比較評価した．実験では，既存の形態素解析器・係り受け解析器の結果を利用した．

DEP 係り受けの構造情報を用いた手法
DEP+R1 DEP に素性 “R1” を加えた手法
DEP+R2 DEP に素性 “R2” を加えた手法
DEP+R1+R2 DEP に素性 “R1” と “R2” を加えた手法

4.1 評価データ

抽出対象となる［X, Y, R］を付与した，日本語の新聞 1,400 記事とブログ 4,800 記事を用いる．本コーパス中には，何らかの関係を有する組 X, Y は 17,075 組あった（但し，X, Y が同文に出現する組に限る）．内訳は，明示的関係が 10,938 組 (約 64%)，暗黙的関係が 6,137 組 (約 36%) であった．この明示的関係のうち，関係表現 R も X, Y と同じ文に出現する場合が 8,907 組で，異なる文に出現する場合は 2,031 組であった．

本稿では，係り受けの構造情報を用いた手法をベースにしているため，X, Y と同じ文に出現する関係表現 R を正解として評価する．つまり，17,075 組を入力とし，8,907 組で関係表現 R を文内から選択し，残りの 8,168 組は関係表現 R なしと返すべき問題として評価する．また，17,075 組を文書単位に 10 分割し，1〜6 を学習データ，7 を開発データ，8〜10 を評価データとして利用した．評価データは，5,470 組が入力で，2,886 組で関係表現 R が文内に出現した．

4.2 実験結果

文内に出現する関係表現 R の同定において，提案する関係名詞らしさを導入することによりどの程度解析性能が向上するかを調査した．実験結果として，選択された候補の，分類器の出力した識別関数の値を動かして描いた再現率-精度曲線を図 3 に示す．なお精度と再現率は次式の通りである．

$$\text{精度} = \frac{\text{システムが同定した正解の関係表現数}}{\text{システムが同定した関係表現数}}$$

$$\text{再現率} = \frac{\text{システムが同定した正解の関係表現数}}{\text{正解の関係表現数}}$$

図 3 より，提案した DEP+R1+R2 が文内に出現する関係表現の同定において有効であることが確認できる．また，識別関数の値が 0 のとき，DEP は精度 61.1% (982/1,608)，再現率 34.0% (982/2,886) なのに対し，DEP+R1+R2 は精度 67.6% (1,178/1,743)，再現率 40.8% (1,178/2,886) と精度が 6.5 ポイント，再現率が 6.8 ポイント向上したことが分かる．

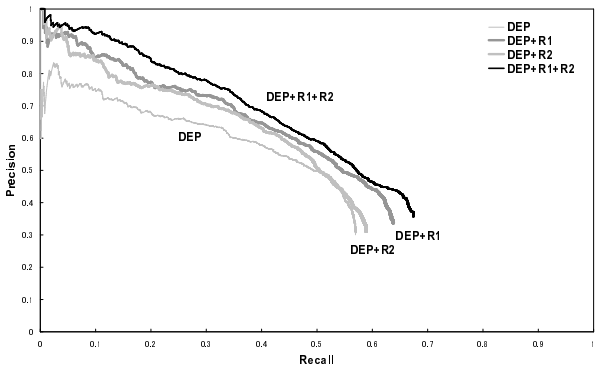

図 3: 明示的関係の同定における再現率-精度曲線

次に，図 3 において，各手法の最大再現率を比較し，関係表現が文内に出現する 2,886 組の X, Y に対して，関係表現 R を適切に選択できたかを評価する (表 1)．DEP の最大再現率は 57.0% であるのに対し，DEP+R1+R2 は 67.4% と 10.4 ポイントも向上しており，ここでも提案手法の有効性が確認できた．

最後に，図 3 の再現率-精度曲線を比較し，関係表現が文内にない 2,584 組の X, Y に対して，選択された候補を関係表現 R でないと適切に棄却できたかを評価する．ここで，DEP と DEP+R2 の曲線を比較すると，再現率 55% あたりでは共に精度 43% だが，再現率 10% あたりでは，DEP は精度 75% であるのに対し，DEP+R2 は精度 85% と 10 ポイントも高い．DEP+R1 と DEP+R1+R2 でも，同様の現象が確認できる．これらは R2 の素性が関係を示さない名詞候補のスコアをディスカウントし，関係表現が文内にない組で選択された候補を適切に棄却できたことを示している．

4.3 誤り分析

関係表現が文内に出現する 2,886 組の X, Y に対して，提案手法がどの程度関係表現 R を適切に選択できたかを見ると，942 事例が誤っていることが分かる (表 1)．このうち，362 事例は正解の関係表現が名詞であり，その約 90%で誤って動詞を選択した．

この誤りを改善するためには，本稿では用いていない格フレームやゼロ代名詞などの動詞の情報を利用する必要がある．例えば，述語項構造解析や照応解析での知見を利用することで，更なる性能改善が期待できる．

別の改善案として，今回は，英語での関連研究と同様，名詞と動詞の関係表現を同時に抽出することを試みたが，名詞と動詞の関係表現を別々に抽出する方法も考えられる．例えば，動詞の関係表現は述語項構造解析や照応解析に任せ，名詞の関係表現だけを対象に本手法を行うことでも性能改善が期待できる．

表 1: 各手法の最大再現率

DEP	DEP+R1	DEP+R2	DEP+R1+R2
57.0% (1,644/2,886)	63.7% (1,839/2,886)	58.8% (1,697/2,886)	67.4% (1,944/2,886)

5 おわりに

本稿では，入力された X, Y のテキスト中に出現する関係表現 R の同定に取り組み，関連研究で注目されていた文の構造情報に加え，新たに，関係を示しやすい語の特徴（関係名詞らしさ）を導入した手法を提案した．評価実験では，提案手法は精度 67.6%，再現率 40.8%と構造情報のみの手法より，精度が 6.5 ポイント，再現率が 6.8 ポイント向上したことがわかり，提案手法の有効性が確認できた．

今後は，文内に出現する関係表現の同定性能を改善するとともに，文間に出現する関係表現の同定や，暗黙的関係の関係表現の推定にも取り組む予定である．

参考文献

[1] Agichtein, E. and Gravano, L.: Snowball: Extracting Relations from Large Plain-Text Collections, *Proceedings of the 5th ACM conference on Digital libraries*, pp. 85–94 (2000).

[2] Banko, M. and Etzioni, O.: The Tradeoffs Between Open and Traditional Relation Extraction, *Proceedings of the 46th Annual Meeting on Association for Computational Linguistics: Human Language Technologies*, pp. 28–36 (2008).

[3] Hasegawa, T., Sekine, S. and Grishman, R.: Discovering Relations among Named Entities from Large Corpora, *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, pp. 415–422 (2004).

[4] Hearst, M. A.: Automatic Acquisition of Hyponyms from Large Text Corpora, *Proceedings of the 14th conference on Computational linguistics*, pp. 539–545 (1992).

[5] 平野徹, 松尾義博, 菊井玄一郎: 文脈的素性を用いた固有表現間の関係性判定, 自然言語処理, Vol. 15, No. 4, pp. 43–58 (2008).

[6] 池原悟, 宮崎正弘, 白井諭, 横尾昭男, 中岩浩巳, 小倉健太郎, 大山芳史, 林良彦: 日本語語彙大系 CD-ROM 版, 岩波書店 (1999).

[7] Kambhatla, N.: Combining Lexical, Syntactic, and Semantic Features with Maximum Entropy Models for Extracting Relations, *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, pp. 178–181 (2004).

[8] 菊池隆典, 白井英俊: 日本語名詞句の意味解釈の検討, 日本認知科学会第 19 回大会発表論文集, pp. 78–79 (2002).

[9] 工藤拓, 松本裕治: 半構造化テキストの分類のためのブースティングアルゴリズム, 情報処理学会論文誌, Vol. 45, No. 9, pp. 2146–2156 (2004).

[10] Pantel, P. and Pennacchiotti, M.: Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations, *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pp. 113–120 (2006).

[11] Sekine, S.: NSF Sponsored Symposium: Semantic Knowledge Discovery, Organization and Use (2008).

[12] 田中省作, 冨浦洋一, 日高達: 意味範疇の散らばりに基づいた名詞の統語範疇の分類, 情報処理学会論文誌, Vol. 40, No. 9, pp. 3387–3396 (1999).

[13] Zelenko, D., Aone, C. and Richardella, A.: Kernel Methods for Relation Extraction, *Journal of Machine Learning Research*, Vol. 3, pp. 1083–1106 (2003).