テキストにおける固有表現間の意味的関係の抽出

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，統計的言語処理技術の発展によりテキスト中の人名や地名，組織名といった固有表現 (Named Entity) を高精度で抽出できるようになってきた．これを更に進めて,「安倍晋三（人名)」は「日本（地名)」の「首相（関係ラベル)」であるといった固有表現間の関係を抽出する研究が注目されている [1, 2, 4, 6, 7, 11]．固有表現間の関係が抽出できれば，より複雑な情報検索，質問応答や要約に有益である．

我々は，入力されたテキストから関係の 3 つ組である［固有表現$_1$，固有表現$_2$，関係ラベル］を抽出する研究を進めている．例えば,「温家宝首相は人民大会堂で日本の安倍晋三首相と会談した。」というテキストから［温家宝，安倍晋三，会談］［温家宝，人民大会堂，会談］［安倍晋三，人民大会堂，会談］［安倍晋三，日本，首相］のような関係の 3 つ組を抽出する．この関係の 3 つ組をテキストから抽出するには，(a) テキストにおける固有表現間の関係の有無を推定（**関係性判定**）する技術と，(b) 固有表現の組みの関係ラベルを推定する技術が必要である．

本稿では，(a) 関係性判定である，テキスト中の二つの固有表現の間の直接的な意味的関係の有無を推定する手法を提案する．ここで二つの固有表現が直接的な意味的関係にあるとは，二つの固有表現が共に出現する意味フレームが想定できることである．例えば「温家宝首相は人民大会堂で日本の安倍晋三首相と会談した。」というテキストにおいて,「温家宝⇔安倍晋三」,「温家宝⇔人民大会堂」と「安倍晋三⇔人民大会堂」は「会談」という意味フレームに,「安倍晋三⇔日本」は「首相」という意味フレームに共に出現すると判断できるため意味的関係にあるとする．一方,「温家宝⇔日本」と「人民大会堂⇔日本」は共に出現する意味フレームがないため意味的関係にないとする．

固有表現間の関係性判定の従来研究は，文構造を素性として用いた機械学習の研究が多い [1, 4, 7, 11]．例えば，Kambhatla らの研究 [7] では，二つの固有表現の関係の有無を判断するのに，係り受け木における二つの固有表現の最短パスを素性として利用した手法を提案している．しかし，3.2 で後述するように，実データ中に存在する直接的な意味的関係のある固有表現の組みのうち，異なる文に出現する場合は全体の約 35 %を占め，従来研究のように文構造などの文に閉じた素性だけを用いた手法では不十分である．そこで，二つの固有表現が異なる文に出現する場合に有用だと考えられる文脈情報などの複数の文をまたぐ素性を用いることで，固有表現間の関係性判定の性能向上が期待できる．

本稿では，文構造などの文に閉じた素性だけでなく，文脈情報などの複数の文をまたぐ素性を導入した機械学習に基づく関係性判定手法を提案し，その有効性について議論する．

以下，2 節で提案手法を説明し，3 節で提案手法の有効性を調査するために行った評価実験の結果を報告する．4 節はまとめである．

2 関係性判定における文脈的素性の利用

提案手法では，異なる文に出現する二つの固有表現の関係性判定のためにセンタリング理論に基づく文脈的素性を利用する．これと従来研究で用いられている文構造に基づく素性を組み合わせた機械学習手法により精度向上を目指す．ここでは，文脈的素性の基本的な考え方と関係性判定に利用する具体的な素性について説明する．

2.1 文脈的素性の基本的な考え方

異なる文に出現する二つの固有表現の間に関係性があるということは，先行する固有表現が後続する固有表現を含む文において「文脈的に参照され易い」ことを意味する．例えば次のテキストにおいて,「小泉首相」と「アメリカ」は前者が後者の文において文脈的に参照され易い（実際にガ格ゼロ代名詞の先行詞である）ため「渡る」という関係を持つが,「胡錦濤国家主席」と「アメリカ」の場合は上述のような関係が成立しないため関係はない．

> 明日、**小泉首相**$_i$ は中国を訪れ，胡錦濤国家主席と会談する。その後 (**φ**$_i$**ガ**) アメリカに渡り ブッシュ大統領と会談を予定している。

以上のことから，異なる文に出現する二つの固有表現の間の関係性を判別する上で,「先行する固有表現が、後側の固有表現が出現する文においてどの程度参照され易いか」という情報を素性として用いることが有用であると考えられる．ある名詞句が後続する文脈において「どの程度参照され易いか」を評価する枠組みとして本研究ではセンタリング理論 [5] [1]を用いる．

2.2 センタリング理論

センタリング理論とは，代名詞や省略の先行詞同定のための経験的優先規則であり，代名詞や省略がテキストに出現した際，既に出現した名詞句を先行詞になりやすい順に並び替えるものである．その並び替え方法は，まず，代名詞や省略が出現するまで，テキストの先頭から順に名詞句を格助詞ごとにスタックする．例において，先頭から「渡る」の省略要素までにある名詞句の「明日」「小泉首相」「中国」「胡錦濤国家主席」を順に，格助詞に基づいてスタックすると，図 1 のような情報が取得できる．そしてスタックされた情報を次のルールによって並び替える．

1. 格助詞は「は＞が＞に＞を＞他」の順序
2. 格助詞内はスタック構造の順序

例えば，図 1 のようにスタックされた情報では，先行詞としてふさわしい順は，上記の 2 つの規則により，1.「小泉首相」，2.「中国」，3.「胡錦濤国家主席」，4.「明日」となる．この並び替え規則によって最上位になった「小泉首相」を「渡る」の省略要素の先行詞に同定することができる．

2.3 関係性判定への適用方法（提案 1）

関係性判定では，上記のセンタリング理論を，関係を判断したい二つの固有表現のうち，先行する固有表現が，後側の固有表現が出現する文において参照され易いか否かを表現するために利用する．ここで，本研究においては明示的な省略補完は行わないことに注意されたい．

具体的には，関係を判断したい二つの固有表現のうち，後側の固有表現が出現するまで，テキストの先頭から順に名詞句を格助詞ごとにスタックし，スタックされた情報を 2.2 の 2 つの規則によって並び替える．そして，並び替え規則によって最上位になった名詞句と前側の固有表現が一致すれば，前側の固有表現は，後側の固有表現が出現する文において参照され易いと判断する．

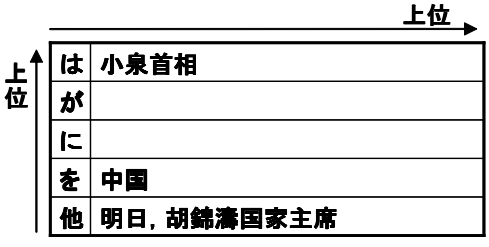

図 1: スタックされた情報

例において,「小泉⇔アメリカ」の関係性を判定する際，テキストの先頭から後側の固有表現である「アメリカ」が出現するまでにある名詞句の「明日」「小泉首相」「中国」「胡錦濤国家主席」を順に，格助詞に基づいてスタックする（図 1）．そして，2 つの規則により，1.「小泉首相」，2.「中国」，3.「胡錦濤国家主席」，4.「明日」と並び替える．ここで，最上位になった「小泉首相」と前側の固有表現「小泉」が一致するので，後側の固有表現が出現する文において参照され易いと判断する．この参照され易いか否かを素性として関係性判断に利用する（CT：Centering Top）．

2.4 スタックされた情報の利用方法（提案 2）

2.2 で述べたセンタリング理論の並び替えは，人名や組織名の主語になりやすい語は上位に，地名や時間の主語になりにくい語は下位になる傾向がある．しかし，固有表現間の関係性判定においては，主語になりやすい人名や組織名だけでなく，場所や時間といった主語になりにくいがテキスト内容の状況（いつ，どこで）を示す固有表現も考慮する必要がある．

そこで，二つの固有表現の関係の有無を判断するために， 2.3 の方法によって**スタックされた情報を，構造を持つ情報と捉え，その構造情報を固有表現間の関係性判定に用いる**（CS：Centering Structure）．ここで，図 1 のようにスタックされている情報を，図 2 に示す構造を持つ情報と捉える．図 2 の構造に変換するには，まず，図 1 が関係を判断したい二つの固有表現「小泉⇔アメリカ」のうち，後側の固有表現「アメリカ」が出現するまでのスタック情報であることから，ルートノードを「アメリカ」とする．そして，図 1 のスタック情報を，格助詞に基づき図 2 へ変換する．この変換された構造において，二つの固有表現の最短パスの構造を，固有表現間の関係性判定の情報として利用する．例えば,「小泉⇔アメリカ」の情報として，図 2 の構造における最短パスの構造である「アメリカ→は：小泉首相」を利用する[2].

このように，センタリング理論によってスタックされた情報の構造を用いることで，場所や時間としてどの程度参照され易いかが取得できると考えられる．

2.5 分類器

分類器には，一般的に構造情報を用いた研究で高精度な分類結果が報告されている構造情報を明示的に利

[1]実際にはセンタリング理論を拡張した亀山の研究 [8] を用いた

[2]「A → B」は A が B に係ることを示す

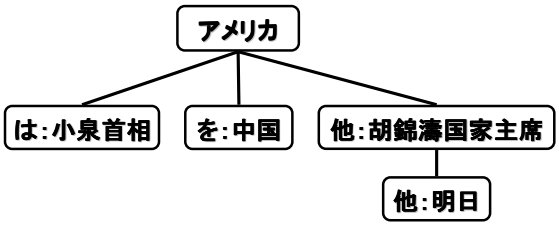

図 2: 構造化されたスタック情報

用した分類手法を用いる．構造情報を明示的に利用した分類手法には，Collins の Tree Kernel[3] や，鈴木らの HDAG Kernel[10] などのカーネル法を利用した手法，工藤らの部分木を素性とするブースティングを利用した手法 [9] などがある．今回の実験では，比較的学習時間が短く実験が容易に行える工藤らのアルゴリズムが実装された分類器 BACT を使用した．

固有表現間の関係性判定では，学習時に抽出された文構造に基づく素性と提案した文脈的素性を，ひとつの大きな木構造で表現し，その木から分類に有効な規則集合を学習する．解析の際には，学習時と同様，解析対象となる二つの固有表現の文構造に基づく素性と提案した文脈的素性を大きな木構造で表現し，学習した規則集合を適用することで二つの固有表現の直接的な意味的関係の有無を判断する．

3 評価実験

提案手法の有効性を調査するために，テキスト中の二つの固有表現の間の関係性判定実験において，次の手法を比較評価した．なお，文構造に基づく素性や文脈的素性の抽出は，既存の形態素解析器・係り受け解析器で得られた結果を利用した．

1. **WD**：二つの固有表現が前後 n 単語内にある場合は意味的関係にあるとする手法
2. **STR**：文構造を用いた手法 [7]
3. **STR-CT**：STR に加え，スタック情報から最上位の名詞句を用いた手法（提案 1）
4. **STR-CS**：STR に加え，スタック情報の構造情報を用いた手法（提案 1 ＋提案 2）

3.1 評価データ

テキスト中の固有表現の組みに人手で意味的関係の有無を判断した日本語の新聞記事とブログの計 1451 テキストを用いる．なお今回の実験では固有表現の組み合わせとして人名と地名の組みに限って実験を行う[3]．評価データ中には，人名と地名の総組み合わせである 236142 組みのうち 5110 組みが意味的関係にある．この評価データを対象に 10 分割交差検定を行った．

[3]その他の固有表現の組み合わせの評価は今後行う予定である

3.2 学習方法

今回の実験では，対象となる二つの固有表現が (A) 同じ文に出現する場合と (B) 異なる文に出現する場合に分けて学習しモデルを作成した．このように対象を 2 つに分ける理由として，(A) には文に閉じた素性が，(B) には複数の文をまたぐ素性が有効であると考えられ，分けずに学習すると各々の特徴が目立たず適切な学習ができないと考えられるからである．また評価データにおける (A) 同文と (B) 異文の内訳は表 1 のようになっており，(A) と (B) で組み合わせ総数に対する意味的関係のある組みの割合が極端に異なることからも (A) と (B) を分けて学習することが考えられる．

表 1: 評価データにおける (A) 同文と (B) 異文の内訳

	関係あり数	組み合わせ総数
(A) 同文	3333	10626
(B) 異文	1777	225516
(A)+(B)	5110	236142

3.3 実験結果

テキストにおける固有表現間の関係性判定実験において，提案した文脈的素性を用い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．実験結果を表 2 に示す．表 2 では，”(A)+(B) 全体”，”(A) 同文”，”(B) 異文”ごとに結果を示している．また，分類器の出力した識別関数の値を動かして再現率-精度曲線を描いた (図 3)．ただし，WD では，単語距離 n を変化させて再現率-精度曲線を描いた．なお精度と再現率は次式の通りである．

$$\text{精度} = \frac{\text{システム出力した正解関係あり数}}{\text{システムが出力した関係あり数}}$$

$$\text{再現率} = \frac{\text{システムが出力した正解関係あり数}}{\text{正解関係あり数}}$$

表 2 から，STR-CS(提案 1 ＋提案 2) は STR(従来研究) より精度が約 4.4 %，再現率が約 6.7 %向上したことがわかる．また，STR-CT(提案 1) も従来研究より精度が約 2.1 %，再現率が約 4.0 %向上したこともわかる．これらの結果より，2 つの提案（提案 1 と提案 2）の有効性が確認できた．

また，表 2 の”(A) 同文”と”(B) 異文”の結果を見ると，提案手法によって関係性判定の性能向上をねらっていた (B) で精度が約 12.6 %，再現率が約 17.0 %向上しているだけでなく，(A) でも精度が約 4.5 %，再現率が 5.4 %向上していることがわかる．これは，提案手法が，並列文における固有表現間の関係性判定に有効だったと考えられる．例えば,「長澤まさみは渋谷で、速水もこみちは新宿で舞台挨拶を行なった。」の並列文において，係り受け木における二つの固有表現の最短パスを

表 2: 関係性判定における各手法の結果比較

	(A)+ (B) 全体		(A) 同文		(B) 異文	
	精度	再現率	精度	再現率	精度	再現率
WD10	0.430 (2501/5819)	0.489 (2501/5110)	0.481 (2441/5075)	0.732 (2441/3333)	0.080 (60/744)	0.034 (60/1777)
STR	0.693 (2562/3696)	0.501 (2562/5110)	0.756 (2374/3141)	0.712 (2374/3333)	0.339 (188/555)	0.106 (188/1777)
STR-CT	0.714 (2764/3870)	0.541 (2764/5110)	0.784 (2519/3212)	0.756 (2519/3333)	0.372 (245/658)	0.138 (245/1777)
STR-CS	0.737 (2902/3935)	0.568 (2902/5110)	0.801 (2554/3187)	0.766 (2554/3333)	0.465 (348/748)	0.276 (348/1777)

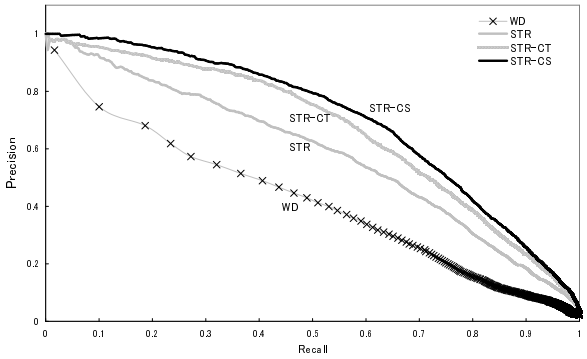

図 3: 関係性判定における再現率-精度曲線

文構造とする従来研究は，関係のある「長澤まさみ⇔渋谷」と関係のない「長澤まさみ⇔新宿」を区別できなかった．一方，提案手法では,「長澤まさみ⇔渋谷」に対し「渋谷→は：長澤まさみ」,「長澤まさみ⇔新宿」に対し「新宿→は：速水もこみち→は：長澤まさみ」が情報として利用され，これらを区別できるようになったと考えられる．

3.4 誤り分析

テキストにおける固有表現間の関係性判定において，誤って推定した事例のうち分類器の出力する識別関数の絶対値が大きいものから 200 事例を分析した結果，次の問題が大半を占めることがわかった．これらの誤りについては今後の課題としたい．

並列構造の文

係り受け木における二つの固有表現の最短パスを文構造とする従来研究では， 3.3 でも述べたように，並列文の並列関係を考慮できない．

特殊な照応

首相や社長といった役職を示す名詞は代名詞と同様に先行詞を持つことがある．また，同日などのように「同 · · ·」といった名詞による照応が存在する．

4 おわりに

本稿では，テキストにおける固有表現間の関係性判定に取り組み，従来の文構造などの文に閉じた素性だけでなく，複数の文をまたぐセンタリング理論に基づいた文脈的素性導入した機械学習に基づく手法を提案した．

人名と地名に対する評価実験では，提案手法は精度 73.7 %，再現率 56.8 %と，従来研究より精度が約 4.4 %，再現率が約 6.7 %向上したことがわかり，提案手法の有効性が確認できた．

なお，提案手法として用いたセンタリング理論は，先行詞同定の古い技術であるにも関わらず，評価実験において，その有効性が確認できた．このことから，先行詞同定に関する最近の研究成果を利用することにより，関係性判定においてさらなる性能向上が期待できる．

今後は，上記の誤り分析で述べた問題に取り組むとともに，人名⇔地名以外の固有表現の組みにおいても提案手法を評価したい．その後，(b) 関係ラベルの推定に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, *5th ACM International Conference on Digital Libraries*, pp. 85–94 (2000).

[2] Brin, S.: Extracting patterns and relations from world wide web, *WebDB Workshop at 6th International Conference on Extending Database Technology*, pp. 172–183 (1998).

[3] Collins, M. and Duffy, N.: Convolution Kernels for Natural Language, *Proceedings of the Neural Information Processing Systems (NIPS)*, pp. 625–632 (2001).

[4] Culotta, A. and Sorensen, J.: Dependency Tree Kernels for Relation Extraction, *Annual Meeting of Association of Computational Linguistics*, pp. 423–429 (2004).

[5] Grosz, B.A.K.Joshi and S.Weistein: Providing a unified account of definite nounphrases in discourse, *In Proceedings of the 21st Annual Meeting of the American Association for Computational Linguistics, Cambridge. MA:ACL*, pp. 44–50 (1983).

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

[7] Kambhatla, N.: Combining Lexical, Syntactic, and Semantic Features with Maximum Entropy Models for Information Extraction, *Annual Meeting of Association of Computational Linguistics*, pp. 178–181 (2004).

[8] Kameyama, M.: Intrasentential Centering: a case study, *In M. Walker, K. Joshiand E. Prince (eds.), Certering theory in discourse, Oxford: Clarendon Press*, pp. 89–112 (1998).

[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] 鈴木潤, 佐々木裕, 前田英作: 階層非循環有向グラフカーネル, 電子情報通信学会論文誌, Vol. 88, No. 2, pp. 230–240 (2005).

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