Web から収集した学習データを用いた人物と実体間の関係の推定

堂前 友貴* 関 洋平† 神門 典子‡
筑波大学 情報学群 知識情報・図書館学類*
筑波大学大学院 図書館情報メディア研究系† 国立情報学研究所‡
s1013143@u.tsukuba.ac.jp* yohei@slis.tsukuba.ac.jp† kando@nii.ac.jp‡

1 はじめに

本研究では，人名と実体名との間の関係を推定する手法を提案する。毎年多くの人が新たに知られるようになるため，人名は，地名などに比べると圧倒的に新語として出現しやすい。また，近年，マイクロブログや SNS が活発に利用されていることにより，インターネット上で，それまで知らなかった人との出会いが増えている。しかし，人と人とをつなぐ上で，その人物についての情報がなければ，適切に人を探すことは困難である。これらの背景を踏まえ，本研究では，人名と，人名と一緒に出現しやすい実体名との間でよく現れる関係を推定することにより，特定の人名についての情報の整理を支援することに取り組む。

人名と周囲に現れる実体名との間に成立する関係は，社会においてよく利用される情報という観点からあらかじめ指定しておくことができる。あらかじめ指定した関係のタイプの推定には，主に学習データを使用した機械学習による手法が用いられる [9]。しかし，人手で大規模なコーパスを作成することはコストが高いため，最近では，初期データから学習データを自動的に拡張する研究が活発に行われている [2][4]。具体的には，ある関係が成立する単語対である関係例 (relation instances[2]) を収集し，その単語対を含む文を更に収集している。この方法は,「単語対がある関係を持っているならば，その単語対を含むすべての文は同じ関係を持っている」という仮定に基づく。

関係例を初期データとし，それを拡張していく方法では，ある特定の関係が成立している多くの単語対の集合を収集することが，有用なラベル付きコーパスの収集につながる。本研究では，高精度かつ汎用性の高い関係例の収集を目指し，拡張固有表現 [3] を利用した手法を提案する。また，その関係例を含む文を，幅広い分野の情報を収集しやすい Web から収集することにより，学習データの拡張を行う。

提案手法の有効性を検証するために，関係推定の精度を評価する。まず，自動収集した学習データから，人手で関係が成立していないと判定されるデータを除き，人手判定に基づく学習データを作成する。次に，人手判定に基づく学習データの割合を確認する。最後に，自動収集および人手判定に基づく学習データによる関係推定の精度を比較し，有意差がないことを確認することにより，本提案である学習データの自動収集にもとづく関係推定の有効性を示す。

2 関連研究

2.1 関係例を用いた関係推定

Mintz ら [2] は，関係とその関係例が記述された大規模なデータベースを初期データとして関係例を収集し，その関係例を含む文を Wikipedia から収集している。また Stijn ら [4] は，低頻度構文パターンからの単語対の収集という点で目的は異なるが，パターンを定義し Web 上から関係例を収集し，それを含む文を用いてさらに Web 上から関係例を収集している。本研究ではパターンから関係例を収集する際，拡張固有表現を利用し，文の収集を Web からおこなう。

2.2 拡張固有表現

固有表現とは，人名，地名などの“ 実世界において何らかの参照物が想定できる言語表現 ”[7] であり，情報抽出において重要になりやすく，単位がはっきりしているものを指す。そのため，固有表現間の関係を推定する研究 [7][11] は，数多く行われている。従来の固有表現のタイプは 10 種類程度であったが，質問応答等に応用されることにより，必要とされるカテゴリが飛躍的に増加している。関根ら [3] は，固有表現のタイプに 3 階層の階層関係を設定し，第三階層でタイプの種類を 200 種類とした拡張固有表現を提案した。

Copyright(C) 2012 The Association for Natural Language Processing.
All Rights Reserved

本研究では，実体のタイプを汎用的かつ詳細に特定できる枠組みとして，拡張固有表現を採用する。

3 関係タイプの定義

人名と周囲に現れる実体名との間に成立する関係は，社会においてよく利用される情報という観点からあらかじめ指定しておくことができる。このような関係を持つ情報は，新聞記事等において，人物の描写に用いられることにより，コーパス中でも人物名と頻繁に共起すると考えられる。本研究では，人物名とその近くに頻出する実体名との組み合わせに対し，分析を行い，関係が成立するものの中から上記の観点をもとに関係のタイプを定義した。また，その際の関係の分析には，拡張固有表現タグ付きコーパス [10] を使用した。

表 1 に，本研究で推定する関係のタイプ 10 種類を示す。学習データの収集や，関係推定器の構築では，この 10 種類の関係を用いる。

表 1: 推定する関係のタイプ (10 種類)

関係名
人 - 地位職業，人 - 所属会社，人 - 罪状
人 - 競技種目，人 - 国籍
人 - 出身都道府県，人 - 出身市区町村
人 - 著作，人 - 受賞 ，人 - 監督映画

4 学習データの自動収集

4.1 関係推定器の概要

本研究では，以下の手順により，単語対を入力とし，その間の関係を出力する。

1. Web 検索で単語対を含む文を収集
2. 分類器を用いて 1 文ずつ関係タイプを推定
3. 分類器の出力の多数決で出力する関係を決定

4.2 関係例を含む文の収集

分類器を構築に使用する学習データの収集について述べる。

まず，拡張固有表現タグ付きコーパス [10] 上の新聞コーパスで，関係例を収集するためのパターンの定義を行った。パターンの定義は，関係を定義した拡張固有表現の組み合わせの文を分析し，人手で 40 程度定義した。またこの際，拡張固有表現を制約として利用した。その結果，1,428 対の関係例が収集できた。

次に，関係例に現れる単語対をクエリとして Web から検索することにより，収集した Web ページを用いて，関係例に現れる単語対両方を含む文を収集し，その文に対して関係ラベルを付与することにより学習データを拡張する。Web 検索には Google[1] を利用し，Web ページは検索結果上位 100 件のものを収集した。そのページから関係例を含む文を抽出し，関係例である各単語（人名と実体名）に対してラベルの付与を行った。

例として,「$<person>$ 井伏鱒二 $</person>$ さんは“ $<work>$ 川釣り $</work>$ ”という作品で，次のように記してる。」のようなデータが得られる。

なお，収集の際，関係例数の違いを考慮し，関係例の数は最も数が少ない「受賞」の 45 に合わせて収集を行った。収集の結果，12,398 件の文[1]の学習データが収集できた。「競技種目」や「著作」,「監督映画」は，記録や紹介ページなどが多いためか多くの文を収集できる傾向にあったが,「罪状」や「出身市区町村」は文が他の関係に比べ収集できる数が少なかった。これは，軽犯罪についてはニュース以外のページで同一文中に関係例が出現しないこと，出身についての記述は，人名と同一文中では都道府県までの表記が多いことが理由としてあげられる。

4.3 負例の収集

定義した関係が成立しない場合の学習データ（負例）を収集する。負例となる文にラベル付けされる単語対は，人名と実体名の組み合わせであるが定義された関係を持たないものである。

学習データの正例として，入力とした人名と実体名との間には関係が成立していると仮定できる。また，文中に出現するそれ以外の人名は，関係例である実体名と関係を持たない可能性がある。

例えば「太宰治は遺書の中で“ $<work>$ 黒い雨 $</work>$ ”などの著者である $<person>$ 井伏鱒二 $</person>$ さんを『悪い人です』と書き残している。」という学習データは，人-著作という関係について,「井伏鱒二」と「黒い雨」にラベルがついている。ここで出現する他の人名「太宰治」は，ラベル付された実体名「黒い雨」と人-著作という関係にはない。

そのため負例は，関係例を入力とし，拡張したラベル付きデータから収集を行う。ただし，同一文中に関係例である実体名が複数あると，他の人名とも関係が成立しやすい。また，関係例である人名と並列関係にある場合は，同一の実体名に対し同じ関係が成立する場合が多い。上記のことなどを考慮し，負例は，関係例を入力とし拡張したラベル付きデータのうち，以下

[1] 10 の関係について，各 45 件，合計 450 の関係例をクエリとして収集した結果，1 関係辺り最大 2,924 件（競技種目），最小 272 件（出身市区町村) の文を収集した。

Copyright(C) 2012 The Association for Natural Language Processing.
All Rights Reserved

の条件などを満たすものを選択し，新たにラベルを付与した。

- ラベル付けされた人名以外を含む。
- ラベル付けされた人名と，ラベル付けされていない人名は並列関係にない。
- ラベル付けされた実体名は複数回出現しない。

その結果，収集できた負例は 1,910 件であった。

5 評価

5.1 自動収集データの収集精度の評価

不適切なデータを人手で削除したデータを作成し，自動収集したデータとの比較を行った。

関係例としては，適切な関係が成立していても，関係例を含む文においては，単語同士について，別の関係を表している場合がある。そのため，以下の条件のいずれかを満たすデータを削除した。同一文中で複数の関係が成立している場合は，付与されている関係が成立していれば適切であるとした。

- 他の関係を表しているデータ
- 実体名が別の人名との間にのみ関係が成立している
- 負例として収集したデータのうち，関係が成立するデータ

人手判定の結果，全体の一致率は 92% であった。「著作」,「受賞」,「監督映画」については，ほとんどのデータが人手判定と一致していた。一致率が最も低かったのは，関係例を使用しなかった「関係が成立しない」であり 74% ，関係例を使用したものでは,「出身都道府県」の 76% であった。「出身都道府県」は，他の関係を表しているデータとして，出身都道府県での講演や訪問などの文が多く含まれていた。

5.2 分類器の評価

5.2.1 実験環境とテストデータ

入力された文を，定義した関係と，定義した関係が成立しない場合との 11 のクラスに分類する分類器を構築した。

分類器には SVM を使用し，カーネルは多項式カーネル，学習方式には SMO(Sequential Minimal Optimisation) を採用した。なお，実装には Weka[5] のライブラリを使用した。

素性は，実体名の品詞，単語対のそれぞれの前後に現れる形態素とその品詞，人名と実体名のどちらが先にくるかの 10 要素 (3,072 次元) である。この際，形態素解析には MeCab[8] を使用した。

学習では，自動収集データのうち 11,798 件の文と，人手判定データのうち 9,671 件の文それぞれについて行った。人手判定データを 10 分割し，1 つをテストデータとし，人手判定データについては残りを学習データ（10 分割の交差検定），自動収集についてはテストデータに用いられた文を取り除いたデータを学習データとして，精度，再現率，F 値を 10 回評価し，その平均値を算出した

5.2.2 結果と考察

自動収集データを用いた分類器の性能は，マクロ平均で，精度 0.86，再現率 0.83，F 値 0.84 であった。一方で，人手判定データを用いた分類器の性能は，精度 0.84，再現率 0.84，F 値 0.84 であった。各関係の分類性能の評価を表 2 に示す。また，自動収集データと人手判定データの F 値に対して，t 検定（両側検定，有意水準 5%）で行ったところ，有意差がないことを確認し，自動収集したデータにより，人手で判定したデータと同程度の精度で関係推定が行えることを明らかにした。

「国籍」「出身都道府県」「出身市区町村」は，いずれも F 値が高くなった。これは，地名の場合，ほぼ辞書に登録されている語であるために，実体名の品詞が正しく与えられることが理由の一つである。また，実体名である都道府県や市区町村の後には “県” など特定の接尾辞が現れやすいことが理由と考えられる。

また，誤り分析の結果，いくつかの傾向がみられた。「著作」と「監督映画」は，お互いに推定を誤りやすいことがわかった。これは，お互いに実体名が作品名であり，実体名の品詞が他の関係に比べて特定のものがでないことや，関係性が似ていることから，“作”など似た形態素が出現しやすい傾向にあることによる。他の関係では,「定義された関係が成立しない」は，負例を作成した際に利用した元データに含まれる関係名に分類を誤ることが多かった。

表 2: 分類器の性能評価

関係名	精度	再現率	F 値	
	自動収集			人手判定
著作	0.84	0.86	0.85	0.84
監督映画	0.85	0.78	0.82	0.81
受賞	0.94	0.76	0.84	0.85
所属会社	0.91	0.87	0.89	0.91
地位職業	0.78	0.72	0.75	0.75
競技種目	0.83	0.90	0.87	0.86
出身都道府県	0.87	0.87	0.87	0.86
出身市区町村	0.88	0.91	0.90	0.86
国籍	**0.98**	**0.99**	**0.99**	**0.99**
罪状	0.94	0.78	0.86	0.85
定義されている関係が成立しない	0.69	0.69	0.69	0.70
マクロ平均	**0.86**	**0.83**	**0.84**	**0.84**

Copyright(C) 2012 The Association for Natural Language Processing.
All Rights Reserved

5.3 関係推定器の評価

5.3.1 実験環境とテストデータ

関係推定器の評価には，前節で構築した分類器を使用し，学習データには，自動収集データのうち 11,798 件の文を使用した。

評価に用いた単語対は，各関係 50 組と，定義している関係が成立しない 50 組[2]との合計 550 組である。単語対は， Wikipedia[6] から収集し，一部の関係についてはニュース記事からも収集した。

5.3.2 結果と考察

評価結果は，マクロ平均で，精度 0.71，再現率 0.76，F 値 0.73 であり，評価結果を表 3 に示す。F 値が最も高かったのは「受賞」の 0.83 であり，最も低いのは「定義されている関係が成立しない」の 0.48 であった。

評価値が高かった,「受賞」,「罪状」,「競技種目」は，どの単語対でも多くの場合，特徴的な表現が出現する文が多く，多数決を取ることにより正しい関係に推定されることが多かった。

全体として，分類器よりも低い評価結果となった。これは，学習データの収集に用いた関係例と，評価に用いた単語対とのジャンルの不整合が原因の一つにあげられる。例えば「国籍」の場合，新聞記事をもとにしたコーパスから関係例の収集をしたため，国際試合の結果などで記述の多いスポーツ選手の国籍が多くなったが，評価に用いた単語対は芸能人や政治家などの国籍なども用いたために，収集された文において周囲の形態素に異なる傾向がみられた。

分類器と同様の傾向としては,「著作」と「監督映画」は，お互いに推定を誤りやすい。最も評価が低くなった「定義されている関係が成立しない」場合は，実体名が共通である関係に，推定を誤る傾向が見られた。たとえば，人名と，その両親の出身地では,「出身都道府県」に推定された。現在の素性は，係り受け関係など構文的な情報を使用していないこともあり，類似した表層的特徴が現れる関係名に推定を誤る傾向がある。この問題に対する解決は，今後の課題とする。

表 3: 関係推定器の評価結果

関係名	精度	再現率	F 値
著作	0.81	0.76	0.78
所属会社	0.65	**0.85**	0.74
地位職業	0.57	0.78	0.66
出身都道府県	0.63	0.85	0.72
出身市区町村	0.74	0.68	0.70
国籍	**0.87**	0.58	0.69
監督映画	0.64	0.87	0.74
罪状	0.75	0.84	0.79
受賞	0.82	**0.85**	**0.83**
競技種目	0.88	0.73	0.79
定義されている関係が成立しない	0.45	0.53	0.48
マクロ平均	**0.71**	**0.76**	**0.73**

また，構築した関係推定器に対し，各関係 50 組，関係が成立しない 50 組の計 550 組の単語対で評価を行ったところ，精度 0.71，再現率 0.76，F 値 0.73 となった。今後の課題としては，素性の改良による分類器の精度向上や，推定する関係数の拡張などがあげられる。

6 まとめ

本研究では，特定の人名についての情報の整理を支援することを目的とし，人名と，人名と一緒に出現しやすい実体名との間の関係の推定に取り組んだ。

学習データの収集では，拡張固有表現を手がかりの制約として関係例を収集した後，Web を利用し，関係例を含む文の収集を行った。人手で判定を行ったところ，割合として 92%の文が適切なものであった。

[2]人名と，10 の関係が対象とする実体名との単語対で，なんらかの関係が成立しているが，定義した 10 の関係は成立しないもの。

参考文献

[1] Google. http://www.google.co.jp/. (参照：2011-11-30).

[2] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant Supervision for Relation Extraction without Labeled Data. In *Proceedings of the 47nd Annual Meeting of The Association for Computational Linguistics*, pp. 1003–1011, 2009.

[3] Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata. Extended Named Entity Hierarchy. *In Proceedings of the 3rd International Conference on Language Resources and Evaluation*, pp. 1818–1824, 2002.

[4] De Saeger Stijn, 鳥澤健太郎, 土田正明, 風間淳一, 橋本力, 山田一郎, Varga, Istvan, 顔玉蘭, 呉鍾勳. 低頻度構文パターンからの意味的関係獲得. 言語処理学会第 17 回発表論文集, pp. 9–12, 2011.

[5] Weka. http://www.cs.waikato.ac.nz/ml/weka/. (accessed 2011-7-28).

[6] Wikipedia. http://ja.wikipedia.org/wiki/. (参照：2011-12-3).

[7] 菊井玄一郎, 松尾義博, 小林のぞみ, 平野徹, 浅野久子. リッチアノテーション：固有表現に焦点をあてた知識抽出の試み. 電子情報通信学会技術研究報告. NLC, 言語理解とコミュニケーション, Vol. 108(141), pp. 73–78, 2008.

[8] 工藤拓. MeCab: Yet Another Part-of-Speech and Morphological Analyzer. http://mecab.sourceforge.net/. (参照：2011-5-27).

[9] 石崎俊 (編). 3.4 節 情報抽出. デジタル言語処理学事典, pp. 348–361. 共立出版, 2009.

[10] 橋本泰一, 乾孝司, 村上浩司. 拡張固有表現タグ付きコーパスの構築. 情報処理学会研究報告. 自然言語処理研究会報告, pp. 113–120, 2008.

[11] 平野徹, 松尾義博, 菊井玄一郎. 関係名詞らしさを用いた固有表現間の関係同定. 言語処理学会 第 15 回年次大会, pp. 921–924, 2010.

Copyright(C) 2012 The Association for Natural Language Processing.
All Rights Reserved