言語的手がかりを用いた固有表現の二項関係知識の分類

高久　陽平†　　鍛治　伸裕††　　吉永　直樹††　　豊田　正史††
† 東京大学大学院 情報理工学系研究科　†† 東京大学 生産技術研究所
{takaku, kaji, ynaga, toyoda}@tkl.iis.u-tokyo.ac.jp

1　はじめに

近年, ウェブなどを知識源として世界知識を獲得する研究 [3, 1, 2] が盛んに行われている. 特にテキストから二つの固有表現とその間の関係を抽出する関係抽出は, 質問応答システム [6] といったアプリケーションを支える重要なタスクの一つとして研究されている.

関係抽出では, 例えば, 「本田圭佑が CSKA モスクワに所属している」という知識を, (1a) のような三つ組として表現し収集する. しかし, 実際に獲得される二項関係には (1a) に加え, (1b) のように過去に成立していたが, 現在は成立していないものが獲得されてしまうことがある.

この問題に対する自然な解決策として, ある固有表現 (本田圭佑) に対して複数の関係知識が獲得された場合には, 最も頻度が高い関係を採用するという方法が考えられる. しかしながら, (1c) と (1d) のように一つの固有表現 (東京) に対して複数の関係が同時に成立することもあるため, そのような単純な方法による解決は難しいと考えられる.

(本田圭佑, X が所属する Y, $CSKA$ モスクワ)　(1a)
(本田圭佑, X が所属する Y, VVV フェンロ)　(1b)
(東京, X を流れる Y, 荒川)　(1c)
(東京, X を流れる Y, 多摩川)　(1d)

そこで本研究では, 関係には「時間において恒久的に成立するか否か」, また「一対一で成立するか否か」という二つの特質 (以降, 前者を恒久性, 後者を一意性と呼ぶ) があることを指摘し、その分類手法を提案することでこの問題の解決を図る.

提案手法では, 時系列ウェブテキストを用いた関係における固有表現の分布と、モダリティや時間表現といった言語手がかりを基に機械学習を用いて関係を分類する手法を考案した. 実験では, 5 年分のブログ記事と新聞記事から獲得された関係を分類し、提案手法の有効性を確認した.

2　関連研究

本研究のように二項関係を精緻化する研究として, Lin ら [4] は一意性に着目し, 二項関係おいて Y に当てはまる固有表現の分布に関する類似度を用いて分類を行った. しかし, 彼らの研究では恒久性を暗黙に仮定している. そのため, 二項関係 (1a), (1b) のように, 時間的に変化するために複数の関係が獲得されたとき, あたかも Y に複数の値が存在するように見えてしまい, 正しく判別することは難しい.

また, Ling ら [5] は情報抽出において, あるイベントが発生している期間を文章中の時間表現 (例えば since 1997) に着目して抽出する手法を提案している. これに対して, 本研究では時間情報そのものを抽出することを目的としているのではなく, 関係知識が時間によって変化しうるかどうかという性質を議論している.

3　恒久性と一意性に基づく関係分類

本章では, 本研究が提案する関係の恒久性と一意性について説明する.

二項関係において, 前項 X に固有表現が代入されたとき, 後項 Y に当てはまる固有表現が時間変化しないとき, その二項関係は恒久性を有するという. また、このときこの二項関係を恒久的関係と呼ぶ. 例えば, 関係 “X が所属する Y” は, (1a), (1b) のように X の所属先が時間変化するので, 恒久的関係ではない. 一方, 関係 “X を流れる Y” は, (1c), (1d) のようにその地域を流れる川は基本的には変化しないので, 恒久的関係である.

次に, 前項 X に固有表現が代入されたとき, 後項 Y に当てはまる固有表現が任意のある時点において常に唯一に決まるとき, 一意性を有するという. また、このときこの関係を一意的関係と呼ぶ. 例えば, 関係 “X が所属する Y” は, (1a) のようにある時点においては所属先は一意に決まるので, 一意的関係である. 一方で,

関係 “X を流れる Y” は, (1c), (1d) のようにその地域を流れる川は複数存在するので, 一意的関係でない.

なお, 関係の恒久性及び一意性という概念は, 後項 Y に固有表現を代入した場合にも論ずることできるが, 本研究では前項 X に固有表現を代入した場合のみを扱うことにする.

本研究では, 入力として関係 (例えば “X が所属する Y”, “X を流れる Y”) が与えられたとき, その関係の恒久性と一意性の有無をそれぞれ個別に判定するという二つの問題を扱う. ただし, 入力となる関係は既存の関係抽出手法 [1, 2] を用いて与えられることを想定し, 関係抽出手法自体については本研究の議論の対象外とする.

4 提案手法

恒久性と一意性のそれぞれに対する分類には教師あり学習を用いる. この際, 学習に用いる特徴量の選択は大変重要である. 本研究では, この特徴量の抽出に時系列ウェブテキストを用いる.

本研究では時系列ウェブテキストとして, 時間情報がメタ情報として付与されたブログ記事集合を用いる. ブログ記事は一月単位で集積して, これを時間の最小処理単位として用いる. このデータの詳細については, 5.1 節で詳しく述べる.

以下, 本章では各関係分類タスクごとに, 実際に時系列ウェブテキストの時間情報と言語的手がかりを用いた特徴量について述べる.

4.1 恒久性の分類

4.1.1 時系列情報基づく特徴量

図 1(a) と図 2(a) は実際にウェブテキストから獲得された二項関係において, それぞれ恒久的関係でない例として “X が所属する Y” の X に “本田圭佑”, 恒久的関係の例として “X を流れる Y” の X に “東京” を代入したときの Y の時系列分布を示している. ここで, 時系列データ上である一定期間の部分のみを切り出した時間移動窓 (Time Sliding Window) を設けたとき, 2009 年 2〜7 月における時間移動窓 (t) の Y の分布を示したものが図 1(b) と図 2(b) であり, 2010 年 2〜7 月における窓 ($t^{'}$) の分布を表したものが図 1(c) と図 2(c) である.

ここから, 図 1(b) と図 1(c) で Y の分布が変化していることが見て取れる. また一方で, 同様に図 2(b) と図 2(c) の変化と見ると, 先程より分布の変化が小さいことが分かる. このように, 恒久的でない関係では, ある 2 つの窓 t, $t^{'}$ において Y の分布が変化することが期待される.

この違いに注目し, 時間移動窓ペアの Y の分布をコサイン類似度で計算し, 最大値 (式 (2)), 最小値 (式 (3)), 及び平均値 (式 (4)) を特徴量として用いた.

$$\frac{1}{n}\sum_{x\in\mathcal{X}_n}\max_{t_x,t'_x\in\mathcal{T}_x,t\neq t'} cos(t_x,t'_x) \tag{2}$$

$$\frac{1}{n}\sum_{x\in\mathcal{X}_n}\min_{t_x,t'_x\in\mathcal{T}_x,t\neq t'} cos(t_x,t'_x) \tag{3}$$

$$\frac{1}{\sum_{x\in\mathcal{X}_n}\binom{|\mathcal{T}_x|}{2}}\sum_{x\in\mathcal{X}_n}\sum_{t_x,t'_x\in\mathcal{T}_x,t\neq t'} cos(t_x,t'_x) \tag{4}$$

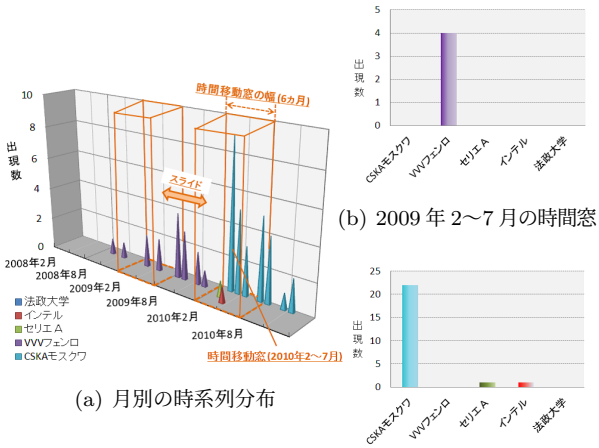
ここで $\mathcal{X}_n$ は X の出現数上位 n 位の固有表現の集合, 固有表現 $x\in\mathcal{X}_n$ を代入したときの時系列分布において少なくとも 1 つの Y が出現している時間移動窓の集合である. なお, 実験では n=5 とした.

最後に, どのような区間で時間移動窓を設定するかについて議論する. 時間移動窓 t, $t^{'}$ の区間は, これを狭くし過ぎるとデータスパースネスの問題が発生し, 広くし過ぎると時間情報の細かな違いが区別できなくなる. そのため, 提案手法ではこの幅を, 1, 3, 6, 12ヵ月の 4 段階設定し, それぞれに対して個別の特徴量を抽出する. これは, 4.2.1 節でも同様である.

4.1.2 言語的手がかりに基づく特徴量

次に, 表 1 に示す言語的手がかりを特徴量として使う方法について説明する. 例えば、所属関係の場合, 「昨年, 本田圭佑が所属していたVVV フェンロは〜」というテキストは, 所属関係が時間によって変化しうることを示唆していると言える.

表の一行目は時間表現に基づく特徴であり, 表中の手がかり表現が関係と同一文中に閾値 θ 回以上出現したとき 1 を, それ以外のときは 0 の特徴量をとる. また, 二行目は関係が時間変化することを示唆する接頭辞に基づく特徴であり, 関係中の名詞の直前にこれらの接頭辞が閾値 θ 回以上出現したとき 1 を, それ以外のときは 0 の特徴量をとる. 最後に, 三, 四行目は「テンス (過去)」と「モダリティ (継続)」を表す助動詞であり, 関係中の動詞の直後にこれらの助動詞が閾値 θ 回以上出現したとき 1 を, それ以外のときは 0 の特徴量をとる.

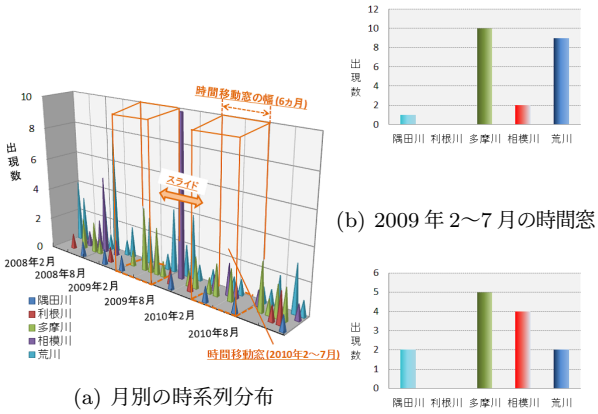


(a) 月別の時系列分布

(b) 2009 年 2〜7 月の時間窓

(c) 2010 年 2〜7 月の時間窓

図 1: “X が所属する Y (X=本田圭佑)” の時系列分布



(a) 月別の時系列分布

(b) 2009 年 2〜7 月の時間窓

(c) 2010 年 2〜7 月の時間窓

図 2: “X を流れる Y (X=東京)” の時系列分布

特徴	手がかり表現
時間表現	#年, #月 (#は数字), 来年, 昨年, 現在 etc.
時間接頭辞	前〜, 現〜, 次期〜, 元〜, 新〜 etc.
テンス (過去)	〜た
モダリティ (継続)	〜ている, 〜てる

表 1: 恒久性の分類に用いる言語的手がかり

特徴	パターン
並立助詞	と, とか, や, やら, だの, たり, だの, なり etc.
複数助詞	など, 等, ら, たち, 達

表 2: 一意性の分類に用いた言語的手がかり

4.2 一意性の分類

4.2.1 時系列情報基づく特徴量

一意性の関係では, 時間移動窓の各時点で Y が唯一出現することに着目し, 特徴量を二種類設定した.

出現数上位一, 二位間の出現数比率 一意的関係では, X に固有表現を代入したとき, Y がとる固有表現は唯一一つになることが期待されるため, 出現頻度一位の固有表現 y_1 を除きその他はノイズであると考えられる. そこで, 出現頻度一位の固有表現 y_1 の出現数 $c(y_1)$ に対し, 出現頻度二位の y_2 の比率を特徴とする. 例えば, 一意的関係である “X が所属する Y” の図 1(b) では y_1=“VVV フェンロ”, y_2=“CSKA モスクワ” であり $c(y_2)/c(y_1) = 0$ である. 同様に, 一意的関係でない “X を流れる Y” の図 2(b) では y_1=“多摩川” の出現数 $c(y_1) = 10$, y_1=“荒川” の出現数 $c(y_2) = 9$ から $c(y_2)/c(y_1) = 0.9$ となり図 1(b) の割合の方が小さい. そこで, 全時間移動窓間におけるこれらの割合の最大値 (式 (5)), 最小値 (式 (6)), 及び平均値 (式 (7)) を用いた.

$$\frac{1}{n}\sum_{x\in\mathcal{X}_n}\max_{t_x\in\mathcal{T}_x}\frac{c_{t_x}(y_2)}{c_{t_x}(y_1)} \tag{5}$$

$$\frac{1}{n}\sum_{x\in\mathcal{X}_n}\min_{t_x\in\mathcal{T}_x}\frac{c_{t_x}(y_2)}{c_{t_x}(y_1)} \tag{6}$$

$$\frac{1}{\sum_{x\in\mathcal{X}_n}|\mathcal{T}_x|}\sum_{x\in\mathcal{X}_n}\sum_{t_x\in\mathcal{T}_x}\frac{c_{t_x}(y_2)}{c_{t_x}(y_1)} \tag{7}$$

種類数 一意性がある関係では出現する固有表現 Y の種類数が少ない. 例えば, 図 1(b) では一種類であるのに対し, 図 2(b) では四種類である. そこで, 全時間移動窓間における Y の種類数の最大値, 最小値, 及び平均を用いた. ぞれぞれの式は, 式 (5)) と式 (6) 及び式 (7)) の比率計算項を時間移動窓 t に出現する Y の種類数で置き換えたものとなる.

4.2.2 言語的手がかりに基づく特徴量

表 2 に示す一意性に関する言語的手がかりを特徴に用いた. 例えば, 「東京を流れる荒川 や 隅田川は有名だ。」といった文章があったとすると, この助詞 “や” は荒川の他にも東京を流れる川があることを意味している. このように, 表 2 の単語群は一意性を持たない関係に多く出現すると思われる.

分類対象となる関係の後項 Y の直後に「並立助詞」が θ' 回以上使われたか否かを特徴量として用いる．また,「複数助詞」についても同様の特徴量を用いる．

5 評価実験

5.1 実験設定

提案手法の入力となる二項関係の抽出に関しては様々な手法 [1, 2] が提案されておりどの手法を用いても支障はないが, 実験では簡略に文章中から 2 つの固有表現 (固有名詞) とその間の最短係り受けパス上の単語を抽出してきて, この中から人手によって二項関係を獲得した．なお, 形態素解析に MeCab[1], 係り受け解析に J.DepP[2]を用いた．また, 語彙パターンについては受身、使役及び否定の情報は残し, その他のモダリティなどについては削除した．

我々が蓄積した時系列ウェブテキスト (2006 年 2 月〜2011 年 9 月, 約 23 億文) 及び, 毎日新聞記事データ[3](2006 年〜2010 年版) から上記で述べた手法によって獲得した二項関係に対して, 獲得数上位から 1,000 個を人手によるラベル付けを行った．なお, 機械学習には線形 SVM による LIBLINER[4]を用いた, 5 分割交差検定を行った．

5.2 実験結果

評価実験の結果は, 表 3 のようになった．なお, 両者ともベースラインを最多クラスへ分類した場合とした．

実験結果より, 時系列情報に基づく特徴によりベースラインに比べ分類精度が向上した．また, 言語的手がかりを用いることでさらに分類精度が向上し、提案手法の有用性が示された．

ウェブ上の記事では話題性のある一部の固有表現間の関係しか獲得されないという問題があり, 時系列情報だけでは十分に分類することができなかった．しかし, 言語的手がかりを用いることでその問題が補われ, 精度が向上したと考えられる．

分類	ベースライン	時系列情報	言語的手がかり	全特徴
恒久性	57.2%	61.1%	67.0%	69.3%
一意性	57.6%	68.7%	58.1%	70.0%

表 3: 実験結果 (精度)

6 おわりに

本稿では, 二項関係を恒久性と一意性の二つの観点から分類することを提案した．さらに, ウェブテキストの時系列情報と言語手がかりに基づく分類手法を提案し, 評価実験により提案手法の有効性を示した．

今後の課題としては, Bollegala ら [2] のように, 同じ意味を表す関係や固有表現をクラスタリングすることを考えている．これにより, 1 関係あたりのサンプル数を増やし, かつ固有表現の表記ゆれを吸収することができるのでさらなる精度の向上が見込まれる．

[1] http://mecab.sourceforge.net/

[2] http://www.tkl.iis.u-tokyo.ac.jp/~ynaga/jdepp/

[3] http://www.nichigai.co.jp/sales/mainichi/mainichi-data.html

[4] http://www.csie.ntu.edu.tw/~cjlin/liblinear/

参考文献

[1] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni. Open information extraction from the web. In Proceedings of IJCAI, pp. 2670–2676, 2007.

[2] D. Bollegala, Y. Matsuo, and M. Ishizuka. Relational duality: Unsupervised extraction of semantic relations between entities on the web. In Proceedings of WWW, pp. 151–160, 2010.

[3] M. A. Hearst. Automatic acquisition of hyponyms from large text corpora. In Proceedings of Coling, pp. 539–545, 1992.

[4] T. Lin, Mausam, and O. Etzioni. Identifying functional relation in web text. In Proceedings of EMNLP, pp. 1266–1276, 2010.

[5] X. Ling and D. S. Weld. Temporal information extraction. In Proceedings of AAAI, 2010.

[6] D. Ravichandran and E. Hovy. Learning surface text patterns for a question ansering system. In Proceedings of ACL, pp. 41–47, 2002.