6S-7

Web を利用したエンティティペア集合の代表的関係名抽出
Accurate relation name extraction from entity pair set using the Web

河崎 翔 † グェン トアン ドゥク ‡ ボレガラ ダヌシカ ‡ 石塚 満 ‡
† 東京大学工学部電子情報工学科 ‡ 東京大学大学院情報理工学系研究科

1 まえがき

21 世紀は「情報爆発」の時代である. 日々増殖する膨大な情報の中から価値あるものを取得する情報抽出は将来不可欠な技術であり, 今後重要な課題である. 中でも, 2 つの単語間に成立する関係情報を抽出する研究は多くなされている. しかし, 単語間に成立する多くの関係の中で最も「代表的」なものは何かという研究は未だ少なく, 適切な関係を抽出できない.

そこで本研究では, ある関係が成り立つエンティティペアの集合から, 「代表的」な関係名を既存手法より高精度で抽出する方法を提案する. なお, 本研究におけるエンティティとは 1 単語以上から成る固有名詞である. また, 人間が最も連想しやすいであろう関係名を「代表的」であると仮定する. 例えば図 1 のように, 3 つのエンティティペアから成る集合 E = {(Tokyo, Japan), (Paris, France), (Berlin, Germany)} の代表的な関係名は “capital” であろう. そこで E の各エンティティペアの第 1 エンティティ, 第 2 エンティティをそれぞれ変数 X, Y で表す時, “X is the *capital* of Y” のような語彙パターンの抽出を目的とする. このように代表的な語彙パターンには, 代表的な関係名が含まれている場合が多い. また, 関係類似度計算 [1] や関係検索の研究 [2, 3] では, このような語彙パターンをベースとしており, 本研究は直接上記の研究に応用できると考えられる.

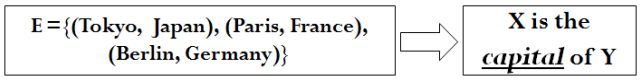

図 1: エンティティペア集合から関係名抽出の例

2 関連研究

テキストデータからエンティティペアを抽出するシステム TextRunner[4] は, 単語間の関係を述語として抽出する. 例えば, “Franz Kafka was risen up in Prague.”, “Franz Kafka was born in Prague.” という文から, エンティティペア (Franz Kafka, Prague) と述語を抜き出し, risen_up(Franz Kafka, Prague) や was_born(Franz Kafka, Prague) を出力する. このように同一のエンティティペアに対し複数の関係名を抽出することができる. しかしどの関係が最も代表的かという問題は解決していない.

Seed となる 1 組以上のエンティティペアと同一の関係を持つエンティティペアを生成する研究 [5] では, 関係をエンティティペア間に成立する語彙パターンとして表現する. しかし, 抽出された語彙パターンは複数あり, どの語彙パターンが最も有効かという問題は解決していない. 本研究も関係を語彙パターンで表し, この問題を解決する.

エンティティペア集合の代表語彙パターンを抽出する先行研究 [6] では, 語彙パターンをクラスタリングし, 似ているクラスタを併合して, 最も適切な語彙パターン集合 cluster core を抽出する. しかし, cluster core は 2 つ以上の類似するクラスタに属す語彙パターンからなるので, “X and Y” のように関係情報を持たないが高頻度な語彙パターンも 「代表的」とされる可能性がある.

3 提案手法

本手法ではまず, コーパスからエンティティペアとその関係を特徴づける語彙パターンを抽出しエンティティペアと語彙パターンの共起頻度を計算する. 次に, 共起頻度の情報を使いエンティティペア集合の各エンティティの代表語彙パターンを χ^2 検定で見つける. 最後に, エンティティペア集合全体に渡って最適な代表語彙パターンを抽出する. 図 2 に提案手法の全体図を示す.

例えば, エンティティペア集合 E = {(Tokyo, Japan), (Paris, France), (Berlin, Germany) } が入力される場合を考える. 各エンティティペア間に成立する語彙パターンを抽出し, Local Scoring によりランキングする. 次に, Global Scoring により E の代表語彙パターンとして “X is the capital of Y” を出力する.

次に, Local Scoring と Global Scoring の詳細を説明する.

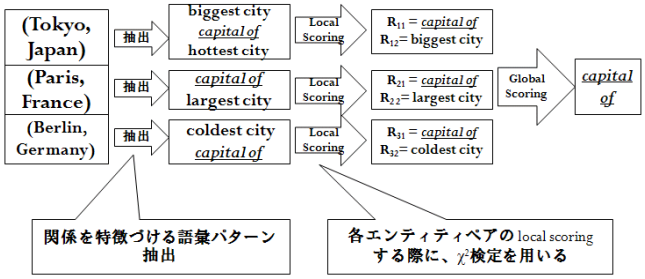

図 2: 代表語彙パターン抽出の過程

3.1 Local Scoring

Local Scoring の段階では, 各エンティティペアに成り立つ関係を表す語彙パターンをランキングする. ランキングの目的はエンティティペアに成り立つ関係を特徴付ける語彙パターンを高く評価することである. そのため, 語彙パターンのランキングを求めたいエンティティペアについて, そのエンティティペアと各語彙パターンがどの程度有意に共起するかを χ^2 検定により計測する.

χ^2 検定プロセスでは表 1 に示した値を計算する. ここで, $f(w,p)$ は語彙パターン p とエンティティペア w の共起する頻度であり, その他も同様である. また $\bar{p}$ は p 以外のコーパス内に出現する全ての語彙パターンであり, $\bar{w}$ は w 以外のコーパス内に出現する全てのエンティティペアである. コーパス内にエンティティペアが N 個, エンティティペア w_i に語彙パターンが m_i 個出現するとする. w と p が共起する場合, w と p がランダムに共起した場合の理論値 ex_a, ex_b を算出し, エンティティペア w における語彙パターン p のローカルスコア $LS(w,p)$ を χ^2 値として定義する.

表 1: χ^2 検定の表

	p	$\bar{p}$
w	$a = f(w,p)$	$b = \sum_{j=1}^{m} f(w,p_j) - f(w,p)$
$\bar{w}$	$c = \sum_{i=1}^{N} f(w_i,p) - f(w,p)$	$d = f(w,p) + \sum_{i=1}^{N}\sum_{j=1}^{m_i} f(w_i,p_j) - \sum_{i=1}^{N} f(w_i,p) - \sum_{j=1}^{m} f(w,p_j)$

$$ex_a = \frac{(a+b)(a+c)}{(a+b+c+d)}, ex_b = \frac{(a+b)(b+d)}{(a+b+c+d)}$$
$$LS(w,p) = \chi^2\text{値} = \frac{(a-ex_a)^2}{ex_a} + \frac{(b-ex_b)^2}{ex_b}$$

3.2 Global Scoring

Global Scoring の段階では, エンティティペア集合全体に渡って最も代表的な語彙パターンを抽出する. そこでローカルスコア

†Sho Kawasaki, School of Engineering, The University of Tokyo
‡Nguyen Tuan Duc, Danushka Bollegala, Mitsuru Ishizuka, Graduate School of Information Science and Technology, The University of Tokyo

$LS(w_i, p)$ を利用し, エンティティペア集合 $W = \{w_1, w_2, ..., w_n\}$ における語彙パターン p の Global Scoring を行う.

語彙パターン p とエンティティペア w_i が共起しない場合, w_i には p が表す関係は全く成立しない. よって, ローカルスコア $LS(w_i, p)$ は語彙パターン p に関する適切な情報を持たない可能性がある. またローカルスコア LS は, あくまで語彙パターンとエンティティペアが共起した場合にどの程度有意に共起しているかの目安であり, 共起しない場合の目安ではない. そこで, 語彙パターン p と共起するエンティティペアの集合 $W' = \{w'_1, w'_2, ..., w'_m\}$ に注目し, 式 (1) に従いグローバルスコア $GS(p)$ を算出する.

$$GS(p) = \prod_{i=1}^{m} LS(w'_i, p) \quad (1)$$

エンティティペア集合の代表語彙パターンは, エンティティペア集合の多くの要素エンティティペアと共起し, 各要素エンティティペアとのローカルスコア LS が高い可能性が高い. 故に, グローバルスコア $GS(p)$ が最も高い語彙パターンを代表語彙パターンとして抽出する.

4 評価

4.1 実験

Duc らが実装した Relational Search エンジン [2] にインデックスされている 12437 個の Web ページをコーパスとして利用した. またテスト用に用いた 4 種のエンティティペア集合を表 2 に示す. テストに用いた関係の種類は関係類似度計算の研究 [1], 関係抽出の研究 [4] や関係検索の研究 [2] でよく用いられる.

表 2: 評価実験に用いたテスト用エンティティペア集合

関係	エンティティペア集合
社長	CEO={(“Steve Jobs”,“Apple”)...} (計 16 組)
所在地	HQ={(“AOL”,“New York”)...}(計 15 組)
誕生地	BP={(“Michelangelo”,“Italy”)...}(計 20 組)
買収	AQ={(“Google”,“YouTube”)...}(計 48 組)

提案手法の有効性を示すために, Bollegala らの先行研究 [1] を比較手法とした. この手法ではまずエンティティペアを, 共起する語彙パターンの出現頻度を特徴とする特徴ベクトルで表す. エンティティペアをクラスタリングした後, それぞれのクラスターにユニークな ID を割り振り, Classias[7] を利用して, L1 正則化ロジスティック回帰で多クラス分類を行う. そこで, 与えられた学習データを利用して重み係数を決定する. 重み係数最大の語彙パターンが代表的である. なお, 評価実験する際はクラスタリング済みの入力を考えているため, Bollegala らの研究のような多クラス分類ではなく二値分類を行う. 例えば, 表 2 のエンティティペア集合 CEO の代表語彙パターンを求める場合, 正例, 負例それぞれに CEO, HQ+BP+AQ の情報を与える.

各手法で得られた代表語彙パターンは表 3 のようになった. なお各語彙パターン併記の数字は, 各エンティティペア集合内の出現頻度の順位である.

表 3: 提案手法と既存手法 [1] で得られた代表語彙パターン

関係	提案手法	既存手法 [1]
社長	X , the chief executive of Y(10)	X , Y(1)
所在地	X headquarters in Y(1)	X headquarters in Y(1)
誕生地	X was born in Y(1)	X was born in Y(1)
買収	X buys Y(1)	X and Y(3)

4.2 考察

提案手法では全ての関係に対し, 関係を特徴付ける語彙パターンを得た. 一方, 既存手法 [1] では, 所在地関係と誕生地関係に対しそれらの関係を特徴付ける語彙パターンを得たが, 社長関係と買収関係に対しては関係名 (ラベル) を含むような語彙パターンを得られなかった.

各手法で得られる上位の語彙パターンを比較すると違いが現れた. 特にエンティティペア集合 BP において違いが明確であった. 誕生地関係に対し各手法で得られる上位の語彙パターンを表 4 に示す. 1 位はともに “X was born in Y” を抽出しているが, 2 位 3 位が異なる. 提案手法では “X was born in Y” が派生した語彙パターンを得た. 一方, 既存手法 [1] では “X : Y” のような出現頻度は高いが関係情報を持たない語彙パターンと, “X was born in salzburg , Y” のような BP 内の 1 組のエンティティペアにのみ共起する特殊な語彙パターンを抽出している.

表 4: 誕生地関係に対し各手法で得られる上位の語彙パターン

順位	提案手法	既存手法 [1]
1 位	X was born in Y	X was born in Y
2 位	X was born in Y in	X : Y
3 位	X was born in Y ,	X was born in salzburg , Y

一般的に, 提案手法では表 3 に示す代表語彙パターンに非常に似た語彙パターンが上位を占めたが, 既存手法 [1] では “X and Y”, “X : Y” のような出現頻度は高いが関係情報を持たない語彙パターンや, “X was born in salzburg , Y”, “X , paris - Y” のような 1,2 組のエンティティペアにのみ共起する特殊な語彙パターンが上位に出現した.

提案手法では, エンティティペア集合内で満遍なくかつ有意に共起する語彙パターンが高く評価されるので, 結果的に望ましい代表語彙パターンを得られた. 提案手法はコーパス全体の情報を利用するのに対して, 既存手法は CEO+HQ+BP+AQ のエンティティペア集合の情報のみ利用する等の理由が挙げられる.

5 結論

本稿では, ある関係が成り立つエンティティペア集合から代表語彙パターンを抽出する手法を提案した. 提案手法ではまず, 集合に属するエンティティペアの関係を特徴づける語彙パターンを χ^2 検定により Local Scoring を行う. 次に, ローカルスコアの積を計算することで Global Scoring を行い, エンティティペア集合全体に渡って最も代表的な語彙パターンを抽出する. 提案手法は既存手法 [1] に比べ高精度な結果を達成した. 今後は, 関係抽出や関係検索システムにおいて代表語彙パターンを利用し, 抽出アルゴリズムや検索ランキングの精度の向上の実現を予定している.

参考文献

[1] D. Bollegala, Y. Matsuo, and M. Ishizuka. Relational Duality: Unsupervised Extraction of Semantic Relations between Entities on the Web. In *Proc. of WWW 2010*, pages 151–160, 2010.

[2] N. T. Duc, D. Bollegala, and M. Ishizuka. Using Relational Similarity between Word Pairs for Latent Relational Search on the Web. In *Proc. of WI 2010*, pages 196–199, 2010.

[3] T. Goto, N. T. Duc, D. Bollegala, and M. Ishizuka. Exploiting Symmetry in Relational Similarity for Ranking Relational Search Results. In *Proc. of PRICAI 2010*, pages 595–600, 2010.

[4] O. Etzioni, M. Banko, S. Soderland, and D. S. Weld. Open Information Extraction from the Web. *Commun. of the ACM 2008*, 51(12):68–74, December 2008.

[5] H. Li, Y. Matsuo, and M. Ishizuka. Relations Expansion: Extracting Relationship Instances from the Web. In *Proc. of WI 2010*, pages 184–187, 2010.

[6] D. Davidov, A. Rappoport, and M. Koppel. Fully Unsupervised Discovery of Concept-Specific Relationships by Web Mining. In *Proc. of ACL 2007*, pages 232–239, 2007.

[7] N. Okazaki. Classias: a Collection of Machine-learning Algorithms for Classification, 2009.