4G1-5

任意の係り受けパターンを対象とした事態間知識の獲得

Acquiring event relation knowledge based on arbitrary dependency patterns

東山 翔平 *1　大西 貴士 *1　渡邉 陽太郎 *1
Shohei Higashiyama　Takashi Onishi　Yotaro Watanabe

*1NEC 情報・ナレッジ研究所
NEC Knowledge Discovery Research Laboratories

Event relation knowledge is one of the important knowledge to realize deep language understanding and reasoning. There exist acquiring methods of event relations targeting to general knowledge from large scale documents. However, efficient methods are needed to extract domain-dependent knowledge from restricted amount of documents. In this report, we propose event relation acquiring methods based on arbitrary dependency patterns between event expressions, which can acquire knowledge broader than existing methods.

1. はじめに

深い言語理解や推論に必要な知識に，事態（イベントを表す表現）の間の関係を表す事態間知識がある．一般的な事態間知識を獲得する従来手法では，大規模なテキスト集合から，事態を含む文節（事態文節）の間に特定の係り受け関係のパターンが成立する事態のペアを抽出することで，知識を獲得している．しかし，一部の形式の係り受け関係を有する事態ペアしか抽出の対象にしていないため，取り逃している事態ペアが多く存在すると考えられる．一方，特定ドメインに依存する知識を獲得する場合には，知識獲得源として必ずしも大規模なテキストは期待できない．そのため，限られた量のテキストから効率的に知識を獲得することが必要となる．

事態間知識獲得の従来手法 [4, 5, 7] は，係り受け関係の木において，2 つの事態が親子関係や先祖-子孫関係（直列の係り受け関係と呼ぶ）にある事態ペアを抽出の対象にしている．しかし，これらの係り受け関係にない事態のペアであっても，前後・因果などの関係を有するケースが存在する．

図 1 は，非直列の係り受けパターンと共起する事態ペアの例である．文節間の係り受け関係を「係り元 → 係り先」の形式で表し，注目する事態を下線で示した．(a) では「(ご飯を) 食べる-(眠く) なる」，(b) では「(傘を) 忘れる-(雨に) 濡れる」という，前後関係（しばしば前後して起こる）を有する事態のペアが文中に存在する．しかし，事態ペアが直列の係り受け関係になく，従来手法で獲得することができない．

抽出される事態ペアに漏れがあることは，一般的知識を獲得する際など，大規模なテキストが利用できる場合には問題とならないこともある．しかし，ドメイン依存の知識を獲得する場合には，必ずしも大規模なテキストは期待できない．そのため，抽出する事態ペアの漏れを少なくし，限られた量のテキストから効率的に知識を獲得することが必要となる．そこで，本稿では，任意の係り受けパターンと共起する事態表現を対象にすることで，事態間知識を広範に獲得できる手法を提案する．

2. 関連研究

本節では，事態間知識の獲得を行っている関連研究を述べる．これらの研究では，動詞句や述語項を事態の単位として採用している．

連絡先: 東山翔平，NEC 情報・ナレッジ研究所，
神奈川県川崎市中原区下沼部 1753，TEL 044-455-8733，
s-higashiyama@ak.jp.nec.com

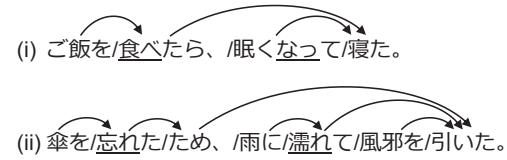

図 1: 非直列な係り受けパターンと共起する事態ペアの例

柴田ら [7] は，事態間の共起度として PMI（Pointwise Mutual Information）[1] を使用し，共起度の高い事態ペアを関連の強い事態ペアとして獲得している．特に，アソシエーション分析の手法である Apriori アルゴリズムを用いることで，膨大な述語と項の組合せに対して効率的に共起度の計算を行っている．事態間の共起度を用いることにより，「拾う-届ける」や「財布を拾う-警察に届ける」といった同一の述語対を含む事態ペアの中から，項として必須であるものを含む適切な述語項の単位を自動的に決定できる．しかし，柴田らの手法では，係り受け関係にある事態ペアを共起度計算の対象としており，その他の事態ペアを獲得できない．

阿部ら [5] は，実体間知識獲得の代表的な手法である Espresso [3] を拡張し，ブートストラップ法に基づく事態間知識獲得手法を提案した．阿部らの手法では，少数のシードとなる事態ペアを入力し，共起パターンの抽出と事態ペアの抽出を反復することで，因果関係を有する事態ペアを獲得している．阿部らは，「前方および後方の事態文節が，係り受け関係の木において先祖-子孫関係にある」場合（直列の係り受けパターンと呼ぶ）に限定して，事態ペアの周辺語句からなる文字列を共起パターンとして使用している．限定された形式のパターンを用いることで，特定の種類の関係を有する事態ペアに絞ることができる一方，非直列の係り受けパターンと共起する事態ペアを獲得することができない．

乾ら [6] は，少量の教師データと多量の教師なしデータに基づき，事態ペアの因果性を判定する半教師あり学習の確率モデルを提案している．しかし，係り受け関係にある事態ペアの中で，因果関係を有する正例および有しない負例を教師データとして用いているため，係り受け関係にない事態ペアについては，正確に因果性を判定できない可能性がある．

鳥澤 [4] は，事態間で共有され得る名詞を手掛かりに，並列句表現から推論規則の知識（「A するなら，しばしば B する」の形式の知識）を獲得している．しかし，2 つの述語が結ばれ

た単純な並列句（たとえば,「ビールを飲み，酔う」）を対象しているため，該当する関係にない事態ペアを獲得することができない．

3. 提案手法

本稿では，柴田らの手法と阿部らの手法を拡張するとともに組み合わせることで，任意の係り受けパターンと共起する事態ペアを獲得可能な手法を提案する．なお，本手法においても，事態に相当する単位として述語項を用いる．本手法は，次の 3 つのステップからなる．

1. **ブートストラップ法によるパターン獲得**
   ブートストラップ法を用いて，シードとなる事態ペアを元に共起パターンを獲得する．阿部らと異なり，非直列を含む任意の係り受けパターンを対象にパターンの獲得を行う．
2. **パターンからの事態ペア候補抽出**
   次に，獲得したパターンを使用し，パターンと共起した事態ペアを，関係を有する事態ペアの候補として抽出する．
3. **共起度スコア計算による関連の強い事態ペアの抽出**
   抽出された事態ペア候補に対し，共起度スコアを計算することで，関連の強い事態ペアを選定する．柴田らと異なり，自動抽出した係り受けパターンと共起する事態ペアが対象となる．また，共起度スコア計算では，パターンの確からしさを表す信頼度スコアを利用し，信頼度の高いパターンと共起する事態ペアのスコアが高くなるように，事態間の PMI を補正した値を共起度スコアとして使用する．

本手法では，任意の係り受けパターンと共起する様々な事態ペアを対象とするため，抽出される候補にノイズが多く含まれる可能性がある．そこで，パターンの信頼度で補正することにより，低信頼度のパターンから抽出された事態ペアのスコアを下げ，獲得される事態ペアに含まれるノイズを低減することを意図している．

3.1 ブートストラップ法によるパターン獲得

ステップ 1. では，関係を有する既知の事態ペア（インスタンス）をシードとして用い，事態ペアと共起するパターンの抽出，パターンと共起する事態ペアの抽出を繰り返すことで，入力テキストから事態ペアとよく共起するパターンを抽出する．抽出するパターンとしては，係り受け関係の木において，注目する二つの事態文節を含む部分木の中で，ノード数が最小であるものを用いる．

阿部らの手法では，事態ペアとパターンの間の共起度（PMI）に基づき，信頼度の高いパターンに支持される事態ペアの信頼度は高くなり，かつ，信頼度の高い事態ペアに支持されるパターンの信頼度が高くなるように，事態ペアの信頼度とパターンの信頼度を相互再帰的に定義している．本手法でも，パターン獲得のためのパターン抽出および事態ペア抽出の処理には同様の信頼度スコアを用いる．

パターン抽出，事態ペア抽出の各ステップでは，抽出されたパターン p に対する信頼度 $r_\pi(p)$，抽出された事態ペア i に対する信頼度 $r_\iota(i)$ をそれぞれ計算し，信頼度の高いパターンおよび事態ペアを抽出する．信頼度は，パターンと事態ペアの PMI を元に，式 (1) および式 (2) で定義される．

$$r_\pi(p) = \frac{1}{Z_\pi}\sum_i \mathrm{PMI}(i,p)\cdot r_\iota(i) \tag{1}$$

$$r_\iota(i) = \frac{1}{Z_\iota}\sum_p \mathrm{PMI}(i,p)\cdot r_\pi(p) \tag{2}$$

Z_π および Z_ι は，それぞれパターン，事態ペアの信頼度の最大値であり，信頼度の値を 0 以上 1 以下の範囲に正規化する係数である．なお，PMI は式 (3) の定義を用いる [*1]．

$$\mathrm{PMI}(x,y) = \frac{P(x,y)}{P(x)P(y)} \tag{3}$$

3.2 パターンからの事態ペア候補抽出

ステップ 2. では，自動抽出したパターンを用いることで，任意の係り受けパターンと共起する事態ペアを，関係を有する事態ペアの候補として抽出する．ステップ 1. のインスタンス抽出ステップとほぼ同等の処理となるが，最終的な事態ペアの選定は事態間の共起度により判定するため，ここでは，式 (2) による事態ペアの信頼度の計算と絞り込みは行わない．

3.3 共起度スコア計算による事態ペアの抽出

ステップ 3. では，柴田らの手法を拡張した方法により，抽出された事態ペア候補に対する共起度スコアを計算する．

柴田らは，事態を構成する述語と項をアイテムとみなし，関連の強いアイテム集合からなるルールを抽出するアソシエーション分析の手法（Apriori アルゴリズム）を応用することで，事態間の共起度計算を効率的に行っている．柴田らの手法では，述語と 0 個以上の項からなる述語項のペアに対し，PMI と等価な尺度である lift 値を共起度スコアとして計算する．そして，共起度の高い事態ペア e，e' をルール $e \to e'$ の形で抽出している．

一方，本手法では，事態ペアと共起したパターンの信頼度に基づくスコア（信頼度補正 PMI）を式 (4) で定義し，事態ペアの共起度スコアとして用いる．式 (4) では，頻度に基づく通常の PMI の代わりに，パターンの信頼度で頻度を重み付けした重み付き頻度 $\theta(\cdot)$ に基づき，PMI を計算している．

$$\mathrm{PMI}_\theta(e,e') = \frac{\frac{\theta(e,e')}{N_\theta}}{\frac{\theta(e)}{N_\theta}\frac{\theta(e')}{N_\theta}} \tag{4}$$

ここで，事態 e および事態ペア $i=(e,e')$ の重み付き頻度と，事態ペアの重み付き頻度の総和 N_θ は，式 (5)〜(7) で定義される．$n(i,p)$ は事態ペア i とパターン p の共起頻度であり，I_e は事態 e を含む事態ペア全体の集合を表す．

$$\theta(i) = \theta(e,e') = \sum_p n(i,p)\cdot r_\pi(p) \tag{5}$$

$$\theta(e) = \sum_{i\in I_e}\theta(i) \tag{6}$$

$$N_\theta = \sum_i \theta(i) \tag{7}$$

4. 評価実験

4.1 実験設定

実験には，Web から収集した社会カテゴリ（事件，事故，判決など）のニュース記事 1.2 万文書，17 万文を入力コーパスとして用いた．なお，入力コーパスの係り受け解析には，CaboCha（version 0.69）[2] を使用した．

実験の設定は次の通りである．

*1 阿部らの手法では，式 (3) を真数とする対数の値を PMI として使用し，-1 以上 1 以下の範囲に信頼度を正規化している．

- 事態文節間距離によるパターンの制限
  使用するパターンを，事態文節に相当するノード間の木構造上の距離が 3 以内であるものに限定した．係り受け木上の離れた箇所で出現する事態ペアを対象にすると，関係を有しないペアの割合が増えると考えたためである．
- シードインスタンス
  前述のコーパスと同一ドメインの文書中で高頻度で出現した事態ペアから，前後関係を有すると人手で判断したもの 10 件を選択した．
- イテレーション数 m
  反復処理の繰り返し回数で，シードインスタンスからのパターン抽出を 1 回行った後，(インスタンス抽出ステップ, パターン抽出ステップ) の対を m 回繰り返す．本実験では，$m = 0$ とし，シードインスタンスと共起するパターンの集合を一度だけ抽出し，これを事態ペア候補の抽出に用いた．反復処理を何度も行わないのは，パターンに距離の制限を設けることで，共起パターンの種類が数件しか存在せず，一度の抽出ステップで獲得できたためである [*2]．なお，獲得されたパターンは，すべて事態ペア候補の抽出に用いた．

上述の設定の下，表 1 に示す 3 つの手法を用いて，それぞれ事態ペアの獲得を行う．(a) は，3. 節のステップ 1. の処理を行わず，事態ペアが係り受け関係を有するパターン（直接係り受けパターン）のみを入力パターンとしてステップ 2. の処理を行い，通常の PMI を共起度スコアとしてステップ 3. を行う手法である．つまり，抽出対象となる事態ペア候補とその共起度スコアが柴田らの手法と同一である．(b) および (c) では，ステップ 1. によりシード事態ペアと共起するパターンを獲得し，ステップ 2. 以降の処理を行う．(b) はステップ 3. の共起度スコアとして通常の PMI を使用する手法であり，(c) は信頼度補正 PMI を使用する提案手法である．

なお，各手法で事態ペアの共起度計算を行う際の処理として，事態ペアを述語と項の集合とみなした場合に，集合として互いに包含関係にある事態ペア群 [*3] のそれぞれに対し，共起度スコアが最大の事態ペアを選択することで，必須の項を含む適切な述語項の単位を決定した．

4.2 パターンおよび事態ペアの抽出結果

シード事態ペアに基づくパターン獲得の結果，手法 (b) および (c) では，直接，直列，非直列を含む 6 個のパターンが得られた．シード事態ペアと最も多く共起した直接係り受けパターンが信頼度最大となり，その他のパターンは，直接係り受けパターンとの比で，0.2〜0.02 倍程度の信頼度となった．

事態ペア抽出の結果としては，表 1 に示すように，手法 (a) で約 7000 件の事態ペアが抽出されたのに対し，手法 (b) および (c) では，その 4 倍弱の件数が抽出された．表中の (a)-被覆率は，各手法で出力された事態ペアのうち，手法 (a) の出力に含まれるものの割合を指す．獲得されたパターンに直接係り受けパターンを含むことから，手法 (b) および (c) は，手法 (a) の出力の大部分を包含する手法となっている．特に，提案手法 (c) は，95%以上の被覆率である．なお，前述した必須の項の選択の処理のため，同じ係り受けパターンを用いていても，共起度スコアの計算結果が異なると，出力される事態ペアは完全に同一とはならない．

表 1: 比較する手法と事態ペア抽出結果

手法	抽出件数	(a)-被覆率
(a) 直接パターン+PMI	6967	100%
(b) 任意パターン+PMI	26781	90.8%
(c) 任意パターン+信頼度補正 PMI	26378	95.5%

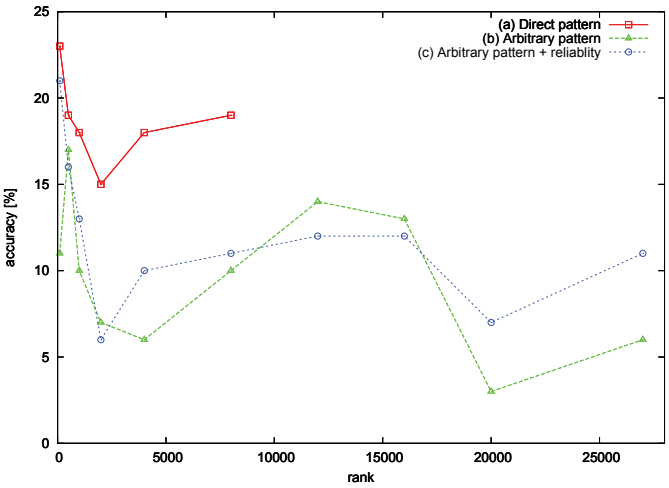

図 2: 各手法の順位の区間ごとの正解率

4.3 正解率の評価

3 つの手法それぞれについて，獲得された事態ペアを共起度スコアの降順に並べ，上位 100 件と，101〜500 位まで，同様に 1000 位まで，2000 位まで，4000 位まで，8000 位まで，12000 位まで，16000 位まで，20000 位まで，20000 位以降の各区間から，ランダムに 100 件ずつ事態ペアを抽出し，各区間における正解率を算出した．抽出された事態ペアについて，前後関係（一方の事態が生じた後に，しばしば他方の事態が生じる），含意関係（一方の事態が成立する際，必ず他方の事態も同時に成立することになる），前提関係（一方の事態が成立する際，必ず他方の事態が事前に成立している）が成立しているか否かを判定し，いずれかの関係を有すると判断された場合に，正解，つまり，関連の強い事態間知識であるとした．

評価の結果を図 2 に示す．まず，抽出された事態ペアの正解率は全体的に低く，正解率が最も高い手法 (a) の上位 100 件でも 23%に留まっている．これは，柴田ら [7] が報告している正解率 96%と比較すると 4 分の 1 に満たない．この主な原因として，入力コーパスの規模の差とフィルタリング条件の違いが考えられる．柴田らが 16 億文からなる大規模なコーパスを用いているのに対し，本実験で用いたのは 17 万文のコーパスである．PMI は，低頻度の場合に値が過度に高くなる問題が知られており，一般的な傾向から外れる少数の事例がノイズとなることが起こる．柴田らは，事態ペアの出現頻度に相当する尺度の閾値と，PMI の上限の閾値を設け，出現頻度が一定値未満かつ PMI が閾値を超える事態ペアをフィルタリングすることで，この問題に対処している．一方，本実験で抽出された事態ペアの中には，出現頻度が 1，2 回の低頻度のものが多くを占めていたが，本研究では小規模なドメインコーパスから広く知識を獲得することを意図しているため，頻度や共起度の閾値による取捨選択を行っていない．出現頻度が少ない事例についても，正しい知識であるか否かを判別可能にし，全体の精度

*2 今回は，何らかの関係を有する事態ペアを抽出する目的で，係り受け構造のみからなるパターンを使用した．一方，関係の種類を限定したい場合は，単語や品詞などの情報を含むパターンが有効であり，反復処理を繰り返すことで様々なパターンが獲得可能となる．

*3 たとえば,「拾う-届ける」,「財布を拾う-届ける」,「財布を拾う-警察に届ける」は述語項の集合として包含関係にある．

表 2: 提案手法で獲得された知識の例

	事態ペア	関係
(1)	運転を 見合わせる / 運行を 再開する	前後・前提
(2)	事件を 起こす / 検挙される	前後
(3)	雪が 降る / 初雪を 観測する	含意
(4)	墓参りに 訪れる / 菓子を 供える	前後
(5)	病院に 入院する / 容体が 回復する	前後

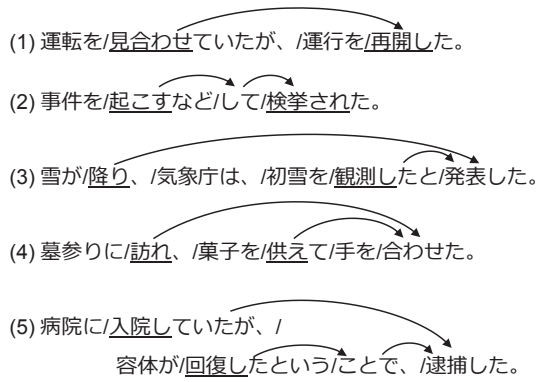

図 3: 獲得された知識の出現文脈

を改善する必要がある．

また，順位の区間ごとの正解率の変化は，3 つの手法とも，下位の区間の方が正解率が低いことが多い傾向があるものの，一貫した推移となっていない．原因としては，使用した共起パターンの抽象度の問題が考えられる．本実験では，共起パターンとして，文節間の係り受け構造のみからなる抽象的なパターンを使用したため，特定の関係を有する事態ペアが出現する文脈の特徴を捉えられなかったと言える．したがって，特定のパターンと共起した事態ペアの中に，関係を有するペアと有しないペアが混在し，共起度の高いペアの中にも関係を有しないペアが多く含まれるようになったと考えられる．

次に，各手法の正解率を比較すると，手法 (a) の正解率が，(b) および (c) の正解率をいずれの区間でも上回り，直接係り受け以外のパターンを用いることで，ノイズが増加することを示している．一方，(b) と (c) の比較では，各区間で正解率に大きな差がないか，提案手法 (c) の方が正解率が高いという特徴が見られる．つまり，パターンの信頼度による PMI の補正が，ノイズの低減に寄与していると判断できる．また，手法 (b) および (c) しか出力が存在しない 8000 位以降の区間でも，提案手法では 10%前後の正解率で推移している．したがって，共起パターンを増やして事態ペアの獲得数を増やすことで，関係を有する妥当な事態ペアが増加していることが見込まれる．

4.4 獲得された知識の例

提案手法 (c) で獲得された事態ペアの例を表 2 に示し，各事態ペアの出現文脈となったテキストを図 3 に示す．図 3 では，事態中の述語に相当する箇所を下線で示し，パターンを構成するノード間の係り受け関係を矢印で示した．つまり，文節をノード，係り受けの依存関係をエッジとする有向グラフが共起パターンとなる．

例 (1) は，直接および非直接のパターンと共起して得られたペアである．手法 (a) でも獲得されたものの，非直接のパターンとの共起もカウントすることで，事態間の共起頻度が増加する効果が得られた．一方，例 (2)〜(5) は，手法 (a) では獲得されず，非直接のパターンを扱うことで初めて獲得可能となった事態ペアである．(2) は直列のパターン，(3)〜(5) は非直列のパターンとの共起により獲得された．ただし，非直接のパターンのみと共起し新規に獲得された事態ペアは，非直接のパターンの信頼度の低さに起因し，共起度スコアが小さくなる結果となった．したがって，扱う共起パターンを増やすことで抽出可能になる事態ペアについては，高スコアの領域で検出可能にするために，パターンあるいはスコアリング方法の改良が必要と考えている．

5. おわりに

本稿では，任意の係り受けパターンと共起する事態を対象にすることで，事態間知識を広範に獲得できる手法を提案した．提案手法では，ブートストラップ法に基づく事態間知識獲得手法を拡張し，任意の係り受けパターンの獲得を行う．続いて，獲得されたパターンを使用することで，限定された係り受けパターンに基づく従来手法よりも，高い網羅性で事態ペアを抽出することが可能となった．また，共起パターンの信頼度を考慮して事態間の共起度を計算することで，獲得される知識に含まれるノイズを低減可能となることを示した．

今回の評価では，文節間の係り受け構造のみからなる抽象的な共起パターンを使用した．そのため，特定の関係を有する事態ペアが出現する文脈の特徴を十分に捉えられず，事態ペアの共起度スコアが，知識としての確からしさを必ずしも直接的に反映しない結果となった．そこで，文節中の単語の品詞や表層といった語彙的な情報をパターンに組み込むなど，適切な抽象度のパターンを定義することで，スコア上位での精度を向上させる予定である．また，提案手法により獲得可能となる知識の量の増分についても，詳細な評価が必要である．

参考文献

[1] K. W. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29, 1990.

[2] T. Kudo and Y. Matsumoto. Fast methods for kernel-based text analysis. In *Proc. of ACL 2003*, pages 24–31, 2003.

[3] P. Pantel and M. Pennacchiotti. Espresso: leveraging generic patterns for automatically harvesting semantic relations. In *Proc. of COLING/ACL 2006*, pages 113–120, 2006.

[4] K. Torisawa. Acquiring inference rules with temporal constraints by using japanese coordinated sentences and noun-verb co-occurrences. In *Proc. of HLT-NAACL 2006*, pages 57–64, 2006.

[5] 阿部修也, 乾健太郎, 松本裕治. 項の共有関係と統語パターンを用いた事態間関係獲得. **自然言語処理**, 2010.

[6] 乾孝司, 高村大也, 奥村学. 因果関係知識獲得のための隠れ変数モデル. **言語処理学会第 *12* 回年次大会発表論文集**, pages 959–962, 2006.

[7] 柴田知秀, 黒橋禎夫. 述語項構造の共起情報と格フレームを用いた事態間知識の自動獲得. **情報処理学会研究報告 自然言語処理研究会**, 2011.