事態含意名詞の利用と共起パターンの学習による事態間関係知識の獲得

阿部修也　乾健太郎　松本裕治
{shuya-a, inui, matsu}@is.naist.jp

奈良先端科学技術大学院大学 情報科学研究科

1 はじめに

人間に近い高度な言語情報処理能力を工学的に実現するには，辞書や文法などの言語知識の他に，大量の世界知識を計算機に与える必要がある．

大規模なテキストデータから事態関係知識を自動的に獲得するという試みが報告されている [1, 2, 4, 7]．例えば，Inui らは，接続助詞「ため」を含む複文（タメ複文）に現れる主節と従属節から事態間の因果関係を獲得する方法を提案しており，獲得した関係を Cause, Precondition，Effect，Means の 4 種類に高い精度で分類できると報告している [4]．ただし,「書店に行く」と「書店で買う」のように，わざわざ言及するまでもない当たり前の関係をタメ複文のようなパターンからの獲得は限界があり，実際 Torisawa の報告によると，Torisawa の方法で獲得した知識の多くがタメ複文では獲得できないものであった [7]．一方，Torisawa の方法は，動詞テ形接続や連用中止接続のように頻度が高く一般的な手がかりと，別途収集した格関係の統計を巧妙に組み合わせることによって，より常識的な事態間関係を獲得することを狙うもので,「モノの用途とその準備の関係」の獲得で成果を上げている．ただし，こうした方法を広く他の事態間関係に適用できるかどうかは今のところ明らかでない．

こうした研究の知見をまとめると，タメ複文やテ形接続のような特定の複文パターンを使って事態間関係のインスタンスを獲得するアプローチには次のような問題がある．

- 特殊なパターンを使うと，より意味的に制約された関係のインスタンスを獲得することができるが，そうしたパターンは頻度が少なく，十分な量のインスタンスを獲得するのは難しい．また，常識的な関係のインスタンスはこうした特殊なパターンに現れにくい．
- より一般的なパターンは，常識的な関係のインスタンスとも共起し，頻度も高いが，所望の関係でないインスタンスも同様に共起するので，ノイズのフィルタリングに工夫が必要である．

すなわち，知識獲得の精度と規模を両立するには，意味的な制限の強い特殊な共起パターンを数多く用意することが望ましい．この問題に対し，本稿では，用言だけでなく体言の中にも事態を表す，あるいは含意するもの（以下，事態含意名詞）が多数あることに着目し，事態含意名詞を含むより広範な共起パターンを利用して，事態間関係のインスタンスを獲得する方法を検討する．

2 事態含意名詞と事態間関係獲得

本稿では，行為者が意志的に行う**行為**とそれ以外の**出来事**（経験や状態，状態変化など）を併せて**事態**とよび，事態間関係の知識を獲得する問題を考える．我々の目標は，因果関係（「運動する」と「汗をかく」）や部分全体関係（「研究する」と「実験する」），上位下位関係（「罪を犯す」と「盗む」）などを含む広い範囲の関係を獲得し，深い言語理解に資する知識ベースを構築することである．

事態を表す言語表現は「汗をかく」のような動詞句に限られるわけではない．次に挙げるように，名詞の中にも事態を直接表現するものや間接的に指すものが少なくない．本稿ではこうした事態を直接的あるいは間接的に参照する名詞を**事態含意名詞**と呼ぶ．

- **動詞形を持つ事態名詞：** 窃盗，外食など
- **動詞形を持たない事態名詞：** 雨，ガス欠，運動会など
- **事態の項を指す名詞：** 犯人（「罪を犯す」という行為の行為者），運転者など
- **特定の用途を持ったモノを指す名詞：** 包丁（「食材を切る」ための道具），研究室（「研究する」ための場所）など

事態を表す表現として上のような事態含意名詞まで含めて考えると，事態間関係獲得の手がかりとなる共

起パターンは，２つの動詞句からなる複文のパターンに比べて飛躍的に広がる．例えば，事態含意名詞と動詞句の共起には,「雨**で**タイヤがすべる」,「食**後に**歯を磨く」,「研究室**で**実験する」,「窃盗**の容疑で**逮捕される」のようなパターンが考えられるし,「入会**時の**登録」のような事態含意名詞どうしの共起も有用な手がかりになる可能性がある．1 節で述べたように，我々は意味的な制限の強い特殊な共起パターンをたくさん集めたいので，事態含意名詞の利用によって共起パターンの候補のプールが拡大することは重要である．

次の問題は，膨大な数の共起パターンの候補からどうやって信頼性の高いものを探すかである．従来の事態間関係獲得の研究は，少数の適当な共起パターンを人手で見つけて利用することを前提にしていたので，パターン発見の問題にはアプローチしてこなかった．一方,〈著作物名〉と〈著者名〉のような実体（entity）間の関係を抽出する，いわゆる関係抽出の研究では，実体の出現パターン（つまり共起パターン）の自動獲得が主要な課題の一つとして認識され，一定の成果を得るに至っている [5, 6].

以上のような背景を踏まえると，まずは，事態間関係知識の獲得を上の意味での関係抽出タスクの一種と見なし，関係抽出のための共起パターン獲得技術を事態間関係獲得に適用することの有効性を調べることが重要である．そこで我々は，共起パターン獲得技術の一例として，Pantel らが最近提案した Espresso と呼ばれるアルゴリズム [6] を取り上げ，これを拡張して事態間関係獲得に応用することを試みる．

本稿では，3 節で Pantel らの手法を概説し，4 節でこれを事態間関係獲得に適用する際に必要となる拡張を述べた後，5 節で現在までに得られている実験結果を報告する．

3 関係獲得アルゴリズム Espresso

Pantel ら [6] は，信頼性の高い実体間関係のインスタンスをシードとしてパターンを獲得し，このパターンを使って新しい実体間関係のインスタンスを獲得するブートストラップ的関係獲得手法を提案している．このシステムを Espresso と称している．

Pantel らのブートストラップ的関係知識獲得手法は，パターンの獲得とインスタンスの獲得次の 4 つの操作を繰り返し適用することで信頼性の高いパターンとインスタンスを集めて，最終的に精度の良いインスタンスを獲得する．

3.1 パターンのランキングと選択

所望の関係のインスタンス $\{x, y\}$ を与えたとき，コーパスから x と y が含む文を抽出し，それらを一般化してパターンとする．例えば，インスタンス “*Italy*, *country*” を与えたとき，テキスト “country **such as** Italy” からパターン “Y such as X” を獲得する．

式 1 を用いてパターンの信頼性を評価する．

$$r_\pi(p) = \frac{\sum_{i \in I} \left(\frac{pmi(i,p)}{\max_{pmi}} \times r_\iota(i) \right)}{|I|} \tag{1}$$

I はインスタンスの集合，p はパターン，$\max_{pmi}$ は全てのインスタンスと全てのパターンにおける PMI の最大値である．インスタンス i の信頼性 $r_\iota(i)$ は式 2 で与えられる．

3.2 インスタンスのランキングと選択

パターン p を与えたとき，コーパスから p を含む文を抽出し，ここからインスタンス $\{x, y\}$ を獲得する．例えばパターン “Y such as X” を用いて，テキスト “country **such as** Italy” からインスタンス “*Italy*, *country*” を獲得する．

式 2 を用いてインスタンスの信頼性を評価する．

$$r_\iota(i) = \frac{\sum_{p \in P} \left(\frac{pmi(i,p)}{\max_{pmi}} \times r_\pi(p) \right)}{|P|} \tag{2}$$

P はパターンの集合．なお，シードインスタンスは $r_\iota(i) = 1$ とする．

4 事態間関係獲得への適用

Pantel らはブートストラップ的関係知識獲得手法を用いて実体間関係知識を獲得したが，我々は事態間関係知識を獲得する．そのための工夫を述べる．

4.1 動詞句

「切符を買う→電車に乗る」という事態間関係を考えた場合,「買う→電車に乗る」は「買う」が一般的過ぎるために不適切な関係である．また「駅で切符を買う→電車に乗る」は事態間関係として間違いではないが「駅で切符を買う」は特殊過ぎるため，この関係を利用するときに問題が生じる可能性がある．

事態が動詞句であるときに適切な格を選ぶことで良い事態を獲得することができる．ブートストラップ的関係獲得手法中のスコア関数は適切な格を持つ事態に高いスコアを与えると仮定できるので，動詞句を格の有無によって展開してスコア関数を用いて適切な格を持つ動詞句を選ぶ．ここでは格の有無による組み合わせが増え過ぎるのを防ぐために，動詞句の格を最大で 1 つとして展開した．例えば「駅で切符を買う」とい

う動詞句を,「駅で買う」「切符を買う」「買う」に展開する.

4.2 事態含意名詞について

サ変名詞と接尾辞の組み合せは事態含意名詞であると考える．例えば，サ変名詞「マッサージ」に接尾辞を付与し,「マッサージ器」「マッサージ師」「マッサージ中」「マッサージ」（接尾辞なし）となる．これらはどれも,「マッサージする」という事態を含意している．実験では，サ変名詞を含む名詞句を事態含意名詞と見なした.

4.3 パターン

我々が用いたパターンは次の要素の組み合わせからなる.

- 係り受け関係に基づく 3 つのパターン
  - 事態と事態が直接係り受け関係になっている
  - 事態と事態が任意の文節要素を介して係り受け関係になっている
  - 事態と事態がそれぞれ共通の任意の文節に係っている
- 事態の接尾辞と助詞
  - 接尾辞：〜者，〜機，〜中，…
  - 助詞：〜が，〜を，〜に，〜ために，…
- 事態が行為か出来事かの区別
  - 行為：走る，食事をする，…
  - 出来事：風邪をひく，事故にあう，…

パターンを展開する．例えば,「煮詰めましょう」という文節を「煮詰めましょう」「煮詰めます」「煮詰る」と展開し，同じ規則をパターンに含まれる全ての文節に適用することでパターンを展開する.

また，あ事態が行為なのか出来事なのか区別するために，LCS 辞書 [11, 10] の意志性の判断に，さらに人手で意志性の有無を追加した．実験に用いたデータは意志性ありは 8968 語，意志性なしは 3597 語，意志性が曖昧な語は 547 語であった．このうち意志性が曖昧な 547 語は実験に用いていない.

4.4 実験条件

関係獲得には，河原ら [8] が収集した「Web 上の 5 億文の日本語テキスト」の約 1/4 を用いた．低頻度の事態や共起パターン，文節内の文字数が 16 文字を越える事例を除いた．さらに，幾つかの事態について格を伴なわなければいけないという制約を加えた[1]．ガ格やヲ格の格要素が事態性名詞となるパターンを削除した.

[1]「ある」「なる」「いる」「する」「とる」「かかる」「かける」「なる」「付く」「できる」「過ぎる」「経る」「経過する」等

事態間関係として，行為-効果の関係（行為の結果事態がおうおうにして起こる．または行為をすることは事態を保つこと）と行為の部分全体関係（ある行為を行なう間にしばしば行なう行為）を対象に実験する．それぞれ 100 組程度のインスタンスをシードとして与えた.

5 結果

獲得したインスタンスを信頼度順に上位 1〜1000 件, 1001〜5000 件，5001〜10000 件の領域に分け，各領域から 100 組の事態対をランダムに抽出し，人手で評価した．行為-効果の関係，行為の部分全体関係の結果を表 1 に示す．行為-効果の関係で獲得した事態対を表 2 に示す.

制約が緩い行為の部分全体関係とよりも制約の多い行為-効果で同程度の精度であった．信頼度が下ってもそれほど精度が低下しなかった．今後獲得する事態対を増やした場合でも軽微な信頼度の低下でより多くの事態対を獲得できる望みがある.

事態含意名詞を用いることで獲得できる事態間関係が増えたことを示すために，人手で評価した正解事例から動詞のみの共起（事態含意名詞を用いない）で獲得できる事態対を抽出した．例えば「開票**の結果**当選した」「戦争**による**破滅」「開票**の結果**当選した」が事態含意名詞を用いて始めて獲得できる事態対である．また事態含意名詞を用いて始めて獲得できる事態対は，行為-効果の関係で 32%，行為の部分全体関係で 35% であった．事態含意名詞を用いることでより多くの事態間関係を獲得できることがわかった.

事態含意名詞の種類 事態含意名詞が含意する事態には,「専用機」の「専用する」のように事態が名詞の種類や性質を表わしている場合と,「通信機」の「通信する」のように事態が名詞の目的や役割を表わしている場合がある．事態含意名詞の事態が表しているものが名詞の性質なのか，目的なのかを判断しないと事態間の関係を誤って獲得することがある.

例えば，目的の関係を表わすパターン “＜名詞; 後件; 意志性あり＞**機で**＜動詞; 前件; 意志性あり＞” は前件の事態含意の事態が目的を表している場合のみ，前件と後件が目的の関係を表わし,「通信機」の「通信する」は名詞の手段を表しているので「通信**機で**呼び出す」から獲得できる「呼び出す→通信する」は目的の関係である．しかし,「専用機」の「専用する」は名詞の種類を表しているので「専用**機で**出発する」から獲得できる「専用する→出発する」は目的の関係ではない.

6 議論

本実験を通して次の 4 つの問題点がわかってきた.

表 1: 行為-効果の関係と部分全体関係の精度

関係	1〜1000 件	1001〜5000 件	5001〜10000 件
行為-効果の関係	0.65	0.64	0.61
部分全体関係	0.67	0.64	0.61

表 2: 行為-効果の関係で獲得したパターンと事態対の例

パターン	事態
〈 後件; 名詞; 意志性なし 〉**直前まで**〈 前件; 動詞; 意志性あり 〉	沸かす→沸騰する，加熱する→沸騰する，温める→沸騰する
〈 前件; 名詞; 意志性あり 〉**による**〈 後件; 名詞; 意志性なし 〉**者数**	推薦する→内定する，自殺する→死亡する，迫害する→死亡する
〈 前件; 名詞; 意志性あり 〉**の結果**〈 後件; 動詞; 意志性なし 〉**ました**	勉強する→合格する，開票する→決まる，検索する→判明する，審査する→入選する
〈 後件; 名詞; 意志性なし 〉**に向けて**〈 前件; 動詞; 意志性あり 〉	頑張る→優勝する，指導する→進級する，努力する→成功する，頑張る→突破する
〈 後件; 動詞; 意志性なし 〉**ほどに**〈 前件; 動詞; 意志性あり 〉	焼く→焦げ目がつく，握り締める→爪が食い込む，求める→渇く，入れる→溢れる
〈 前件; 動詞; 意志性あり 〉**たい**〈 後件; 動詞; 意志性なし 〉**たい**	戦う→死ぬ，傷つける→傷つく，勉強する→理解する
〈 後件; 動詞; 意志性なし 〉**たい**〈 前件; 動詞; 意志性あり 〉**たい**	勉強する→知る，自殺する→死ぬ
〈 前件; 名詞; 意志性あり 〉**不足で**〈 後件; 動詞; 意志性なし 〉**ません**	調査する→分かる，勉強する→知る，勉強する→分かる
〈 前件; 動詞; 意志性あり 〉**ないと**〈 後件; 動詞; 意志性なし 〉**ません**	調査する→分かる，治療する→治る
〈 後件; 動詞; 意志性なし 〉**ほど**〈 前件; 動詞; 意志性あり 〉**た**	飲む→二日酔いになる，笑う→腹がよじれる，見慣れる→見飽きる，煮る→崩れる，

第 1 に，1 億文強のデータを用いたが表現の多様性を無視できるほどに十分な共起データを集められなかった.「病院に通う」と「通院する」,「仕事を始める」と「仕事する」,「スーパーで買い物をする」と「スーパーで買う」等，同じ事態だが異なる表現になっている場合がある．また実験結果はデータがまだまだスパースであることを示している．今後は，実験の規模を拡大すると同時に，意味的に等価なパス，言い換え表現／機能動詞構文を考慮し表現の多様性を吸収する方法を試みたい.

第 2 に，シードを増やすと精度が良くなる傾向にあった．より簡単に精度よくシードを作成することが重要になる．Inui らや Torisawa の手法を用いてシードとなる事態対を作成して実験することを試みたい.

第 3 に，事態の抽象度を適切に決めることの重要性と難しさである．従来の関係抽出は実体を表わす固有表現間の関係の抽出が中心であり，そこでは固有表現そのものの抽象度の選択は問題にならなかった．一方，4.1 で述べたように，事態間関係抽出では事態表現の抽象度を決める問題が顕現する．今回は事態間関係を抽出するときのスコアが最適な抽象度の事態を自動的に選択すると期待して実験を行ったが必ずしも期待通りの結果が得られたとは言えない．スコア関数の工夫や事態を汎化することでこの問題に対処しようと考えている.

第 4 に，パターンへ導入する素性が不足しているために対処できない事例があることがわかった．事態含意名詞の種類や否定形の取り扱いの問題がある．他に，テキストからより高い精度で事態の共起を獲得するためには，省略された格を補ったり，2 つの事態の項の間の共参照関係を同定したりする述語項構造解析が欠かせない．これについては，現在開発中のゼロ照応解析技術 [3][9] を利 用することを検討している.

謝辞

「Web 上の 5 億文の日本語テキスト」の使用許可を下さった情報通信研究機構の河原大輔氏と京都大学大学院の黒橋禎夫氏に感謝いたします.

参考文献

[1] Timothy Chklovski and Patrick Pantel. Global path-based refinement of noisy graphs applied to verb semantics. In *In Proceedings of Joint Conference on Natural Language Processing (IJCNLP-05)*, 2005.

[2] R. Girju, A. Badulescu, and D. Moldovan. Automatic discovery of part-whole relations. *Computational Linguistics*, Vol. 32, No. 1, pp. 83–135, 2006.

[3] Ryu Iida, Kentaro Inui, and Yuji Matsumoto. Anaphora resolution by antecedent identification followed by anaphoricity determination. *ACM Transactions on Asian Language Information Processing (TALIP)*, Vol. 4, No. 4, pp. 417–434, 2005.

[4] Takashi Inui, Kentaro Inui, and Yuji Matsumoto. Acquiring causal knowledge from text using the connective marker tame. *ACM Transactions on Asian Language Information Processing (TALIP)*, Vol. 4, No. 4, pp. 435–474, 2005.

[5] D. Lin and P. Pantel. Dirt - discovery of inference rules from text. In *In Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining 2001*, 2001.

[6] Patric Pantel and Marco Pennacchiotti. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, pp. 113–120, 2006.

[7] Kentaro Torisawa. Acquiring inference rules with temporal constraints by using japanese coordinated sentences and noun-verb cooccurrences. In *In Proceedings of Human Language Technology Conference/North American chapter of the Association for Computational Linguistics annual meeting (HLT-NAACL06)*, pp. 57–64, 2006.

[8] 河原大輔, 黒橋禎夫. 高性能計算環境を用いた web からの大規模格フレーム構築. 情報処理学会 自然言語処理研究会 NL-171-12, pp. 67–73, 2006.

[9] 小町守, 飯田龍, 乾健太郎, 松本裕治. 共起用例と名詞の出現パターンを用いた動作性名詞の項構造解析. 言語処理学会第 12 回年次大会論文集, 2006.

[10] 竹内孔一, 乾健太郎, 藤田篤. 語彙概念構造に基づく日本語動詞の統語・意味特性の記述. レキシコンフォーラム, Vol. 2, pp. 85–120, 2006. ひつじ書房.

[11] 竹内孔一, 乾健太郎, 藤田篤, 竹内奈央, 阿部修也. 分類の根拠を明示した動詞語彙概念構造辞書の構築. 情報処理学会 自然言語処理研究会 NL-169-18, pp. 123–130, 2005.