文内共起パターンと格要素共有情報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

1 はじめに

含意関係や因果関係などの事態間関係は，人間に近い高度な言語処理能力を工学的に実現する上で欠かせない知識のひとつであり，質問応答，情報抽出，対話，要約など，幅広い言語処理アプリケーションに役立つことが期待できる．

本稿では大規模コーパスから事態間関係を自動獲得する手法について論じる．本稿で対象とする事態間関係とは，「X=電話;X をかける→ X が通じる（行為-効果の関係）」のように，(a) 事態表現間に成り立つ関係が「行為-効果関係」や「行為-前提関係」のように分類されており，さらに (b)「X（=電話）」のように 2 つの事態間で同一の要素からなる項（「かける」のヲ格と「通じる」のガ格）があればその情報が特定されているという条件を満たす関係を云う．したがって，事態間関係を自動獲得するには，(a) 事態表現対の関係の識別および (b) 項の共有関係の同定が必要である．事態間関係獲得についてはすでにいくつかの手法が提案されている [1, 2, 4, 6, 7, 8, 9] が，いずれの手法も上の (a), (b) の問題を同時に解決するものにはなっていない．以上の背景を踏まえ，本稿では事態表現の文内共起情報と文章内共起情報を組み合わせることによって，関係の識別と項の共有情報の同定をともに実現する手法を提案する．

2 事態間関係獲得手法

既存の手法を格の重なりに基づく手法と文内共起パターンを利用する手法に分けて紹介する．

格の重なりに基づく手法 事態間の格の重なりに基づく手法がある．DIRT[4] と TE/ASE[8] は格要素集合の分布が似ている事態対は類義/同義/含意関係にあるという仮説に基づく手法であり，Pekar の手法 [7] は同一談話関係にある事態対は一連の状況で一緒に起りやすい事態対になっているという仮説に基づいている．これらは項の共有を伴なう含意関係を獲得できるという利点はあるが，関係を推定する証拠が少ないために前提/結果/手段関係等のより詳細な区分が必要な関係を識別することが困難である．

文内共起パターンを利用する手法 類義/同義/含意関係をより区別した前提/結果/手段関係等を獲得するためにはより多くの関係を表わす証拠が必要である．そこで文内で事態対が共起したときの共起パターンを関係を表わす証拠とすることで，前提/結果/手段関係のような事態間関係を認識することができる [1, 2, 6]．例えば，Inui らは因果関係を表わす共起パターンとして接続助詞「ため」を利用した．しかし，文内で事態対が共起する場合はしばしば片方の事態の格が省略されるため，事態間の項の共有がわらかなくなる欠点がある．例えば，「お茶を**淹れて飲む**」では，「飲む」のヲ格の「お茶」が省略されている．

ここまでの手法を要約すると，格の重なりに基づく手法は項の共有情報を認識できるが，類義/同義/含意関係よりも詳細な区別が必要な関係を認識することが困難である．一方，文内共起パターンに基づく手法では前提/結果/手段関係のような関係を認識できるが，項の共有情報を認識できない．

この問題に対し，Torisawa[9] は格の重なりに基づく手法と文内共起パターン手法を組み合せたような手法を用いることで，項の共有情報を認識しつつ「モノの用途とその準備の関係」を獲得することができた．この手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．

こうした研究の流れを踏まえ，本稿では広く他の事態間関係に適用可能でありながら，項の共有を認識可能な手法を提案する．本手法では，Pekar の談話関係解

析による手法を参考にして，文章内共起を用いて，一連の状況で一緒に起りやすい事態対を項の共有情報と共に獲得する．この項の共有情報と，文内共起パターンを用いて認識した事態間関係を組み合わせて，項の共有情報を伴なう事態間関係を獲得する．言い換えると，文内共起パターンで認識した事態間関係に，文章内共起を用いて認識した項の共有情報を割り当てる手法である．この手法は，関係認識手法に文内共起パターンを用いているため，広く他の事態間関係を認識することが可能であり，Torisawa の手法の欠点を克服している．同時に，文章内共起を用いて項の共有情報を認識しているため，項の共有情報を付与することが難しいという文内共起パターンの欠点を克服している．詳細は 4 節で説明する．

3 文内共起パターン

文内共起パターンに基づく獲得手法は Abe[1] と同様の手法を用いた．この手法は Pantel ら [6] が提案した Espresso と呼ばれる実体間関係獲得手法を，事態間関係獲得に拡張した手法である．Espresso 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である．

事態を表わす表現 コーパス中では様々な形式で事態が表現されている．本研究では動詞句の他に事態を含意する名詞句からも事態間関係を獲得する．事態を含意する名詞句は様々あるが，本実験においてはサ変名詞と接尾辞の組み合せだけを対象とする．例えば，「**掃除**機で**きれい**にする」から「掃除する→きれい」という事態対（行為-結果の関係）を獲得できる可能性がある．

共起パターンの表現 Espresso は実体間のテキストを一般化したものを共起パターンに用いたが，事態間関係の場合は前述したように項が存在するため，係り受け関係に基づく共起パターンを用いた．

事態対が直接係り受け関係になっている場合と，事態対が任意の文節要素を介して係り受け関係になっている場合だけを事態間関係獲得の対象にした．共起パターンは事態の接尾辞（～者，～機，～中，…）や助詞（～が，～を，～に，～ために，…）や事態の間の任意の文節を含み，これを「日本語機能表現一覧」[5] を用いて一般化した．さらに，事態が行為（走る，食事をする，…）か出来事（風邪をひく，事故にある，…）かの区別も共起パターンに含ませた．この区別は後述の意志性辞書を利用した．

例えば「**電話をかけ**たけれども**通じ**ない」では，「電話をかける」は動詞で行為，「通じる」は動詞で出来事なので，「<verb;action>たけれども<verb;effect>ない」という共起パターンを獲得する．

意志性辞書 事態が行為なのか出来事なのか区別するために，人手で意志性の有無を追加した辞書を作成した．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 文章内共起

項の共有情報を伴なう事態間の含意関係を獲得するために，Pekar[7] を参考に，同一文章内で登場人物や対象物，場所が等しければ，それを項とする事態間に一連の状況で一緒に起りやすい関係があるという仮説を立てた．また，このとき項の間で共有される実体をアンカーと呼ぶ．図 1 の文章内共起では，アンカー「電話」を用いて「かける」と「通じる」が一連の状況で一緒に起りやすい関係にあることを認識している．

アンカーは自身や物を指す代名詞を除く名詞（非アンカーリストは人手で作成した 219 語）とし，文節の主辞を表わす 1 形態素とした．ただし，接尾辞の場合は直前の形態素を含めて 2 形態素とした．さらに対象とする格をガ格，ヲ格，ニ格，デ格，カラ格，ヘ格，ト格，ヨリ格に制限した．

文内共起で認識した事態間関係と文章内共起で認識した項の共有を伴う事態対を組み合わせて，項の共有を伴う事態間関係を獲得する．

5 実験

文内共起パターンで認識した事態対の精度を評価する．次に，文内共起パターンで認識した評価済みの事態対に，文章内共起を用いて認識した項の共有情報を付与し，項の共有情報を伴う事態間関係が適切に獲得されたことを確認する．

5.1 文内共起パターンの評価

河原ら [3] が収集した「Web 上の 5 億文の日本語テキスト」から文内共起パターンを用いて事態間関係を獲得した．事態対とパターンの共起頻度が 20 回未満の事例，ガ格やヲ格の格要素が事態性名詞となる事例を除いた．さらに，「ある」「なる」「いる」「する」等の語義が曖昧になりやすい動詞は直前格を伴なうという制約を与えた．

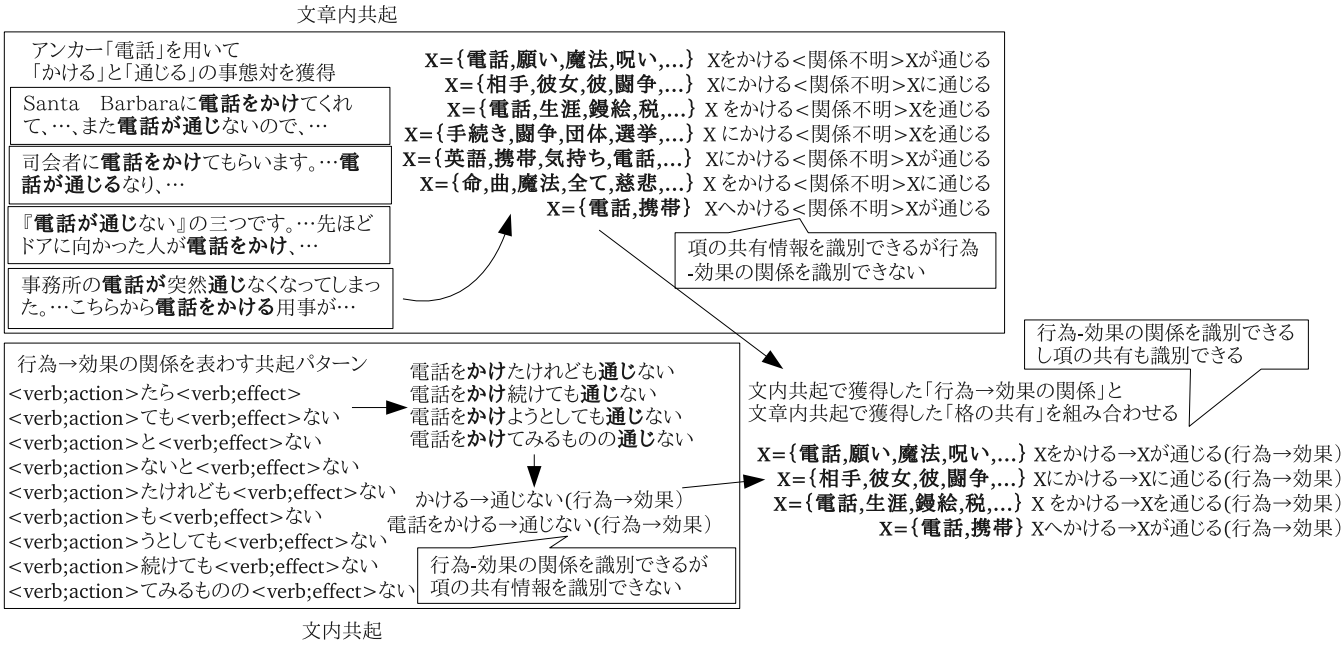

図 1: 文内共起と文章内共起の組み合わせによる事態間関係獲得

表 1: 文内共起パターンの精度

評価方法＼システム	文内共起	ベースライン
strict	60%	21%
lenient	86%	42%

事態間関係として，行為-効果関係（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）を対象に実験する．1000 組程度の事態対をシードとして与えた．

ベースラインは，行為-効果関係を表す接続表現「たため」「たから」「て」を用いて，接続表現と共起する事態対を PMI の順に並べた．

獲得した事態対を信頼度順に上位 1〜500 件，501〜1500 件，1501〜3500 件，3501〜7500 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評価した．同様にベースラインもサンプリングし評価した．格を含めて正しい関係にある事態対を strict，もし適切な格を埋めることができれば正しい関係になるであろう事態対を lenient と称する．例えば，事態対「かける→通じる」は不適切な事態対だが，「電話をかける→電話が通じる」であれば行為-効果の関係である．結果を表 1 に示す．獲得した事例を図 1 の文内共起に示す．

strict に評価した結果はベースラインよりも良い精度であり，文内共起パターンを用いることで上手く事態間関係を獲得できたことがわかる．一方で strict と lenient の差の 26%は，正確に格を付与することができれば 26%の精度向上が望めることを示している．

5.2 文内共起パターンと項の共有情報の評価

文内共起パターンに用いた実験と同様に，「Web 上の 5 億文日本語テキスト」から文章内共起を用いて項の共有情報を伴なう含意関係にある事態対を獲得した．係り受け誤りを含む事例を除くために，格と事態の PMI が 0 未満の事態を含む事例を除いた．同様に名詞と格の共起と事態の PMI が 0 未満となる事例も除いた．

文内共起パターンを用いて獲得した行為-効果の関係にある事態対中で評価済みの 400 事例に対して，談話関係に基いて獲得した項の共有情報を伴なう含意関係にある事態対を結び付ける．これによって，行為-効果の関係にある事態対に項の共有情報が付与される．

評価済みデータ 400 事態対中 356 事態対（89%）に項の共有情報を付与することができた．一つの事態対に対して複数の項の共有情報が付与され，項の共有情報が付与された 356 事例について合計で 1621 個の項の共有情報が付与された（1 事例につき平均 4 つの項の共有情報が付与された）．項の共有情報が付与された事態対の例を図 1 に示す．

項の共有情報を付与できた 356 事例を人手で評価した．結果を表 2 に示す．

各事態対において頻度の高い最大 3 つの項の共有情報について，3 つ全て正しければ事態対を正解とした場合の精度を arg-all と称する．最も頻度の高い項の共有情

表 2: 格共有情報の精度

項＼アンカー	anc-all	anc-any	anc-none
arg-all	29%	48%	53%
arg-top1	62%	80%	83%
arg-any	80%	91%	92%

報だけを評価した場合の精度を arg-top1 と称する．頻度の高い最大 3 つの項の共有情報について，最低 1 つが正しければその事態対を正解とした場合の精度を arg-any と称する．

本実験ではアンカーを共有される項の具体的な事例とみなし，これを含めて評価した．アンカーを頻度順に最大 3 つ示し，全て正しければ事態対を正解とした場合を anc-all，最大 3 つのアンカーうち一つでも正しいアンカーがあれが事態対を正解とした場合を anc-any，アンカーの存在を無視して評価した場合を anc-none と称する．

6 議論

lenient の精度 86% は格が適切に付与されていた場合の精度を予測した数字である．それにもかかわらず，arg-any かつ arg-none の精度は 92%で，予測値よりも高い精度であった．この理由は lenient の評価の難しさにある．評価者は lenient の評価をするときに，格を想起することができれば正解，できなければ不正解と判断する．しかし，実際は適切な格があるにも関わらず格を想起することができない事例があった[1]．こういった事例，lenient の精度と arg-any かつ arg-none の精度の差の原因である．

arg-all の精度と anc-all の精度は他の精度と比較して著しく低い．これは適切な項の共有情報の選択，適切なアンカーの選択がまだ不十分であるということを意味している．しかしながら，arg-any の精度と anc-any の精度は高いので，正解を含む集合を選択できているが，正解だけを選べていない状況にあると考えられる．候補集合には正解があるので適切なフィルタリング方法を開発することで今後の精度向上が期待できる．

7 まとめ

文内共起に基づく手法と文章内共起に基づく手法を組み合わせることで高い精度で事態間関係を獲得することができた．しかし，項の共有情報の選択と共有される格要素の選択がまだ不十分であり，今後の課題である．

[1]評価者は「投げる→間に合う」に格を追加しても行為-効果の関係にならないと判断した．提案手法で項の共有を付与したところ「X=一塁;X に投げる→ X に間に合う」となり，この事態対は行為-効果の関係になっている．

文章内共起と組み合わせることで高い精度を実現したが，文内共起のみ精度はまだ低いと考えている．今後は文内共起のみの精度を向上させることで，文章内共起と組み合わせたときの精度を向上させたいと考えている．

今回の評価方法には幾つか不十分な点があった．今後は評価事例を増やして複数人で評価することでより信頼性のある評価結果を示したい．今回は文内共起だけの結果と文内共起と文章内共起を組み合わせたときの結果を評価したが，文章内共起だけの結果は評価していない．この評価も今後の課題とする．

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の黒橋禎夫氏に感謝いたします．

参考文献

[1] Shuya Abe, Kentaro Inui, and Yuji Matsumoto. Acquiring event relation knowledge by learning cooccurrence patterns and fertilizing cooccurrence samples with verbal nouns. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing*, pp. 497–504, Hyderabad, India, January 2008.

[2] 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.

[3] Daisuke Kawahara and Sadao Kurohashi. A fully-lexicalized probabilistic model for japanese syntactic and case structure analysis. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pp. 176–183, 2006.

[4] Dekang Lin and Patrick Pantel. Dirt: discovery of inference rules from text. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 323–328, New York, NY, USA, 2001. ACM.

[5] Suguru Matsuyoshi, Satoshi Sato, and Takehito Utsuro. Compilation of a dictionary of japanese functional expressions with hierarchical organization. In *Proceedings of the 21st International Conference on Computer Processing of Oriental Languages*, pp. 395–402, 2006.

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

[7] Viktor Pekar. Acquisition of verb entailment from text. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pp. 49–56, New York City, USA, June 2006. Association for Computational Linguistics.

[8] Idan Szpektor, Eyal Shnarch, and Ido Dagan. Instance-based evaluation of entailment rule acquisition. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 456–463, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

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