事態オントロジー：言語に基づく推論のためのコトに関する基本知識

乾 健太郎
奈良先端科学技術大学院大学 情報科学研究科
inui@is.naist.jp

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．そうした世界知識には，モノだけでなくコト，すなわち事態に関する上位下位関係や部分全体関係，因果関係などの知識が含まれる．

1.1 深い言語理解のための推論

我々が事態に関する知識を使って実現したい推論は大きく 2 種類ある．一つは，近年研究者の注目を集めつつあるテキスト間の含意関係を認識する推論 [5] である．例えば，次の文 (1a) が成り立つならば，文 (1b) が成り立つ．

(1) a. 夏目漱石は明治 39 年の春に『坊ちゃん』を雑誌「ホトトギス」に発表した。
b. 『坊ちゃん』の著者は夏目漱石である。

このようなテキスト間の含意関係[1]を認識する推論は，質問応答や情報抽出，機械翻訳など，広範囲の言語処理アプリケーションにおいて重要な役割を果たすと考えられ，近年，含意関係認識や言い換えに関する研究が急速に広がりを見せている [5, 9]．(1) の含意関係の認識には，例えば (2) のような述語項構造間の関係に関する知識があれば実現できると考えられる．

(2) a. 「X が著作物 Y を雑誌 Z に発表した」なら，それ以前に「人 X は著作物 Y を書いている」はずである．
b. 「人 X が著作物 Y を書いた」ならば,「人 X は著作物 Y の著者である」.

もう一つの推論は，人間の行動やそのプランに関する推論である．例えば,「庭に洗濯物を干したとたんに雨が降ってきた」という文が望ましくない事態であることを推論するためには，

(3) a. 「洗濯物を干した」行為の目的が「洗濯物を乾かす」ことであって，
b. その目的が「雨」のために達成されなかった，

といったことを認識する必要がある．このような推論はプラン認識と呼ばれ，対話プランニングや人の行動支援に有用である．こうした推論を実現するには，人の行動に関する目的，手段，条件などの知識が必要と考えられる．(3) の例では,〈ものを干す〉と〈ものが乾く〉が行為-効果の関係にあることや,〈雨が降る〉と〈ものが乾く〉が負の因果関係にあるといった事態間の関係に関する知識が必要である．

1.2 コーパスからの事態知識の獲得

こうした高度な言語理解の研究は 1980 年代に遡るが，当時の研究はそれに必要な大量の言語知識や常識的世界知識を用意する方法論を持たなかったために，実用規模には発展しなかった．しかし，現在は次の 2 点で状況が大きく異なる．

- 統計的言語処理の発展によって，形態素・係り受け解析や照応・省略解析などの基盤技術が実用的な水準に達してきた．
- 情報技術が日常生活に浸透しブログのような情報発信手段が普及したことにより，人間の日常的行動を記述した文書が大量に入手可能になった．この傾向は今後さらに顕著になると予測される．

こうした背景から，一定の水準に達した基盤的言語解析技術を用いて大量の日常行動の記録から実用規模の常識的世界知識を自動獲得し，獲得した知識を使ってより高度な推論を必要とする言語理解にアプローチするという，いわば「コーパスベース知識工学」とでも呼びうる新しいパラダイムの可能性が拓けてきた [15]．実際，大規模コーパスからこうした関係を自動獲得する研究が進んでおり，一定の成果を上げている [11, 17, 4]．ただし，こうしたパラダイムを現実のものとするには課題も多い．最大の問題はデータの過疎性である．これら既存の知識獲得手法はいずれも,「〜したため〜した」のような特定の共起パターンに基づく述語間の共起情報に頼るものであり，知識獲得を補助する基本的な知識がなければ，Web サイズのデータを以てしても過疎性の問題から逃れ得ない．その最大の原因は言語表現の多様性にある．これを解決するには，例えば「〜を生産する」,「〜を作る」,「〜の製造」といった表現が近似的に同義であることを認識し，表現の些末な差を吸収する仕組みが必要である．

我々は，こうした事態に関する知識の総体を広く**事態オントロジー**と呼び，その構築に (a) 動詞語釈文の構造化，(b) 語彙概念構造に基づく事態上位オントロジーの開発，(c) コーパスからの知識獲得の 3 方向からアプローチする研究プロジェクトを進めている．以下，その概要と現状を報告し，今後の方向性を展望する．

[1] ここで考える含意関係は，形式論理における論理的含意関係を指すものでは必ずしもなく，一方が成り立てば他方が成り立つ可能性が高まるという程度の弱い関係も含む [5].

2 知識をどのように表現しておくか

前節の議論からも明らかなように，事態に関して必要な知識の中心は事態間の関係である．事態間関係は，関係に立つ２つの事態（タイプ）と関係を表すラベルで表現できる．関係を表すラベルについては次節に譲り，ここではまず事態を表す表現について議論する．

事態を表す言語表現は,「雨」や「運動会」のような語から,「友人に手紙を送る」のような述語項構造,「酒を飲みすぎて酔っぱらう」のような複数の述語項構造の組み合わせ，そしてそれらをさらに組み合わせた談話構造に至るまで多岐にわたる．このうち事態を表す単位と見なせるのは,「雨」のような非動詞由来の事態性名詞と「友人に手紙を送る」のような述語項構造である．さらに，非動詞由来の事態性名詞についても，例えば「雨」は「雨が降る」という述語項構造で表現される事態を指すといったように述語項構造で表現できる．そこで，事態を表す表現の単位として述語項構造を用いることを考える．これには次の利点がある．

第１に，述語項構造で事態を表現するということは，自然言語に用意された動詞や名詞など，オープンクラスの語彙を基本的にそのまま利用することを意味する．このように自然言語の語彙を知識表現に用いるアプローチは，個々の事態概念やモノ概念に記号を与えて人工的で形式的なオントロジーを用意するアプローチに比べ，形式性が確保できないという欠点はあるものの，オントロジーの開発コスト，および新しい概念に対する頑健性の面で優れて有利である．また，コーパスからの知識獲得においても，獲得した知識を利用した言語理解においても，知識を自然言語の語彙で表現しておく方が技術的な負荷が軽い．

第２に，述語項構造の表現レベルでは，動詞句の名詞句化や連体節化，照応/省略などによって表層的な統語構造に生じる差異を吸収し，標準的な表現に統一することができる．例えば,「夏目漱石によって発表された『坊ちゃん』」や「夏目漱石による『坊ちゃん』の発表」は統語構造では異なるが，述語項構造では「夏目漱石ガ『坊ちゃん』ヲ発表する」に統一される．また,「事件がマスコミの注目を集める」や「EU が CO_2削減を推進する」のように名詞化された事態の項にギャップが存在する場合でも，述語項構造のレベルでは「マスコミガ事件ニ注目する」,「EU ガ CO_2 ヲ削減する」のように述語と項の関係が特定される．

第３に，述語項構造解析の技術が，英語では PropBank や NomBank，また日本語でも京都テキストコーパス第 4.0 版や GDA コーパス，NAIST テキストコーパス [7] などの資源の蓄積とともに近年急速に発展してきており [8, 13, 14]，述語項構造を処理の単位とするアプローチに技術面からも現実性が出てきた．

3 どのような事態間関係を対象にするか

一口に事態間関係と言っても，我々が世界について意識的に，あるいは無意識に持っているあらゆる知識をやみくもに収集することを目指すのは現実的とは思われない．どのような種類の知識の収集に注力すべきかをガイドする何らかの方針が必要であろう．1.1 で述べたように，我々の当面の目的は，事態間の含意関係を認識したり，人間の行動やそのプランについて推論する計算モデルと事態間関係知識を構築することである．この目的に照らせば，少なくとも次の点に注意するべきであると考えられる．

- 含意関係認識に直接的に有用なのは，事態 E が事態 F を含意するという関係に立つ事態対 $E \rightarrow F$ の知識であろう．我々はそうした含意関係にある事態対を優先的に収集したい．
- 上の含意関係の中には，一方が他方を必然的に含意する関係もあれば，帰結が蓋然的でしかないものもある．含意関係認識の目的に照らせば，必然的含意関係は他の蓋然的含意関係と区別して収集しておくのが望ましい．
- 人間の行動やそのプランに関する推論では，まず人が意志的に行う行為に関連する知識が必要になる．プラン認識の研究では，行為に関する知識として，行為の目的，手段，前提条件などの知識を仮定するのが一般的である．こうした知識は，それが無ければ行為の達成が困難になる，あるいはそもそも行為が意図されないという意味で，人の行為に関する推論という我々の目的にも不可欠な情報と考えられる．
- 行動やプラン関する推論では，個々の事態が意志的な行為，すなわち条件が整えば行為主体が意志的に行える行為（以下単に「行為」）であるか，それ以外の種類の事態（以下「出来事」）であるかを知っている必要がある．事態間関係の分類おいても行為と出来事が区別されていることが望ましい．
- 含意関係認識や人の行動に関する推論では，過去または現在の「事実としての事態」と未来の「可能性としての事態」を区別する必要が出てくる．前者は変えることができないのに対して，後者は蓋然的でしかなく，人の行動によって変わる可能性があり，そのことが人の行動を動機づける場合もある．したがって，事態間関係は，関係に立つ事態対の時間的な前後関係を区別するものであることが望ましい．

以上を勘案すると，我々がまず集めたいと考える事態間関係は例えば次のようなものである．ただし，ここに記した関係の分類は上述の観点に沿って例を整理するための仮の分類である．事態間関係にどのような分類を認めるのが適当かは，実際の推論における正確性と頑健性，さらに知識獲得の現実性に基づいて今後検討していく必要がある．

同義（near-synonymy）

(4) a. X ガ喘ぐ→ X ガ息を切らす
　　b. X ニ足を運ぶ→ X ニ立ち寄る

上位（troponymy[6]）

(5) a. X ガ激増する→ X ガ増える
　　b. X ヲ Y ニ置く→ X ヲ Y ニ動かす

行為の必然的結果として起こる出来事

(6) a. X ヲ増やす→ X ガ増える
　　b. X ヲ Y ニ動かす→ X ガ Y ニある

手段である行為の目的

(7) a. X ガ新聞を広げる→ X ガ新聞を読む
b. X ガ店で $Y_{料理}$ を注文する→ X ガ Y ヲ食べる
c. X ガ $Y_{映画館}$ ニ行く→ X ガ Y デ映画ヲ観る

行為の蓋然的効果として起こる出来事

(8) a. 照明を付ける→明るくなる
b. X ガ運動する→ X ガ汗をかく

出来事の蓋然的結果として起こる出来事

(9) a. X ガ汗をかく→ X ガやせる
b. 雨が降る→洗濯物が乾かない

行為の準備として事前に行う行為

(10) a. X ガ $Y_{本}$ ヲ読む→ X ガ Y ヲ買う
b. X ガ $Y_{商品}$ ヲ売る→ X ガ Y ヲ製造する

行為を実行するための前提

(11) a. X ガ Y ニ $Z_{事実}$ ヲ教える→ X ガ $Z_{事実}$ ヲ知っている
b. X が $Y_{会社}$ ヲ退職する→［その前に］X ガ Y ニ勤めていた

出来事が成り立つための前提

(12) a. $X_{試験}$ ニ合格する→［その前に］X ヲ受けた

4 事態オントロジー構築の試み

1 節で述べたように，大規模なコーパスからの知識獲得の道が開けてきたとは言え，Web 規模のコーパスを以てしてもデータの過疎性は深刻な問題であり，それを埋める補助的な知識が必要である．これに対し，我々は次の 3 つの方向からアプローチし，それらの組み合わせによって事態オントロジーの構築をめざす．

4.1 動詞語釈文の構造化

第 1 のアプローチは，国語辞典の語釈文を利用するもので，語釈文に対し述語項構造と意味関係の情報を注釈付けすることによって，述語項構造によって表現される事態間の上位下位関係だけでなく，手段-目的関係や行為-前提関係など，多様な関係を収集する．例えば，岩波国語辞典によると，動詞「倒す」の意味は (13a) の語釈文で与えられており，この記述から (13b) の上位下位関係だけでなく，(13c) の手段-目的関係や (13d) の行為-前提関係など，多様な関係を収集することができる．

(13) a. 倒す ＝ 立っている物に力を加え傾け、横にする（岩波）
b. X が Y を倒す －上位→ X が Y を横にする
c. X が Y を倒す －手段→ X が Y に力を加える
d. X が Y を倒す －前提→ （行為前は）Y が立っている

文献 [2] で別途報告するように，現在「岩波国語辞典コーパス 2004」として入手可能な岩波国語辞典第 5 版の収録動詞（11469 語，17104 語義）について語釈文から述語項構造を抽出し，見出し語の述語項構造との意味関係（上位・同義，結果状態，前提条件，付帯状況，手段，目的，反義，不可分の 8 種類），および項の対応関係を記述する作業を進めている．これまでに得られた結果から，作業者 1 人週 30 時間あたり約 1000 語義のペースで作業進んでおり，コスト面でも十分見合うこと，また関係の分類は上述の 8 種類であれば作業者間でゆれなく安定して記述できることがわかっている．こうした作業により，事態間の含意関係に関する基本的な知識がノイズの少ない状態で得られるとともに，手段や前提といった世界知識の一部を収集することができる．後者は 4.2 で述べる事態間関係獲得のシードにも利用できると考えている．また，岩波国語辞典コーパス 2004 には，語釈文中の各語に同辞書の語義情報が付与されており，この情報と上述の構造化情報を組み合わせれば，含意関係計算に有用な基礎資源になるものと期待している．

辞書の語釈文から述語間の関係を抽出する研究は，MindNet[18] や鍜治ら [12] の言い換えモデルなど，すでにいくつか報告がある．こうした研究はいずれも語釈文を自動解析し，全自動で知識を抽出することを試みている点では我々の試みに比べコスト面で有利である．ただし，扱っている関係は専ら上位下位関係あるいは同義関係であり，手段や目的，前提といった多様な関係を抽出する試みは報告がない．

述語間の意味関係を人手で記述した代表的な資源に WordNet[6] と FrameNet[3] がある．WordNet には我々の分類に近い意味関係が記述されており，FrameNet もフレーム間に同様の意味関係が記述されている．ただし，WordNet には項の対応関係の記述は一切なく，FrameNet でもフレーム間をまたぐ述語間については項の対応関係は原則として特定されていない．これら既存の資源に対し，我々が作成中の資源は，述語項構造の対を関係の単位として項の対応関係を記述するとともに，従来よりも詳細に意味関係を分類している点に資源としての価値があると考えている．

4.2 語彙概念構造に基づく事態上位オントロジー

第 2 に，語彙概念構造（LCS）と呼ばれる述語意味記述を導入し，これを事態の上位オントロジーと見なして，述語の集合をその下に分類する．これにより，各述語には「意識的な行為か」,「状態変化を含むか」といった意味的な特性が付与される．例えば「X ガ Y ヲ Z ニ置く」は「X の意志的な活動が Y に作用し，Y の位置変化を引き起こす」という述語タイプに分類され，(5b) や (6b) の関係も得られる．これまでに日本語基本動詞約 4 千語について意味記述を行っており [16]，最終的には約 1 万語の動詞意味記述辞書を開発する予定である．LCS に基づく述語分類には次のような利点がある [10].

- LCS は述語を意志性や状態変化性のような意味的な特性で多元的に分類する．このことは複雑な述語の意味を分類するのに都合がよく，またある意味特性を固定したまま別の特性を細分化したり，定義修正することができるため．辞書の漸進的な洗練が可能になり資源開発に有利である．
- LCS では，語の統語的振る舞いをその語が持つ意味的な分類から予測する．ある語がある意味特性を持つかどうかは，その意味特性に対応する統語的振る舞いの観察によって客観的に判断できるので，辞書の開発過程に作業者の恣意性が入りにくい．
- 含意関係計算にとって重要な述語の統語的振る舞い（アスペクトや態に関する特性，動詞交替，複合語

など）と意味特性の関係が LCS の研究者らによってすでに活発に調査されており，その成果を利用することができる．

4.3 コーパスからの事態間関係獲得

「〜したため〜した」のような特定の共起パターンを使って大規模なテキストデータから事態間関係知識を自動的に獲得する試みにはすでにいくつかの報告がある [4, 11, 17]．こうした研究から明らかになったのは，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のが望ましいということである．「〜ため〜」のような一般的なパターンは出現頻度が高く，多様な事態と共起するが，知識獲得のノイズも大きい．一方，特殊なパターンを使うと，より意味的に制約された事態対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，獲得する知識のカバレージを確保するのが難しい．この問題に対し，我々は，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を獲得する方法を研究している [1]．(14) は行為-効果関係の獲得に寄与した共起パターンと獲得できた関係の例である．

(14) a. 〜を経て〜します，〜して〜を招く，〜で〜するのか，〜させるために〜します，〜しておくと〜する，〜を目指して〜する，〜による〜が
b. 入院する→全快する，休む→回復する，酌む→和む，飲む→酔う，活動する→自分が納得する，デートする→会う，口説く→成功する，目指す→完成する，学ぶ→分かる，聞く→分かる，戦闘する→死ぬ，運動する→息切れする，説明する→安心する，勉強する→合格する

今後は，4.1，4.2 で述べた資源を用いて事態表現の多様性を吸収し，また事態の分類情報を有効に活用することによってデータの過疎性に対処する方法を検討する予定である．

5 おわりに

言語処理技術の進化をはかるには，含意関係認識やプラン認識など，事態に関する推論を実現する計算機構と事態に関する膨大な知識が必要である．本稿では，知識の問題に対し，(a) 動詞語釈文の構造化，(b) 語彙概念構造に基づく事態上位オントロジーの開発，(c) コーパスからの知識獲得の 3 方向からアプローチする我々の試みを報告した．開発中の資源はいずれも順次研究用途に広く公開していく予定である．詳細は http://cl.naist.jp/nldata の該当項目を参照されたい．

謝辞

本研究は，文科省科研費基盤研究 (B)「語彙意味論に基づく言い換え計算機構の工学的実現と言い換え知識獲得への応用」（17300047，代表：乾健太郎），科研費基盤研究 (A)「円滑な情報伝達を支援する言語規格と言語変換技術」（16200009，代表：佐藤理史）の支援を受けている．ニューヨーク大学の関根聡氏，北陸先端科学技術大学院大学の鳥澤健太郎氏，奈良先端科学技術大学院大学の松本裕治氏，名古屋大学の佐藤理史氏，京都大学の黒橋禎夫氏には本研究に対し貴重な助言をいただいた．また，岡山大学の竹内孔一氏，名古屋大学の藤田篤氏，言語アナリストの竹内奈央氏，甲南大の中谷健太郎氏，奈良先端大の阿部修也氏，青山桜子氏，大西良明氏には多大な協力をいただいた．記して深く感謝する．

参考文献

[1] 阿部修也, 乾健太郎, 松本裕治. 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. 言語処理学会第 13 回年次大会発表論文集, 2007.

[2] 青山桜子, 阿部修也, 乾健太郎, 松本裕治. 事態間関係の獲得のための動詞語釈文の構造化. 言語処理学会第 13 回年次大会発表論文集, 2007.

[3] Collin F. Baker, Charles J. Fillmore, and John B. Lowe. The berkeley framenet project. In *Proceedings of COLING-ACL*, pp. 86–90, 1998.

[4] Timothy Chklovski and Patrick Pantel. VerbOcean: Mining the Web for fine-grained semantic verb relations. In *Proc. of EMNLP*, pp. 33–40, 2004.

[5] Ido Dagan, Oren Glickman, and Bernardo Magnini. The pascal recognising textual entailment challenge. In *Proc. of the PASCAL Challenges Workshop on Recognising Textual Entailment*, 2005.

[6] Christiane Fellbaum. A semantic network of English verbs. In *Christiane Fellbaum (ed.), WordNet: An Electronic Lexical Database*. The MIT Press, 1998.

[7] 飯田龍, 小町守, 乾健太郎, 松本裕治. 述語項構造と共参照関係のアノテーション: Naist テキストコーパス開発の経験から. 言語処理学会第 13 回年次大会発表論文集, 2007.

[8] Ryu Iida, Kentaro Inui, and Yuji Matsumoto. Exploiting syntactic patterns as clues in zero-anaphora resolution. In *Proc. of COLING-ACL*, pp. 625–632, 2006.

[9] 乾健太郎, 藤田篤. 言い換え技術に関する研究動向. 自然言語処理, Vol. 11, No. 5, pp. 151–198, 2004.

[10] 乾健太郎, 藤田篤. 言い換え計算モデルの工学的実現への語彙意味論的アプローチ. レキシコンフォーラム, Vol. 2, pp. 27–56, 2005.

[11] 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.

[12] 鍜治伸裕, 河原大輔, 黒橋禎夫, 佐藤理史. 格フレームの対応付けに基づく用言の言い換え. 自然言語処理, Vol. 10, No. 4, pp. 65–81, 2003.

[13] Daisuke Kawahara and Sadao Kurohashi. A fully-lexicalized probabilistic model for japanese syntactic and case structure analysis. In *Proc. of HLT-NAACL*, pp. 176–183, 2006.

[14] 小町守, 飯田龍, 乾健太郎, 松本裕治. 事態性名詞の項構造解析における共起尺度と構文パターンの有効性の分析. 言語処理学会第 13 回年次大会発表論文集, 2007.

[15] 関根聡, 乾健太郎, 鳥澤健太郎. 提唱「コーパスベース知識工学」. 言語処理学会第 13 回年次大会併設ワークショップ「大規模 Web 研究基盤上での自然言語処理・情報検索研究」, 2007.

[16] 竹内孔一, 乾健太郎, 藤田篤, 竹内奈央. 語彙概念構造に基づく事態上位オントロジーの構築. 言語処理学会第 13 回年次大会発表論文集, 2007.

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

[18] L. Vanderwende, G. Kacmarcik, H. Suzuki, and A. Menezes. Mindnet: An automatically-created lexical resource. In *Proc. of HLT/EMNLP 2005 Interactive Demonstrations*, 2005.