

# 事態間関係知識の整備と含意文生成への応用

大西良明 乾健太郎 松本裕治

{yoshiaki-o, 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

The PASCAL Recognising Textual Entailment(RTE) Challenge[1]に代表されるように、含意関係の認識は近年研究者の注目を集めている。含意関係<sup>1</sup>とは(1)の例文において文(1a)が成り立つときに文(1b)も成り立つような関係を指す。

(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。  
含意関係を認識する技術は、質問応答や情報抽出、機械翻訳など多くの言語処理アプリケーションで重要な役割を果たすことが期待される。

(1)のような含意関係を認識するためには、モノに関する知識の他にコト、すなわち事態についても(2)のような上位下位関係や部分全体関係、因果関係など様々な知識が必要である。

(2) a. XがYヲZニ発表する→XがYヲ書く

b. XがYヲ書く→Yの著者はXだ

そこで本稿では、このようなコトに関する知識の構築に関して、これまでに行ってきた国語辞典からの事態間関係の抽出とその知識を用いた頑健な照合を行うための整備について述べる。また、構築した事態間関係を含意文生成の規則として利用することで、構築した事態間関係そのものの質を評価する。

以下、2節で関連研究を述べ、3節で我々が行った事態間関係の整備について説明する。4節でその知識を用いた含意文生成の評価について報告し、最後に5節でまとめと今後の課題を述べる。

## 2 関連研究

事態間関係を辞書から抽出したリソースとしては、MindNet[6]がある。Mindnetは辞書と百科事典の解析を行い、意味関係構造に変形することで、見出し語と定義文に関係を付与する。動詞語積文を用いた言い換え手法として鍛治らの言い換えモデル[3]がある。鍛治らの言い換えモデルでは国語辞典の見出し語を定義文の主辞で置き換える際に、大規模コーパスから格フレームを学習し、見出し語と定義文の主辞の格フレームを対応付けし、用言の言い換えを行う。MindNetや

<sup>1</sup>ここでいう含意関係とは論理的含意ではない。このような形式的でない定義は含意関係が一般的な人間の言語理解に基づいている為である[1]

鍛治らの手法は語積文を自動解析し、知識を抽出することを試みている。ただし、MindNetでは事態間の関係は分類されているものの、項の対応関係については捉えられておらず、事態間の含意関係を認識するために必要な項の照合を行うことができない。また、鍛治らの手法は項の対応関係も扱っているが、関係は上位関係・言い換え関係に限定されている。これらに対し、我々は人手で国語辞典の動詞語積文を用いて見出し語と定義文中に出現する動詞の項との対応をつけ、8つの関係を付与した[7]。このようにして作成した資源は、人手で作成しているためノイズが少なく関係も多岐に渡るため、含意関係認識のための推論だけでなくプラン認識のための推論など多くの推論に必要な基礎知識となる。

## 3 事態間関係の整備

2節で述べたように、項の対応関係を考慮し広く事態間の関係を扱っている資源を整備することは有用であると考えられる。そこで本節では、まず事態を表現する単位について述べ、国語辞典における動詞語積文から事態間関係を抽出した手法を説明する。また、竹内らの動詞意味分類[9]を用いた格交替情報の付与について述べる。

### 3.1 事態表現の単位について

事態表現は「雨」や「運動会」のような語、「コンビニで本を買う」のような述語項構造、「毛布がなくて寒い」のような述語項構造を組み合わせたもの、さらにそれらを組み合わせたものなど多岐に渡るが、我々は事態を表す単位として述語項構造を用いることを考える。

述語項構造を用いた表現は、様々な事態を表す統語構造間に生じる差異を吸収し、正規化された表現が可能となる。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」や「夏目漱石は『坊ちゃん』を発表した」は統語構造が大きく異なるが、これらを述語項構造による表現にすれば[[夏目漱石ガ][『坊ちゃん』ヲ]発表する]に統一される。より詳しい説明は文献[8]を参照されたい。

### 3.2 国語辞典の動詞語釈文を用いた事態間関係の抽出

「岩波国語辞典タグ付きコーパス 2004<sup>2</sup>」を用いて事態間関係の抽出を行った。岩波国語辞典タグ付きコーパスでは、国語辞典の語釈文に出現する名詞と述語に語義情報が付与されており、この情報を利用することで、動詞見出し語と語釈文中の述語の間に語義レベルでの対応を付けることができる。我々はこのコーパスを用いて以下の手順で事態間関係の抽出を行った。

(i) 見出し語に必須と思われる項をつけ、それらの項に変数を代入する。

(ii) 語釈文中の述語(定義語)に関しても(i)と同様の作業を行うが、その際に見出し語の項構造に代入した変数と、新たに作成する項構造間に代入する変数の対応関係をつける。

(iii) 作成した述語項構造間に関係を付与する。(ii)で作成した述語項構造間につける関係を、1. 述語項構造間の時間的関係 2. 含意関係が必然的であるか、蓋然的であるか 3. 事態の種類(出来事か行為か)の観点から分類し、表2に示すものを含めて8つの関係を定義した。より詳しい関係の定義は文献[7]を参照されたい。

### 3.3 動詞意味分類を用いた格交替情報付与

語釈文から抽出した事態間関係は表層格で事態間の項の対応が表現されているため、実際に入力された述語項構造をこの知識に適用する際に、格交替などの現象により適切に照合できない可能性がある。

- (3) a. 入力：彼が技を師匠に学ぶ  
b. 事態間規則：XガYヲZカラ学ぶ→ZガXニYヲ教える

そこで我々は竹内らの動詞意味分類[9]を用い、事態間関係に格交替の情報を付与する。竹内らの動詞意味分類には、動詞の各語義がどのような意味役割をとり、それらがどのような表層格で言語化されているかが記述されている。また、この動詞意味分類には、それぞれの動詞に対して深層格が付与されている。この資源を用いて格交替情報を追加した事態間関係の例を表1に示す<sup>3</sup>。

表1の「罰を与える」のように、関係を保持するために必要な項は埋め込んで記述している。また変数の対応関係が間接的な場合は表1の「思い付く」の「[Yの心<sub>対象</sub>]ニ」のように記述している。そのほか、作成した資源には今までに述べた変数の対応関係、事態間関係、格交替情報以外にも見出し語事態の各項には意味役割、代表的な名詞を記述している。また各事態対には代表的な例文が付与されている。

### 3.4 現在の事態間関係の整備状況

現在、国語辞典に登録されている動詞、サ変名詞14,082語、23,955語義から合計33,770個の事態間関係を抽出した。それらの内、竹内らの動詞意味分類によ

<sup>2</sup>http://www.gsk.or.jp/catalog.html

<sup>3</sup>表にある、見出し語事態とは、(i)で作成した見出し語の項構造を指し、定義語事態とは(ii)で作成した定義語の項構造を指す

表 1: 事態間関係の一例

意味関係	前提条件
見出し語事態	[X <sub>動作主</sub> ]ガ[Y <sub>対象=運転</sub> ]ヲ[Z <sub>起点</sub> ](ニ カラ)学ぶ (49000-0-0-1-0)
定義語事態	[Z <sub>動作主</sub> ]ガ[X <sub>着点(人)</sub> ]ニ[Y <sub>対象</sub> ]ヲ教える (5441-0-0-1-0)
代表例文	生徒が心理学を教授から学ぶ
意味関係	同義・上位
見出し語事態	[X <sub>動作主</sub> ]ガ[Y <sub>人</sub> ]ヲ罰する (41486-0-0-0-0)
定義語事態	[X <sub>経験者</sub> ]ガ[Y <sub>対象</sub> ]ニ罰を与える (755-0-0-3-0)
代表例文	大学が学生を罰する
意味関係	同義・上位
見出し語事態	[X <sub>経験者</sub> ]ガ[Y <sub>生成物=名案</sub> ]ヲ思い付く (21024-0-0-2-0)
定義語事態	[Y <sub>生成物</sub> ]ガ[Xの心 <sub>対象</sub> ]ニ浮かぶ (3396-0-0-2-0)
代表例文	漫才師がネタを思い付く

り整備されている8,846語義に対しては意味役割と格交替の情報が付与されている。また、述語項構造で表現することができる慣用表現1,765個についても同様に事態間関係として抽出した。

## 4 含意文生成への応用

この節では、入力された述語項構造と含意関係にある文を生成するタスクに事態間関係を適用することで構築した知識そのものの品質を評価すると同時に頑健な含意文生成の為に何が必要となるかを考察する。

### 4.1 含意文生成タスクの概要

本稿で述べる含意文生成のタスクは、入力された述語項構造に事態間関係を適用することでどの程度の規模の含意文を生成することができ、生成された文の適格性がどの程度なのかを調査する。また事態間関係に適用することで得られた述語項構造のうち、不適切な表現に関しては選択選好の情報を用いて棄却する課題にも取り組む。例えば、

- XガYヲZニ与える→XガYヲZニ渡す  
に対し、以下の二つの述語項構造を考える。

- (4) a. 母がおもちゃを息子に与える  
b. 母が母乳を息子に与える  
これらを上記の事態間関係に適用すると以下の文が生成されるが、文(5b)のように不適切な文を生み出す場合がある。

- (5) a. 母がおもちゃを息子に渡す  
b. \*母が母乳を息子に渡す  
これは特に代入される変数に制約を加えていないため、「母乳を渡す」のような不適切な表現がされてしまったことによる。つまり、正しい含意文のみを生成するためにはこのような不適切な文を何らかの形で棄却する必要がある。

この問題に対してPantelら[5]は、動詞の選択選好性を適格さの尺度と捉え、推論規則から出力された文の制御を行った。我々もこれに従い、動詞の選択選好性を用いて生成された文を棄却するモデルを用いる。

表 2: 文献 [7] でまとめた事態間関係と例

関係名	時間的關係	必然性	見出し語事態の種類	定義語事態の種類	例
上位・同義	同時	必然的	出来事/行為	出来事/行為	X が Y を使う → X が Y を使う X が漂流する → X が流れる
付帯状況	同時	蓋然的	出来事	出来事/行為	X が流行る → X が次々と伝わる X が Y を無視する → X が Y の価値を認めない
手段	同時	蓋然的	行為	行為	X が稼ぐ → X が働く X が Y を炙る → X が Y を火に当てる
結果 (状態)	定義語事態が後	必然的	出来事/行為	出来事	X が Y を傾ける → Y が傾く X が Y を起こす → Y が立つ
目的	定義語事態が後	蓋然的	行為	出来事/行為	X が Y をメモする → X が Y を忘れない X が Y を味見する → X が Y の味加減を知る
前提条件	定義語事態が前	必然的	出来事/行為	出来事	X が言い逃れる → X が問い詰められる X が Y を迎え撃つ → Y が攻める

#### 4.2 適格性による不適格な文の棄却

大規模なコーパスから係り受け関係にある〈名詞 ( $n$ ), 格助詞 ( $c$ ), 動詞 ( $v$ )〉の3つ組の頻度を用いて, 文の適格さを求めるモデル (適格性モデル) を構築する. 具体的には,  $n$  と  $\langle c, v \rangle$  の自己相互情報量 ( $PMI$ ) を動詞の選択選好性の尺度として, 述語項構造のそれぞれの格要素の適格性を求める. しかし,  $\langle v, c, n \rangle$  の組み合わせが広大なため,  $P(\langle v, c, n \rangle)$  を推定する際にデータスパースネスの影響があると考えられる. そこで pLSI[2] を用いて, 確率のスムージングを行った. これにより推定された  $P(\langle v, c, n \rangle)$  を用いて求めた  $PMI(\langle v, c, n \rangle)$  が 0 未満の項を含む文を不適格な文とし棄却する.

#### 4.3 評価実験

述語項構造の入力に, 事態間関係を適用することで生成された文に対し適格性モデルを用いて適格さを求め, 不適格な文を棄却する. 棄却されなかった文を含意関係のある文と認め, その評価を適合率, 再現率, F 値により行う.

我々は新聞記事を基にした NAIST テキストコーパス [10] の一部から, 項が埋められている述語項構造 214 事例を入力とし, 生成された 573 事例 (以下, 新聞データ) に対し適格性による棄却を試みた. また代表例文から選んだ 1,520 事例<sup>4</sup> の述語項構造に関しても事態間関係を適用し, 生成された 6,564 事例 (以下, 作例データ) に対し適格性モデルによる棄却を試みた. さらに, それぞれの事態間関係には語義の情報が付与されているので, 入力文の語義曖昧性が既に解消されている場合は対応する語義の事態間関係にのみ用いて生成することも考えられる. この場合にどのように含意文生成の結果が変化するか調査する為に, 入力文の語義 ID を参照し, 対応する事態間関係のみを適用した場合についても, 生成した 2921 事例 (以下, 曖昧性解消作例データ) に対し適格性モデルによる棄却を試みた.

適格性モデルは 5 億文の Web テキスト [4] より作成した. また, ベースラインは適格性モデルによる不適

<sup>4</sup>動詞は語義数が 5 個前後の動詞, かつ 5 億文の Web データで頻度が 1,000~10,000, 10,000~100,000, 100,000~1,000,000 からそれぞれランダムに選んだ

表 4: 作例データにおける含意文生成の結果の例

入力文	出力文	モデルの判定	真の正解
僕が酒を味わう	僕が酒を楽しむ	○	○
旅館が夕食を用意する	旅館が夕食に備える	○	○
ウサギが穴に隠れる	ウサギが穴にいる	○	○
警察が遺留品を見つける	警察が遺留品を探す	○	○
彼女がカーテンを閉じる	彼女がカーテンを終わりにする	○	×
僕が論文を手入れる	僕が論文を保存する	○	×
店員が宝石を扱う	店員が宝石を操作する	○	×
カメラマンが女優の顔をアップする	女優の顔が上がる	○	×
ピアニストがショパンを演奏する	ピアニストがショパンを奏する	×	○
師匠が弟子を育てる	弟子が一人前になる	×	○
僕が会議を閉じる	僕が会議を終わりにする	×	○

切な文の棄却は行わず, 事態間関係に適用して生成された全ての文を含意文とみなした場合である. 結果を表 3 に示す.

新聞データ, 作例データ共に適合率, F 値でベースラインを上回った. 作例データより新聞データの方が適合率, F 値共によいスコアとなったが, これは作例データは語義数が 5 個前後のものを中心に作成していることで, 新聞データより語義に曖昧性があることに起因すると考えられる. 曖昧性解消作例データでは適合率は上がったものの F 値はベースラインを下回った. これは曖昧性が解消されていることで, 不適格な文がそもそも少ないことに起因すると考えられる. 表 4 に含意文生成の結果の例を示す. 作例データでは含意関係がないにも関わらず, 含意関係があると判定された事例が 1547 事例存在した. また含意関係があるにも関わらず, 含意関係がないと判断された事例も 601 事例存在

表 3: 適格性モデルを用いた含意文生成の実験結果

対象データ	モデルの適用	適合率	再現率	F 値
新聞データ	なし	58.1%(333/573)	100%(333/333)	0.735
新聞データ	あり	74.2% (270/364)	81.0%(270/333)	0.775
作例データ	なし	48.0%(3151/6564)	100%(3151/3151)	0.649
作例データ	あり	62.2%(2550/4097)	81.3%(2550/3151)	0.704
曖昧性解消作例データ	なし	73.7%(2153/2921)	100%(2153/2153)	0.849
曖昧性解消作例データ	あり	84.6%(1711/2023)	79.5%(1711/2153)	0.819

した。そこで我々は誤り事例の中からそれぞれ 150 事例ずつサンプリングし、誤り分析を行った。

含意関係はないにも関わらず、含意関係にあると判定された事例によく出現した誤りについて述べる。誤り事例のうち、49 事例が表 4 の「店員が宝石を操作する」の例のように入力に使われた述語項構造に付けられた語義と参照した事態間関係に付けられている語義が異なっているものから生成されたものだった。また「論文を踏み込む」のような本来ならば適格性モデルで棄却されるべき事例が 30 事例存在した。他にも「事故が起きる」が入力され「事故が床から出る」と明らかに不適格な文にも関わらず、「事故が出る」「床から出る」で見ると適格とみなされ、含意文となってしまう事例があった。

含意関係はあるが、含意関係にないと判断された事例に関しては特定の誤りの原因は見られず、ほとんどが適格性モデルで棄却されるべきでない事例であった。中でも動詞「する」が 28 事例「なる」が 10 事例存在した。これらは多くの格・名詞と共起するために適切な見積もりがなされていないと考えられる。このような機能語相当の動詞に関しては、直前格を含ませて一つの述語と考えるなど、より正確な見積もりが行えるような方法を検討していく必要がある。また動詞の表記が平仮名になっているために正確に見積もれない場合もあった。これに対しても今後、表記の標準化を行っていく必要がある。

## 5 まとめ

本稿では、国語辞典の動詞語釈文から抽出した事態間関係を基にした事態間関係の整備について述べた。整備した事態間関係を用いて含意文生成のタスクに適用し知識の品質の評価を行うとともに、頑健な含意文生成のために何が必要となるかを考察した。自己相互情報量  $PMI((v, c), n)$  による適格さを用いての不適格な文の棄却は、語義の曖昧性を考慮しない入力に対しては有効である事が分かったが、曖昧性が解消された入力に対しては F 値を下げる結果となった。

今後の予定は、事態間関係の整備に関しては、竹内らの動詞意味分類のカテゴリ情報を用いることで、上位下位関係に関しての事態間関係を追加していく予定である。また含意文生成のタスクにおいては機能語相

当が含まれている文に対する適格さをより正確に見積もることを考えている。

## 謝辞

本研究は、文化省科研費基盤研究 (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への応用」(17300047, 代表: 乾健太郎) の支援を受けている。ま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「5 億文の Web テキスト」の使用許可をいただいた。記して深く感謝する。

## 参考文献

- [1] Ido Dagan, Oren Glickman, and Bernardo Magnini. The PASCAL recognising textual entailment challenge. In *Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment*, 2005.
- [2] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual international ACM SIGIR conference on Research and Development in information retrieval*, pp. 50–57, 1999.
- [3] 鍛冶伸裕, 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.
- [4] Daisuke Kawahara and Sadao Kurohashi. Case frame compilation from the web using high-performance computing. In *the 5th International Conference on Language Resources and Evaluation*, 2006.
- [5] Patrick Pantel, Rahul Bhagat, Bonaventura Coppola, Timothy Chklovski, and Eduard Hovy. ISP: Learning inferential selectional preferences. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pp. 564–571, April 2007.
- [6] Stephen D. Richardson, William B. Dolan, and Lucy Vanderwende. MindNet: acquiring and structuring semantic information from text. In *Proceedings of the 17th international conference on Computational linguistics*, pp. 1098–1102, 1998.
- [7] 青山桜子, 阿部修也, 大西良明, 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*, pp. 286–289, 2007.
- [8] 大西良明, 乾健太郎, 松本裕治. 言語表現された推論知識による帰結文生成の技術的課題. *情報処理学会 自然言語処理研究会 NL-180-7*, pp. 55–60, 2007.
- [9] 竹内孔一, 乾健太郎, 竹内奈央, 藤田篤. 意味の包含関係に基づく動詞項構造の細分類. *言語処理学会第 14 回年次大会*, 2008.
- [10] 飯田龍, 小町守, 乾健太郎, 松本裕治. NAIST テキストコーパス: 述語項構造と共参照関係のアノテーション. *情報処理学会 自然言語処理研究会 NL-177-10*, pp. 71–78, 2007.