述語項構造の共起情報と節間関係の分布を用いた事態間関係知識の獲得

大友 謙一　　柴田 知秀　　黒橋 禎夫
京都大学大学院情報学研究科
{ken_ichi, shibata, kuro}@nlp.kuee.kyoto-u.ac.jp

1　はじめに

RTE(Recognizing Textual Entailment: テキスト含意関係認識) が自然言語処理の分野で注目されている。RTE が高い精度で行われるためには様々な技術や知識が必要である。例えば、構文解析や共参照解析、照応解析などの技術や、体言に関する知識、用言と体言の間の知識、事態間に関する知識などが挙げられる。本研究では事態間の関係に関する知識 (以降、**事態間関係知識**と呼ぶ) を獲得する。

従来の事態間関係知識の獲得手法には大きくわけて 2 つある。1 つ目は述語項構造の項の分布類似度を指標に知識獲得を行うものである [1]。例えば「X ヲ焼く ⇒X ガ焦げる (X：パン、肉…)」という項が共通である知識が獲得される。しかし、この手法では、項が共有されない知識、例えば「晴れる ⇒ 天気が良い」という知識は獲得することができない。2 つ目は、共起パターンを用いて事態間の関係の分類に行うものである [4]。共起パターンとは「PA_1 ため PA_2 (行為-効果関係)」といった関係獲得を目的に設計されたパターンであり、「パンを焼いたため焦げた」という文に対して上記のパターンを用いると、「焼く ⇒ 焦げる (行為-効果関係)」という知識が獲得できる。しかし、共起パターンで獲得できる知識の量は限られており、カバレッジが低いことが問題となる。

これらの問題を解決するために、本研究では述語項構造の項と用言の共起情報と節間関係の分布を用いて事態間関係知識を獲得する。本研究は、項が共有される知識だけでなく、項を共有しない知識も対象とした事態間関係知識の獲得を行う。本研究の概要を図 1 に示す。まずコーパスから係り受け関係にある述語項構造を抽出する。次に、高頻度で出現する「順接」の節間関係にある述語項構造ペアに対して、述語項構造が行為か出来事かによって 4 つにあらく分類する。そして、それぞれの分類において述語項構造の共起度を計算する。この時に例えば用言「刺される」と「腫れる」の関係においては述語項構造 1 のニ格である「蚊」は必須であるが述語項構造 1 のヲ格である「足」や述語項構造 2 のガ格である「腕」などは必須でないと判断することができる。また、述語項構造ペア間で項が共通であるという仮定をおいていないため、項を共有しない知識も獲得することができる。最後に、「順接」以外の「条件」や「理由」といった節間関係の分布を用いて時間経過、手段、因果関係などといった事態間関係に分類する。

2　関連研究

人手による事態間関係の構築として LifeNet がある [2]。これは Web 上で人手によって作成された知識と、2 語がどのような関係にあるかが記述された OMCSNet を用いて、事態間の関係をグラフの形で表現したものである。しかしこの事態間関係は主に時間経過に関するものであり、因果関係、手段などは含んでいない。

コーパスからの事態間関係知識の獲得手法には大きくわけて 2 つある。1 つ目は述語項構造の項の分布類似度を用いた知識獲得が提案されている [1]。しかし、項の分布類似度を用いて獲得した知識には、同義、含意、類義など様々なタイプの関係が含まれており、その関係まで分類することはできない。

2 つ目は、特定の事態間関係知識を獲得するために、共起パターンを用いた知識獲得が行われている。乾らは接続標識「ため」を用いた知識獲得および分類を行っている [4]。

また、項の共有情報と共起パターンを併用することにより、事態間関係を獲得する手法が提案されている。鳥澤は動詞テ形接続や連用中止形といった頻度の高い共起パターンと、項の共有情報を用いることで「時間的な前後関係のある推論知識」を獲得している [5]。阿部らは項の共有情報と事態間関係を示す様々な共起パターンを用いることで事態間関係の分類と項の共有情報を同時に獲得する知識獲得を行っている [3]。

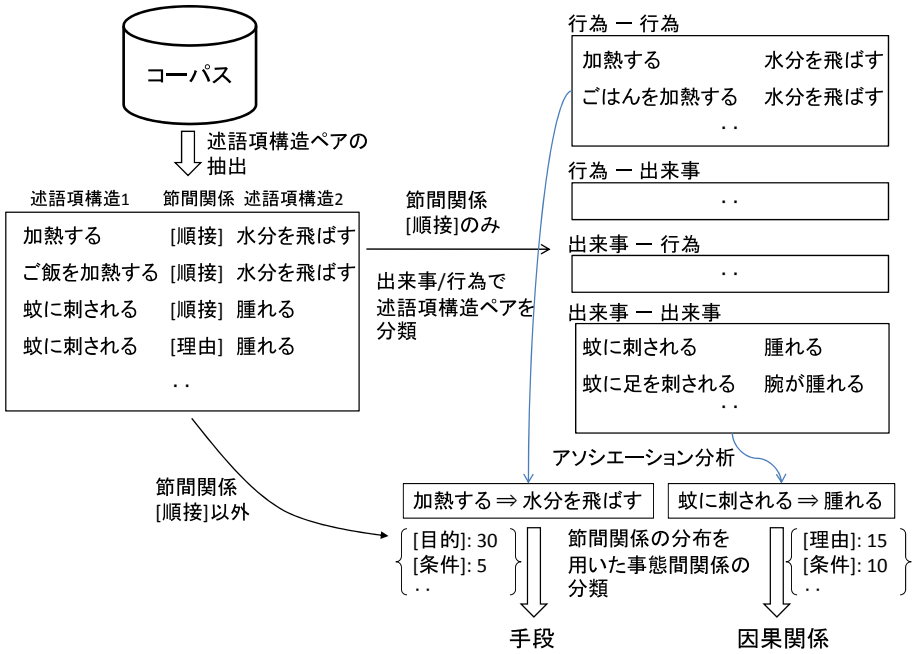

図 1: 本手法の概要

表 1: 節間関係と表層表現

節間関係	表層表現
順接	～て、～(連用中止形)
理由	～ので、～から、～せいで
条件	～と、～ならば、～ば
目的	～ために、～のに、～べく
逆接	～けれど、～が
同時	～ながら

3 述語項構造ペアの抽出

3.1 係り受け関係にある述語項構造ペアの抽出

コーパスに対して構文解析を行ない、係り受け関係にある述語項構造ペアを抽出する。従属節を構成する述語項構造を述語項構造 1、主節を構成する述語項構造を述語項構造 2 とする。獲得する項はガ格、ヲ格、ニ格の 3 つとする。否定、使役や受身などの用言に関する情報があれば付与する。また、述語項構造ペア間の節間関係を獲得する。表 1 に示す節間関係を獲得対象とする[1]。

3.2 項の汎化

項を単語クラスへ汎化し、表現は異なるが単語クラスが同じものを同一視することで、同一の事態間関係を認識し、データスパースネスの影響を軽減する。項の汎化には大規模類似語リスト [6] を用いる。大規模類似語リストとは、係り受けの大規模なクラスタリング結果を用いて作成されたもので、100 万語という大規模な語彙を対象としている。大規模類似語リストの全 2000 クラスに対して各々の上位 400 位を獲得する。名詞 n に対して、最も出現確率 $P(c|n)$ が高いクラス c を取得し、単語を「$\langle c \rangle$」と置き換える。例えば、「蚊に刺される ⇒ 腫れる」、「蜂に刺される ⇒ 腫れる」という述語項構造ペアにおいて、「蚊」、「蜂」ともに単語クラス「77」が最も出現確率が高いクラスであるので「$\langle 77 \rangle$ に刺される ⇒ 腫れる」と汎化され、同一視することができる。

[1]順接は逆接と対立する概念とされるが、本研究では「～て」「～(連用中止形)」などを順接と定義し、理由、条件などは別途定義する。

4 行為/出来事による述語項構造ペアの分類

分類を行ないたい事態間関係には、例えば、「手段」の場合、述語項構造 1,2 ともに行為である必要があり、「因果関係」の場合は述語項構造 1 が行為または出来事で述語項構造 2 は出来事である必要がある。ここで、行為は主体の意志が伴うもの、出来事は意志が伴わないものとする。

そこで、抽出した述語項構造ペアに対して、述語項構造 1,2 が行為か出来事かの判断を行い、行為/出来事の組み合わせで分類する。述語項構造ペアは、述語項構造 1,2 がそれぞれ行為か出来事によって表 2 のように分類することができる。

述語項構造の行為/出来事の分類は以下の基準を用いて行う。

表 2: 行為/出来事の組み合わせによる分類

		述語項構造 2	
		行為	出来事
述語項構造 1	行為	[**手段**] アルバイトをする ⇒ 生計を立てる 加熱する ⇒ ⟨861:水分, 湿気, ...⟩ を飛ばす [**時間経過**] 失恋する ⇒ ⟨1513:前髪, 髪 ⟩ を切る ⟨1247:川, 河, ...⟩ に行く ⇒ 泳ぐ	[**因果関係**] 麻酔をする ⇒ 痛くない 冷蔵庫に入れる ⇒ 冷える [**時間経過**] 手を伸ばす ⇒ 手が届く ⟨1428:練習, 訓練 ⟩ を重ねる ⇒ 上手になる
	出来事	[**前提条件**] ⟨618:計算, 集計 ⟩ が間違う ⇒ 修正する 声が聞こえる ⇒ 後ろを振り返る [**時間経過**] 沸騰する ⇒ 火を弱くする 会場に入る ⇒ ⟨880:アリーナ席, 席, ...⟩ に着く	[**因果関係**] ⟨796: 太陽光, 日差し, ...⟩ が強い ⇒ 暑い ⟨1359:クーポン, チケット, ...⟩ が付く ⇒ 得だ [**時間経過**] 冷める ⇒ 不味くなる 日焼けする ⇒ 皮が剥ける

1. 形容詞、受動態、可能動詞を**出来事**とする。
2. 使役態、他動詞を**行為**とする。
3. ガ格をとり、ガ格の JUMAN カテゴリが「人」、もしくは「組織・団体」であれば**行為** (例：父が寝る) とし、それ以外を **出来事**(例:蜂が刺す) とする。
4. 格フレームにおいて、ヲ格が必須格[2]であるものを**行為**とする。
5. 格フレームにおけるガ格において、カテゴリが「人」、「組織・団体」である割合や、固有表現認識において「PER」、「ORG」と判断されたものの割合の合計値が閾値以上を**行為**(例:働く) とし、閾値以下を**出来事**(例:産卵する) とする。

述語項構造 1,2 がそれぞれ行為か出来事によって、4 つにあらく分類する。

5 述語項構造ペアの共起度計算

前節で分類した 4 つの分類それぞれにおいて、共起度の高い述語項構造ペアを獲得する。ここでは、どの事態間関係の場合でも存在し、コーパス中に高頻度で出現する節間関係「順接」であるもののみを用いて共起度計算を行なう。

述語項構造の共起度計算には、**アソシエーション分析**を用いる。アソシエーション分析とは大量の入力データ（トランザクションデータ）より、Xが起きた際にYが起こりやすいというアソシエーションルール X⇒Y を獲得するものである。アソシエーションルールの獲得には図 2 の **support 値、confidence 値、lift 値**を指標として用いる。support 値は X、Yがトランザクションデータ中に同時に出現する確率である。confidence 値はトランザクションデータにて、Xが出現した際にYが出現する条件付き確率である。lift 値は、Y の X に対する依存性を見る指標である。1 以上であれば Y は X に対して依存関係にある。実験では、support 値が 1.0×10^{-7}、confidence 値が 1.0×10^{-3} 以上の述語項構造ペアに対して、lift 値が 100〜10,000 の間のものを抽出する。

$$supp(X \Rightarrow Y) = \frac{C(X \cap Y)}{M} \quad (1)$$

$$conf(X \Rightarrow Y) = \frac{C(X \cap Y)}{C(X)} = \frac{supp(X \Rightarrow Y)}{supp(X)} \quad (2)$$

$$lift(X \Rightarrow Y) = \frac{conf(X \Rightarrow Y)}{supp(Y)} \quad (3)$$

M : 観測データの総数 $C(X)$: 事象 X の観測数

図 2: アソシエーション分析に用いる評価指標

共起度計算の結果、項が伴うものと伴わないものの lift 値を比較し、lift 値が高いものを採用する。表 3 の例では、「修める ⇒ 卒業する」という関係においては ⟨362:研究科, 課程 ⟩ が必須であり、「産む ⇒ 育てる」という関係においては項が必須ではないと判断される。

6 節間関係の分布を利用した分類

前節で得られた述語項構造ペアを事態間関係に分類する。事態間関係によって節間関係の出現分布が異なると考えられ、例えば、因果関係であれば節間関係「理

[2]必須格とは、用言の取る項の総数に対して、対象とする項の出現割合が一定量を超えたものをさす。

表 4: 事態間関係の獲得数と分類精度

		述語項構造 2			
		行為		出来事	
述語項構造 1	行為	手段	2,869 (32/50, 64%)	因果関係	3,389 (33/50, 66%)
		時間経過	6,596 (37/50, 74%)	時間経過	4,428 (39/50, 78%)
	出来事	前提条件	4,947 (31/50, 62%)	因果関係	6,761 (37/50, 74%)
		時間経過	2,671 (41/50, 82%)	時間経過	2,725 (39/50, 78%)

表 3: lift 値によって項が必要かどうかを判断した例

述語項構造ペア		lift 値
⟨362:研究科, 課程⟩を修める	**⇒ 卒業する**	**6,026**
修める	⇒ 卒業する	2,226
産む	**⇒ 育てる**	**292**
子供を産む	⇒ 育てる	199

由」がよく出現し、手段であれば節間関係「目的」がよく出現する。そこで、各事態間関係それぞれにおいて少数の正例を作成し、それぞれにおいて節間関係の分布を算出し、未知の述語項構造ペアの節間関係の分布と最も類似している事態間関係に分類する。

用いる節間関係は理由、条件、逆接、同時、逆接否定、反転目的とする。逆接否定とは、節間関係が「逆接」で、述語項構造 2 の用言が否定表現を含むものを表し、反転目的とは、述語項構造ペアの述語項構造 1 と 2 を反転させた際に出現する節間関係「目的」を表す。また、分布の比較には cosine 類似度を用いる。

7 実験

本手法の有効性を実証するために Web コーパスから事態間関係の獲得を行なった。Web コーパスとして、日本語約 6.5 億ページからなるコーパスを利用した。これは約 416 億文からなる。ウェブにはミラーページなどの重複ページが多数存在することから、約 416 億文から重複を除いた約 69 億文を実験に利用した。また、事態間関係の正例はそれぞれ 10 例ずつ作成した。

各事態間関係の分類精度はランダムに 50 個選び、評価を行なうことにより算出した。表 4 に結果を示す。時間経過については 70%から 80%程度の精度が得られ、その他の関係については 60%から 75%の精度を得ることができた。獲得された数は合計で約 34,000 個であった。誤り要因を以下にまとめる。

行為/出来事の分類誤り 例えば、風邪を引く, 成分を含む, 体調を崩すなどは他動詞であるため行為と分類されたが、正しくは出来事である。

必須項が欠如している 前提条件と分類されたものに、「⟨1840:ななめ, 横, ... ⟩に生える ⇒ 抜く」があるが、これは述語項構造 1 のガ格に歯が必要と思われる。省略解析などを行なうことによって述語項構造抽出時に直接係っていない項も含める必要がある。

8 おわりに

述語項構造における項と用言の共起情報と節間関係の分布を用いて事態間関係知識の獲得を行った。今後の課題としては、分類精度を向上させるとともに、RTE タスクでの利用を行なう予定である。

参考文献

[1] Dekang Lin and Patrick Pantel. Discovery of infernce rules for question answering. In *Natural Language Enginnering 7*, 2001.

[2] Push Singh and William Williams. Lifenet: A propositional model of ordinary human activity. In *Proceedings of Workshop on Distributed and Collaborative Knowledge Capture*, 2003.

[3] 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, Vol. 17, No. 1, pp. 121–139, January 2010.

[4] 乾孝司, 乾健太郎, 松本裕治. 接続標識「ため」に基づく文書集合からの因果関係知識の自動獲得. 情報処理学会論文誌, Vol. 45, No. 3, pp. 919–932, 2004.

[5] 鳥澤健太郎. 「常識的」推論規則のコーパスからの自動抽出. 言語処理学会第 9 回年次大会予稿集, pp. 318–321, 2003.

[6] 風間淳一, Stijn De Saeger, 鳥澤健太郎, 村田真樹. 係り受けの確率的クラスタリングを用いた大規模類似語リストの作成. 言語処理学会第 15 回年次大会発表論文集, pp. 84–87, 2009.