推論ルール学習による単語間の意味的関係推論法

土田 正明 †¶　　鳥澤 健太郎 †　　De Saeger Stijn†
呉 鍾勲 †　　風間 淳一 †　　大和田 勇人 ‡
† 情報通信研究機構 MASTAR プロジェクト 言語基盤 グループ
‡ 東京理科大学 理工学部 経営工学科
¶ 東京理科大学 理工学研究科 経営工学専攻

1　はじめに

大規模な単語間の意味的関係の知識ベースは，イノベーション支援やリスク発見などに有用である [8]．既存の意味的関係獲得法の多くは，特定の関係を持つ2つの単語 (関係インスタンスと呼ぶ) を構文パタンによってコーパスから抽出している．そのため，特定の意味的関係を表現する構文パタンで書かれていない関係インスタンスの獲得は困難であった．

本研究では，大規模なコーパスにも明示的に書かれていない，非明示的な関係インスタンスの獲得を目指す．提案法は，既存の構文パタンに基づく方法 [2] で獲得した関係インスタンス（シードインスタンスと呼ぶ）とコーパス中の言及を組み合わせ，シードインスタンスと同じ種類の関係インスタンスを推論するルールを学習する．以下は，因果関係のルールの例である．

$$\text{因果}_{\text{SEED}}(Y,Z) \wedge Y \text{ を豊富に含む } X \rightarrow \text{因果}_{\text{HYPO}}(X,Z)$$

この推論ルールは「Y が Z の原因で，X が Y を豊富に含んでいるならば，X が Z の原因である」と解釈できる．すなわち，シードインスタンス (因果$_{\text{SEED}}$(Y,Z)) と構文パタン (Y が豊富な X) を用いて抽出したコーパス中の言及を組み合わせ，仮説 (因果$_{\text{HYPO}}$(X,Z)) を生成する．これは，X と Z を含む構文パタンの存在を仮定せずに，それらの関係を推論可能なため，非明示的な関係インスタンスをも獲得できる．本稿では，提案法について説明し，評価実験で，実際に約6億のウェブページにも明示的に書かれていない可能性が高い関係インスタンスを獲得できたことを報告する．

2　関連研究

既存の関係獲得法の多くは，特定の意味的関係を表す構文パタンを学習し，それらの構文パタンを用いて関係インスタンスを獲得している [5, 2]．これらの方法（以降，パタンベース法と呼ぶ）は高い精度で関係インスタンスを獲得できるが，コーパス中に頻出する構文パタンと共起しない，すなわち明示的に書かれていないインスタンスの獲得は困難である．

近年，一種の推論によってコーパス中に明示的に書かれていない関係インスタンスを獲得する研究がなされ始めている [7, 9]．Tsuchida ら [9] は，似た意味の単語は，似た関係を持ちやすいと仮説に基づき，シードインスタンスの単語を，その分布類似語で置換して仮説を生成する．Schoenmackers ら [7] は，TextRunner [1] のインスタンスを入力に，帰納論理プログラミング [6] の枠組みで1つ以上の関係から他の関係を推論するルールを学習する．次に，それらのルールを制約，TextRunner のインスタンスを事実（grounding fact）とし，Markov Logic Network(MLN) を用いて，各制約の重みを学習し，ルールから生成される仮説の確率を計算する．

本研究は，Schoenmackers らの研究 [7] と類似するが，単一の関係の間の含意関係を扱うか否かで大きく異なる．Schoenmackers らは，“can cause(A,B) → cause(A,B)” といった単一の関係の間の含意関係のルールを中心に学習[1]しているが，本研究ではこのようなルールは扱わない．なぜなら，既存のパタンベース法でも，獲得したい関係の言い換えや含意関係となる構文パタンを学習し，インスタンスを抽出していることから，このようなルールは，パタンベース法と比べて本質的なカバレッジの向上にはつながらないと考えられるためである．実際，本研究でシードインスタンスの獲得に用いる De Saeger らの方法 [2] でも，目標の関係を表現するいくつかの構文パタン（シードパタン）を入力に，それらのシードパタンの言い換えパタンを学習し，関係インスタンスの抽出を行っている．また，他の異なる点として，Schoenmackers らは，様々な種類の関係のインスタンスを静的に与えてルールを学習しているが，本研究は，目標とする関係の推論に有用な他の関係を動的に発見してルールを学習する．

[1] 文献 [7] によれば，実験で獲得された新しいインスタンスの約 70%は，単一の関係間の含意ルールから生成されていた．また，Schoenmackers らは，ルールの長さに応じて異なる事前分布のパラメタを用いることで，MLN でルールの重みを学習する際に，短いルール (body 部の述語数が少ない) が高い重みを持つようにバイアスをかけている．

3 提案手法

提案法は，シードインスタンスを入力に，まず可能な推論ルールを列挙し，各ルールのスコアを計算する．次に，それらのルールで仮説を生成し，仮説のスコアをそれを生成したルールのスコアから計算する．シードインスタンスは，De Saeger らの方法 [2] で獲得する．

本研究では，以下の 2 つの形のルールを対象とする．

タイプ 1: "X *pattern* Y" ∧ R_{SEED}(Y,Z) → R_{HYPO}(X,Z)

タイプ 2: "X *pattern* Y" ∧ R_{SEED}(Z,Y) → R_{HYPO}(Z,X)

R は関係の種類を表し，例えば因果関係などである．このように，本研究の推論ルールは，関係 R のシードインスタンス（R_{SEED}）とコーパス中でパタンを伴って書かれた言及 ("X *pattern* Y") を組み合わせて，同種の関係の仮説 (R_{HYPO}) を生成する．これらの形は，仮説の X と Z が Y と何らかの関係を持つことを保証するため，無意味なルールが大量に学習される危険性を減らすことができると考えられる．また，これらの形のルールは，入力されたシードインスタンス集合から容易に列挙できる．

提案法は，各シードインスタンス自身を生成されるべき仮説と見なし，あるシードインスタンスから他のシードインスタンスを推論できるルールを学習する．以下では，タイプ 1 の推論ルールを想定して説明するが，タイプ 2 に対しても自明な拡張で対応できる．具体的には，まず，シードインスタンス集合から第 2 項 (Z) に同じ語を持つシードインスタンスのペアを生成する．次に，それぞれのペアの一方を R_{SEED}(Y,Z)，もう一方を R_{HYPO}(X,Z) と見なし，それらのペアの第 1 項の単語同士を結ぶ構文パタンを "X *pattern* Y" として列挙する．最後に，学習された各構文パタンを用いて，ルールの形に従い R_{SEED}，R_{HYPO} の各単語を変数化する．例えば，第 2 項を共有するシードインスタンスのペアが「因果$_{SEED}$(アルコール, 眠気)」「因果$_{HYPO}$(ワイン, 眠気)」の場合,「アルコール」と「ワイン」を結ぶ「アルコールをワインから抽出」「ワインに含まれるアルコール」などの構文パタンが発見される．それぞれの構文パタンをルールの形に当てはめアルコールを Y，ワインを X，眠気を Z と変数化すると，下記のような推論ルールが得られる．

"Y を X から抽出" ∧ 因果$_{SEED}$(Y,Z) → 因果$_{HYPO}$(X,Z)

"X に含まれる Y" ∧ 因果$_{SEED}$(Y,Z) → 因果$_{HYPO}$(X,Z)

シードインスタンスのペアのどちらを R_{SEED} とするかには任意性があるので，構文パタンの Y と X の場所が入れ替わったルールも獲得される．

構文パタンには，De Saeger らによって提案されたクラス依存パタン [2] を用いる．クラス依存パタンとは，変数に取れる単語の意味クラスに制約を掛けた構文パタンである．構文パタンにクラス制約を書けることで，表現する関係の曖昧性を解消できる．例えば「X による Y」という構文パタンは「[X:化学物質] による [Y:病気]」のように X が化学物質で Y が病気のクラスの単語の場合は，X と Y の因果関係を表すが,「[X:組織] による [Y:製品]」の場合は，会社と商品の関係と表す．このように意味クラスによる制約によって，表す関係の曖昧性が解消される．このような意味クラスは，De Saeger ら [2] と同様，Kazama らの方法 [4] によって自動獲得できる．具体的には Kazama らの方法で，隠れクラスの集合 C を用いて名詞の隠れクラスへの事後確率の分布を求め，$P(c|n) \geq 0.2$ もしくは $\max_{c \in C} P(c|n)$ を満たす c を名詞 n の意味クラスとして獲得する．例えば，上記の例で，アルコールのクラス ID が 273 で，ワインのクラス ID が 287 の場合,「アルコールからワインを抽出」というパタンは「Y:273 から X:287 を抽出」とクラス制約が掛かる．

"Y:273 を X:287 から抽出" ∧ 因果$_{SEED}$(Y:273,Z) → 因果$_{HYPO}$(X:287,Z)

以降，単純化のため，ルールのクラス制約は表記しない．最後に，非生産的なルールを除去するため，M 個以上 (実験では 10) のシードインスタンスを生成できるルールのみを残す．

推論ルールを学習した後は，シードインスタンスのみを正例，その他全てを負例と仮定し，各ルールで生成された仮説集合から F-measure を計算し，それをルールのスコア (r_score) とする．ただし，F-measure の計算時には，各ルールの仮説の再現率として，最も多くのシードインスタンスを生成したルールの生成数を分母とした相対再現率を用いる．

最後に，多くの信頼できる推論ルールから生成される仮説が確からしいと考え，仮説のスコアをその仮説を生成したルールのスコアの和として計算する．

$$h\_score(h) = \sum_{r \in \mathbf{Irules}(h, \mathbf{Seeds}, \mathbf{Rules})} r\_score(r)$$

h は仮説，r はルール，**Seeds** はシードインスタンス集合，**Rules** は推論ルール集合，**Irules** は **Rules** の中で仮説 h を **Seeds** から生成可能なルールの集合である．

提案法は基本的に上記の通りであるが，精度向上のために 2 つのヒューリスティクスを適用している．

妥当な推移律の逆方向のルールの除去　前述したが，提案法は，同じ単語を共有するシードインスタンスペアのうち，どちらを R_{SEED} と見なすかに任意性がある．そのため，以下のように，パタン内の変数の場所のみが異なる 2 つのルールが学習される．

A: "X が Y の原因" ∧ 因果$_{SEED}$(Y,Z) → 因果$_{HYPO}$(X,Z)

B: "Y が X の原因" ∧ 因果$_{\text{SEED}}$(Y,Z) → 因果$_{\text{HYPO}}$(X,Z)

多くの場合，あるルールの中のパタンが表す方向の推論が妥当な場合，その逆方向のパタンを持つルールは妥当でない．上記例の場合，因果関係を矢印で表すとすると，ルール A は「X → Y, Y → Z」と因果関係の推移律を表すので妥当であるが，ルール B は「Y → Z, Y → X」と X が Z の原因とは言えず，妥当でない．

また，妥当な推移律 (ルール A) の逆方向のルールは，「因果 (眠気, 眠気)」のような同じ単語からなる仮説を生成する傾向にある．これは，推移律の逆方向のルールでは，Y に関して X と Z が同じ関係 (ルール B では X と Z は Y の結果) となるためである．すなわち，各シードインスタンスの第一項の語 (Y) に関して，コーパスから X の単語を抽出すると，その X の単語にはシードインスタンスの第二項の語 (Z) と同じ単語が含まれる可能性が高く，結果として同じ単語からなる仮説が生成される．

以上に基づき，パタンの変数の場所のみが異なる 2 つルールについて，同じ単語からなる仮説の生成数を比較し，大きい方を推移律の逆方向，すなわち妥当でないルールと見なし除去する[2]．

曖昧語除去による推論誤りの軽減　文書中で指示対象が曖昧な単語を Y に用いると，しばしば誤った仮説を生成してしまう問題がある [7]．例えば「X:脳梗塞により Y:病気が起こる」と「因果$_{\text{SEED}}$(Y:病気,Z:肺炎)」から「因果$_{\text{HYPO}}$(X:脳梗塞,Z:肺炎)」を生成した場合，「病気」が指す対象が曖昧であるため，この仮説は不確かである．そこで，本研究では文書頻度が一定以上の語 (実験では 40 万) を曖昧と考え，X,Y,Z に該当する単語のストップワードとする．

4　評価実験

提案法を 3 種類の関係獲得タスクで評価した．

因果関係 (X,Y): X が Y の直接・間接的な原因になる．
予防関係 (X,Y): X が Y を直接・間接的に防止できる．
材料関係 (X,Y): X が Y の材料・原料である．

本実験では，文献 [9] で用いられている評価法を採用した．具体的には，評価対象の各仮説に関して，Yahoo!API[3]で獲得した仮説の 2 語を含む最大 20 個のテキストを 3 人の評価者に提示し，2 人以上が正しいと判断した場合のみ，その仮説を正解と判定する．本実験では，約 6 億ウェブページから仮説を生成したが，本評価法は，Yahoo!API からアクセス可能な膨大なウェブページを検証に使用できるため，元の情報源の 6 億

[2]同数の場合は適用しない．基本的には，推移律の逆方向のルールでない限り，同じ単語からなる仮説は生成されないため，推移律の逆方向のルール以外にはほぼ適用されない．

[3]http://developer.yahoo.co.jp/webapi/search/

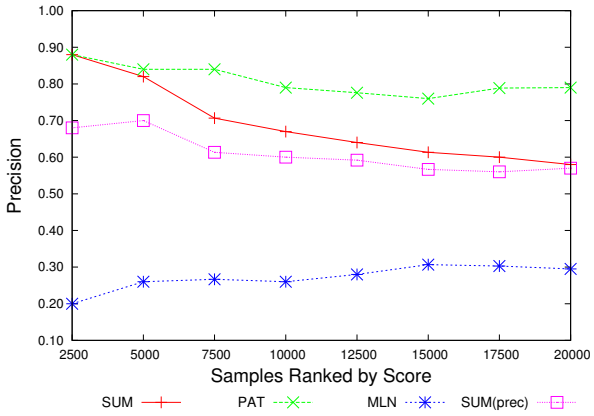

図 1: 因果関係の上位 2 万仮説の累積精度

ページには明示的に書かれていない関係インスタンスの一部が検証可能になると考えられる．

各関係のシードインスタンスには，De Saeger らの方法 [2] の上位の出力に対して，[4]，文献 [2] に示されている低コストなクリーニングを行って獲得した 2 万個の関係インスタンスを用いた．推論ルールは，因果で 24,044 個，予防で 17,868 個，材料で 14,978 個獲得された．生成された仮説からはシードインスタンスを除去し，全て新規の関係インスタンスとした．

各関係に関して，上位 2 万仮説からの 200 サンプルを評価した．評価者間の判定の kappa 値は，因果で 0.58，予防で 0.57，材料で 0.60 であった．図 1 に因果関係の精度を示す．横軸は順位で，縦軸はその順位までに含まれるサンプルの精度である．提案法は「SUM」で，比較として，シードインスタンスの精度 (PAT)，ルールスコアに F-measure でなく精度 (Precision) を用いる方法 (SUM(prec))，Markov LogicNetwork [3] でルールと仮説をスコアリングする方法 (MLN) の精度も記載している．MLN は，Schoenmackers ら [7] のスコアリング法を単純化したものと考えられる．具体的には，我々の MLN では，文献 [7] で提案された確からしいインスタンスに高い重みを付与してルールのスコアを学習する方法を用いていない．提案法の「SUM」は，「SUM(prec)」「MLN」と比べて精度が高いことから，提案のスコアリング法の有効性が確認できた．図 1 を見ると，提案法は，パタンベース法で獲得したインスタンスの精度 (PAT) より約 15%低かった．提案法は，関係抽出において強力な手掛かりとなる 2 語の直接的な構文パタンを用いていないため，これは当然の結果ともいえる．代わりに，提案法は，パタンベース法では獲得不可能な，非明示的なインスタンスを獲得できる利点がある．上述の全ての傾向は，予防関係で「SUM」が「PAT」と上位 2 万でほぼ同じ精度を達成

[4]ただし，文献 [2] と構文パタンが異なる．[2] では係り受け木上の最短パスを構文パタンとしているが，本実装は単語列からなる表層パタンを用いた．

表 1: 上位 2 万個の仮説の精度と推定した正解の仮説数

条件	因果関係	予防関係	材料関係
全サンプル	58% (116/200), 11600	58% (115/200), 11500	44% (88/200), 8800
NS(一文内に共起がない)	37% (28/75), 2800	41% (30/74), 3000	32% (25/77), 2500
N4S(隣接する 4 文内にも共起がない)	24% (7/29), 700	23% (7/30), 700	19% (6/32), 600

表 2: 仮説とそれを生成したルールの例．表記の単純化のためルール内のクラス制約は省略している．“NS” は一文内に，“N4S” は隣接する 4 文内に共起がない，“S” は一文内に共起がある，“*” は誤った仮説をそれぞれ表す．

	ラベル	仮説	仮説を生成したルールのサンプル
因果関係	N4S	因果$_{\text{HYPO}}$ (Z=ＳＯ２ガス, X=アレルギー症状), Y= 喘息	因果$_{\text{HYPO}}$ (Z,X) ← ”X は Y を引き起こす” ∧ 因果$_{\text{SEED}}$ (Z,Y). 因果$_{\text{HYPO}}$ (Z,X) ← “Y で X が悪化” ∧ 因果$_{\text{SEED}}$ (Z,Y).
	S	因果$_{\text{HYPO}}$ (X=農薬, Z=大腸がん), Y=有害物質	因果$_{\text{HYPO}}$ (X,Z) ← “X に含まれる Y” ∧ 因果$_{\text{SEED}}$ (Y,Z). 因果$_{\text{HYPO}}$ (X,Z) ← “X と言った Y” ∧ 因果$_{\text{SEED}}$ (Y,Z).
	*	因果$_{\text{HYPO}}$ (X=ビリルビン, Z=大腸がん) Y=胆汁	因果$_{\text{HYPO}}$ (X,Z) ← “Y に含まれる X” ∧ 因果$_{\text{SEED}}$ (Y,Z). 因果$_{\text{HYPO}}$ (X,Z) ← “Y から出る X” ∧ 因果$_{\text{SEED}}$ (Y,Z).
予防関係	N4S	予防$_{\text{HYPO}}$ (X=青魚, Z=脳血栓症), Y= EPA	予防$_{\text{HYPO}}$ (X,Z) ← “X から摂った Y” ∧ 予防$_{\text{SEED}}$ (Y,Z). 予防$_{\text{HYPO}}$ (X,Z) ← “Y に富む X” ∧ 予防$_{\text{SEED}}$ (Y,Z).
	NS	予防$_{\text{HYPO}}$ (Z=ひまわり油,X=心臓病), Y=高血圧, Y1=リノール酸	予防$_{\text{HYPO}}$ (Z,X) ← “Y が X を起こす” ∧ 予防$_{\text{SEED}}$ (Z,Y). 予防$_{\text{HYPO}}$ (Z,X) ← “Z に含まれる Y1” ∧ 予防$_{\text{SEED}}$ (Y1,X).
	*	予防$_{\text{HYPO}}$ (Z=ナイアシン, X=腎臓病), Y=高血圧	予防$_{\text{HYPO}}$ (Z,X) ← X に伴う Y ∧ 予防$_{\text{SEED}}$ (Z,Y). 予防$_{\text{HYPO}}$ (Z,X) ← X が Y の原因 ∧ 予防$_{\text{SEED}}$ (Z,Y).
材料関係	NS	材料$_{\text{HYPO}}$ (Z=甜菜, X=水素), Y=エタノール	材料$_{\text{HYPO}}$ (Z,X) ← “Y から X を抽出” ∧ 材料$_{\text{SEED}}$ (Z,Y). 材料$_{\text{HYPO}}$ (Z,X) ← “Y から X に変換” ∧ 材料$_{\text{SEED}}$ (Z,Y).
	S	材料$_{\text{HYPO}}$ (Z=とうもろこし, X=エチレン), Y=*ethanol*	材料$_{\text{HYPO}}$ (Z,X) ← “Y から作られる X” ∧ 材料$_{\text{SEED}}$ (Y,Z). 材料$_{\text{HYPO}}$ (Z,X) ← “Y を原料とする X” ∧ 材料$_{\text{SEED}}$ (Z,Y).
	*	材料$_{\text{HYPO}}$ (Z=ブルーベリー, X=プレーンヨーグルト) Y=ブルーベリージャム	材料$_{\text{HYPO}}$ (Z,X) ← “Y を X に混ぜる” ∧ 材料$_{\text{SEED}}$ (Z,Y). 材料$_{\text{HYPO}}$ (Z,X) ← “Y を入れた X” ∧ 材料$_{\text{SEED}}$ (Z,Y).

したことを除き，全関係で同様であった．

表 1 は，提案法で獲得した上位 2 万の仮説の精度 (ALL) と，その中に含まれる非明示的なインスタンス (NS,N4S) の精度とその推定数を表す．「非明示的なインスタンス」に関して，実際に 6 億ページを確認することは困難であるため，2 つのレベルに分けて調査した．

NS: 6 億ページの一文内で共起がない 2 語の仮説

N4S: 6 億ページの隣接 4 文内で共起がない 2 語の仮説

NS のインスタンスは，一文内での直接的な言及がないため，パタンベース法では原理的に獲得不可能である．N4S は，2 語の関係が 6 億ページ中に書かれていない可能性が高いインスタンスである．表 1 より，例えば，一文内に共起のない 2 語の因果関係を約 2,800 個獲得でき，さらには，隣接 4 文内にすら共起のない因果関係を約 700 個獲得できていることが分かる．すなわち，提案法によって，6 億ページという大規模なコーパスにも明示的に書かれていない可能性が高いインスタンスを獲得されていることが確認できた．実際の仮説と推論ルールの例を表 2 に示す．

5 おわりに

本稿では，自動獲得されたシードインスタンスから推論ルールを学習し，それらのルールによって 2 つの関係インスタンスを組み合わせ，非明示的な関係インスタンスを推論する方法について述べた．実験によって，約 6 億のウェブページ中のいかなる文にも直接言及されていない，すなわち明示的に書かれていない可能性の高いインスタンスを獲得できること示した．

参考文献

[1] M. Banko et al. The Tradeoffs Between Open and Traditional Relation Extraction. In *Proc. of the 46th ACL-08:HLT*, pp. 28–36, 2008.

[2] S. De Saeger et al. Large Scale Relation Acquisition Using Class Dependent Patterns. In *Proc. of the 9th ICDM*, pp. 764–769, 2009.

[3] T. N. Huynh et al. Discriminative Structure and Parameter Learning for Markov Logic Networks. In *Proc. of the 25th ICML*, pp. 416–423, 2008.

[4] J. Kazama et al. Inducing Gazetteers for Named Entity Recognition by Large-scale Clustering of Dependency Relations. In *Proc. of the 46th ACL*, pp. 407–415, 2008.

[5] P. Pantel et al. Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations. In *Proc. of the COLING-ACL06*, pp. 113–120, 2006.

[6] J. R. Quinlan. Learning Logical Definitions from Relations. *Machine Learning*, 5(3):239–266, 1990.

[7] S. Schoenmackers et al. Learning First-Order Horn Clauses from Web Text. In *Proc. of EMNLP2010*, pp. 1088–1098, 2010.

[8] K. Torisawa et al. Organizing the Web’s Information Explosion to Discover Unknown Unknowns. *New Generation Computing*, 28(3):217–236, 2010.

[9] M. Tsuchida et al. Large Scale Similarity-based Relation Exapnsion. In *Proc of the 4th IUCS*, pp. 140–147, 2010.