単語の意味クラスを用いたパターン学習による大規模な意味的関係獲得

Stijn De Saeger　鳥澤健太郎　風間淳一　黒田航　村田真樹

情報通信研究機構　MASTAR プロジェクト　言語基盤グループ

{stijn,torisawa,kazama,kuroda,murata}@nict.go.jp

I. はじめに

本研究ではウェブから因果関係・予防策関係など高度な意味的関係を獲得する minimally supervised な手法を提案する。これらの意味的関係は例えば様々なトラブルの意外な原因や予防策を含み、有用なリソースとなる。主な貢献は大規模な単語クラスタリング結果を用いて、単語の意味クラスによる制限のついたパターンを少量のシードパターンから自動的に学習することである。実際に因果関係、予防策関係、材料関係を獲得し、提案手法の有用性を示す。

本研究で用いるクラス制限付きパターンは、そのパターンと共起する単語対を特定の意味クラスに限定したパターンであり、抽出ターゲットの意味的関係の強い手がかりとなる。抽出候補の語を単語クラスを用いて制限することで、長年問題となっている “generic pattern” [3], [4]、すなわち、「X の Y」、「X による Y」など高度に多義で出現頻度の高いパターンも活用可能となる。

こういった多義なパターンの取り扱いは困難な問題である。こうしたパターンは、多くの関係インスタンス、すなわち単語対と共起するため、関係獲得の再現率を向上させるが、関係のインスタンスでない単語対ともよく共起するので適合率には悪影響を与える。本研究では、パターンの単語の意味クラスを限定することで、こうした多義性に対処する。例えば、「X の Y」という多義なパターンを多数のクラス制限付きのバージョン「c_i の c_j」、「c_k の c_l」、...（c_i が意味クラスを意味する）に分割すると、それぞれの単語クラスの組み合わせが多義なパターンのユニークな意味的解釈と一致し、多義性を大幅に回避できる。

例えば、“インフルエンザの熱” のように、X と Y が「病気」と「症状」のクラスに属するとすれば、「X の Y」が因果関係を指す可能性が高いであろう。一方、“京都の清水寺” のように「地名」と「名所」のクラスだったらむしろ「所在地関係」になる。「人物」と「作品」のクラスの場合（例えば、“Stephen King の小説”）はおそらく「作者・作品関係」もしくは「所有関係」を指す。

II. 提案手法：概要

従来、意味的関係の獲得にはブートストラップ法が用いられることが多かった [4]。ブートストラップ法に基づいた関係抽出法はパターンの自動学習とコーパスからのインスタンス抽出という二段階を交互に繰り返しながら、コーパスから大量に関係を獲得する。本手法はこれらと以下の二点において異なる。

まず、本手法の入力は関係の種となる名詞対ではなく、パターンである。これらのパターンを以下ではシードパターンと呼ぶ。ただし、シードパターンは単語クラスの制限はついていない。シードパターンはコーパスで大量の名詞対と共起すると期待され、ターゲットとなる関係を表す代表的なパターンをシードパターンとすることで、それらと共起する大量の名詞対の統計から確からしい手がかりを入手できる。例えば，ターゲットとなる意味的関係を持ちやすい単語対が典型的に持つ単語クラスの対を見つけることができる。本手法では、こうした手がかりをもとに、新たなクラス制限付きパターンを学習する。

また、パターン学習と単語対抽出を区別しないというのも本手法の重要な特徴である。獲得過程全体は、コーパスに見つかる全名詞対のランキングをワンステップで行うと見なされ、クラス制限付きパターンの学習は、このランキングのスコア時に暗黙裡に行なわれるのみである。

より具体的に、提案手法では、各名詞対に対して、それと共起するすべてのクラス制限付きパターンを、ターゲットとなる関係の証拠として考慮し、それぞれのパターンの重みを計算する。この重みは、最終的な各名詞対のスコアにおいて考慮され、最終的な名詞対のランキングが行なわれる。次節でこのスコアについて具体的に説明する。

III. スコアリング法

本研究では意味クラス c を名詞の集合として扱い、名詞 n_i, n_j に対して、$n_i \in c_i$ と $n_j \in c_j$ なら、名詞対 (n_i, n_j) が $c_i \times c_j$ というクラス組に属するものとする。パターン p がある文 S で名詞対 (n_i, n_j) と共起するというのは、S の構文木で p が終端節点である名詞 n_i と n_j を結ぶ係り受けパスに含まれる単語から成っているということであると仮定する。

抽出対象となるデータとして、ウェブコーパス (TSUBAKI, [5]) の約 5000 万文書を関係抽出の対象にし、パターンと名詞対の共起マトリクスを構築した。このマトリクスを $\mathcal{M}$ と呼び、名詞対 (n, n') とパターン p が 0 以上の

共起出現頻度があることを $(n, p, n') \in \mathcal{M}$ で表す。その共起出現頻度自体は $\|(n, p, n')\|$ と書き、あるパターン p の共起を $c_i \times c_j$ というクラスの名詞対に限定したバージョンは $p_{c_i \times c_j}$ と書く。提案手法は次のスコア関数によってコーパスに見つかる全名詞対 (n_i, n_j) をランキングする。ここで、SP は入力となるシードパターンの集合である。

$$Score(n_i, n_j, SP) = \max_{c_i \in class(n_i),\ c_j \in class(n_j),\ (n_i,p,n_j) \in \mathcal{M}} \{ CScore(c_i, c_j, SP) \cdot Para(p_{c_i \times c_j}, SP) \cdot Assoc(n_i, p, n_j) \} \quad (1)$$

ここで $class(n_i)$ は n_i が属する意味クラスを指す。名詞対 (n_i, n_j) の最終スコアは、$CScore$、$Para$ と $Assoc$ という 3 つの部分スコアの積の最大値である。この最大化におけるパラメータは意味的クラス c_i,c_j とパターン p であり、基本的に、これらの全ての組み合わせに関してスコアの値を計算することになる。ここでパターン p は、クラス制限無しのパターンであるが、この最終的なスコアを最大化する意味クラス c_i,c_j によって共起単語が制限されたクラス制限付きパターンと見なすことができる。

A. 部分スコアの定義

$CScore$ はクラス対用のスコアで、ある意味クラス対に属する名詞対とシードパターン SP と共起する名詞対の重複から、獲得する関係とこのクラス対の相性を推定する。例えば、(“高熱”,“けいれん”) などのように、「病気」と「病状」のクラスを含むクラス対には因果関係のインスタンスが多いであろう。一方で「人物」と「場所」からなるクラス対に因果関係はまれだと思われる。定義は下記のとおりである。

$$CScore(c_i, c_j, SP) = \begin{cases} \dfrac{\sum_{(n_i,n_j) \in c_i \times c_j} \|(n_i, SP, n_j)\|}{\sum_{(n_i,n_j) \in c_i \times c_j} \|(n_i, *, n_j)\|} & \text{if condition } \alpha \text{ holds} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

$\|(n, SP, n')\|$ は (n, n') のシードパターンとの共起頻度である ($\|(n, SP, n')\| = \sum_{p \in SP} \|(n, p, n')\|$)。$\|(n, *, n')\|$ は (n, n') の文内共起である $\left(\sum_{(n,p,n') \in \mathcal{M}} \|(n, p, n')\|\right)$。Condition α はシードパターンとクラス組の重複をチェックし、一定個以上のシードパターンが観測されるクラス組のみを考慮するためにある。Condition α の定義は $\|\{p \in SP \mid \exists (n_i, n_j) \in c_i \times c_j, (n_i, p, n_j) \in \mathcal{M}\}\| \geq \beta$ である。実験では β を 3 とした。

$Para$ という部分スコアはクラス制限付きパターン $p_{c_i \times c_j}$ に関してシードパターンとの “言い換え度” を測定する。基本的には、$c_i \times c_j$ に属する名詞対の中で共起する名詞対の重複が大きければ大きいほど、$p_{c_i \times c_j}$ とシードパターンが意味的に類似し、$p_{c_i \times c_j}$ をシードパターンの妥当な言い換え表現と見なす。例えば、“X による Y” というパターンは、トラブルと災害を表すようなクラス対で制限された場合（例：“交通事故による死亡”）は因果関係の手がかりになり、例えば、「X で引き起こされる Y」というようなシードパターンの言い換えと見なせるが、他のクラスではそうと限らない（例：“電子メールによる登録”）。

一方で、与えられたクラス対に属する名詞対のみを考慮した場合、クラス対による制限によってデータスパースネスの影響を受けやすくなる。こうした状況を避けるため、実験で用いた $Para$ はクラス依存のパターン類似度とクラスを限定しないパターン類似度の両方を考慮する。ここで $\mathcal{I}(p_{c_i \times c_j}) = \{(n_i, n_j) \in c_i \times c_j \mid (n_i, p, n_j) \in \mathcal{M}\}$ を $p_{c_i \times c_j}$ と共起する名詞対の集合としよう。同じく、$\mathcal{I}(SP_{c_i \times c_j}) = \cup_{q \in SP} \mathcal{I}(q_{c_i \times c_j})$。クラス依存の $Para$、“$Para_C$”、は下記のように定義される。

$$Para_C(p_{c_i \times c_j}, SP) = \frac{\|\mathcal{I}(p_{c_i \times c_j}) \cap \mathcal{I}(SP_{c_i \times c_j})\|}{\|\mathcal{I}(p_{c_i \times c_j}) \cup \mathcal{I}(SP_{c_i \times c_j})\|} \quad (3)$$

つまり、$Para_C$ は、クラス組 $c_i \times c_j$ の名詞対に限定した、パターン p とシードパターン SP のジャカード係数である。

また、前述したデータスパースネスを回避するため、$Para_C$ を全名詞対の共起を反映したグローバルなパターン類似度で補い、最終的に $Para$ を次の通りに定義する。($\mathcal{I}(p)$ はコーパス全体でパターン p と共起する名詞対の集合)

$$Para(p_{c_i \times c_j}, SP) = Para_C(p_{c_i \times c_j}, SP) \cdot \frac{\|\mathcal{I}(p) \cap \mathcal{I}(SP)\|}{\|\mathcal{I}(p) \cup \mathcal{I}(SP)\|} \quad (4)$$

最後に、$Assoc$ という部分スコアはパターンと名詞対の関連度を示す。

$$Assoc(n, p, n') = log \frac{\|(n, p, n')\|}{\|(n, *, n')\|\|(*, p, *)\|} \quad (5)$$

B. 単語クラスタリングによる意味クラス情報

名詞を意味クラスに分類するには様々な方法がある。本研究では風間らが導入した確率的クラスタリング法 [2] を用い、50 万名詞を 500 クラスにクラスタリングした。

IV. 評価・比較実験

提案手法の有効性を示すため、因果関係、材料関係、予防策関係の 3 つの関係抽出タスクによって、提案手法、2 つのベースライン手法と既存のブートストラップ法（Espresso, [4]）と比較した。これ以降、提案手法は **CD** (“*Class Dependent*”) と略す。また、以下に、比較した手法の説明を行なう。

1. Espresso (ESP) Pantel らは、Espresso [4] というブートストラップ法を提案しており、generic pattern に対する

表 I
獲得結果の実例。“*” は正しくない例を指す (LENIENT)

	クラス組	ランク	インスタンス
因果関係	$c_{471} \times c_{290}$	22	チロチナーゼ - そばかす
	$c_{468} \times c_{290}$	62	かび - ニオイ
	$c_{468} \times c_{290}$	274	だに - 皮膚トラブル
	$c_{471} \times c_{290}$	394	残留塩素 - かゆみ
	$c_{475} \times c_{1}$	5889	日本酒 - 肥満
	$c_{290} \times c_{290}$	6523	虫歯 - 口臭
	$c_{471} \times c_{1}$	17135	タウリン - 動脈硬化*
材料関係	$c_{176} \times c_{475}$	614	麦芽 - ウィスキー
	$c_{176} \times c_{227}$	1128	コラーゲン - ゼリー
	$c_{176} \times c_{227}$	5032	カカオ豆 - チョコ
	$c_{252} \times c_{270}$	34971	サトウキビ - 自動車燃料
予防策関係	$c_{172} \times c_{49}$	820	サーモスタット- 加熱
	$c_{333} \times c_{49}$	831	発泡スチロール - 蒸発
	$c_{212} \times c_{191}$	11856	手段 - 洪水被害*
	$c_{240} \times c_{191}$	17627	会議 - 事故*

解決を提案した。「正しい関係インスタンスは多義でないパターンとの関連度 (相互情報量) が高い」という仮定に基づき、名詞対の信頼度のスコアを計算し、これが一定のしきい値を超えない場合、その名詞対を捨てる。

Espresso はシードパターンではなく、名詞対の集合を入力としているため、フェアな比較を行なうために提案手法 **CD** と同じく、シードパターンを入力とするように変更した。今回特別に付与した、初期化ステップでまずシードパターンと共起する名詞対を集め、Espresso の通常のスコアリング法でそれらをランキングし、[4] と同じく上位 200 個を次のイテレーションの入力とする。

2. シングルクラス (SC) は、クラス依存性の有用性を評価するため、名詞全体を一つの意味的クラスのものと見なして提案手法 **CD** を適用するクラス非依存の手法である。

3. シードパターン (SP) は、シードパターンと共起する名詞対から、ランダムで 100 対を抽出するものである。このベースラインはシードパターンから生成される名詞対の精度を示す。興味深いことに、提案手法 **CD** は、今回実験した三種の関係の獲得において、このベースラインを大幅に上回る性能を達成した。IV-A 節を参照のこと。

A. 実験設定

以上に述べた、提案手法 **CD**、比較手法を因果関係、予防策関係、材料関係という 3 つの意味的関係の獲得タスクによって評価した。表 I は、生成したクラス組と出力でのランクと共に提案手法に獲得された名詞対を表示する。

各手法には、同じシードパターンが入力として与えられた。評価基準としては、出力された名詞対 (n, n') に関して、一番スコアが高いパターン p を選び (Espresso の場合、(n, n') と一番相互情報量が高い p)、(n, p, n') が共起する文をコーパスから検索し、最も出現頻度が高い 3 文を評価者 3 人に評価してもらった。その 3 文のうち、n と n' の間にターゲットとなる関係が成立する証拠が見つかれば、(n, n') を正解とした。

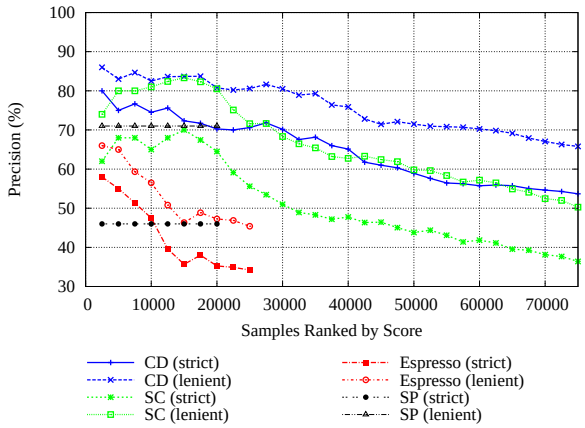

図 1. 因果関係獲得の精度グラフ

また、精度を測定する際に 2 つの評価基準を用いた。strict な基準では評価者全員の一致を求め、lenient な評価基準では評価者 2 名以上が一致したものを正しいと見なす。

1. 因果関係 図 1 に各手法の出力をスコア順にランキングし、上位からの累積の精度をプロットしたグラフを示す。提案手法 **CD** が比較手法 **ESP** とベースライン手法 **SC** と **SP** を大幅に上回り、lenient な評価基準では上位 3 万対に関して 80%、上位 6 万対に関して 70%の精度を得ることが分かる。本実験の入力としては「X が Y の原因となる」、「X が Y を引き起こす」というようなシードパターン 20 個を使用した。評価者間の一致率 (カッパ) は 0.67 で、substantial な一致 (Landis&Koch, 1977) を示した。

提案手法の出力をみると、上位がほとんどシードパターンのクラス制限付きのバージョンにより生成されたものである。しかし、上位 15,000 対から 35,000 対までは一番出現頻度が高いパターンは「X による Y」という generic pattern であった (評価サンプルの 22%を生成)。このパターンにより生成された名詞対の精度は 86.4%(strict) と 100%(lenient) であった。さらに、上位 35,000 対から 75,000 対も同様に「X による Y」が一番有効なパターンとなった (評価サンプルの 15%、strict な精度が 93.3%、lenient な場合は 100%)。この結果は本研究が仮定していた、パターンの多義性を解消するため、単語意味クラスによる制限を用いることの有用性を証明し、提案手法の有効性を示している。

また、興味深いことに、シードパターンと共起する名詞対の精度 (ベースライン手法 **SP**) が提案手法の出力と比べて lenient で 10%以上低い。同じシードパターンが提案手法の唯一の入力となるので、**CD** はまるでシードパターンを与えた側の意図を推測できたように見える。今のところ、厳密な検証はしていないが、こうした結果が得られた理由は以下のようなものだと考えられる。大雑把に言って、あるクラス対は、その *CScore* の値が大きければ、それに属するする名詞対の多くがシードパターンと共起するが、複数のシードパターンとクラス中の名詞対が共起している可能性が高い。ここで、複数のシードパターンと

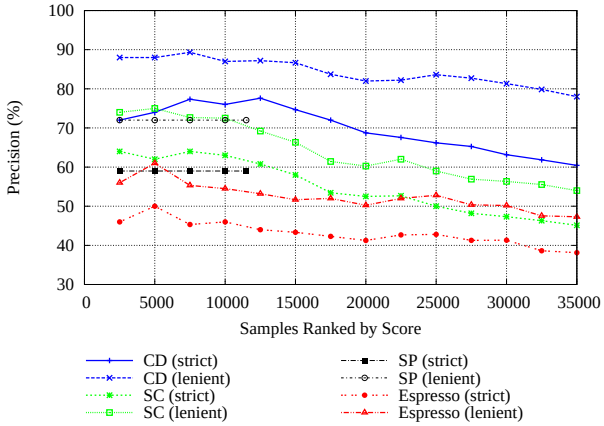

図 2.　材料関係獲得の精度グラフ

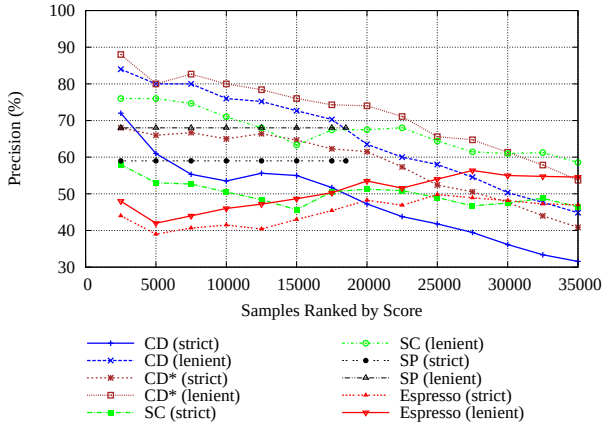

図 3.　予防策関係獲得の精度グラフ

共起する名詞対の集合、別の言い方をすると、複数のシードパターンについて、その各々と共起する名詞対の積集合を考えると、*CScore* が大きなクラス対はそうした名詞対の積集合に近いものとなっている可能性が高い。ここで、仮に、シードパターンの名詞対の積集合が、ターゲットとなる関係の典型的な名詞対を多く含んでいるとすれば、*CScore* が大きなクラス対はまさに、そうした典型的な名詞対、つまり、ターゲットとなる関係の典型例を多く含むことになる。

提案手法では、この *CScore* が独立した部分スコアとなっており、実際に出力する名詞対の上位は、仮にパターンが generic pattern のようなスコアが必ずしも高くないパターンであっても、*CScore* が高いクラス対からの名詞対である傾向がある。言い換えれば、パターンが証拠として、仮に当てにならなくても、シードパターンと共起する名詞対、しいては、ターゲットとなる関係の典型例に意味的に近い名詞対であれば、上位に出力されるということである。今回の実験では、３種の関係いずれにおいても、こうした名詞対が正解と見なされることが多かったということであると考えている。

2. 材料関係　図 2 が示すように、この関係の獲得においても、因果関係と同様の傾向が示され、提案手法が最高の精度を示し、ベースラインを上回る性能を示した。

3. 予防策関係　予防策関係の精度グラフは図 3 にある。予防策関係は全体的に精度が低かった。しかしながら、結果を見たところ、誤りと見なされた単語対の多くが、ごく少数の単語クラス対に属する単語対であることがわかり、この単語クラス対を候補から取り除いたところ、簡単に性能が向上した。この手法の精度は **CD*** としてプロットされている。なお、**CD** の出力を見て、削除すべき単語クラス対を特定するのに要した時間は１０分程度であり、十分に実用性のあるトリックであると考えている。

V. 関連研究

本研究に一番類似している研究は Gliozzo ら [1] である。彼らの手法では、ドメインに対応した単語クラス（例えば,「医師」と「薬品」が同じクラスに属する）を利用するものであり、「人」「薬品」などがそれぞれ別のクラスに属するものと仮定される我々の手法とは大きく異なる。Gliozzo らの手法は、Espresso の出力結果に対するフィルタリングであり、基本的に Espresso が出力できる以上の単語対を出力することはできない。我々の提案手法は、実験において、いずれも Espresso よりも大量の名詞対をより高い精度で獲得しており、いずれにせよ、Gliozzo らの手法に対する優位性は示されたものと考えている。

VI. まとめ

本論文では、単語の意味クラス情報を有効に使用する、非ブートストラップ的な関係抽出法を提案した。クラス制限付きのパターンを導入し、関係抽出の手がかりとなる言語パターンの多義性を解消することで、多様なパターンをフルに活用でき、大量な関係インスタンスを獲得できることを示した。

References

[1] A. M. Gliozzo, M. Pennacchiotti, and P. Pantel. The domain restriction hypothesis: Relating term similarity and semantic consistency. In *Proc. of HLT/NAACL*, pages 131–138, 2007.

[2] J. Kazama and K. Torisawa. Inducing gazetteers for named entity recognition by large-scale clustering of dependency relations. In *Proc. of ACL/HLT'08*, pages 407–415, 2008.

[3] M. Komachi, T. Kudo, M. Shimbo, and Y. Matsumoto. Graph-based analysis of semantic drift in espresso-like bootstrapping algorithms. In *Proc. of EMNLP'08. Honolulu, USA*, pages 1011–1020, 2008.

[4] P. Pantel and M. Pennacchiotti. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proc. of COLING/ACL'06*, pages 113–120, 2006.

[5] K. Shinzato, T. Shibata, D. Kawahara, C. Hashimoto, and S. Kurohashi. TSUBAKI: An open search engine infrastructure for developing new information access. In *Proc. of IJCNLP*, pages 189–196, 2008.