超低頻度構文パターンからの意味的関係獲得

De Saeger Stijn　鳥澤健太郎　土田正明　風間淳一　橋本力
山田一郎　呉鍾勳　Varga István　顔玉蘭
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,m-tsuchida,kazama,ch,iyamada,rovellia,istvan,yulan}@.nict.go.jp

1　はじめに

本論文では auto-learning を用いた高度な意味的関係の獲得手法を提案する。本手法の特徴は、大規模なウェブコーパスでもまれにしか出現しない複雑なパターンのみと共起する意味的関係のインスタンス、つまり単語対が獲得できる点にある。これらの単語対は、論理的にも実質的にも、パターン学習に基づいた従来手法では獲得困難，もしくは不可能なインスタンスである。多くのパターンベース法は共起統計に基づき、構文パターン (*lexico-syntactic patterns*, [5]) の類似度を計算するため、データの過疎性により、低頻度パターンとしか共起しない単語対は獲得困難である。さらに極端な場合、コーパスに一回しか出現しない構文パターンは論理的に扱うことができない。提案手法ではこうした構文パターンからもターゲットの意味的関係を持つ単語対が獲得できることを実験により示す。

従来の関係獲得法 [2, 1, 7] は単語間の二項関係を特徴づける主な手がかりとして構文パターンを用いる。これらの構文パターンは半自動的に学習され、ターゲットの意味的関係のインスタンスをコーパスから抽出するテンプレートとして利用される。こうした技術は因果関係等の高度な意味的関係の網羅的なデータベースを構築する際にきわめて重要である。

抽出用の構文パターンは多数の単語対との共起統計を元に学習されるので、ある程度の出現頻度で観測される構文パターンしか活用できない。以下の例のような、複雑で出現頻度が極端に低いパターンとしか共起しない単語対は実質的に獲得困難と思われる。

> 高血圧 を治療することは、腎機能低下速度を緩和し 頭蓋内出血 の危険因子を低下させる。

上記の例文からは，人間であれば下線部の単語対の因果関係を読み取ることができるが、この単語対を結ぶパターン（「X を治療することは、腎機能低下速度を緩和し、Y の危険因子を低下させる」）は非常に複雑で、億単位のウェブページコーパスにでもまれにしか観測されないため、従来のパターンベース手法で利用するのは困難である。一方、提案手法ではこういった文からでもに関係を獲得できる。上記の例文は本手法で実際に獲得できたインスタンスである。

本論文はこのような獲得困難な関係インスタンスの獲得を狙いとし、二つの関係獲得器で構成される２ステップからなる手法を提案する。第一の獲得器は [3, 4] で提案されたパターンベースの手法を 6 億ウェブページのコーパスに適用し、その獲得結果を教師ありデータとして第二の獲得器に提供する。第二獲得器はその獲得結果を用い、上記のような文から意味的関係を抽出する分類器を学習する。より具体的には、まず第一獲得器の出力から得られる二つの情報を用いるヒューリスティクスに基づいて、獲得対象となる関係を含む可能性が高い文をコーパスから選択し、そうした文と単語対のペアを分類器で関係を表すものとそうでないものに分類する。第一獲得器の出力から第二分類器の学習データを自動生成するので、手法全体として、人手作業が発生するのは第一獲得器に与えるシードパターン（後述）の作成のみである。

実験では、本手法を 6 億ウェブページからの関係獲得タスクに適用し、超低頻度のパターンからも実質的に関係インスタンスが獲得できることを示す。さらに本手法は、全コーパスに一回しか出現しないパターンパターンのみと共起するインスタンスに関しても有意な数獲得する事が出来た。これらの単語対はパターンベースの従来法では論理的に獲得できないものである。

2　提案手法

本手法は第一獲得器と第二獲得器の二つのステップからなる。以下では両者について詳しく説明する。

2.1　第一獲得器

本手法は、２つの関係獲得器を順に適用することで、超低頻度パターンから意味的関係を獲得する。

第一獲得器は [3, 4] で提案されたパターンベース手法である。この関係獲得法はターゲットの意味的関係を表すパターン (「シードパターン」と呼ぶ) を入力とし、その言い換え表現と見なせるパターンを大量に学習する。また、パターンの曖昧性解消のためシード

パターンと意味的に類似した意味的クラス制約付きパターンを学習する。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 が意味クラスに該当する）に分割すると、それぞれの意味的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する可能性が高まり、多義性を大幅に回避できる。“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」のクラスであればむしろ「所在地関係」になる。このように単語の意味的クラス情報を元に、高精度で意味的関係が抽出できる。実験では 100 万語を風間らの手法 [6] で 500 クラスに分類し、意味的クラスとして使用した。

最終的に、第一獲得器は、学習したクラス制限付きパターンで抽出した単語対にスコアを付け、スコア順で並べられた単語対をその単語対を抽出したクラス制限付きパターンと共に出力する。

2.2 第二獲得器

第二獲得器は「学習データ作成モジュール」、「関係分類器モジュール」と「候補生成モジュール」というモジュールからなる。以下で各モジュールを説明する。

候補生成モジュール このモジュールはヒューリスティクスにより関係候補として有望な単語対を含む文をコーパスから選択する。まず、第一獲得器の獲得獲得インスタンスをスコア上位 N 対に限定し (本実験では $N = 25,000$)、それらを抽出したクラス制限付きパターンから意味的クラス対を獲得する。次に、それらの意味的クラス対に属する単語対が共起する文をコーパスから抽出する。最後にこの文集合から獲得対象となっている意味的関係の部分的証拠を含む文にターゲットを絞り込む。そのために上記のクラス制限付きパターンを部分パターンに分割する。クラス制限付きパターンは文内で単語対を結びつける係り受け関係のパスであり、構文木の部分木となる。部分パターンは単語対中の一単語からその部分木の主辞となる単語 (動詞又は形容詞に限定) までのパスと定義する。例えば、「X が Y を引き起こす」というパターンは「X が引き起こす」と「Y を引き起こす」という部分パターンに分割される（X、Y のクラス制限は無視）。

候補生成モジュールは、「ターゲットの意味的クラス対に属する単語対を含む」、かつ、「単語対のいずれかが部分パターンのどちらか一つとマッチする」という 2 つの条件を満たす文をウェブコーパスから収集し、単語対とその文からなる三組を候補として関係分類器モジュールに提供する。

学習データ作成モジュール 本モジュールは以下の関係分類器モジュールが用いる学習データを候補生成モジュールと同様に自動的に第一獲得器の出力とウェブコーパスから作成する。そのために第一獲得器の獲得結果上位 N 対を正例と見なし、それらの単語対が共起する文をコーパスから収集する。候補生成モジュールと同様に、部分パターンをマッチする文に限定し、正例データをこの単語対と文の組からランダムで選択する。一方、負例は、正例の文に含まれるが第一獲得器の獲得結果に含まれない、任意の単語対とその文から作成する (文自体は正例の文と同じ)。本実験で正例・負例をそれぞれ 25,000 に設定した。

関係分類器モジュール 本モジュールは学習データ作成モジュールが自動生成したデータを用いて SVM 分類器を学習し、候補生成モジュールが提供する候補の単語対とそれを含む文の三組を分類する。

素性は、単語対の周辺にある形態素の unigram と bigram 、候補の単語対と他に文に出現する名詞の意味的クラス、文がマッチした部分パターン、学習データで単語対間によく現れる形態素である。SVM としては TinySVM(3 次元の多項式カーネル) を利用した。

最終的には関係分類器モジュールが ⟨ 単語 1 , 単語 2 , それを含む文 ⟩ の三組を SVM スコア（分離超平面からの距離）と共に出力し、各単語対に関してスコアを最大化する三組のスコアを最終スコアと見なし、単語対をそのスコア順に出力する。

3 評価

獲得対象のウェブコーパスとしては、KNP で解析された 6 億ウェブページ群を使用した。第一獲得器で使われる構文パターンは係り受け関係のパスに相当する構文木の部分木であり、最大 8 文節からなる部分木に限定した。ただし、すべての部分木を無制限で使用するとパターンの数が膨大になり、計算が困難になるため、コーパスの一部の約 5 千万文書（全コーパスの 1/12）でパターン頻度を概算し、共起する（抽出できる）名詞対の異なり数が 10 個以下のものを取り除いた。パターンベース法と提案手法を比較する際にこの頻度の閾値に関しては微妙な問題があるが、それについては後ほど議論する。

また、実験の目的は提案手法が第一獲得器で獲得できないインスタンスを多数獲得することを示すためであるため、第一獲得器の獲得結果上位 N 対を抽出したパターン（クラス制限無し）と共起する単語対を第二獲得器の出力から取り除いた。つまり、第一獲得器でも獲得が可能な単語対は評価対象から除外した。

提案手法を「因果関係」、「予防策関係」の獲得タスクによって評価した。提案手法の出力となるスコア順で並べられた単語対とそれが共起する文の三組のラン

ダムサンプルを評価者３人に評価してもらい、単語対が獲得対象の関係を持つ証拠が文中に見つかれば、そのサンプルを正解とした。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が正解とした場合、lenient な評価基準では評価者２名以上が正解とした場合に正解と見なす。全評価実験で評価者間の一致率 (kappa) が約 0.6 であった。

3.1 評価結果

以下では (*i*) 第二獲得器が第一獲得器では獲得できないインスタンスを高精度で獲得できたことと、(*ii*) 提案手法が全コーパスで一回しか出現しないという超低頻度パターンからも正しいインスタンスを獲得できることを示す。

まず、第一獲得器の獲得結果の上位 25,000 対の精度は，Lenient 基準で因果関係に関して 80.8%で、予防関係に関しては 72.8%であった。因果関係のシードパターンは「X が Y を引き起こす」、「X が Y の原因となる」というようなパターンを、合計で 5 個用いた。予防関係のシードパターンには「X が Y を防ぐ」の言い換えとなるパターン 11 個を利用した。

これによって生成された候補の数は、因果関係の場合、⟨ 単語対、文 ⟩ が 554,452 組、ユニークな単語対 247,438 対であった。予防関係の場合は、⟨ 単語対、文 ⟩ が 125,734 組、ユニークな単語対が 77,139 対であった。

(*i*) **提案手法の精度**　図 1 と 2 は獲得した関係（ユニークな単語対とそのスコアを最大化する文）をスコア順に並べたランダムサンプルの精度を示す。サンプル数は因果関係が 300 個で、予防関係が 200 個である。Lenient 基準では、因果関係の場合、第一獲得器では獲得できない名詞対 10 万対を 75%以上の精度で獲得できた。

(*ii*) **超低頻度パターンの取り扱い**　まず、超低頻度パターンの極限のケースの取り扱いについて評価した。単語対の共起統計に基づいてパターン学習を行う手法では、全コーパスに一回しか出現しないパターンは学習が困難である。そういったパターンを SO パターン (“*Single Occurrence*”) と呼ぶ。SO パターンとしか共起しない関係インスタンスはパターンベース手法にとって獲得不可能である。なぜならば、単語対の共起統計に基づいたパターン学習法は新しいパターンの信頼性を推定するために、そのパターンと reliable なパターン（シードパターンを含む）と共起する単語対の重複を調べ、重複が大きいパターンを reliable なパターンの集合に追加する。重複が小さいパターンは新しいインスタンスを抽出できるテンプレートとして考慮されない。SO パターンと共起する単語対が reliable なパターンとも共起している場合には、既に獲得され、

図 1: 獲得した因果関係の精度（スコア順）。

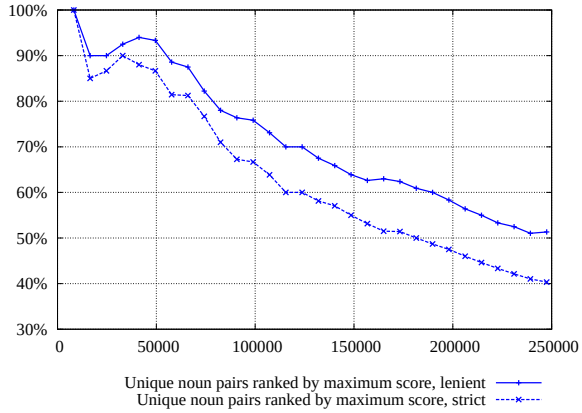
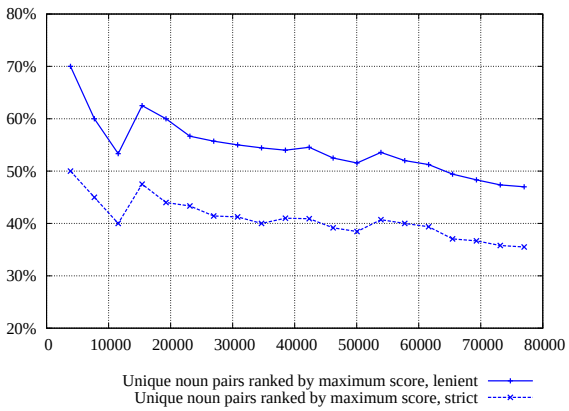


図 2: 獲得した予防関係の精度（スコア順）。



SO パターンをその獲得に使用する必要がない。一方、単語対が reliable なパターンと共起しない場合には、SO パターンは reliable なパターンとの共通集合が空で、学習されない。どちらにしても、SO パターンはパターンベース手法では活用できない。

本手法で獲得されたインスタンスのうち、SO パターンとしか共起しないものは、因果関係の場合は 8,716 対、予防関係の場合は 7,369 対あった（約 1 割）。

表 1 にその例を挙げる（因果関係のみ）。SO パターンとしか共起しない因果・予防関係の候補、それぞれ 200 サンプルを評価した。Lenient 基準で上位 25%の精度が 56%(因果関係) と 52%(予防関係) であった。精度が出力全体より低いとは言え、従来のパターンベース法では獲得不可能なインスタンスを獲得できることが確認できた.

また、提案手法は SO パターンに限らず超低頻度なパターンを含む文を活用できることを示す。図 3 は第一・第二獲得器に関して、獲得されたインスタンスの何割がどれぐらいの頻度のパターンによって抽出されたかを示す。パターン頻度はパターンが共起する単語対の異なり数として定義し、log scale で X 軸に表示される。Y 軸が出力全体に占める割合を表す。赤いグラフがパターンベース手法に基づく第一獲得器に関する

表 1: SO パターンの文から獲得された因果関係と出力全体の 25 万対での順位。⟨⟩ と [] が ⟨ 原因 ⟩ と [結果] を示す。

- 理由は，⟨ 食べカス ⟩ がキーボードの間に挟まったりして，[故障] の原因になるかもしれないからです。(131,880 位)
- ⟨ 頻脈発作 ⟩ が始まりましてキシロカインを静注しましたところ、患者さんが突然意識をなくして [全身痙攣] を起こしたということです。(176,506 位)
- ⟨ カテコラミン ⟩ が心拍数の急上昇をもたらすことから，血管内の血行状態の変化が血管内障害につながり，[血栓形成] を促進する。(234,464 位)

その割合であるが、出力の大半のインスタンスが数千から数万単語対と共起するパターンによって獲得されたことが分かる。一方、提案手法の出力の 48.18%が全コーパスでユニークな単語対 10 個以下と共起するパターンで獲得された。つまり、第二獲得器がパターンベースの従来手法より実際に低頻度パターンに対処できることが分かる。

なお、節 3 で前述したように、第一獲得器の実装時固有事項としてパターンを少なくとも単語対 10 個以上と共起するものに限定した。このパターン頻度の閾値を取り払うことで、パターンベース手法であっても超低頻度パターンに対処できる可能性について議論する。図 3 が示すように、パターンベース手法の出力のほとんどが 1000 個以上の異なる単語対と共起するパターンによって抽出された。もし，パターンベース手法が低頻度パターンを学習し，活用できるのであれば，少ない単語対としか共起しないパターンによって獲得されたインスタンスが多数存在するべきである。仮に 1/12 の部分コーパスで用いた頻度閾値を全コーパスまで拡大すれば、$10 * 12 = 120$ 単語対と共起するパターンも一定量で含まれるはず。しかしながら，図 3 では 1000 以下の単語対しか共起しないパターンがほぼ活用されていないことから，そのような傾向は見られない。そのため，頻度閾値を下げて，低頻度なパターンを考慮したとしても，パターンベース手法でそれらが活用されるとは考え難い。

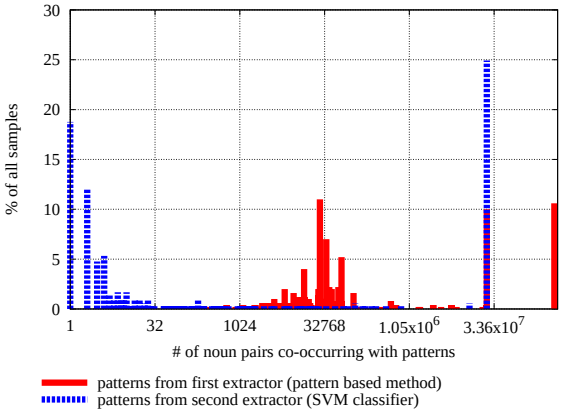
従って、第一獲得器と第二獲得器が扱えるパターンの頻度の差がパターン頻度閾値に起因する可能性が低いと考えられる。つまり、パターンベース手法は中頻度・高頻度のパターンしか学習できない傾向にあり，低頻度パターンの活用は困難であると考えられる．

また、単語対を獲得したパターンが複雑かどうかは定量化し難いが、仮に単語対間の形態素数によって測定すれば、因果関係でその平均値は第一獲得器が 3.30 であり、第二獲得器が 7.54 であり、後者の方が複雑なパターンから関係を獲得していることが分かった。

4 おわりに

本論文では、２つの異なる獲得器から成る意味的関係獲得手法を提案した。一つは既存のパターンベース手法で、もう一つは前者のシステムの出力で自動的に学習された分類器である。本手法の特徴は、大規模なウェブコーパスでも複雑でまれにしか出現しない構文パターンのみと共起する意味的関係のインスタンス、つまり単語対が獲得できる点にある。極端な場合には、全コーパスに一回しか出現しない構文パターンからも、ターゲットの意味的関係を持つ単語対が獲得できることを実験により示した。そうした long tail の意味的関係は網羅性重視の意味的関係データベースを構築する際に非常に重要である。

図 3: 第一獲得器・第二獲得器にとって最も有効なパターンの頻度（因果関係）。



参考文献

[1] E. Agichtein, L. Gravano, J. Pavel, V. Sokolova, and A. Voskoboynik. Snowball: a prototype system for extracting relations from large text collections. In *Proc. of the 2001 ACM SIGMOD international conference on Management of data*, SIGMOD '01, 2001.

[2] S. Brin. Extracting patterns and relations from the world-wide web. In *Proc. of the 1998 International Workshop on Web and Databases (WebDB' 98)*, 1998.

[3] S. De Saeger, K. Torisawa, J. Kazama, K. Kuroda, and M. Murata. Large Scale Relation Acquisition Using Class Dependent Patterns. In *Proc. of the 9th International Conference on Data Mining (ICDM)*, pages 764–769, 2009.

[4] S. De Saeger, 鳥澤健太郎, 風間淳一, 黒田航, and 村田真樹. 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　味的関係獲得. In **言語処理学会第** *16* **回年次大会** *(NLP2010)*, pages 932–935, 2010 年 3 月.

[5] M. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proc. of the 14th International Conference on Computational Linguistics (COLING'92)*, pages 539–545, 1992.

[6] 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.

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