類推による単語間の意味的関係獲得法

土田正明 †¶　De Saeger Stijn†　鳥澤健太郎 †
村田真樹 †　風間淳一 †　黒田航 †　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
{m-tsuchida,stijn,torisawa,murata,kazama,kuroda}@nict.go.jp
ohwada@ia.noda.tus.ac.jp

1　はじめに

単語間の意味的関係は，情報検索，質問応答など，様々な応用が考えられる．例えば，意味的関係として上位〈薬，タミフル〉，効果〈タミフル，インフルエンザ〉は，質問応答システムで,「インフルエンザに効く薬は?」という質問に対して，適切な推論によって「タミフル」と答えるための知識となる．また，機械が自動的に推論するだけではなく，情報検索の支援として，様々な意味的関係によってキーワードの想起を支援することで，ユーザが知らなかった意外な情報も提供できる [5].

単語間の意味的関係は，特定の意味的関係に限定しても量が膨大であるため，人手による作成は現実的でない．そのため，大規模コーパスから単語間の意味的関係を自動的に獲得するための方法が研究されている．しかしながら，多くの先行研究 [2, 3, 6, 1] は，2 語が一文内に存在することを仮定して意味的関係を獲得しているため，獲得可能な単語ペアが限定されていた．これは，良い手がかりとなる言語的パターンで書かれる意味的関係以外は，獲得困難であることを意味する．

本研究の主な狙いは，一文内で書かれている関係，もしくは，一文で書かれているとしても書かれ方が間接的であるため従来のパターンベースの手法では獲得困難な関係も含め，大量の意味的関係を獲得することにある．そのためのアプローチとして，類推の考え方を用いる．類推とは「似ている点を元に他の事を推し量ること」で，例えば，類推によって，銀イオンと銅イオンの類似性から「銀イオンは殺菌効果を持つので，銅イオンも殺菌効果を持つだろう」といった仮説を生成できる．本論文では，こうした類推のプロセスを適切，かつ，自動で適用することで，大量の意味的関係が獲得可能になることを示す．本研究の最終目的は，現在知られていない，もしくは関係性が弱いなどの理由で明文化されていないが成立していそうな関係までも獲得することにある．そのような関係は正否の評価が難しいため，本論文では評価していないが,「未知の関係」は，組織や個人にとって，発想支援やイノベーションにもつながる可能性があるので，今後非常に重要なタスクである．また，この試みが可能になれば，既存の情報を抽出する「情報抽出タスク」を「テキストに書かれていない情報を算出する手段」に拡張することが可能になり，大きな飛躍となる可能性がある．

本論文では，類推をシステマティックに行うことで，意味的関係を獲得する一手法を提案し，実際に従来のパターンベースの方法で困難であった意味的関係を獲得できることを示す．提案法は，1）ある単語ペアとその単語ペアを構成する 2 語それぞれが意味的に似ている単語ペアは，元と同じ意味的関係を持つ可能性が高い，2）何らかの意味的関係を持つ単語ペアは,「テキスト中の近接した箇所に共起しやすい」という 2 つの仮定に基づく．まず，シードとなる意味的関係の単語ペアを入力に，各単語ペアのそれぞれの類似語の組み合わせのうち，一定の頻度で共起する単語ペアを仮説として生成する．次に，得られた仮説を，シードのそれぞれの単語との類似度に基づいてスコアリングする．類推に用いる類似語データには，分布仮説に基づき大規模コーパスから獲得した係り受け関係の類似性を利用する風間らの方法 [9] で作成されたデータを用いる．具体的には，ALAGIN フォーラムで公開されている，日本語の約 50 万名詞に対する類似度付き類似語リスト語[1]を用いる．単語共起頻度は，ALAGIN フォーラムで公開されている，約 1 億文書で上記の約 50 万名詞の全ペアに対して，近接 4 文内で共起する文書頻度を計算した単語共起頻度データベース[2]を用いる．本論文の主な貢献は以下の 3 点である．

- 先行研究で獲得困難であった意味的関係を，単語間の文脈類似度と単語間の共起頻度に基づく類推によって獲得できることを示す．
- 提案法は，先行研究との組み合わせが容易である．例えば，提案法に入力するシードとして，先行研究で自動獲得された関係を用いることができる．また，パターンベースとは異なる視点で，関係の確からしさを評価できるため，従来法で獲得された結果のリランキングにも利用可能と考えられる．
- 提案法は，大量の意味的関係を獲得するために，軽量に動作するようになっている．提案法は，単語間の類似度と共起頻度のみを用いるため，特定の意味的関係やシードに依存する情報を必要としない．この性質から，可能な単語の集合を決めることで，必要な全情報を容易にデータベース化できるため，大量の意味的関係でも高速に獲得できる．

[1] 文脈類似語データベース Version1 の old.500k-2k.data.
[2] 単語共起頻度データベース Version1 の 500k-500k.100m-docs.w4.data

2 類推による意味的関係の獲得法

提案法は，2つの仮定に基づいて，シードとなる意味的関係からの類推によって仮説を生成する．シードには，人手で用意した意味的関係や従来法の獲得結果などが利用できる．

1. ある単語ペアとそれぞれ意味的に似ている単語ペアは，元と同じ意味的関係を持つ可能性が高い
2. 何らかの関係を持つ単語ペアは共起しやすい

すなわち，仮定1に基づき，同じ意味的関係を持っているであろう候補を列挙して，仮定2に基づき，関連のなさそうな候補をフィルタリングすることで，仮説を生成する．ただし，上記の通りに生成された仮説には，正しくないものも含まれるため，各仮説の確からしさを評価して，正しそうな順番に出力できるようにする．

上述の通り，提案法は，大きく2つの処理からなる．

1. **類推による仮説生成**: シードの各関係について，各単語とその類似語の組み合わせのうち，一定以上共起する単語ペアのみを仮説として生成する．
2. **仮説スコアリング**：正しい意味的関係を優先的に獲得するため，各仮説の良さを評価する．「正しい仮説は，多くのシードから高い類似性をもって導出される」という仮定に基づきスコアリングする．

2.1 類推による仮説生成

類推による仮説生成は，意味的関係の集合 $R_{given} = \{r_1 = \langle f_1, s_1\rangle, \ldots, r_n = \langle f_n, s_n\rangle\}$ の各単語ペア r_i に対して，以下のプロセスを行う．ここで，f は単語ペア r の第1項，s は第2項である．

1. 単語ペア $r_i = \langle f_i, s_i\rangle$ の第1項 f_i，第2項 s_i のそれぞれの類似語を取得する．それぞれの類似語集合を，$F_{analogy}, S_{analogy}$ とする．
2. $F_{analogy} \cup f_i$ と $S_{analogy} \cup s_i$ の全組み合わせからなる仮説候補集合 $R^{cand}_{hypothesis} = \{\langle f, s\rangle | f \in (F_{analogy} \cup f_i), s \in (S_{analogy} \cup s_i)\}$ を生成する．
3. $R^{cand}_{hypothesis}$ の各候補のうち，2語の共起頻度が閾値 T_{cooc} 以上の仮説候補のみを $R_{hypothesis}$ に追加する．

本論文の上記1で用いる類似語は，風間らの方法 [9] を用いて獲得する．風間らは，大量コーパスから各名詞 n の ⟨ 助詞, 動詞 ⟩,⟨ の, 名詞 ⟩ の大きく2種類の係り受け関係 dep を収集し，Torisawa の手法 [4],

$$p(n, dep) = \sum_{c_i \in C} P(c_i)P(n|c_i)P(dep|c_i)$$

に基づき，EM アルゴリズムで $P(c), P(n|c), P(dep|c)$ を推定する[3]．これによって，dep をそのまま素性とする場合と比べてスムージング効果が期待できる．次に，上記パラメタから $P(c|n)$ を計算し，名詞 $n1, n2$ の類似度を $P(c|n1), P(c|n2)$ の Jensen-Shannon(JS) ダイバージェンスとして求める．JS ダイバージェンスは確率分布間の距離の一種で，以下の式で計算する．

[3]確率モデルとしては PLSI と等価である

$$JS(P1||P2) = \frac{1}{2}(KL(P1||P_{mean}) + KL(P2||P_{mean}))$$

P1,P2 は確率分布，$KL(P1||P2)$ は KL ダイバージェンス，P_{mean} は P1,P2 をベクトルとしてみた場合の平均である．JS ダーバージェンスは 0 から 1 を取り，小さいほど類似していることになる．そのため，単語 n1,n2 の類似度は $sim(n1, n2) = 1 - JS(P(c|n1)||P(c|n2))$ とする．最終的に，可能な単語集合の中の全ペアについて，1)$sim(n1, n2)$ が閾値 T_{sim} 以上，2) お互いの類似度トップ M 単語に含まれる，の2つの条件を満たす単語ペアを類推のための類似語として獲得する．

仮説生成のイメージを説明するために「原因 ⟨ 脳梗塞，急死 ⟩（脳梗塞が急死の原因となる）」からの類推を考える．「脳梗塞」の類似語として $F_{analogy}$={ 心筋梗塞, 脳卒中，うつ病 },「急死」の類似語として $S_{analogy}$={ 死亡，病死 } が獲得されたとする．シードの各語も含め，全組み合わせを仮説候補とするので，本例では，$R^{cand}_{hypothesis}$ ={⟨ 脳梗塞，急死 ⟩⟨ 心筋梗塞，急死 ⟩⟨ 脳卒中，急死 ⟩⟨ うつ病，急死 ⟩...⟨ うつ病，死亡 ⟩, ⟨ うつ病，病死 ⟩} となる．このように，類似語同士の組み合わせは，同じ意味的関係を持つ可能性が高い．しかしながら，この段階では ⟨ うつ病，死亡 ⟩ など直感的に不適切な仮説も含まれる．このような関連が低い単語ペアは，共起しにくいと考えられるため，T_{cooc} 以上の頻度で共起する単語ペアの仮説のみを仮説集合 $R_{hypothesis}$ に加える．

2.2 仮説スコアリング

仮説スコアリングでは $R_{hypothesis}$ の各仮説の確からしさを「正しい仮説は多くのシードから高い類似性で導出される」と仮定して計算する．つまり，様々なシードから導かれた仮説ほど確からしいという考え方である．

仮説と各シードとの関係は，第1項のみを類似語に置き換えた**第 1 項類推**，第2項のみを類似語に置き換えた**第 2 候類推**，両方置き換えた**全類推**の3種に大別される．全類推は他の2つに比べて，信頼性が低くなると考えられる．

仮説 $\langle f_h, s_h\rangle$ の仮説スコア $S(f_h, s_h)$ は，第1項類推スコア $S_{FA}(f_h, s_h)$，第2項類推スコア $S_{SA}(f_h, s_h)$，全類推スコア $S_{FULL}(f_h, s_h)$ を計算し統合する．各スコアは，以下の式で計算する．

$$S_{FA}(f_h, s_h) = \sum_{f_i \in FA(s_h)} sim(f_h, f_i)$$

$$S_{SA}(f_h, s_h) = \sum_{s_i \in SA(f_h)} sim(s_h, s_i)$$

$$S_{FULL}(f_h, s_h) = \sum_{\langle f_i, s_i\rangle \in R_{given}} sim(f_h, f_i)sim(s_h, s_i)$$

$FA(s)$ は R_{given} で第2項が s の関係の第1項の集合，$SA(f)$ は R_{given} で第1項が f の関係の第2項の集合，$sim(.,.)$ は，前節で説明した単語間の類似度である．式を見てわかるとおり，各スコアは多くのシードの類似語から構成される仮説ほど高くなる．

各スコアの統合法は，3 種のスコアの和である $S^{sum}(f_h, s_h)$ と，積である $S^{prod}(f_h, s_h)$ の2種類が考

えられる．S^{sum} は，類推の種類に関係なく，いずれかのスコアが高い場合に高くなる．S^{prod} は，全種類のスコアが高い場合に高くなる．つまり，S^{prod} は，「バランスよく両方の項に基づき類推される仮説が良い」と考える点で，和と異なる．S^{prod} の計算では，0 になることを回避するため，各スコアに十分に小さい値を足す．

この仮説スコアを全シードを用いて計算し，仮説をランキングすることで，確からしい仮説が優先的に獲得できるようになると期待できる．

3 評価実験

本節では，提案法の有効性を評価する．具体的には，「原因 ⟨X，Y⟩(X は Y の原因となる)」の獲得タスクで，提案法の精度と，従来のパターンベースの方法で困難であった関係を獲得できているかを評価する．

評価は，3 人の評価者によって 1）常識的に正解，2）常識で判断できない場合は，ウェブに正しいと支持するエビデンスが 1 つ以上見つかった場合を正解として，2 名以上一致 (lenient)，3 名一致 (strict) で精度を測定した．2）では，1 つの関係に関して，YahooAPI を用いて「X,Y，原因」の AND 検索で 10 ページ獲得し，各ページから「X,Y，原因」が 200 文字以内に存在するテキストセグメントを最大 3 つ抽出して，最大 30 個 (=10 × 3) のセグメントを提示した．本実験では，各評価者で合計 400 個の評価を行ったが，評価者間の kappa 値は，平均で，0.629 であった．一般的に，kappa 値が 0.6 以上ならば「かなり良い一致率」と言われていることから，評価者間の判定の一致率は概ね良いと言える．

類似語獲得には，風間らの方法 [9] で作成された約 50 万名詞に対する類似度付き類似語リスト[4]を用いた．単語共起頻度は，約 1 億文書で上記と同じ約 50 万名詞の全ペアに対して，近接 4 文内で共起する文書頻度を計算したデータ[5]を用いた．

類似語は，類似度閾値 $T_{sim} = 0.7$ を超え，相互のトップ $M = 20$ 語に含まれてる 2 語を獲得した．仮説候補フィルタリングのための共起頻度の閾値は，$T_{cooc} = 20$ とした．これらパラメタは，経験的に，シードの約 10 倍の仮説が生成されることを目安に設定した．

提案法のシードは，De Saeger らの方法 [3] で獲得した関係から，明らかに不適切な関係をクリーニングした上で，トップ 1 万個を用いた．[3] による関係獲得でパターン学習に用いたデータは約 5 千万文書で，対象の単語集合は，本実験と同様である．方法の詳細は，[3] に譲るが，シードパターンを入力し，それら**シードパターンと同じ 2 語を抽出できる全パターンを用いて関係を再獲得**してランキングするため，パターンベースの方法では，最高レベルの網羅性と考えられる．

本評価法と同様の基準でシードの精度を測定したところ lenient で 0.80,strict で 0.70 であった．つまり，ノイズが含まれるシードからの類推となる．ただし，本評価は [3] と方法が異なり，[3] と比較するとやや低めの値となる傾向にある点に注意されたい．類推の結果，1 万のシードから 102290 個の新しい仮説が生成された．

提案法は，仮説スコアの計算法で，1) 統合法が和 (sum),2) 統合法が積 (prod) の 2 種類が存在するため，それぞれでランキングした結果を評価した．精度は，シードの関係を除いた上で，各スコアのトップ 1 万から 100 個，1 万から 3 万の 100 個の 200 個を評価した．

結果を図 1 に示す．図 1 の 15,000 位以降の精度は，トップ 1 万までの精度とトップ 1 万から 3 万までの精度を用いて，数に応じた加重平均で計算した．lenient を正解とすると，prod はトップ 1 万の精度が 0.63,sum は 0.53 であった．シードの精度が lenient で 0.80 であることを考えると，提案法は，精度が低下しているが，文中での書かれ方を用いずこの精度を達成したという見方もできると考えられる．図 1 の通り，sum と prod を比較すると，prod の方が上位の精度が高いため，良いスコアとなっていると考えられる．

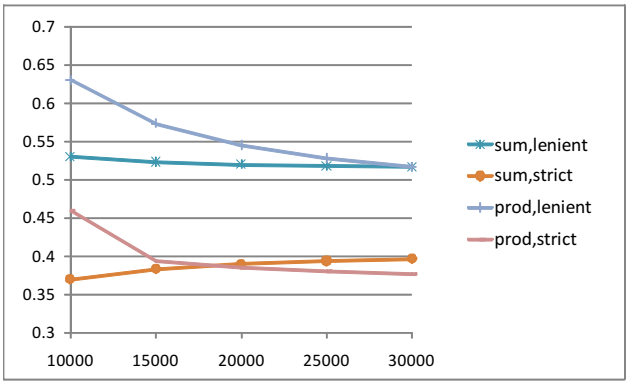

図 1: トップ n の精度

また，パターンベースで獲得困難である関係が獲得できたかを調べた．具体的には，提案法のトップ 1 万の lenient を正解と考え，De Saeger らの方法 [3] で，順位が 100 万位以下である関係の数の割合を調査し，その割合を用いてトップ 1 万に含まれる正解の関係数を推定した．ただし，De Saeger らの方法では，5 千万文書を用いていることに対して，提案法は 1 億文書での共起頻度を用いているので，フェアな比較とは言えない．厳密には，文書集合を揃えた上で比較すべきで，De Saeger らとの比較は，予備的なものである．結果を表 1 に示す．表 1 より，実際に，パターンベースの従来法では獲得困難であった関係が獲得されていることが確認できた．

最後に，提案法で獲得された中で正解と判定された関係とその類推に用いられたシードを表 2 に示す．⟨ミネラル不足，花粉症⟩は，シードの両方の項を置き換えた全類推のみから生成された仮説である．全類推のみからの仮説は，元の関係との関連が弱くなるため，不適切な仮説が多くなると考えられるが，この問題に対して，提案法は，一定以上共起する仮説のみを採用しているため，精度を維持できていると考えられる．今後，共起頻度によるフィルタリングの効果を評価する予定である．また，⟨食習慣，ニキビ⟩は，常識的に確からしい仮説と考えられるが，1 億文書中の 4 文内共起が 35 回と低頻度である点が興味深い．一般的とは言えない⟨ミネラル不足，花粉症⟩が 105 回共起していることから「常識的，もしくは一般的なことはあまり言及されない」という可能性を示唆している．ただし，

[4]ALAGIN フォーラムで公開されている文脈類似語データベース Version1 の old.500k-2k.data.

[5]ALAGIN フォーラムで公開されている単語共起頻度データベース Version1 の 500k-500k.100m-docs.w4.data

これは用いた文書集合（ウェブ文書）の性質による可能性もあるため，この点も今後調査して行きたい．

表 1: トップ 1 万中，従来法で獲得困難な関係の推定数

	[3] の 100 万位以下
sum	約 3100 個 (31/100)
prod	約 3300 個 (33/100)

表 2: 類推で獲得された例

獲得された関係	**類推元のシード**
〈 ミネラル不足, 花粉症 〉	〈 カルシウム不足, アトピー 〉, など
〈 食習慣, ニキビ 〉	〈 生活習慣, ニキビ 〉, など

4 関連研究

関係獲得の研究には，関係を直接獲得する研究 [2, 3, 7] と与えられた 2 つの関係の類似度を測る研究 [6, 1] に分けることができる．

どちらも含めて，多く先行研究は，関係の獲得や類似度計算の手がかりを 2 語を結ぶパターンとしている [2, 3, 6, 1]．これは，文内での 2 語の書かれ方は，関係を特定する有効な手がかりとなるため当然とも言える．本研究は，パターンではなく単語間の類似性と共起に基づき関係獲得を行っている点で，これらの手法と異なる．

加藤ら [7] の研究は，パターンを用いずに関係を獲得している点で本研究と類似している．加藤らは単語ペア X,Y の関係の手がかりとなる「関係接続語」を獲得して，関係のスコアリングに用いている．具体的には，ウェブ検索で 1)X,Y が含まれる文書，2)X のみ含まれる文書，3)Y のみ含まれる文書を取得し，1 と 2，1 と 3 を用いて，X,Y が含まれる文書で有意に出現確率が高くなる語を「関係接続語」として獲得する．加藤らの手法は，シードが増える度に関係接続語の計算が必要であるため，計算量が大きく，大規模な関係獲得が困難と考えられる．また，可能な単語の集合を限定しても，事前に全組み合わせに「関係接続語」を事前にデータベース化するのは容易でない．一方，本研究は，大規模な関係獲得を想定している．本研究は，可能な単語の集合を決めることで，容易に全単語の組み合わせに対して，類似度と共起頻度をデータベース化できるため，大規模な関係獲得では本研究が優位であると考えられる．

また，発想支援を目的とした類推による仮説生成の研究に，石川ら [8] の研究がある．シードとなる関係の語を置き換えて仮説を生成するという点では同じであるが，置き換える語が語基を共有する語（例：ペプチドと抗菌ペプチド）のみである点，仮説のスコアリングが考慮されていない点で，本研究とは異なる．

5 まとめ

本論文は，一文内で書かれている関係，もしくは，一文で書かれているとしても書かれ方が間接的であるため従来のパターンベースの手法では獲得困難な関係も含め，大量の意味的関係を獲得することを目的として，類推に基づく意味的関係の獲得法を提案し，評価を行った．

提案法は，1）ある単語ペアとそれぞれ意味的に似ている単語ペアは，元と同じ意味的関係を持つ可能性が高い，2）何らかの意味的関係のある単語ペアは共起しやすい，という 2 つの仮定に基づき，まず，シードとなる意味的関係の単語ペアを入力に，各単語ペアのそれぞれの類似語の組み合わせのうち，一定の頻度で共起する単語ペアを仮説として生成する．次に，得られた仮説を，シードのそれぞれの単語との類似度に基づいてスコアリングする．

評価実験では，パターンベースの方法 [3] で自動獲得した 1 万の因果関係をシードに，新たな意味的関係のトップ 1 万を精度 0.63 で獲得できた．さらに，従来のパターンベースの手法では獲得困難な意味的関係が獲得できていることを確認した．

参考文献

[1] B. Danshuka. Measuring the similarity between implicit semantic relations from the web. In *Proc. of the 18th WWW*, pp. 651–660, 2009.

[2] P. Pantel and M. Pennacchiotti. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proc. of the 21st COLING and 44th ACL(COLING-ACL-06)*, pp. 113–120, 2006.

[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 ICDM*, pp. 764–769, 2009.

[4] K. Torisawa. An Unsuperviesed Method for Canonicalization of Jpanese Postpositions. In *Proc. of the 6th NLPRS*, pp. 211–218, 2001.

[5] K. Torisawa, Stijn De Saeger, Y. Kakizawa, J. Kazama, M. Murata, D. Noguchi, and A. Sumida. TORISIKI-KAI, An Autogenerated Web Search Directory. In *Proc. of the 2nd IUCS*, pp. 179–186, 2008.

[6] P.D. Turney. A Uniform Approach to Analogies, Synonyms, Antonyms, and Associations. In *Proc. of the 22nd COLING*, pp. 905–912, 2008.

[7] 加藤誠, 大島裕明, 小山聡, 田中克己. 共起に基づく Web からの類似関係のブートストラップ抽出. 日本データベース学会論文誌, Vol. 8, No. 1, pp. 11–16, 2009.

[8] 石川大介, 石塚英弘, 藤原譲. 特許文献における因果関係を用いた類推による仮説の生成と検証–ライフサイエンス分野を対象として–. 情報知識学会誌, Vol. 17, No. 3, pp. 164–181, 2007.

[9] 風間淳一, De Saeger Stijn, 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, pp. 84–87, 2009.