情報処理学会論文誌 Vol. 52 No. 4 1761–1776 (Apr. 2011)

単語分布類似度を用いた類推による単語間の意味的関係獲得法

土田正明[†1,†2] デ・サーガ ステイン[†1] 鳥澤健太郎[†1]
村田真樹[†3] 風間淳一[†1]
黒田航[†4,†5] 大和田勇人[†2]

情報爆発の時代に入り，大規模コーパスと計算機パワーの増大を背景に，構文的パターンに基づいて「因果関係」などの単語間の意味的関係の知識を獲得する研究が進められている．しかしながら，それらの研究は，文書中に直接的かつ明示的に書かれた知識を獲得するにとどまり，人間であれば解釈可能な間接的記述から獲得することや，文書に書かれていない知識を過去に蓄積された知識からの推論によって大規模に獲得することは行われていない．このような知識の獲得は，より大量の関係を獲得するためだけではなく，人類のイノベーションの加速にとっても重要である．本稿では，既存の構文的パターンに基づく方法で獲得された単語の意味的関係のデータベース，すなわち，特定の意味的関係を持つ単語対の集合を，類推によって大規模に拡張する方法を提案する．提案法は，入力された単語対の中の語を，ウェブから自動獲得した類似語に置換して大量の仮説を生成し，さらに単語間の類似度に基づいて仮説をランキングする．提案法は，従来法では困難な間接的記述からの意味的関係獲得を可能にして，さらには，そもそも文書に記述されている可能性が低い知識を獲得できる．約1億ページのウェブ文書を用いた実験によって，これらを検証するとともに，いくつかの意味的関係に関して，提案法で上位にランキングされた仮説では，最新の構文パターンに基づく獲得法とほぼ変わらない精度を達成できることを示す．

Analogy-based Relation Acquisition Using Distributionally Similar Words

Masaaki Tsuchida,[†1,†2] Stijn De Saeger,[†1]
Kentaro Torisawa,[†1] Masaki Murata,[†3]
Jun'ichi Kazama,[†1] Kow Kuroda[†4,†5]
and Hayato Ohwada[†2]

With the advent of terabyte scale corpora in this information explosion age, extracting high-level semantic relations like causality using lexico-syntactic patterns has come of age. While such knowledge acquisition methods have matured greatly, they are necessarily limited to extracting relations mentioned explicitly in some text collection. Until now, inference-based methods for acquiring "indirect" or "implicit" relational knowledge from a corpus have never been investigated on the same scale as pattern-based methods. In this work we propose a method for extending a database of semantic relations acquired by existing pattern-based methods using analogical reasoning. This method uses lexical word similarities acquired automatically from the Web to generate and rank new relation instance candidates from its input. Not only can it acquire semantic relations from indirect descriptions in the corpus, which is exceedingly difficult for pattern-based methods, our method can acquire valid relational knowledge that is unlikely to be written down before. We validate these claims experimentally using a 10^8 Web page corpus, and show that for some relations our method exhibits precision figures indistinguishable from state-of-the-art pattern-based methods in top-ranked relation instances.

1. はじめに

情報爆発の時代に入り，ウェブに代表されるような大規模コーパスにアクセス可能になったことと，計算機パワーの増大を背景に大規模コーパスに基づく統計的自然言語処理技術が進歩したことで，自然言語処理の適用可能性は劇的に増大している．本稿では，人工知能の教科書で必ずといってよいほど取り上げられてきた，いわゆる「意味ネットワーク」に相当する知識を，類推によって大量に獲得する方法を提案する．

この意味ネットワークをかつて人工知能が目指していた常識を解する知的なソフトウェアで活用しようとすれば，巨大なものとなることはいうまでもない．実は，すでに，統計的自然言語処理技術をウェブに適用することで，一個人の知識量ではカバーできない，100万語オーダの意味ネットワークを蓄積することが可能となり始めてきている[8),29),30),33)]．さら

†1 情報通信研究機構
National Institute of Information and Communications Technology
†2 東京理科大学
Tokyo University of Science
†3 鳥取大学
Tottori University
†4 京都工芸繊維大学
Kyoto Institute of Technology
†5 早稲田大学総合研究機構
Comprehensive Research Organization, Waseda University

に実際に一般ユーザが，彼ら自身の観点からして「意外でありながら有用な情報」を，そうした意味ネットワークから取得することが可能となっており，また，一部の企業ではそうした意味ネットワークの実用化に向けた動きが始まっている[38]．

しかしながら，こうした技術は，ほとんどの場合，いわゆる構文パターンを用いる方法（以降パターンベース法と呼ぶ）である．より具体的には，たとえば，因果関係に関する知識を獲得することを考えると「X は Y の原因である」といった構文的パターンを大量のテキストに適用し，X および Y の位置に来る単語の対を因果関係を持つ単語の対として抽出するといった手法である．こうした手法は，テキストに直接的かつ明示的に書かれた知識のみを抽出，蓄積するにとどまっており，たとえば，後に例示するような，人間であれば解釈可能な非明示的な記述からの知識の獲得はいまだ困難な課題である．さらに，明文化されたことはないが，人間であれば，過去に蓄積された知識からの推論によって導出できるような知識についても，大規模に獲得，蓄積する研究は行われていない．一部，生物医学のテキストマイニングの分野では，複数の文献の情報を組み合わせて，有望な仮説を発見する研究[23)–25)]が行われているが，これらは，MEDLINE や MeSH ⋆1 のメタデータなど，専門家が長い時間をかけて高度に整備した情報の存在を前提としているため，そのような情報がない分野や情報源に適用することは困難である．

本稿では，このような問題意識に基づき，ウェブからすでに獲得，蓄積された意味ネットワーク，すなわち，単語の意味的関係の集合を類推によって大規模に拡張する手法を提案する．本稿は，獲得対象のコーパス中に書かれていない単語間の意味的関係も含めて，正しい単語間の意味的関係を獲得することを目的としている．すなわち，獲得対象のコーパスに書かれているかとは関係なく，そのコーパスから正しい関係を得ることを「獲得」とする．ここでは単語の意味的関係とは，特定の意味的関係を持つ名詞の対の集合を指すものとする．たとえば，「因果関係 ⟨ウィルス，病気⟩（ウィルスが病気の原因になる），材料関係 ⟨ぶどう，ワイン⟩（ぶどうがワインの材料になる）」などである．類推とは本来精度が低い推論の形式であると見なされがちであるが，知識のタイプによっては，大量のウェブ文書（約 1 億ページ）の情報をうまく活用することで，最新のパターンベース法[8)]によって獲得，蓄積された知識とそれほど遜色ない精度で知識の拡張を行えることを実験によって示す．さらに，少なくとも処理対象となっているコーパス，つまり，ウェブ文書 1 億ページに間接的にしか書かれていない，もしくは書かれていない可能性が高い知識までもが獲得できることを実験により示す．これは，処理対象となるコーパスが今後増大し，仮想的にウェブ全体と見なせるような状況になったとしても，提案法によって，そこに書かれていない知識を獲得できる可能性があることを示唆している．大規模なコーパスに書かれていない知識の中には，未来に発見され，正しいと検証されるものも含むと考えられるため，提案法は，イノベーションを支援するエンジンとして期待できる．また，本稿には，単に意味ネットワークを構築するだけでなく，これまで人でなければ実施不可能であった「類推」のプロセスを，従来研究[27)]のような人手による作りこみを避けつつ，大規模に計算機上で実現し，「類推」の経験論的な検討を行う試みの第一歩としての意義がある．そもそも「類推」はおそらく人類が誕生して以来，日常的に使われているもので，これまで様々なイノベーションにおいて活用されてきたが，経験論的，情報学的な検討が十分に行われてきたとはいい難い．この検討をさらに進めて，システマティックにイノベーションを加速させる手がかりを見つけることは非常に重要であると考えている．

類推とは「似ている点をもとにして他のことを推し量ること」であるが，我々は「似ている点」を見つける手がかりとして「似ている文脈で出現する語は意味的に似ている」とする分布仮説[11)]を採用する．すなわち，提案法は，ターゲットとなる意味的関係を持つ単語対の集合（以降，シードインスタンスと呼ぶ）に現れる語を，大規模コーパスから分布仮説に基づいて収集された類似語に置換することで，新しい仮説を生成する．たとえば，因果関係の獲得で，シードインスタンスとして「因果関係 ⟨老廃物，耳鳴り⟩（老廃物が耳鳴りの原因である）」が与えられ，「耳鳴り」と「めまい」が類似語として獲得されている場合は，仮説として，因果関係 ⟨老廃物，めまい⟩ を生成する．このようなシードインスタンスは，人手で用意することも考えられるが，既存のパターンベース法[7)–9),19),20)]を用いることで，低コストに用意できる．実際，本稿の評価実験では，既存のパターンベース法[8)]の出力に簡単なクリーニングを行うことでシードインスタンスを作成している．

提案法の具体的なアルゴリズムは以下のとおりである．まず，上で述べたように，各シードインスタンスの語を類似語に置き換えて仮説を生成する．たとえば，シードインスタンスとして「因果関係 ⟨老廃物，耳鳴り⟩（老廃物が耳鳴りの原因である）」が与えられ，「耳鳴り」が「めまい」の類似語として獲得されている場合，「耳鳴り」を類似語の「めまい」に置き換えて「因果関係 ⟨老廃物，めまい⟩」という因果関係の仮説を生成する．同様に，全シードインスタンスと類似語に関して上記処理を行い「確からしい仮説は，複数のシードインスタンスから生成される」という考えに基づき，各仮説のスコアを計算することで，確からしい順に仮説を取得する．たとえば，「因果関係 ⟨老廃物，頭痛⟩」というシードインスタ

⋆1 Medical Subject Headings（http://www.nlm.nih.gov/mesh/）

ンスと「耳鳴り」と「頭痛」が類似語であれば，前述の仮説である「因果関係 ⟨老廃物，耳鳴り⟩」は，確からしさを増す．また，さらなる精度向上のため「何らかの意味的関係のある 2 語は近接共起する」という考えに基づき，2 語の共起頻度が閾値以下の場合にフィルタリングする方法も試みる．実際に，評価実験では，共起頻度として，約 1 億の日本語のウェブページを対象に，近接 N 文内の共起頻度を用いる（実験では $N = 4$ とした）．

以下に，最後に述べた近接 N 文内の共起頻度を用いる効果の端的な例を示す．

> “**老廃物**は，様々な健康上の問題を引き起こします。薬 X は，あなたの体の中から老廃物を除去します。これにより，**めまい**，耳鳴り，頭痛，冷え性などが改善します。”

人間ならば，この文書から「老廃物」が「めまい」の原因と読み取れるが，パターンベース法で文の境界をまたがるような長い範囲のパターンを利用することは，現在の最新の技術でも現実的ではない．このように，パターンベース法では獲得が難しい場合でも，関係のある 2 語は近い範囲に共起する傾向にあると考えられる．一方，提案法は，類推と上記のように近い範囲に共起しているという手がかりとを併用することで，いっさいのパターンを用いることなく，上述のような「因果関係 ⟨老廃物，耳鳴り⟩」を獲得できる．

評価実験では，約 1 億の日本語ウェブページを用いた 3 種類の意味的関係の獲得タスクによって，提案法の効果を示す．3 種類の意味的関係とは，因果関係 ⟨X,Y⟩（X が Y の原因である），材料関係 ⟨X, Y⟩（X が Y の材料・原料である），予防関係 ⟨X,Y⟩（X が Y を抑制・予防する）である．特に，因果関係の獲得では，既存のパターンベース法[8] で獲得した約 10,000 のシードインスタンスを入力に，約 70%の精度で新たな 10,000 の関係を獲得できた．ちなみに，これは，獲得に用いたパターンベース法[8] の精度（73%）とほぼ変わらない精度で，同数の因果関係を獲得できたことになる．

また，因果関係獲得の実験で得られた約 17 万の仮説の中に，パターンベース法では実質的に獲得不可能といえる，1 億ウェブページ内で文内共起のない単語対でありながら，正しく因果関係を持つと思われるものが 1 万個以上含まれていた．さらには，4 文内でも共起がなく，そもそも対象となっているウェブページ内で因果関係の存在が記述されていない可能性が高い単語対でありながら，正しく因果関係を持つと思われるものが約 1,700 個含まれていることを確認した．これは，間接的な記述しかない，もしくは，そもそも記述がない可能性が高い意味的関係を獲得できていることを示す．この結果は，2 語の直接的な言及を手がかりにすることなく，大規模コーパスから計算した意味的類似度が，意味的な関係をある程度推測できたことを示しており，非常に興味深い．

以降，本稿は，以下のように構成される．2 章で，提案法の詳細を説明する．3 章では，評価実験について述べる．4 章では，関連研究を通して本研究の位置づけを説明し，5 章で結論を述べる．

2. 提案方法：類推による関係獲得

提案法は，目的となる関係を持つ名詞の対（シードインスタンス）を入力として，類推によって仮説を生成し，生成された各仮説の確からしさのスコアを計算して，ランキングする．本研究では，シードインスタンスを既存のパターンベース法[8] で獲得する．本手法では，各シードインスタンスの各語を意味的な類似語に置換することで，類推による仮説生成を行う．次に「多くのシードから高い類似度の類似語によって生成された仮説ほど良い」という仮定に基づいて，各仮説のスコアを計算し，仮説をランキングする．また，さらなる精度向上のため，「何らかの関係のある 2 語は近接共起する」という仮定に基づき，仮説の 2 語の近接 N 文内の共起頻度を用いて，閾値以下の共起頻度の仮説をフィルタリングする方法も提案する．以降，それぞれの詳細を説明していく．

2.1 シードインスタンスの獲得

本研究では，既存のパターンベース法[8] を用いてシードインスタンスを獲得する．De Saeger らの方法[8] は，獲得したい単語間の関係を表現する少数の構文パターン（シードパターンと呼ぶ）を入力として，まず，シードパターンと同じ意味的関係を表現可能な一種の言い換えと見なせる構文パターン（言い換えパターンと呼ぶ）を大量に学習する．言い換えパターンとは，シードパターンのいずれかで獲得できるインスタンスを獲得できる構文パターンである．次に，それらすべての言い換えパターンを用いてインスタンスを獲得し，各言い換えパターンの言い換えスコアなどを用いて，各インスタンスをスコア付けしてランキングする．各言い換えパターンの「言い換えスコア」は，シードパターンで獲得できるインスタンスの集合と，言い換えパターンで獲得できるインスタンスの集合の Jaccard 係数として計算する．すなわち，シードパターンと同じインスタンスを獲得できる言い換えパターンほど，言い換えスコアが高くなる．

また，言い換えパターンの学習では，X や Y にとれる単語の意味（意味クラス）に制約をかけることで，意味的に曖昧な構文パターンの曖昧さを解消している．たとえば，単語の意味クラスを [意味クラスを表すラベル] と表すと，「X による Y」という構文パターンは「[組織] による [製品]（例：Apple による iPad）」ならば会社と製品の関係，「[道具] による [作業]（例：パソコンによる編集）」ならばツールと目的の関係，「[化学物質] による [症状]（例：

アセトアルデヒドによる二日酔い)」ならば因果関係といったように，様々な関係を表すことから，曖昧な構文パターンといえる．一方で，上記の例のように，曖昧な構文パターンでも，X と Y にとる単語の意味クラスに制限をかけることで，表現される関係が区別できる．

単語の意味クラスは，名詞の確率的クラスタリング法[14] を用いて自動獲得する．Kazama ら[14] は，Torisawa[28] と同様に，名詞 n，助詞 r，動詞 v の同時確率 $P(n,r,v)$ を，隠れクラス c_i を導入し，以下のように定式化している．

$$P(n,\langle r,v\rangle)=\sum_{c_i\in C}P(c_i)P(n|c_i)P(\langle r,v\rangle|c_i)$$

各パラメータ $P(c)$，$P(n|c)$，$P(\langle r,v\rangle|c)$ は，下記の対数尤度を目的関数とした EM アルゴリズムによって推定できる．具体的な EM アルゴリズムに基づくパラメータの更新式は，文献 14)，28) を参照されたい．

$$L=\sum_{n_k\in N}\sum_{\langle r_l,v_m\rangle\in Rel\times V}n(n_k,\langle r_l,v_m\rangle)\log\left(\sum_{c_i\in C}P(c_i)P(n_k|c_i)P(\langle r_l,v_m\rangle|c_i)\right)$$

ここで，N は名詞の集合，Rel は助詞の集合，V は動詞の集合，C は隠れクラスの集合，$n(n_k,\langle r_l,v_m\rangle)$ は，コーパス中で名詞 n_k と助詞 r_l からなる文節が動詞 v_m に係る回数である．$\langle r_l,v_m\rangle$ は，名詞 n_k の文脈情報と見なせるため，本定式化によって，似た文脈情報を持つ名詞が同じ隠れクラスから生成されやすくなることが分かる．最後に，推定された $P(n_k|c_i)$，$P(c_i)$ を用いて，$P(c_i|n_k)=\frac{P(n_k|c_i)P(c_i)}{\sum_{c_j\in C}P(n_k|c_j)P(c_j)}$ を計算することで，名詞 n_k の各クラスへの所属確率を表す事後分布を計算する．文献 8) では，$|C|=500$ として，$P(c_i|n)\geq 0.2$ のとき，名詞 n がクラス c_i に属すると見なしている．我々の実験でもこれと同様の設定を用い，名詞，助詞の文節と動詞の文節の係り関係は，約 1 億の日本語ウェブページからなる TSUBAKI コーパス[22] から獲得した．

我々の De Saeger らの方法[8] の実装では，TSUBAKI コーパスから，以下の基準によって構文パターンを網羅的に導出した．具体的には，同コーパスで頻出する約 32 万の名詞に対して，各文の係り受け解析木上の 2 つの名詞の最短パスを構文パターン候補とし，以下の基準を満たす候補のみを構文パターンとした．

- 2 つの名詞を結ぶ係り受け構造上の最短パスが 5 文節以下である．
- 10 種類以上の名詞ペアを抽出できる．

結果，約 1.56×10^8 種類の構文パターンが導出された．この数は，単語の意味クラスの制約を用いない，すなわち，構文パターンの表層文字列の異なり数である．De Saeger らの方法[8] では，これらの構文パターンの中から，シードパターンの言い換えパターンを学習し，それらのすべてを用いて名詞ペアの関係を抽出する．

2.2　類推による仮説生成

以下，n，m は，名詞を表す変数とする．また，シードインスタンスの集合を $R_{seed}=\{(n_1,m_1),\ldots,(n_i,m_i)\}$ とし，(n_i,m_i) は，関係のインスタンスとなる名詞ペアとする．$n'\in SW(n)$ は，n' が n の類似語の集合 $SW(n)$ に属することを示し，言い換えると，n' と n が類似語であることを表す．また，$n\in SW(n)$ とする．

提案法は，関係のインスタンスとなる仮説の集合 $R_{hypo}=\{(n',m')\mid(n,m)\in R_{seed},\ n'\in SW(n),\ m'\in SW(m)\}$ を，R_{seed} の各インスタンスの名詞 n と m をそれらの類似語 $n'\in SW(n)$ と $m'\in SW(m)$ でそれぞれ置換することで生成する．$n\in SW(n)$ であるため，(n,m') など，シードインスタンス (n,m) の片方のみを置き換えた仮説も生成される．

多くの仮説を生成するためには，大規模に類似語を獲得する必要がある．WordNet など既存のシソーラスから獲得する方法[2),4)] も考えられるが，登録されている名詞の語義が約 146,000 程度[⋆1] とやや少ない．

大量の類似語を獲得するために，本研究では，大量のウェブ文書から分布仮説[11] に基づき類似語を獲得する方法[35] を用いる．分布仮説とは「意味的に似ている語句はその出現文脈の分布も似ている」という考え方である．Hagiwara ら[10]，風間ら[35] では，2.1 節で述べた確率的クラスタリングによって得られた各名詞の隠れクラスへの事後確率の分布を用いて，2 つの名詞 n と m の類似度 sim を $P(c|n)$ と $P(c|m)$ の Jensen-Shannon（JS）ダイバージェンス[6),16)] という確率分布間の距離に基づき計算することが行われている．ただし，2.1 節とは異なり，隠れクラス数 $|C|$ は風間ら[35] と同様に 2,000 とした．

$$JS(P(c|n)||P(c|m))=\frac{1}{2}(KL(P(c|n)||P_{ave})+KL(P(c|m)||P_{ave}))$$
$$sim(n,m)=1-JS(P(c|n)||P(c|m))$$

ここで，$KL(.||.)$ は，Kullback-Leibler ダイバージェンスで，$P_{ave}=(P(c|m)+P(c|n))/2$ である．JS ダイバージェンスは，距離の尺度で 0 から 1 をとり，近いほど 0 に近づく．類似度は距離と逆の概念であるため，風間ら[35] は，$sim(n,m)=-JS(P(c|n)||P(c|m))$ を類似度としている．本研究では，類似度を正の値にする目的で $sim(n,m)=1-JS(P(c|n)||P(c|m))$ としているが，ある単語から見た 2 つの単語の類似度の順序関係やそれらの類似度の差は

⋆1 WordNet 3.0 の場合

風間らの類似度と等価である．また，Hagiwara ら[10] によって[⋆1]，上記の JS ダイバージェンスに基づく類似度計算が，他の基本的な方法と比べて比較的良い性能であることが実験で示されている．また，風間ら[35] は，計算コストを回避する近似計算を用いながら，100 万語の大規模な語彙に対して類似度の高い語を 500 個列挙することを行っており，実際に計算した類似語リストが ALAGIN フォーラム[⋆2]で公開されている．実験では 2.1 節で述べた De Saeger らの方法[8] で獲得対象となる約 32 万語間の類似語を公開されている類似語リスト[3] から取得した[⋆3]．

我々は，この類似度を用いて，名詞 n の類似語 $n' \in SW(n)$ を以下の 2 つの条件を満たすように獲得する．

- $sim(n, n') \geq T_{sim}$
- 各 n に関して，類似度順に最大 M 個

T_{sim} は，よく似た語のみを類似語と認定するためのパラメータで，M は 1 語あたりの類似語を同程度の量にするためのパラメータである．

このように獲得した意味的な類似語を用いて，各シードインスタンスの各語を類似語に置き換えて，新たな仮説を生成する．仮説生成のイメージを説明するために「因果関係 ⟨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 {心筋梗塞，脳卒中，うつ病}，「急死」の類似語として {死亡，病死} が獲得されたとする．元のシードの語も含め，全組合せを仮説候補 R_{hypo} とするので，本例では，$R_{hypo} =$ {⟨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．

2.3　仮説スコアリング法

提案するスコアリング法は，2 つの仮定に基づく．1 つは「確からしい仮説は，多くのシードインスタンスから生成される」であり，もう 1 つは「より高い類似度の語での置換で生成される仮説ほど確からしい」である．提案のスコアリング法は，これら仮定を反映した式で計算する．

類推によって得られた仮説とシードインスタンスの関係は，以下の 3 種に分類できる．

- 第 1 項類推：シードインスタンスの第 1 項の置換によって生成された仮説
- 第 2 項類推：シードインスタンスの第 2 項の置換によって生成された仮説
- 全類推：シードインスタンスの両方の置換によって生成された仮説

仮説 (n', m') のスコアは，上記 3 種の類推に関するサブスコアをもとに計算する．S_{FA} が第 1 項類推スコア，S_{SA} が第 2 項類推スコア，S_{FULL} が全類推スコアである．

$$S_{\mathrm{FA}}(n', m') = \sum_{(n,m') \in R_{seed}, n \neq n'} sim(n, n')$$

$$S_{\mathrm{SA}}(n', m') = \sum_{(n',m) \in R_{seed}, m \neq m'} sim(m, m')$$

$$S_{\mathrm{FULL}}(n', m') = \sum_{(n,m) \in R_{seed}, n \neq n', m \neq m'} sim(n, n') \times sim(m, m')$$

全類推は，両方の語が置換された仮説であるため，第 1 項類推，第 2 項類推のような片方の語の置換の仮説と比べて，信頼度が低い類推である．

図 1，**図 2**，**図 3** に，因果関係を例とした各種類の類推とそのスコアの説明を示す．ノードは，関係を構成する名詞，実線の矢印はシードインスタンス，破線の矢印は仮説，破線は類似語のペアをそれぞれ表す．これらの図は，仮説「因果関係 ⟨脳梗塞，突然死⟩」が，様々なシードインスタンスからの第 1 項類推（図 1），第 2 項類推（図 2），全類推（図 3）で生成されていることを表す．図 1，図 2 では，破線の sim の和がそれぞれ S_{FA}，S_{SA} となる．図 3 では，各シードインスタンスと仮説に関して，対応する項を結ぶ sim の積の和が S_{FULL} となる．

我々は，上記サブスコアを用いて，ランキングのための 2 種類の仮説スコア[⋆4]を提案する．

$$S^{\mathrm{SUM}}(n', m') = S_{\mathrm{FA}}(n', m') + S_{\mathrm{SA}}(n', m')$$

$$S^{\mathrm{SUM}}_{\mathrm{FULL}}(n', m') = S^{\mathrm{SUM}}(n', m') + S_{\mathrm{FULL}}(n', m')$$

S^{SUM} は，S_{FA} か S_{SA} が高い場合に高くなる．S^{SUM} で，S_{FA} と S_{SA} の両方が 0 の場合は，その仮説は出力しない．これは，両方が 0 の場合は「全シードインスタンスを考慮しても，両方を置換しないと生成できない仮説」と言い換えられることから，元のシードインスタンスとの類似性が低く，結果として精度が低くなると考えられるためである．一方，$S^{\mathrm{SUM}}_{\mathrm{FULL}}$ では，S_{FA} と S_{SA} が 0 であっても $S_{\mathrm{FULL}} > 0$ である仮説，つまり得られた全仮説のスコアを計算

⋆1 類似度の式は $sim(n, m) = e^{-\lambda JS(P(c|n)||P(c|m))}$ と我々と異なるが，ある単語から見た 2 つの単語の類似度の順序関係は同じである．

⋆2 http://www.alagin.jp/

⋆3 評価実験で，De Saeger らの方法[8] と提案法を比較するために，仮説生成に用いられる単語は 2.1 節で述べた我々の De Saeger らの実装で獲得対象となる 32 万語に限定している．

⋆4 さらに，スコア $S^{\mathrm{PROD}}(n', m') = S_{\mathrm{FA}}(n', m') \times S_{\mathrm{SA}}(n', m')$ を検討したが，実験の結果 S^{SUM} と性能の差がなかったため，紙面の都合上割愛する．

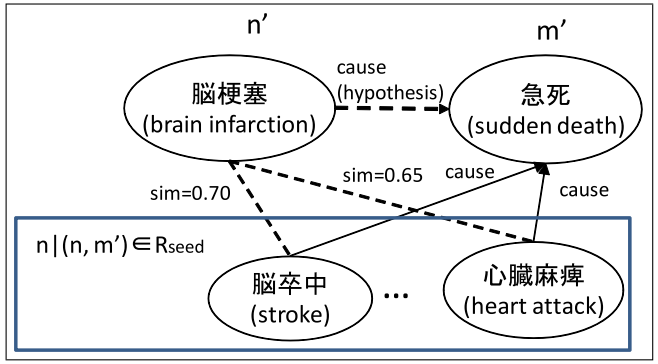

図 1　第 1 項類推とそのスコア
Fig. 1　Illustration of First Argument Analogy and its score.

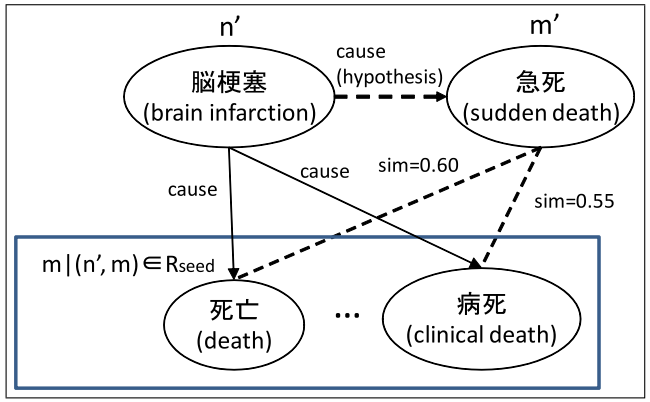

図 2　第 2 項類推とそのスコア
Fig. 2　Illustration of Second Argument Analogy and its score.

して，ランキングする．$S^{\mathrm{SUM}}_{\mathrm{FULL}}$ は，S^{SUM} と比べて精度が低くなると考えられるが，より多くの関係を獲得できる．

これらのスコアは，式から分かるとおり，1) 多くのシードインスタンスから，2) 高い類似度を持つ類似語による置換で生成された仮説のスコアが高くなるため，我々のスコアリングの 2 つの仮定を反映している．そのため，これらのスコアでランキングした上位の仮説の精度が高い場合に，我々の仮定が正しいといえる．

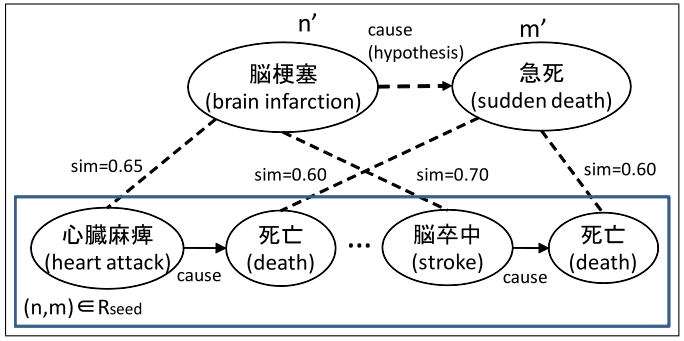

図 3　全類推とそのスコア
Fig. 3　Illustration of Full Analogy and its score.

さらに「確からしい仮説は，多くのシードインスタンスから生成される」という仮定を検証するために，スコア S^{MAX} を提案する．$S^{\mathrm{MAX}}(n', m')$ は，仮説 (n', m') に対して，各シードインスタンス (n, m) との $sim(n, n')$ と $sim(m, m')$ の中で最も高い sim をスコアとする．具体的には以下の式となる．

$$S^{\mathrm{MAX}}_{\mathrm{FA}}(n, m) = \max_{(n', m) \in R_{seed}, n \neq n'} sim(n, n')$$

$$S^{\mathrm{MAX}}_{\mathrm{SA}}(n, m) = \max_{(n, m') \in R_{seed}, m \neq m'} sim(m, m')$$

$$S^{\mathrm{MAX}}(n, m) = \max(S^{\mathrm{MAX}}_{\mathrm{FA}}(n, m), S^{\mathrm{MAX}}_{\mathrm{SA}}(n, m))$$

つまり「1 つでも高類似度の類似語で置換されて生成された仮説が確からしい」という考え方に基づく．そのため，仮に S^{MAX} の上位の精度が S^{SUM} と同等の場合は「確からしい仮説は多くのシードインスタンスから生成される」という仮定は成り立たないことになる．この結果は，3 章の評価実験で述べる．

2.4　共起頻度フィルタリング

精度向上を目的とするフィルタリング法を提案する．基本的なアイディアは「何らかの関係のある 2 語は，文書中の近い範囲で共起しているだろう」という仮定に基づく．具体的には，大規模コーパス内の近接 N 文内の共起頻度を用いて，仮説の 2 語の共起頻度が T_{cooc} 未満の場合は，その仮説をフィルタリングする．T_{cooc} は，パラメータである．これによって，類似語での単純な置換によって関係のない仮説が生成されてしまった場合にでも，それらの仮説を除去できるため，精度向上が期待できる．

3. 評 価 実 験

評価実験では，以下の 3 点を検証する．

(1)　提案のスコアリング法の有効性．
(2)　共起頻度フィルタリングの有効性．
(3)　提案法によって，獲得対象のコーパスからパターンベース法で獲得困難な関係，獲得対象のコーパスには明示的に書かれていない関係，さらには書かれていない可能性が高い関係を獲得できているか．

3.1 実験の設定

我々は，以下の 3 種類の意味的関係獲得のタスクで提案法を評価した．

- **因果関係**：第 1 項が直接的，もしくは間接的に第 2 項の原因である関係．例：⟨ウィルス，病気⟩
- **材料関係**：第 1 項が第 2 項の材料・原料・材質である関係．例：⟨グレープ，ワイン⟩
- **予防関係**：第 1 項が直接的，もしくは間接的に第 2 項を抑制・予防する関係．例：⟨コーヒー，眠気⟩

これら 3 つの意味的関係は，パターンベース法として提案法と比較する De Saeger らの方法[8] で実験されているため，本稿もこれに合わせた．具体的には，3 人の評価者が，提案法で獲得された上記関係のランダムサンプルの正否を評価した．評価の基準として，2 人以上正解と判断した「lenient」，3 人一致で正解と判断した「strict」の 2 種類を用いる．

関係の正否の評価は，評価者の知識や感覚に依存しないように，ウェブ検索を用いて証拠となりうる短いテキスト（最大で 200 文字）をいくつか取得し，それらのテキストのどれかに提示された語の対が目標とする意味的関係を持つことが書かれていれば正解とし，そうでなければ不正解とした．具体的には，Yahoo!API ⋆1 を用いて，各仮説の 2 語と各関係を表す代表的な語（因果関係なら「原因」，材料関係なら「原料」，予防関係なら「防ぐ」）が含まれるテキストを最大 30 個取得した．本方法による評価者間の判定結果の一致度を調べるため，本実験で行った評価結果を用いて，各評価者のペアの kappa 値の平均を測定した．結果は，因果関係で 0.71，材料関係で 0.74，予防関係で 0.58 であった．kappa 値は，0.60 以上ならば “substantial agreement [15]” といわれているため，本方法でおおむね妥当な基準で評価できているといえる．

本評価法は，ウェブをドメイン非依存の知識ベースと考え，1) 十分に大きいコーパス（Yahoo!API でアクセス可能な文書）ならば，多くの正しい意味的関係が書かれている，2) ある程度の大規模コーパス（約 1 億の日本語ウェブページからなる TSUBAKI コーパス[22]）に書かれていない正しい意味的関係でも，それよりも十分に大きいコーパス（Yahoo!API でアクセス可能な文書）にならば書かれているという 2 つ仮定に基づいている．参考として，TSUBAKI コーパスと Yahoo!API でアクセス可能な文書数の比を，「今日」「明日」「企業」「情報」など高頻度語のヒット件数の比から推定したところ，Yahoo!API からアクセス可能な文書数は TSUBAKI コーパスの 100 から 200 倍であった．Yahoo!API からアクセスできる文書数は，TSUBAKI コーパスに比べて十分に多いと考えられるため，TSUBAKI コーパスに書かれていない意味的関係の一部は，Yahoo!API を用いて検証可能であると考えられる．すなわち，本評価法で，TSUBAKI コーパスには書かれていない関係を仮想的に「未知の知識」と見なすことで，未知の知識を発見できたか否かを，ある程度定量的に評価可能になると考えられる．しかしながら，本評価法でも，人類にとって真に未知の関係や，既知であっても書かれていない関係は検証できないという問題は残される．また，本評価法は，ウェブ上の文書に書かれていれば正しいと見なしているが，仮説の真の正否は科学的な実験で注意深く検証する必要があるという点で，本評価法は仮説が真に正しいか否かではなく「ウェブに書かれている関係か否か」を評価していることに注意されたい．

次に，実験データについて説明する．類似語とフィルタリングに用いる共起頻度の取得は，約 1 億の日本語ウェブページからなる TSUBAKI コーパス[22] を用いた．関係獲得のターゲットとなる名詞は約 32 万語である．共起頻度は，近接 4 文（$N = 4$）内の共起頻度を用いた．類似語獲得では，約 32 万の名詞に対する類似度データをいくつか見ながら，経験的に $T_{sim} = 0.5$，$M = 20$ に設定し，結果として，約 214 万の類似語ペアが得られた．

シードインスタンスは，TSUBAKI コーパスを用いて，同じターゲットの名詞に関して既存のパターンベース法[8] によって獲得した．具体的には，De Saeger らの方法[8] で獲得したインスタンスから，手動で，各関係に対して，約 30 分のフィルタリング⋆2 を行ったうえで，上位 1 万インスタンスを用いた．上記の方法で獲得された各意味的関係のシードインスタンスを，本研究と同じ方法で評価したところ，基準が「lenient」の場合，因果関係では

⋆1 Yahoo!API（http://developer.yahoo.co.jp/webapi/search/）

⋆2 De Saeger ら[8] により提案されているクラスペアに基づくフィルタリング法を行った．De Saeger らは，各意味的関係を構成しやすい名詞の意味的なクラスのペアを関係獲得の手がかりの 1 つに用いているため，人手で不適切な意味クラスのペアを見つけることで，そのクラスペアを持つ関係を容易に削除できる．

79%，材料関係で 74%，予防関係で 57%であった[⋆1]．

上記データを用いて，共起頻度フィルタリングを用いず（$T_{cooc} = 0$）に，提案法を実行したところ，S^{SUM} では，約 17 万の因果関係，約 16 万の材料関係，約 17 万の予防関係が獲得された．また，$S^{\mathrm{SUM}}_{\mathrm{FULL}}$ では，因果関係で約 87 万，材料関係で約 79 万，予防関係で約 94 万の仮説が獲得された．これらの関係は，すでにシードインスタンスを除去しているため，すべて新規の仮説である．

3.1.1 比　較　法

本評価では，提案法と以下の方法とを比較した．

- **ランダム法**：共起頻度フィルタリングで用いるのと同じ共起頻度データを用いて 10 回以上共起する 2 語のペアをランダムに生成して仮説とした．このランダム法による 100 個の仮説を評価したところ，3 種類の意味的関係は 1 つも含まれていなかった．これはターゲットとしている関係の獲得が自明なタスクでないことを示している．紙面の都合上，以降では特に言及しない．
- S^{MAX}：2.3 節で説明した「多くのシードインスタンスから生成された仮説が確からしい」という仮定を確認するための比較手法である．グラフの凡例は「(S^{MAX})」である．
- **パターンベース法**：比較のためのパターンベース法として De Saeger らの方法[8] を用いた．以降のグラフの凡例を「(De Saeger et al.)」として表す．

3.2 スコアリングによるランキングの効果

本節では，比較手法も含めて，上位にランキングされた仮説の精度を測定することで，関係獲得の精度と提案するスコアリングの効果を評価する．具体的には以下を示す．

- いくつかの意味的関係で，提案法の出力のランキング上位ではパターンベース法とほぼ変わらない精度を達成する．
- 「高い類似度の語で置換される仮説が確からしい」というスコアリングの仮定が妥当である．
- 「多くのシードから生成される仮説が確からしい」というスコアリングの仮定が妥当である．

上記を示すため，各関係について，S^{SUM} の出力の上位 5,000 からの 100 サンプル，上位 5,000 から 20,000 までの 150 サンプルを評価した．また，各関係について，S^{SUM} で生成した全仮説からの 300 サンプルと，$S^{\mathrm{SUM}}_{\mathrm{FULL}}$ で生成された全仮説の 1,000 サンプルも評価した．

各関係の上位 20,000 の結果を**図 4**，**図 5**，**図 6** に示す．各図は，横軸がランクで，縦軸が各ランクまでの評価済みサンプルから推定した累積精度を示している．図 4 から図 6 は，サンプリングの密度を補正するため各サンプルに重みを付けて計算している[⋆2]．凡例「$T_{cooc} = 1$」は S^{SUM} に対して T_{cooc} を 1 に設定して共起頻度フィルタリングを行った場合，「S^{MAX}」は，S^{MAX} でランキングして，同じサンプリング法で評価した結果である．図 4，図 5，図 6 を見ると，ランクが下がるに従って，精度も低下する傾向にあった．また，後に示す各関係の全体の精度（図 7 から図 9 の Lenient (No cooc filtering) の末端の精度）と比較しても，上位の精度が高いことが分かった．S^{SUM} の共起頻度フィルタリングなし（no cooc filtering）と S^{MAX} を比較すると，その精度に大きな差が確認できた．提案法（図 4 から図 6 の $T_{cooc} = 1$）の性能を表す 1 つの指標として，提案法でシードインスタンスと同量獲得した場合の精度を調べると，上位 10,000 の精度は，因果関係で 71%，材料関係で 54%，

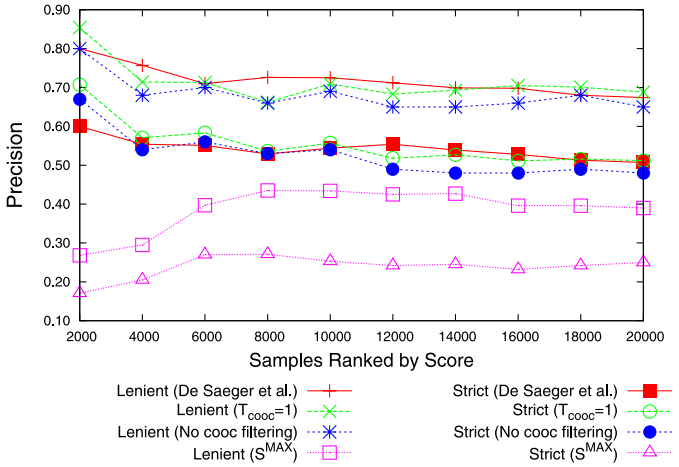

図 4　因果関係：上位 2 万の精度
Fig. 4　Causation: top 20,000 results' precision.

⋆1 本評価法は，De Saeger ら[8] の方法と比べると，精度が低くなる傾向にある．De Saeger らは，評価者に提示する証拠として，良い手がかりとなる単語ペアを抽出した構文パターンを含む文をコーパス中から取得しているため，正しいと思われる証拠を提示できる可能性が高い．一方，我々は，獲得にパターンを用いていないため，同じ方法をとることができない．

⋆2 5,000 から 20,000 の 150 サンプルは，5,000 あたり 50 個のサンプルとなるため，2 回観測された（2 倍の重みをかける）と考えると，擬似的に上位 5,000 からの 100 サンプルと同じ密度でサンプリングしたと見なせる．

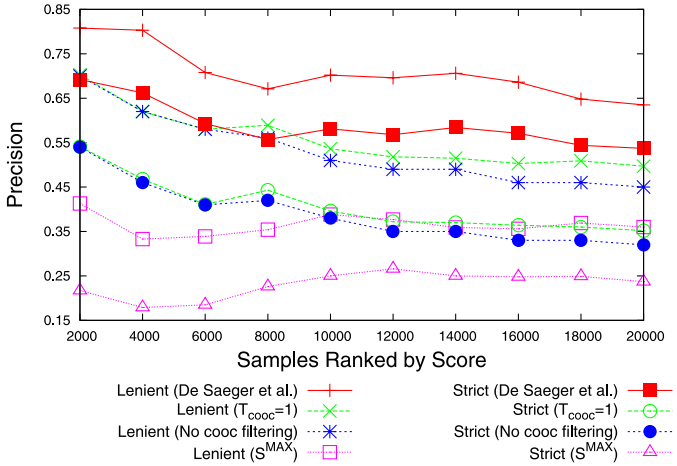

図 5　材料関係：上位 2 万の精度
Fig. 5　Material: top 20,000 results' precision.

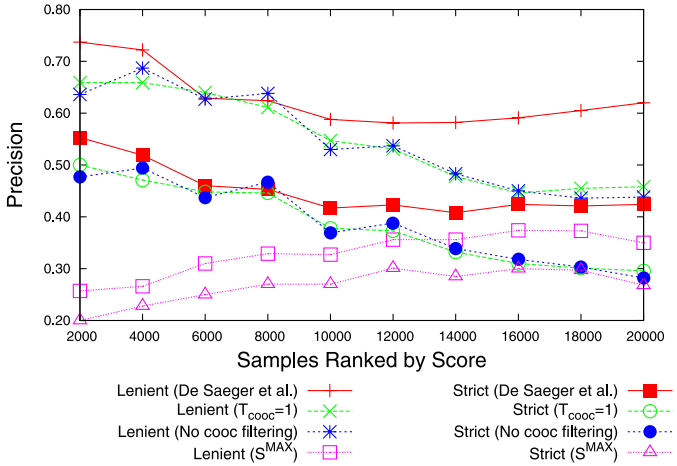

図 6　予防関係：上位 2 万の精度
Fig. 6　Prevention: top 20,000 results' precision.

予防関係で 55%であった．提案法は，上位 10,000 の因果関係，予防関係でパターンベース法（因果関係：73%，予防関係：59%）に比較的近い精度を達成している．これは，構文パターンをいっさい用いていないことを考えると，注目に値する結果である．一方，材料関係では，パターンベース法（70%）と大きな差があった．

同様に，各関係の S^{SUM} の仮説全体からの 300 サンプルから推定した lenient 基準による精度を**図 7**，**図 8**，**図 9** に示す．図 7 から図 9 で，各関係の S^{MAX} と仮説全体の精度[⋆1]を比較すると，S^{MAX} の上位の精度が高いことから，我々のスコアリングのための仮定の 1 つである「高い類似度の語で置換される仮説が確からしい」が妥当であると確認できる．また，上記仮定に加えて「多くのシードから生成される仮説が確からしい」という仮定も反映した提案スコアである S^{SUM} は，S^{MAX} よりも上位で良い精度を達成している．このことから，「多くのシードから生成される仮説が確からしい」というスコアリングの仮定も妥当であると確認できる．

また，我々は，S^{SUM} のサブスコアである S_{FA}（第 1 項類推スコア），S_{SA}（第 2 項類推スコア）の単独の効果も調べた．図 7 から図 9 の凡例「S_{FA}」「S_{SA}」がそれぞれを表す．「S_{FA}」

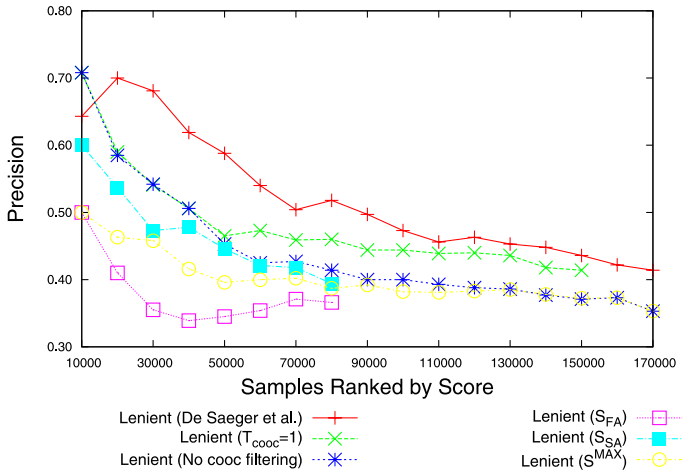

図 7　因果関係の精度（lenient）
Fig. 7　Precision of all causality (lenient).

⋆1 S^{MAX} も S^{SUM} もランキング手法なので，全体の精度は同じであることに注意．

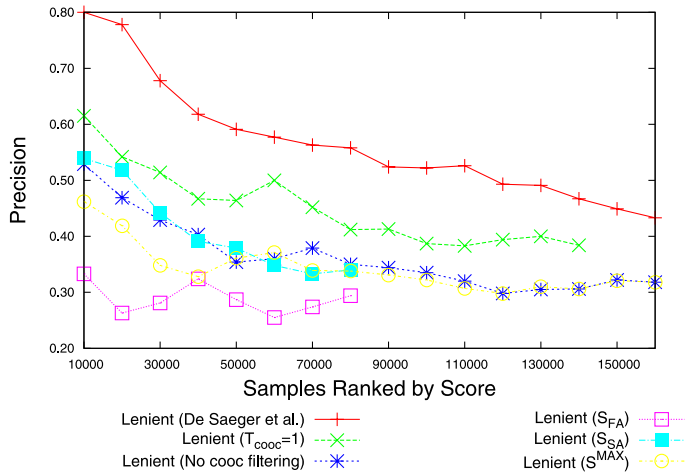

図 8　材料関係の精度（lenient）
Fig. 8　Precision of all material (lenient).

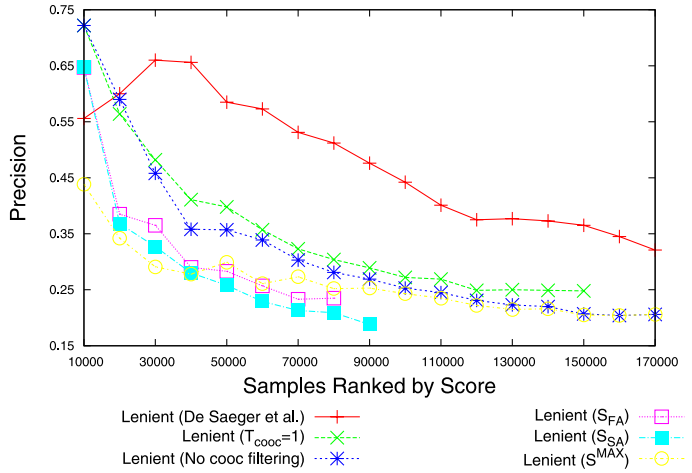

図 9　予防関係の精度（lenient）
Fig. 9　Precision of all prevention (lenient).

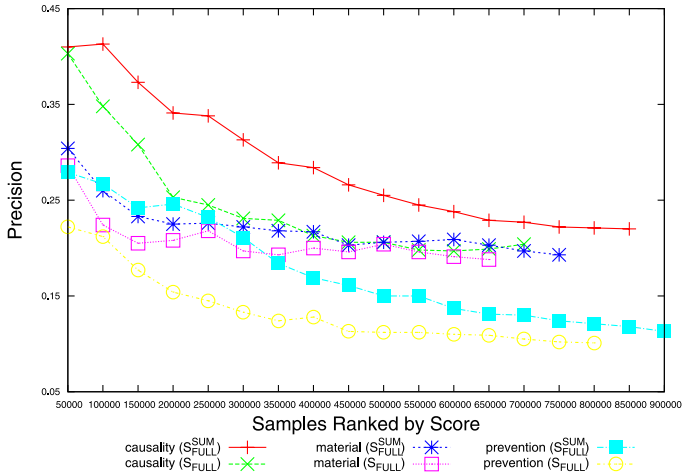

図 10　因果，材料，予防関係の $S^{\mathrm{SUM}}_{\mathrm{FULL}}$ と S_{FULL} の精度（lenient）
Fig. 10　Precision of causality, material and prevention of $S^{\mathrm{SUM}}_{\mathrm{FULL}}$ and S_{FULL} (lenient).

と「S_{SA}」では，それぞれの値が 0 となる仮説もあり，それらを除去したためその線は途中で途切れている．図 7 から図 9 で，共起頻度フィルタリングを用いていない S^{SUM} である「No cooc filtering」と「S_{FA}」や「S_{SA}」を比較すると，その精度は因果関係（図 7），予防関係（図 9）で S^{SUM} が高く，材料関係（図 8）で S^{SUM} と S_{SA} は同等であった．総合的に見ると S_{FA}，S_{SA} の和である提案法スコアの S^{SUM} は，それぞれのサブスコアを単独で用いるよりも上位の精度向上に有効であることを示している．

次に，**図 10** に各関係の全類推 $S^{\mathrm{SUM}}_{\mathrm{FULL}}$ による全仮説の lenient 基準による精度を示す．凡例「S_{FULL}」は，全類推スコアのサブスコア S_{FULL} でランキングしたもので，$S^{\mathrm{SUM}} = 0$，すなわち，シードインスタンスの両方の語を置換しないと生成できない仮説が除去されている．図 10 の「causality」は因果関係，「material」は材料関係，「prevention」は予防関係をそれぞれ表す．図 10 を見ると，各関係で，$S^{\mathrm{SUM}}_{\mathrm{FULL}}$ のほうが S_{FULL} よりも精度が高くなっていることから，$S^{\mathrm{SUM}}_{\mathrm{FULL}}$ のサブスコアである S^{SUM} が上位の精度向上に有効であることを示している．

最後に，因果関係，材料関係，予防関係に関して，S^{SUM} の全仮説の 300 サンプルの中から，誤りの仮説の 50 サンプルに対してエラー分析の結果を示す．具体的には，エラーの原因を，1) シードインスタンスの誤りによるもの，2) シードインスタンスは正しいが類義語置換によって誤った仮説を生成したものの 2 種類に分けた．原因 1 は，たとえば，因果関

表 1　S^{SUM} の因果関係，材料関係，予防関係の 50 サンプルのエラー分析
Table 1　Error anaysis of 50 samples of causality, material and prevention by S^{SUM}.

	シードのエラー	類似語置換によるエラー
因果関係	13	37
材料関係	7	43
予防関係	17	33

係ならば，「因果関係 ⟨紫外線吸収剤，ソバカス⟩」「因果関係 ⟨薬用炭，ニキビ⟩」など，原因にその結果を防ぐものがきている場合など，シードインスタンスがそもそも誤っている場合である．原因 2 は，シードインスタンスは正しい場合でも，置換したことで誤った仮説が導かれてしまう場合である．たとえば，因果関係のシードインスタンスに「因果関係 ⟨カビ，悪臭⟩」があり，「悪臭」と「尿臭」が類似語として獲得されている場合，仮説「因果関係 ⟨カビ，尿臭⟩」が生成されるが，「カビ」は「尿臭」の原因にはならない．**表 1** に，この 2 つの原因の割合を示す．表 1 が示すように，大半は類似語置換によって起こるエラーであった．提案法は，目的とする意味的関係やそのシードインスタンスとは関係なく，同じ類似尺度で獲得した類似語を用いている．しかしながら，先の因果関係の例のように「悪臭」と「尿臭」など，「いやな臭い」という意味で高い類似性が認められるとしても，「カビ」は「尿臭」の原因にはならない．一方で，仮にシードインスタンス「因果関係 ⟨アンモニア，悪臭⟩」ならば「因果関係 ⟨アンモニア，尿臭⟩」と正しい仮説を生成できるため，必ずしも「悪臭」と「尿臭」の類似が類推にとって不適切なわけではない．

他の例としてはシード「材料関係 ⟨柿，干し柿⟩」と「柿」と「イチゴ」の類似性から仮説「材料関係 ⟨イチゴ，干し柿⟩」が生成されていたが，「干し柿」は明らかに「柿」が原料で，他の果物は材料になりえない．一方，これもシード「材料関係 ⟨柿，フレッシュジュース⟩」の場合は，仮説「材料関係 ⟨イチゴ，フレッシュジュース⟩」のように正しい仮説が生成される．このように，類似語として置き換えてよいか否かは，シードインスタンスに依存していることが分かる．他にも，たとえば因果関係では，「目やに」と「抜け毛」や「中絶」と「喫煙」など，一見するとあまり類似していないと考えられる類似語置換によるエラーも存在した．ただし，これらの類似語も，あるシードインスタンスでは，適切な仮説を生成する場合もあると考えられる．このように，より精度向上を目指すには，シードインスタンスや目的とする意味的関係の種類ごとに，適切な単語の類似尺度を使用することが重要と考えられる．

3.3　共起頻度フィルタリングの効果

本実験では，共起頻度フィルタリングの効果を評価する．具体的には以下を示す．

表 2　各ランクにおける T_{cooc} の値と精度
Table 2　Precisions for several T_{cooc} at each rank.

	因果関係			材料関係			予防関係		
	30,000	**60,000**	**90,000**	**30,000**	**60,000**	**90,000**	**30,000**	**60,000**	**90,000**
no cooc filter	54%	43%	40%	43%	36%	34%	46%	34%	27%
$T_{cooc}=1$	54%	47%	44%	**51%**	**50%**	**41%**	**48%**	**36%**	**29%**
$T_{cooc}=5$	**55%**	48%	**48%**	44%	43%	39%	44%	32%	27%
$T_{cooc}=10$	53%	**51%**	**48%**	49%	45%	39%	45%	30%	28%

- 共起頻度フィルタリングは下位のほうが効果が高い．
- 共起頻度の閾値（T_{cooc}）の値よりは，共起のある (1)，なし (0) のほうが有効である．

上位での精度を表す図 4 から図 6 では，共起頻度フィルタリングの効果は小さいことが分かる．一方で，図 7 と図 8 を見ると，下位では共起頻度フィルタリングの効果が顕著となった．たとえば，材料関係の $T_{cooc}=1$ では，共起頻度フィルタを用いない場合（no cooc filter）と比べて 60,000 位で約 14%の精度向上が見られた（図 8）．一方，図 9 の予防関係では効果が小さかった．

さらに，我々は T_{cooc} を 5，10 とした場合も試した．結果を**表 2** に示す．表 2 のとおり，値が大きいほど精度が高くなるといった傾向は見られなかった．この結果は T_{cooc} の値よりも，近接共起が観測されるか否かという情報が重要ということを示唆していると考えられる．

3.4　コーパス中に直接的に書かれていない関係を獲得できているか？

本実験では，共起頻度フィルタリングを用いない（$T_{cooc}=0$）場合を対象に，パターンベース法では，獲得困難な関係をどの程度獲得できているかを評価する．提案法は，高精度で獲得することとは別に，パターンベース法では獲得困難な，明示的に書かれていないような 2 語の関係をも獲得することを目指している．具体的には，以下を示す．

- 提案法によって高頻度パターンと共起がない，文内，4 文内で共起がない 2 語の関係を獲得できる．すなわち，非明示的にしか書かれていない可能性が高い（文内の共起がない），あるいは，そもそもコーパスに書かれていない可能性が高い（4 文内の共起がない）関係をも獲得できる．
- ランキング上位の仮説にも，文内で共起のない 2 語の関係が含まれる．
- シードインスタンスと同義でない，すなわち質的に新しいインスタンスを仮説として得られている．

我々は，従来のパターンベース法では獲得困難な条件として，以下のクラスに分けて調査した．

- **NP**：2.1 節で述べた我々の De Saeger らの方法[8] の実装で学習可能なパターンと共起がない（獲得困難）．
- **NS**：文内で共起がない（間接的に書かれている可能性がある）．
- **N4S**：4 文内で共起がない（関係が書かれていない可能性が高い）．

結果を**表 3**，**表 4** に示す．表 3 は，S^{SUM} の仮説全体の精度と仮説全体に含まれる正しい関係の推定数を示す．表 4 は，$S^{\mathrm{SUM}}_{\mathrm{FULL}}$ を対象とした同様の表である．

“NP” のパターンとは，2.1 節で述べた，我々の De Saeger らの方法[8] の実装で導出された，約 1.56×10^8 種類の構文パターンである．すなわち，“NP” の仮説は，De Saeger らの我々の実装では，いかにシードパターンを設定したとしても学習可能な言い換えパターンが存在しないため獲得不可能といえる．表 3，表 4 の “NP” を見ると，パターンベース法では困難な関係が多く獲得できていることが分かる．また，“NS”（文内で 2 語の共起がない）の仮説は，任意の 1 文内のパターンを用いるパターンベース法で獲得不可能なインスタンスである．たとえば，De Saeger らの方法[8] でも，2.1 節で述べたパターン導出のパラメータ，すなわち，1) パターンを構成する文節数が閾値以下，2) パターンによって獲得できる単語ペアの異なり数が閾値以上という 2 つの制約をどのように緩めても，“NS” の仮説を獲得できるパターンは導出できないため，原理的に獲得できない．表 3，表 4 を見ると提案法によって，“NS” の仮説を獲得できていることが分かり，さらに，獲得対象のコーパス中に 2 語の関係がそもそも書かれていない可能性が高い “N4S”（4 文内で 2 語の共起がない）の仮説を獲得できていることが分かる．

上述の結果は仮説全体の調査なので，S^{SUM} の上位ランク（10,000）における「NS（文内共起なし）」の精度と正しい関係の数を調査した．結果，3 種類の関係で，いずれも 10%以上の文内共起なしの仮説が含まれ，それらの仮説の 100 サンプルから評価した精度は，20%を超えていた．たとえば，因果関係は，上位 10,000 に約 500 個の文内共起のない仮説，すなわちパターンベース法では実質的に不可能な関係が獲得できていた．

表 3，表 4 を見ると「NS（文内共起なし）」「N4S（4 文内共起なし）」の仮説の精度は比較的低いが，十分に大きなコーパスでも “NS” や “N4S” となる仮説には，大規模なコーパスにも書かれていない，すなわち，まだ知られていない関係が含まれる可能性がある．このような関係をある程度の精度で獲得できていれば，専門家が有望そうな候補を選び検証することで，何も候補がない状態と比べれば，効率的に新しい発見が可能になるため，イノベーションの加速に貢献できると考えられる．

次に，シードインスタンスとの同義語による置換で得られた仮説は，厳密には新しい知識を表すインスタンスとはいい難いため，表 3，表 4 の各関係の正解の仮説について，同義語置換で生成された仮説の割合を調べた．具体的には，各仮説が同義語置換で生成されたか否かを 3 人が人手で判定し，2 人以上が同義語と判定した場合，同義語置換と見なした．同義語置換の判定の一致度（kappa 値）は 0.66 であった．kappa 値は 0.60 以上ならば “substantial agreement [15]” といわれているため，信頼できる結果といえる．結果を**表 5** に示す．表 5 に示すように，どの関係に関しても，同義語置換で生成された仮説，すなわちシードインスタンスと同義のインスタンスの割合は低いことが分かる．たとえば，S^{SUM} では，因果関係で 12%，材料関係で 16%，予防関係で 21%であり，$S^{\mathrm{SUM}}_{\mathrm{FULL}}$ では，因果関係で 7%，材料関係で 3%，予防関係で 7%であった．さらに，S^{SUM}，$S^{\mathrm{SUM}}_{\mathrm{FULL}}$ の因果，材料，予防

表 3　S^{SUM} の全仮説からの 300 サンプルによる精度とその精度から推定した正しい関係の数．評価基準は「lenient」
Table 3　Precision and estimated # of correct samples of all results of S^{SUM}(under lenient evaluation).

	因果関係	材料関係	予防関係
A：全サンプル	35% (106/300) 60,400	31% (92/300) 50,300	21% (62/300) 37,000
NP：パターンなし	29% (56/196) 31,700	23% (45/200) 24,000	18% (40/218) 23,700
NS：文内共起なし	18% (20/111) 11,400	15% (16/116) 8,740	12% (16/134) 9,540
N4S：4 文内共起なし	6% (3/51) 1,710	12% (7/57) 3,830	5% (3/62) 1,790

表 4　$S^{\mathrm{SUM}}_{\mathrm{FULL}}$ の全仮説からの 1,000 サンプルによる精度とその精度から推定した正しい関係の数．評価基準は「lenient」
Table 4　Precision and estimated # of correct samples of all results of $S^{\mathrm{SUM}}_{\mathrm{FULL}}$(under lenient evaluation).

	因果関係	材料関係	予防関係
A：全サンプル	22% (220/1,000) 191,000	19% (191/1,000) 151,000	12% (115/1,000) 108,000
NP：パターンなし	18% (152/836) 132,000	16% (126/783) 99,300	9% (77/850) 72,200
NS：文内共起なし	18% (77/579) 66,700	12% (62/469) 48,900	6% (34/598) 31,900
N4S：4 文内共起なし	10% (34/344) 29,400	7% (20/288) 15,800	3% (11/363) 10,300

表 5　S^{SUM} と $S^{\mathrm{SUM}}_{\mathrm{FULL}}$ の正解の仮説中のシードインスタンスと同義の関係の割合
Table 5　Ratio of correct hypotheses synonymous with seed instances.

	因果（S^{SUM}）	材料（S^{SUM}）	予防（S^{SUM}）	因果（$S^{\mathrm{SUM}}_{\mathrm{FULL}}$）	材料（$S^{\mathrm{SUM}}_{\mathrm{FULL}}$）	予防（$S^{\mathrm{SUM}}_{\mathrm{FULL}}$）
A：全サンプル	12% (13/106)	16% (15/92)	21% (31/62)	7% (16/220)	3% (6/191)	6% (7/115)
NP：パターンなし	7% (4/56)	9% (4/45)	18% (7/40)	6% (9/152)	2%(3/126)	6% (5/77)
NS：文内共起なし	10% (2/20)	6% (1/16)	6% (1/16)	3% (2/77)	2% (1/62)	9% (3/34)
N4S：4 文内共起なし	33% (1/3)	14% (1/7)	0% (1/3)	0% (0/34)	5% (1/20)	9% (1/11)

表 6　獲得された関係のサンプル．“*” は不正解を表す
Table 6　Acquired relation instances. “*” marks incorrect relations.

	生成された仮説	オリジナルのシードインスタンスのサンプル
因果関係	⟨アセトアルデヒド，心房細動⟩	⟨アセトアルデヒド，肝障害⟩，⟨アセトアルデヒド，肝機能障害⟩
	⟨眼精疲労，腰痛⟩	⟨眼精疲労，肩こり⟩，⟨眼精疲労，頭痛⟩
	⟨硫化水素，体臭⟩	⟨硫化水素，口臭⟩，⟨硫化水素，悪臭⟩
	*⟨乳酸，にきび⟩	⟨毒素，にきび⟩，⟨過酸化脂質，にきび⟩
材料関係	⟨重曹，石鹸⟩	⟨重曹，洗剤⟩，⟨ココナッツミルク，石鹸⟩
	⟨キウイ，アルコール⟩	⟨イチジク，アルコール⟩，⟨ブルーベリー，アルコール⟩
	⟨アロエ，ジュース⟩	⟨アロエ，ドリンク⟩，⟨ザクロ，ジュース⟩
	*⟨大麦，ワイン⟩	⟨大麦，ビール⟩，⟨大麦，アルコール⟩
予防関係	⟨パントテン酸，心臓病⟩	⟨ナイアシン，心臓病⟩，⟨セレニウム，心臓病⟩
	⟨ビタミン，尿管感染症⟩	⟨ビタミン，悪性貧血⟩，⟨ビタミン，高血圧⟩
	⟨β カロチン，高血圧⟩	⟨タウリン，高血圧⟩，⟨β カロチン，動脈硬化⟩
	*⟨鎮静作用，口臭⟩	⟨殺菌効果，口臭⟩，⟨抗菌効果，口臭⟩

関係に関して「NP：パターン共起なし」「NS：文内共起なし」「N4S：4 文内共起なし」の仮説も，同義語置換で生成された割合が少ないことが分かる．このことから，提案法によって，パターンベース法で獲得困難なインスタンスやコーパス中で明示的に書かれていない可能性が高いインスタンスについても，シードインスタンスとは異なる，質的に新しいインスタンスを獲得可能であることが確認できた．

表 6 に，提案法で得られた仮説の例を示しておく．

4. 関 連 研 究

本稿の関連研究には，大きく 2 つの種類がある．1 つは，関係獲得のための一般的なフレームワークの研究である[1),5),8),9),19),20),34)]．もう 1 つは，関係の分類やクラスタリングなど様々なタスクの基盤技術として，与えられた 2 つの意味的関係の類似度の測定を主目的とする研究である[7),12),18),31),32)]．

現在のところ，どちらの種類の研究も，主な手がかりとして構文パターンを用いている．たとえば，Paşca ら[19)] は，2 語の間に存在する語をパターン（infix パターン）を用いて，関係の候補を抽出している．また，Turney[31),32)] は，2 語を抽出するパターンの類似性から 2 つの関係の類似度を計算している．Ó Séaghdha ら[18)] は，対応する語どうしの語彙的な類似性（lexical similarity）と，前述の Turney のような抽出されたパターンの類似性（relational similarity）を組み合わせて，関係の類似度を計算する方法を提案している．一方，我々の提案法は，いっさいのパターンを用いていない点で大きく異なる．

我々のスコアリング法の一部は，Paşca ら[19)] と類似している．Paşca らは，獲得したインスタンスの候補のスコアリングの一部に，各シードインスタンスの語と候補の語の類似度が高いほどスコアが高くなるように計算している項がある．ただし，Paşca らの方法は，関係の候補をパターンで抽出しているため，パターンベース法であることに変わりはない．

加藤ら[36)] は，与えられたシードとの類似関係をブートストラップ的に獲得する方法を研究している．加藤らは，関係獲得の手がかりに，言語的なパターンではなく，2 語の「関係接続語」を用いている．関係接続語とは，関係を表すような 2 語との共起語であるため，単語ベクトルで表現されるパターンと解釈できる．構文的なパターンと比べると，制約が緩いため，従来の構文的パターンと共起しない関係をも獲得できる利点がある．制約が緩まるため，精度が低下する恐れがあるが，加藤らは，関係全体の第 1 項もしくは第 2 項の語の意味クラスが同じである（文献 36) では同位語と呼んでいる）ことを仮定し，関係をフィルタリングすることで，精度の低下を防いでいる．ただし，この仮定は，関係獲得タスク一般で考えるとやや強い制約である．たとえば，因果関係ならば，⟨ウィルス，病気⟩，⟨不況，デフレーション⟩，⟨徹夜，眠気⟩ など，材料関係でも ⟨ぶどう，ワイン⟩，⟨重曹，石鹸⟩ など，第 1 項，第 2 項の語の意味クラスは様々である．一方，提案法は，関係全体での意味的クラスの語の制約はなく「多くのシードから生成される仮説が良い」と考え，関係のスコアリングを行っている．また，「関係接続語」のような制約を緩めたパターンさえも用いていないため，加藤らの研究と比べて，さらに獲得可能な範囲が広いと考えられる．

生物医学のテキストマイニング分野では，文献からの知識発見の研究が行われている[13),23)–26)]．これらの研究は，複数の文献の情報を組み合わせることで有望な仮説を発見する試みである．明示的に書かれていない関係を発見するという点で，本稿の目的と似ている．ただし，これらは MEDLINE や MeSH のメタデータなど，専門家が高度に整備した情報の存在を前提としているため，そのような情報がない分野や情報源への適用は困難である．たとえば，Swanson ら[25)] は，単語 A と単語 B，単語 B と単語 C の MEDLINE のタイトル中での共起関係を用いて，単語 B を介して単語 A と単語 C の関係の発見を支援するモデルを提案している．また，Srinivasan[23)] は，Swanson らのモデルを自動化するため，MEDLINE の各文献に付与された MeSH のメタデータを用いて，文献間の関連の強さを計算することで，良い単語 B の候補を見つける方法を提案している．Hiristovski ら[13)] は，Swanson らのように，単語の共起ではなく，単語 A と単語 B，単語 B と単語 C の意味的な関係をとらえたうえで，仮説を生成する方法を提案している．具体的には「薬 A が病気 C を治療する」という仮説を，薬 A と物質 B の変化の関係，病気 C と物質 B の変化の

関係などを組み合わせて生成するルールを人手で作成し，それらのルールと，SepRep[21]，BioMedLee[17] と呼ばれる生物医学分野の自然言語処理エンジンによって MEDLINE の文献から獲得した関係を用いて，これまでの Swanson らの方法で発見された正しい知識が再発見できることを示している．これらの方法は，仮説生成に用いる情報が本研究と異なるため，互いの性能を高めるための相補的な技術になりうると考えられる．

最後に，発想支援を主目的とした類推による仮説生成の研究には，石川ら[37] によるものがある．シードとなる関係の語を置換して仮説を生成するという点では同様であるが，置換対象の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，本研究と異なる．

5. 結　　論

本稿では，類推に基づく意味的関係の獲得法について述べた．提案法は，シードインスタンスを入力に，シードインスタンスの語を類似語に置換して仮説を生成する．類似語は，大量の文書から文脈の類似性に基づき自動的に大量に獲得する．また，得られた仮説は「多くのシードインスタンスから高い類似性を持つ語での置換により生成された仮説が確からしい」と考えてスコアリングを行い，ランキングする．さらに，精度を向上させるため「何らかの関係のある 2 語は文書中の近い範囲で共起する」と考え，近接 N 文内の共起頻度を用いて仮説をフィルタリングする．

評価実験では，提案法によって，いくつかの意味的関係のタスクで，意味的関係を精度良く獲得できることと，従来のパターンベース法では獲得困難である文内で共起しない，4 文内でも共起しない 2 語の関係を獲得できることを示した．特に後者は，少なくとも処理対象となっているコーパス，つまり，ウェブ文書 1 億ページに間接的にしか書かれていない（1 文内の共起がない），もしくは書かれていない可能性が高い（4 文内の共起がない）知識までもが獲得できていることを示している．これは，処理対象となるコーパスが今後増大し，仮想的にウェブ全体と見なせるような状況になったとしても，そこに書かれていない知識を提案法によって獲得できる可能性があることを示唆している．今回は，ウェブ文書 1 億ページという十分に大きな規模のコーパスで実験したが，さらに大きな規模のコーパスで実験した場合，パターンベース法では獲得困難な関係や，そこに書かれていない関係を，同じような割合で獲得できるかは必ずしも自明でない．入力のコーパスの規模と提案法の知識発見としての能力の関係の調査は今後の課題である．

我々は，本稿が，直接的かつ明示的な知識を獲得する既存のパターンベース法の限界を突破する第一歩となり，「既知の知識の獲得」から「未知の知識の生成」へと発展させるための橋渡しとなるものと期待している．

参 考 文 献

1) Agichtein, E. and Luis, G.: Snowball: Extracting Relations from Large Plain-Text Collections, *Proc. 5th ICDL*, pp.85–94 (2001).
2) Agirrey, E., Alfonsecaz, E., Hallz, K., Kravalovazx, J., Paşca, M. and Soroa, A.: A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches, *Proc. 13th NAACL-HLT'09*, pp.19–27 (2009).
3) ALAGIN フォーラム：文脈類似語データベース Version 1, old.500k-2k.data. http://alaginrc.nict.go.jp/images/documents/SW_ALAGIN_V1_README.pdf
4) Budanitsky, A. and Hirst, G.: Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures, *Workshop on WordNet and Other Lexical Resources, the 2nd NAACL*, pp.29–34 (2001).
5) Cafarella, M.J., Downey, D., Soderland, S. and Etzioni, O.: KnowItNow: Fast, Scalable Information Extraction from the Web, *Proc. HLT-EMNLP'05*, pp.563–570 (2005).
6) Dagan, I., Lee, L. and Pereira, F.C.N.: Similarity-Based Models of Word Cooccurrence Probabilities, *Machine Learning*, Vol.34, pp.43–69 (1999).
7) Danshuka, B., Matsuo, Y. and Ishizuka, M.: Measuring the Similarity between Implicit Semantic Relations from the Web, *Proc. 18th WWW*, pp.651–660 (2009).
8) De Saeger, S., Torisawa, K., Kazama, J., Kuroda, K. and Murata, M.: Large Scale Relation Acquisition Using Class Dependent Patterns, *Proc. 9th ICDM*, pp.764–769 (2009).
9) Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D.S. and Yates, A.: Web-Scale Information Extraction in KnowItAll (Preliminary Results), *Proc. 13th WWW*, pp.100–110 (2004).
10) Hagiwara, M., Ogawa, Y. and Toyama, K.: PLSI Utilization for Automatic Thesaurus Construction, *Proc. 2nd IJCNLP*, pp.334–345 (2005).
11) Harris, Z.S.: Distributional structure, *Word*, Vol.10, pp.146–162 (1954).
12) Herdağdelen, A. and Baroni, M.: BagPack: A General Framework to Represent Semantic Relations, *Proc. Workshop on GEMS, the 12th EACL*, pp.33–40 (2009).
13) Hristovski, D., Friedman, C., Rindflesch, T.C. and Peterlin, B.: Literature-Based Knowledge Discovery using Natural Language Processing, *Literature-based Discovery, Information Science and Knowledge Management*, Vol.15, pp.133–152 (2008).
14) Kazama, J. and Torisawa, K.: Inducing Gazetteers for Named Entity Recognition by Large-scale Clustering of Dependency Relations, *Proc. 46th ACL*, pp.407–415

(2008).
15) Landis, J.R. and Koch, G.G.: The Measurement of Observer Agreement for Categorical Data, *Biometrics*, Vol.33, No.1, pp.159–174 (1977).
16) Lin, J.: Divergence Measures Based on the Shannon Entoropy, *IEEE Trans. Infromation Theory*, Vol.37, No.1, pp.145–151 (1991).
17) Lussier, Y., Borlawsky, T., Rappaport, D., Liu, Y. and Friedman, C.: PhenoGO: Assigning Phenotypic Context to Gene Ontology Annotations with Natural Language Processing, *Pacific Symposium on Biocomputing*, Vol.11, pp.64–75 (2006).
18) Ó Séaghdha, D. and Copestake, A.: Using Lexical and Relational Similarity to Classify Semantic Relations, *Proc. 12th EACL*, pp.612–629 (2009).
19) Paşca, M., Lin, D., Bigham, J., Lifchits, A. and Jain, A.: Names and Similarities on the Web: Fact Extraction in the Fast Lane, *Proc. COLING-ACL'06*, pp.809–816 (2006).
20) Pantel, P. and Pennacchiotti, M.: Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations, *Proc. COLING-ACL'06*, pp.113–120 (2006).
21) Rindflesch, T.C. and Fiszman, M.: The interaction of domain knowledge and linguistic structure in natural language processing: Interpreting hypernymic propositions in biomedical text, *Journal of Biomedical Informatics*, Vol.36, No.6, pp.462–477 (2003).
22) Shinzato, K., Kawaraha, D., Hashimoto, C. and Kurohashi, S.: A Large-Scale Web Data Collection as a Natural Language Processing Infrastructure, *Proc. 6th LREC*, pp.2236–2241 (2008).
23) Srinivasan, P.: Text mining: Generating hypotheses from MEDLINE, *Journal of the American Society for Information Science and Technology*, Vol.55, No.5, pp.396–413 (2004).
24) Swanson, D.R.: Undiscovered public knowledge, *Library Quarterly*, Vol.56, No.2, pp.103–118 (1986).
25) Swanson, D.R. and Smalheiser, N.R.: An interactive system for finding complementary literatures: A stimulus to scientific discovery, *Artificial Intelligence*, Vol.9, No.1, pp.183–203 (1997).
26) Tanja, B.: Conceptual biology, hypothesis discovery, and text mining: Swanson's legacy, *Biomedical Digital Libraries*, Vol.3, No.2, pp.133–152 (2006).
27) Thagard, P., Holyoak, K.J., Nelson, G. and Coghfeld, D.: Analog Retrieval by Constraint Satisfaction, *Artificial Intelligence*, Vol.46, pp.259–310 (1990).
28) Torisawa, K.: An Unsuperviesed Method for Canonicalization of Jpanese Postpositions, *Proc. 6th NLPRS*, pp.211–218 (2001).
29) Torisawa, K., De Saeger, S., Kakizawa, Y., Kazama, J., Murata, M., Noguchi, D. and Sumida, A.: TORISIKI-KAI, An Autogenerated Web Search Directory, *Proc. 2nd IUCS*, pp.179–186 (2008).
30) Torisawa, K., De Saeger, S., Kazama, J., Sumida, A., Noguchi, D., Kakizawa, Y., Murata, M., Kuroda, K. and Yamada, I.: Organizing the Web's Information Explosion to Discover Unknown Unknowns, *New Generation Computing*, Vol.28, No.3, pp.217–236 (2010).
31) Turney, P.D.: Measuring Semantic Similarity by Latent Relational Analysis, *Proc. 19th IJCAI*, pp.1136–1141 (2005).
32) Turney, P.D.: A Uniform Approach to Analogies, Synonyms, Antonyms, and Associations, *Proc. 22nd COLING*, pp.905–912 (2008).
33) 鳥澤健太郎，中川裕志，黒橋禎夫，乾健太郎，吉岡真治，藤井　敦，喜連川優：キーワードサーチを超える情報爆発サーチ—自然言語処理で価値ある未知をマイニング，情報処理，Vol.49, No.8, pp.890–896 (2008).
34) 阿部修也，乾健太郎，松本裕治：共起パターンの学習による事態間関係知識の獲得，自然言語処理，Vol.16, No.5, pp.79–100 (2009).
35) 風間淳一，De Saeger, S.，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).
36) 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).
37) 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).
38) 野田雄也，高橋哲郎，橋本　力，鳥澤健太郎：WWW から獲得した知識による検索語拡張とレシピ検索タスクにおける評価，言語処理学会第 16 回年次大会，pp.138–141 (2010).

(平成 22 年 3 月 15 日受付)
(平成 23 年 1 月 14 日採録)

土田　正明（正会員）

2005 年東京理科大学大学院理工学研究科経営工学専攻修士課程修了．同年 4 月より NEC に入社．2009 年 4 月に独立行政法人情報通信研究機構に出向し現在に至る．また，2009 年 4 月より東京理科大学大学院理工学研究科経営工学専攻博士後期課程に在籍中．テキストからの情報抽出，知識獲得に関する研究に従事．2008 年人工知能学会大会優秀賞を受賞．言語処理学会，人工知能学会，日本データベース学会各会員．

デ・サーガ　ステイン

2006 年北陸先端科学技術大学院大学知識科学研究科博士課程修了．博士（知識科学）．北陸先端科学技術大学院大学研究員を経て，2007 年に情報通信研究機構に入所．2008 年に NICT MASTAR プロジェクト言語基盤グループに専攻研究員として着任．自然言語処理を用いた知識獲得の研究に従事．

鳥澤健太郎（正会員）

東京大学大学院理学系研究科博士課程専攻中退後，同研究科助手．その後，科学技術振興事業団さきがけ研究 21 研究員（兼任），北陸先端科学技術大学院大学助教授，准教授を経て，2008 年 NICT に，MASTAR プロジェクト言語基盤グループ，グループリーダーとして着任．自然言語処理，特に Web 上の言語処理，Web からの知識獲得，獲得された知識の活用方法の研究に従事．けいはんな連携大学院教授を兼務．博士（理学）．

村田　真樹（正会員）

1970 年生．1993 年京都大学工学部電気工学第二学科卒業．1997 年同大学院工学研究科電子通信工学専攻博士課程修了．博士（工学）．同年京都大学にて日本学術振興会リサーチ・アソシエイト．1998 年郵政省通信総合研究所入所．独立行政法人情報通信研究機構主任研究員を経て，現在，鳥取大学大学院工学研究科情報エレクトロニクス専攻教授．自然言語処理，情報抽出の研究に従事．2005 年 FIT2005 論文賞受賞．共著書に『事例で学ぶテキストマイニング』（共立出版）等がある．言語処理学会，人工知能学会，電子情報通信学会，計量国語学会，ACL 等の会員．

風間　淳一（正会員）

独立行政法人情報通信研究機構知識創成コミュニケーション研究センター MASTAR プロジェクト言語基盤グループ主任研究員．2004 年東京大学大学院情報理工学系研究科コンピュータ科学専攻博士課程修了．博士（情報理工学）．北陸先端科学技術大学院大学情報科学研究科助教を経て，2008 年より情報通信研究機構．自然言語処理の研究に従事．

黒田　　航

元独立行政法人情報通信研究機構知識創成コミュニケーション研究センター MASTAR プロジェクト言語基盤グループ短時間研究員．現在，京都工芸繊維大学（非常勤），早稲田大学総合研究機構（客員研究員）．京都大学から人間・環境学博士を取得．言語学と自然言語処理を融合する研究に従事．

大和田勇人（正会員）

東京理科大学理工学部経営工学科教授．1988 年東京理科大学大学院理工学研究科経営工学専攻博士後期課程修了．同年東京理科大学理工学部経営工学科助手．その後，講師，助教授を経て，2005 年同教授，現在に至る．工学博士．主に，データマイニング，機械学習，帰納論理プログラミングに興味を持つ．人工知能学会正会員，日本ソフトウェア科学会正理事．