情報処理学会論文誌 Vol.55 No.4 1428–1437 (Apr. 2014)

三段論法的パターンに着目した解釈容易な仮説の生成規則獲得と順位付け

関 和広[1,a)] 上原 邦昭[1]

受付日 2013年7月1日, 採録日 2014年1月8日

概要：生物医学文献は日々増大しており，個人がすべての情報を把握・利用することは現実的には難しい．専門家の知識でさえも生物医学の特定分野に限定されていることから，疾病や遺伝子といった概念間の重要な関係が，複数の文献から暗示されながらも，大量の情報に埋もれている可能性がある．このような潜在的な関係を発見する研究は，仮説発見と呼ばれる．本稿では，学術文献から抽出した関係の閉じた連鎖に注目し，仮説発見を行う新しい枠組みについて述べる．具体的には，関係の連鎖を構成する述語を同定し，これらの組合せを仮説生成のための潜在的な規則と見なす．そして，これらの規則によって自動生成された仮説について，学術文献から抽出した知識と照合することでその妥当性を検証する．妥当性が確認された仮説は，仮説の信頼性を推定するための回帰モデルの正例として用いる．そして，得られた回帰モデルの出力に基づき，新たに生成した仮説の順位付けを行う．提案する仮説発見の枠組みの有効性を評価するため，現実の生物医学文献を用いた実験を行ったところ，自動生成された仮説に妥当な仮説が含まれていること，また，回帰モデルの利用によって妥当な仮説がより上位に順位付けられることが確認できた．

キーワード：仮説生成規則，半教師あり学習，意味的類似度，ランキング

Generating and Ranking Interpretable Hypotheses Based on Syllogistic Patterns

Kazuhiro Seki[1,a)] Kuniaki Uehara[1]

Received: July 1, 2013, Accepted: January 8, 2014

Abstract: The ever-growing literature in biomedicine makes it virtually impossible for individuals to grasp all the information relevant to their interests. Since even experts' knowledge is limited, important associations among key biomedical concepts may remain unnoticed in the flood of information. Discovering those hidden associations is called *hypothesis discovery*. This paper reports our approach to this problem taking advantage of a triangular chain of relations extracted from published knowledge. We consider such chains of relations as implicit rules to generate potential hypotheses. The generated hypotheses are then compared with newer knowledge for assessing their validity and, if validated, they are served as positive examples for learning a regression model to rank hypotheses. This framework, called *supervised hypothesis discovery*, is tested on real-world knowledge from the biomedical literature to demonstrate its effectiveness.

Keywords: hypothesis generating rules, semi-supervised learning, semantic similarity, ranking

1. まえがき

生物医学分野の文献情報は増大し続けており，そのすべての研究動向を 1 人の研究者が把握することは事実上不可能である．たとえば，同分野最大の書誌情報データベースである Medline[*1]は，現在 1 千 9 百万以上の文献の書誌情報からなり，その数は毎日 2,000〜4,000 のペースで増加し続けている．このような大量のテキスト集合から必要な情報を獲得し効果的に利用するため，情報検索や情報抽

[1] 神戸大学
Kobe University, Nada, Kobe 657–8501, Japan
[a)] seki@cs.kobe-u.ac.jp

[*1] http://www.ncbi.nlm.nih.gov/entrez

出 [4]，テキストデータマイニング（TDM）[14] などの知的な情報処理技術の重要性が増している．

情報検索や情報抽出が文書中に明示された既存の情報を対象にするのに対して，TDM は既存のテキスト情報を自動的に解析することで，これまでに認識されていない未知の知識の発見を目指す [11]．TDM，あるいは「仮説生成」の先駆的な研究は，1980 年代，Swanson によって行われた．Swanson は，論理的な関係を持ちながらいまだその関係が認識されていない 2 つの前提（premise）が大量の文献中に独立に存在すると考え，これらを結びつけることで新しい知識を発見あるいは生成することが可能だと主張した．これを裏付けるため，Swanson は数十の論文を人手で分析し，魚油がレイノー病の治療に効果的であるという仮説を導いた [22]．なお，この仮説の正当性は後に臨床的にも示されている [6]．

本研究は Swanson の一連の研究 [22], [23], [24] に着想を得たものであり，仮説生成の新たな枠組みを提案し，既存の研究における 2 つの問題の解決を図る．1 つの問題は，生成された仮説の解釈が困難なことである．既存研究の多くは，単に 2 つの語や概念が潜在的に何らかの関係にあるということを示すだけであり，その意味や解釈はユーザに委ねられていた．この仮説の不明瞭さは，仮説生成の有用性を著しく制限するものであり，実利用の障害となっている．この問題に対処するため，本稿では，文献から抽出した大量の既知の関係を利用して，仮説生成規則を獲得する．獲得された規則では，概念（名詞句）N_1 と N_2 間の関係を述語 V として明示的に示すことで，「N_1 V N_2」という容易に解釈可能な仮説を生成する．言い換えると，本研究における「仮説」は，(1) 名詞句で表現される 2 つの概念 N_1，N_2 と (2) 述語で表現される概念間の関係 V から構成され，かつ (3) その概念間の関係がこれまで報告されていないものと定義する．

既存研究の 2 つの目の問題は，生成される仮説の量である．通常，生成された仮説の大部分は誤りであり，ごく一部の仮説のみがさらなる検討・調査に値する．しかしながら，前者の誤った仮説の数は後者よりもはるかに多いため，有益な仮説を見つけることは難しい．そのため，既存研究では，注目する概念の意味的なクラスを制限するなどの方法で，生成される仮説の量を抑制したり，仮説の尤もらしさを定量化したりするなどして，信頼性の高い仮説を発見するなどの試みが行われている．本稿では，獲得した仮説生成規則によって自動的に生成した仮説から真の仮説を同定し，これを正の教師事例とすることで，半教師ありの回帰モデルによって信頼性の高い仮説の特徴を学習する．また，効果的なモデルの学習には素性の選択が重要であるため，仮説生成規則の信頼性や概念の意味的類似度，概念の限定性などの素性の有効性を検討する．

提案する仮説発見の枠組みの有効性を評価するため，生物医学分野の書誌情報データベース Medline を用いた実験を行う．その結果，提案手法によって自動的に獲得された仮説生成規則によって，妥当な仮説（真の仮説）が生成されること，また，生成された仮説に関する様々な特徴を素性として回帰モデルを学習することで，従来研究と比較して，より高精度に仮説の順位付けが行えることを示す．

2. 関連研究

2.1 仮説生成

Swanson [22] は，潜在的には存在しながら見過ごされてきた新しい知識を発見するための情報源として，学術文献の持つ可能性を主張してきた．Swanson の知識発見の方法は三段論法に基づき，たとえば「A であれば B である」，「B であれば C である」という 2 つの前提から，「A であれば C である」という潜在的な関係を導く．前述の魚油とレイノー病のように，このような関係は，実験や検証によって正否を判断すべき「仮説」と考えることができる．レイノー病の例の場合，Swanson は Medline の検索によって得られた魚油に関する論文とレイノー病に関する論文を手作業で分析することで，「レイノー病は高血小板凝固性，高血液粘性，血管収縮によって特徴付けられる」および「魚油は血中脂質，血小板凝固性，血液粘性，血管反応性を減少させる」という関係（前提）を同定し，これらの前提からレイノー病の治療における魚油の潜在性を導いた．この研究が基礎となり，以降，Swanson や他の研究者が，仮説生成を行う，あるいは補助するアルゴリズムを開発している．以下，これらのうち代表的な例についてまとめる．

Weeber ら [27] は自然言語処理ツールを利用し，DAD-system というシステムを構築した．このシステムの特長は，UMLS メタシソーラスを知識表現と推論の枝刈りに用いることである．文章中に現れる語を抽出して知識表現に用いるという従来の研究に対し，DAD-system は MetaMap プログラム [2] を用いることで入力文を当該メタシソーラス中で定義される概念集合に変換する．この方法の利点は，異なる表記の同一概念を自動的に同一概念に縮退できることにある．こうして得られた概念とそれらの文章中での共起関係から，概念間の関連を連鎖的に推定し，概念間の未知の関係，すなわち仮説を得る．なお，メタシソーラスで定義される各概念に付与されている意味タイプと呼ばれるクラスを用いることで，ユーザの興味と無関係な概念に関する（潜在的）関係を除外することができる．これによって，生成される仮説の数を大幅に減らすことが可能になる．また，Pratt ら [18] のシステム LitLinker は，Weeber らのシステムと同様に UMLS メタシソーラスを用いているものの，相関ルールマイニング [1] を用いて 2 つの関連する概念を同定している．

Srinivasan [21] は Manjal と呼ばれるシステムを構築した．先行研究との主要な違いは，Manjal は知識資源として

標題や要約などのテキストを利用せず，MeSH 索引語[*2]のみを利用することである．入力となる概念（たとえば病名）を受け取り，Manjal は Medline を検索し，検索されたレコードに付与されている MeSH 索引語を抽出する．続いて，あらかじめ定められたマッピング表に基づいて，それらの MeSH 索引語を UMLS メタシソーラスの意味タイプに関連付ける．これによって，DAD-system と同様に，特定の意味タイプのみに限定した仮説の発見が可能となる．また，インタフェース向上のため，ユーザに結果を提示する際にも MeSH 索引語は意味タイプごとに分類されて表示される．

より新しい研究としては，Liu ら [16] が，ハイパーグラフを用いた仮説生成の手法を提案している．この手法では，概念をノード，概念間の共起関係をエッジとし，概念間の直接的あるいは間接的な関係を酔歩モデルに基づく通勤時間（commute time），ムーア–ペンローズの擬似逆行列に基づく内積で定義する．Liu らは，少量の人口データや商品購入データ，医療データなどを用いて評価実験を行い，Swanson のレイノー病に関する仮説生成も再現できたと報告している．

このように継続的に研究が行われてきているものの，仮説生成の研究はいまだ発展途上にあり，改善の余地が大きい．先行研究の大部分は，単に 2 つの概念が潜在的に関係していることを示すのみであり，その関係が何であるのかを明示しない．また，評価実験は Swanson が提示した少数の仮説によっていることが多い．一方，本研究では，明示的な意味を示す仮説を生成・順位付けするための新しい枠組みを提案し，Medline から自動的に抽出・同定した多数の真の仮説を用いた定量的な評価実験について報告する．

2.2 含意認識

自然言語処理分野における関連研究として，含意認識（Recongnizing Textual Entailment; RTE）があげられる．RTE では，一対の節が与えられたときに，一方が他方の記述を含意するか否かを判定する．たとえば，「SCO won a lawsuit against IBM」という関係は，「SCO sued IBM」という関係を含意する [25]．このような含意関係を認識するためには，たとえば「x win lawsuit against y」→「x sue y」という含意認識規則が必要であり，このような規則を文書集合（コーパス）から自動的に獲得するための手法が提案されている [3], [25]．

類似の研究として，因果関係にある節のペアの同定 [7], [10]，および矛盾した関係にある節ペアの同定がある [10]．前者の例は，「increase in crime」→「heighten anxiety」のような関係であり，後者の例は，「increase in crime」と「decrease in crime」のような関係である．このようなペアを収集して組み合わせることで，前者の例について，論理学でいうところの裏（reverse）のような因果関係「decrease in crime」→「deminish anxiety」を仮説として生成できる [10]．また，2 つの節（たとえば「*Food* is made from *Ingredient*」，「*Ingredient* contains *Chemical*」）によって示される含意関係（「*Food* contains *Chemical*」）を推論する規則を獲得する研究も行われている [19]．このように，複数節の組合せによって新しい関係を推論するという研究は，本研究の考え方に近い．ただし，これらの研究が，既知の（ただし暗黙的な）関係の推論や言語表現の言い換えを目的としているのに対し，本研究は，真偽が定かではない科学的仮説の生成を目的としている．この目的の違いから，たとえば Schoenmackers ら [19] の手法では，概念を *Food* や *Ingredient* のような意味クラスに抽象化したうえで，意味クラスに共通に成り立つ信頼性の高い規則の獲得を行っている．一方，本研究では，特定の概念（固有表現）において成り立つ科学的仮説を生成することが重要であるため，意味クラスのような概念の抽象化は行わず，Swanson の三段論法的パターンに着目することで仮説生成規則の獲得・仮説の生成を網羅的に行う．そのうえで，仮説生成規則と仮説の特徴を基に生成された仮説の妥当性を推定し，順位付けを行うという枠組みを提案する．

3. 仮説生成と順位付けの枠組み

3.1 概要

提案する枠組みは，仮説生成と仮説の順位付けに大別できる．前者は，大量のテキストデータから述語項関係として抽出した既存知識に基づいて仮説生成規則を生成し，これらの規則を既存知識に適用することで潜在的な仮説を生成する．後者の順位付けは，真の仮説を正例として用いることで回帰モデルを学習し，これを生成された仮説に適用することで，「真の仮説らしさ」を推定し，その尤もらしさによって仮説を順位付けする．この枠組みの概要を**図 1** に示す．以降の節で，図中の個々の要素について詳述する．

3.2 仮説生成規則の獲得

既存研究の多くでは，2 つの語あるいは概念の共起関係を基に，仮説を生成する．この方法でも妥当な仮説を生成できるものの，より多くの，意味のない誤った仮説を生成してしまう．その結果，ユーザにとっては，真に重要な仮説を見つけることが困難になる．本研究では，単なる共起関係ではなく，2 つの概念間の関係の意味を考慮し，既存知識を基に，より合理的な仮説のみを生成する．より具体的には，テキストから抽出した個々の既存知識（関係）は，述語項関係「N_1 V N_2」として表現される．ここで，V が述語，N_1 と N_2 が主格と目的格の項である．そして，仮説生成規則を獲得するため，以降で述べるように，同一の項を基にこれらの関係の統合を行っていく．

[*2] 生命科学分野の文献を索引付けするために利用される統制語彙．

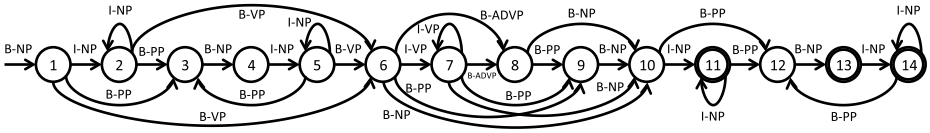


図 2 述語項関係の抽出に用いたオートマトン

Fig. 2 Automaton to extract predicate-argument structures.

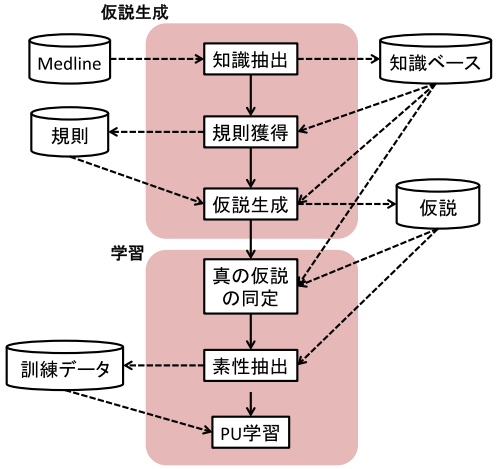


図 1 提案する仮説生成の枠組み．実線は処理の流れ，破線はデータの流れを示す

Fig. 1 Overview of the supervised hypothesis discovery framework. Solid and dotted lines show the flow of the processes and the flow of the data, respectively.

3.2.1 知識の抽出

テキストから既存知識を抽出するため，句抽出器（チャンカ）と固有表現抽出器を用いる．本研究では，両方の機能を持つ Genia タガー [26] を採用する．そして，「NP VP NP」で表される述語項関係を受理する単純なオートマトンでチャンカの出力を解析し，得られた述語項関係を既存知識とする．述語項関係（NP VP NP）の抽出に用いたオートマトンの状態遷移図を**図 2** に示す．オートマトンへの入力はチャンクタグであり，「`B-NP`」（NP の先頭の語）や「`I-NP`」（NP の 2 語目以降）といった記号で表現される．太線の状態（11, 13, 14）は最終状態を示している．

たとえば，入力文が「the guideline is being reexamined currently by the NCRP committee」であった場合，まず，三つ組みの述語項関係 ⟨the guideline, is being reexamined currently by, the NCRP committee⟩ が抽出される．続いて，以下の後処理を施し，表記の正規化を行う．

(1) すべての語を原形に変換．
(2) すべての冠詞を除去．
(3) 否定的な意味を持つ副詞（例：barely）を「not」で置換．
(4) すべての副詞を除去（「not」を除く）．
(5) 不明確な助動詞「may」「might」を持つすべての関係（そのもの）を削除．
(6) すべての助動詞を除去．

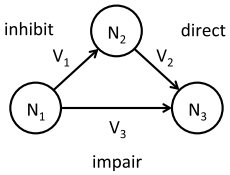


図 3 同一の項で結合した仮説生成規則の基となる関係の連鎖

Fig. 3 A chain of relations leading to a hypothesis generation rule.

(7) 現在形，過去形，未来形の時制を除去．
(8) 受動態を能動態に変換．

この場合，抽出された関係は最終的に ⟨NCRP committee, reexamine, guideline⟩ に正規化される．

一方，固有表現抽出器は，タンパク質や RNA などの生物医学分野の固有表現を同定する．本研究では，生物医学的に意味を持つ仮説だけに注目するために，固有表現抽出器によって，少なくとも 1 つの名詞句が固有表現だと同定された関係だけを保持し，他の関係は削除する．本研究で用いる Genia タガーは，固有表現の意味クラスとして Protein，DNA，RNA，Cell line，Cell type を同定する．以降では，抽出された既存知識の集合を知識ベースと呼び，**K** で示す．

3.2.2 仮説生成規則の獲得と仮説生成

前項で構築した知識ベース **K** から，仮説生成規則 r を 3 つの述語の組 V_1，V_2，V_3 として得る．この基本的なアイデアは，述語項関係を同一の項で結合し，Swanson の三段論法における 2 つの前提と 1 つの結論（2 章参照）に相当する 3 つの関係からなる三段論法的パターンを同定することである．たとえば，2 つの関係「N_1 inhibits N_2」と「N_2 directs N_3」が知識として抽出されていたとする．前者の目的格と後者の主格の項は同一の N_2 であるため，この項によって 2 つの関係を結合することで，「N_1-inhibit-N_2-direct-N_3」という関係が得られる．仮説生成の既存の研究においても，この時点で N_1 と N_3 がなんらかの潜在的な関係を持っていることは分かる．これに対し，本研究ではもう一歩進んで，知識ベース **K** から，「N_1 impairs N_3」のように N_1 と N_3 を項として含む別の述語項関係を見つける．この場合，主格と目的格の項が前述の 2 つの関係とそれぞれ同一である．そこで，これらの項によって関係を統合し，**図 3** に示すような三角形型の関係の連鎖を得る．

以上の関係と図 3 により，次のように一般化した規則 r

が成り立つ可能性が示唆される．

仮説生成規則 r：If "x inhibits y" and "y directs z", then "x impairs z" [*3].

ここで，x，y，z は名詞句を示す．なお，この規則は 1 つの可能性を示すだけであり，必ずしも正当な規則とはいえない点に注意を要する．しかしながら，このような規則によって得られる潜在的な関係は，規則の基となった三段論法的パターンを踏襲しているため，従来の共起関係のみに基づく方法よりも合理的であると考えられ，結果的に誤った仮説を生成することが少なくなるものと期待できる．なお，図 3 と名詞句だけが異なる「N_a inhibit N_b」，「N_b direct N_c」，「N_a impair N_c」という 3 つの関係が知識ベースに存在した場合，こちらのパターンからも図 3 と同一の規則（inhibit, direct, impair）が得られる．本稿では，このように元々は異なる関係の組合せから同一の規則が得られることを，「複数のパターンが同一の規則に帰着した」という．

これらの仮説生成規則（以降では $\mathbf{R} = \{r_1, r_2, \cdots\}$ で表す）は，まず目的格と主格が同一（上の例では N_2）の 2 つの述語項関係を同定し，続いて他の 2 つの項（N_1 と N_3）を主格と目的格として含む関係を見つけることで，容易に獲得できる．なお，規則 r は，基本的に 3 つの述語の組 V_1，V_2，V_3 で構成されるものの，3.3.2 項で述べる理由により，N_1，N_2，N_3 に関する情報も保持しておく．知識ベース $\mathbf{K}$ から，このような規則の集合 $\mathbf{R}$ が網羅的に同定されたのち，$\mathbf{R}$ を $\mathbf{K}$ に適用し，仮説 $\mathcal{H} = \{h_1, h_2, \cdots\}$ を生成する．ここで，h は生成された個々の仮説を示す．この方法によって生成された仮説は，単に 2 つの概念間の潜在的な関係ではなく，述語（上の例の場合は「impair」）によって関係の意味を明示的に示すことができる．なお，この仮説生成手続きによって，知識ベース $\mathbf{K}$ に元々存在する知識も生成されてしまう．このような既存の知識は，知識ベース $\mathbf{K}$ と照合することで同定・破棄し，知識ベースに存在しない新しい知識だけを仮説として出力する．

3.3 真の仮説を用いた学習

前項で述べた仮説生成規則 $\mathbf{R}$ によって生成される仮説の数 $|\mathcal{H}|$ は，共起関係のみを利用した従来方法よりも少なくなるものの，生成されるすべての仮説を吟味・検証することは難しい．実応用を考えた場合，限られた人的・研究資源で仮説の検証を行うため，生成される仮説の尤もらしさを考慮し，仮説を順位付けすることが重要である．そこで，本研究では，真の仮説の特徴を学習することで，生成された個々の仮説 h の尤もらしさ，あるいは確信度を推定する．このステップは，以下で述べる真の仮説の同定，素性の設計，PU 学習からなる．

3.3.1 真の仮説の同定

真の仮説の特徴を学習するためには，ラベル付きの学習事例，つまり真の仮説（正例）と偽の仮説（負例）が必要になる．通常，このような学習事例は人手で作成されることも多い．仮説生成の場合，生成された仮説の真偽を人が判断するためには，専門知識や関連文献の調査，あるいは実験までも必要とするため，現実的ではない．そこで，本研究では，仮説生成規則の獲得と仮説生成に用いる知識ベースよりも新しい文献を活用する．すなわち，古い文献から生成された仮説が新しい文献で報告されていれば，その仮説は正当性が認められた真の仮説であると見なす．

より正確には，まず知識ベース $\mathbf{K}$ を 3 つの部分集合 $\mathbf{K}_1$，$\mathbf{K}_2$，$\mathbf{K}_3$ に分割する．ここで，$\mathbf{K}_i$ は i 番目に古い知識であるとする．そして，$\mathbf{K}_1$ を仮説生成規則 $\mathbf{R}_1$ の獲得に用い，続いて仮説 $\mathcal{H}_1$ の生成にも用いる．より新しい知識 $\mathbf{K}_2$ を基に，生成された $\mathcal{H}_1$ 中の真の仮説を同定する．なお，最も新しい知識 $\mathbf{K}_3$ は，のちの評価実験で用いる．

この方法の問題点の 1 つは，$\mathbf{K}_2$ に現れない仮説を単純に偽の仮説（負例）とは見なせないことである．つまり，そのような仮説は実は真の仮説であり，まだ文献には現れていないという可能性も否定できない．言い換えれば，生成された仮説が真に偽の仮説であると断定することは難しい．この問題に対処するため，本研究では，このような真偽不明の仮説を真偽の仮説が混在したノイズと考え，正例とラベルなし事例からの学習（PU 学習）手法を用いる．具体的には，既存手法である Lee ら [15] の手法を用いる．この手法では，PU 学習をノイズを含んだ学習問題ととらえ，すべてのラベルなし事例を負例として線形回帰モデルを学習する．

3.3.2 素性の設計

学習手法を適用する際には，それぞれの事例（仮説）を正例・負例の特徴を示すような素性の組で表現する必要がある．規則 r から生成された仮説 h の尤もらしさを推定するために利用可能な特徴としては，大別して 2 種類の特徴がある．1 つは，規則 r に関連したものであり，もう 1 つは規則 r と仮説 h に関連したものである．以降では，前者を「規則依存」の素性，後者を「規則・仮説依存」の素性と呼ぶ．なお，仮説生成規則は述語の組（V_1, V_2, V_3）で表されているものの，基となった関係のパターンに現れる名詞句（N_1, N_2, N_3）の情報も保持している．

規則依存の素性として，本研究では以下を用いる．

- 同一の規則（V_1, V_2, V_3）に帰着した三段論法的パターンの数．もし複数のパターンが同じ規則を示していれば，その規則はより信頼性が高いと考えられる．なお，2 つの規則（V_a, V_b, V_c）と（V_d, V_e, V_f）が同一であるとは，$\mathrm{V}_a = \mathrm{V}_d$ かつ $\mathrm{V}_b = \mathrm{V}_e$ かつ $\mathrm{V}_c = \mathrm{V}_f$

[*3] 実際に「actinomycin D *inhibits* mRNA」，「mRNA *directs* protein synthesis」，「actinomycin D *impairs* protein synthesis」という関係が Medline から抽出され，何のドメイン知識も与えることなく，この規則が獲得された．

の場合をいう．

- 述語の詳細度．より詳細な意味を持つ述語は，より詳細で有意義な仮説を生成する可能性がある．この直感に従い，V_1，V_2，V_3 のそれぞれについて，次の 2 つの素性を抽出する（図 3 参照）．
  - Medline における文書頻度 DF．情報検索において，DF の逆数はしばしば，語の詳細度を示す指標として用いられる [20]．
  - 類義語の数．より意味の広い語はより多くの類義語を持つという仮定に基づく．英語のシソーラス WordNet [9] を利用して得る．
- 名詞の詳細度．前述と同様の仮定による．規則 r に関連する個々の名詞句 N_1，N_2，N_3 について，以下の 2 つの素性を利用する．もし，規則 r が複数のパターンから得られた場合は，それぞれのパターンに関して異なる名詞句 N_1，N_2，N_3 が存在するため，便宜上，各パターンを別の規則として扱う．言い換えると，この場合，名詞の詳細度に関して異なる素性値を持つ同一の仮説が複数（パターンの数だけ）出力される．
  - Medline における文書頻度 DF．
  - 類義語の数．

規則・仮説依存の素性としては，以下を用いる．

- 同じ仮説 h を生成した規則の数．複数の規則から同一の仮説が生成された場合，その仮説はより信頼性が高いと考えられる．
- 仮説 h を生成した 2 つの関係「N_4 V_1 N_5」と「N_5 V_2 N_6」に含まれる名詞句の詳細度（**図 4** の右側参照）．それぞれの名詞句（N_4, N_5, N_6）について，Medline における DF 値を用いる．
- 仮説 h 生成時の規則 r の妥当性．規則 r を適用して仮説 h を生成する際，r が妥当であれば生成された h の信頼性も高いと考えられる．そこで，規則 r を適用する際のその妥当性を，規則 r と仮説 h に関する三段論法的パターン間の「アナロジ的類似性」として定義する．詳細は次節で述べる．

3.3.3 アナロジ的類似性

図 4 にアナロジ的類似性の考えを示す．左側の三角関係は規則獲得の基となった三段論法的パターンであり，右側の三角関係は，仮説「N_4 V_3 N_6」とその生成の基となった 2 つの関係「N_4 V_1 N_5」と「N_5 V_2 N_6」を示す．点線は，V_1，V_2，V_3 からなる三段論法的規則において，2 つのパターン間で同一の役割を果たす概念のペアを接続している．これら概念のペアがそれぞれ意味的により類似していれば，同じ関係が成り立つ可能性が高く，N_4，N_5，N_6 からなる右側の三角関係にも規則の適用性が高いと仮定する．

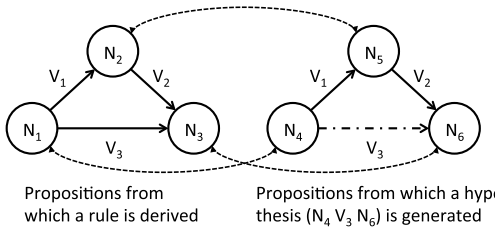


図 4 2 つの三段論法的パターン間のアナロジ的類似性．点線は，同一の役割を持つ概念ペアを接続する

Fig. 4 Analogical resemblance b/w two chains of relations. A dashed line connects two concepts playing the same role.

概念間の意味的類似性を測る方法は，コーパスに基づく方法や辞書に基づく方法など，数多く提案されている [17]．本研究では，その適用範囲の広さから，コーパスに基づく方法として正規化グーグル距離（NGD）[5] を用いる．NGD は正規化情報距離の近似であり，その定式化におけるコルモゴルフ複雑性をグーグル検索のヒット回数で置き換える．概念 x と y の間の NGD は，次のように定義される．

$$\mathrm{NGD}(x,y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x,y)}{\log N - \min\{\log f(x), \log f(y)\}} \quad (1)$$

ここで $f(x)$ は x をクエリとしたときのグーグル検索におけるヒット数であり，N はグーグルで索引付けされた総文書数である．$\mathrm{NGD}(x,y)$ は 0 から ∞ の値をとり，$\mathrm{NGD}(x,y)$=0 は，x と y が同一であることを意味する．

ただし，本研究で扱う対象は生物医学文献のため，本研究ではグーグルの代わりに Medline を用いる．すべての知識（概念）は Medline から抽出されるため，Medline を NGD の算出に用いることで，すべての x について $f(x)$ が存在する（非ゼロである）ことも保証される．定式化は同一であるものの，以降では，Medline を用いて算出した距離を正規化 Medline 距離（NMD）と呼ぶ．仮説 h 生成に関する規則 r の適用性推定のため，図 4 の 3 組の名詞句について意味的距離 NMD(N_1,N_4)，NMD(N_2,N_5)，NMD(N_3,N_6) を算出する．

4. 評価実験

4.1 実験の手順と設定

提案する仮説生成の枠組みの妥当性を検証するため，以下のように評価実験を行った．既存知識としては，TREC ゲノムトラック [12] で提供された Medline の部分集合を用いた．このデータセットは 1994 年から 2003 年までの 10 年間分の書誌情報であり，4,591,008 件の文献情報からなる．このデータセットの論文タイトルと抄録から，Genia タガーを用いて，既存知識を述語項関係として抽出した．3.2 節で述べた正規化処理後，17,904,002 の関係が得られ，これを知識ベース $\mathbf{K}$ とした．続いて，$\mathbf{K}$ をおおよそ同サイズの 3 つの部分集合 $\mathbf{K}_1$，$\mathbf{K}_2$，$\mathbf{K}_3$ に分割した．そして，最も古い知識 $\mathbf{K}_1$ を仮説生成規則獲得と仮説生成に用いた．獲得された規則の数は 12,180 であり，生成された仮説の数は，重複を含め 346,424 であった．

生成された仮説は，まず知識 $\mathbf{K}_2$，続いて $\mathbf{K}_3$ と照合し

た．もし，ある仮説が $\mathbf{K}_2$ に含まれた場合，その仮説は真の仮説として訓練データに加えた．もし $\mathbf{K}_2$ には含まれず，$\mathbf{K}_3$ には含まれた場合は，真の仮説としてテストデータに加えた．$\mathbf{K}_2$ と $\mathbf{K}_3$ のいずれにも含まれない仮説は，ラベルなしの事例とし，訓練データとテストデータのいずれかにランダムに加えた．

このように構築された訓練データを学習に用い，テストデータは仮説生成と順位付けの性能評価に用いた．訓練データは 226 の正事例と 169,060 のラベルなし事例，テストデータは 88 の正事例と 169,059 のラベルなし事例から構成されていた．テストデータ中のラベルなし事例は，将来的には正事例であることが示される可能性があり，必ずしも負事例とはいえないものの，評価のうえでは負事例と見なした．この問題については，次節で評価指標について述べる際に再度議論する．

学習には，Lee ら [15] の PU 学習アルゴリズムを用いた．PU 学習では，正例のみのラベル付き事例とラベルなし事例を基に分類モデルを学習する．Lee らの手法は，PU 学習をノイズを含んだ学習問題ととらえ，真のエラー率の近似関数を最小化するようにロジスティック回帰モデルを学習する．具体的には，次式により各素性の重み w を最急降下法によって反復的に推定する．

$$w_{j,t} = (1-c)w_{j,t-1} + \eta(\Delta_{j,t} + \gamma\Delta_{j,t-1}) \quad (2)$$

ここで，j は素性のインデックス，t はエポック，c は減衰率，η は学習率，Δ は損失関数の勾配，γ は慣性項である．なお，この手法では，正例と負例（実際にはラベルなし事例）の数を考慮した目的関数を用いているため，正負の偏りに対して頑健な予測が可能である．Lee らの報告をもとに，本節の実験では，エポック数 t，減衰率 c，学習率 η，慣性項 γ はそれぞれ 500，0.05，0.00001，0.99 とした．本稿の報告ではパラメータの調整を行っていないため，検証データによるパラメータの最適化などを行えば，より良い結果が得られる可能性がある．

4.2 実験結果と議論

4.2.1 生成された仮説の例

生成された仮説のうち，正例と一致しなかった仮説と一致した仮説のそれぞれ 20 件を**表 1** と**表 2** に例示する．表 1 を見ると，「datum *be escape from* expression of cox 2」のように一見して妥当でない仮説も含まれているものの，表 2 のように正例と一致する仮説も生成されていた．両者を比較すると，表 1 の仮説は比較的長い名詞句を含むのに対して，表 2 の仮説は短いものが多かった．これは，現在は名詞句の完全一致で仮説の同一性を比較しているため，長い名詞句は一致しにくいことによると考えられる．将来的には，2 章で述べた言い換えや RTE の研究成果を応用して，表層的には一致しないものの本来は同義であるような仮説についても，その同一性を判断できるようにすることが望ましい．

表 1 正例と一致しなかった仮説の例
Table 1 Generated hypotheses not found in ground truth.

cross recognition of lps *play* predominant role in salmonella pathogenesis.
ppt injection *be escape from* relaxation in control.
second embolization *regulate* sclerosis of adjacent laryngeal cartilage.
dysplastic epithelium *be inject into* apical portion of luminal cell.
occlusion by intraluminal filament technique *trigger* proliferation of thyroid cell.
homozygote *play* important role in development of common disease among northern indigenous people.
introduction of neomycin resistance gene cartridge in cod region *inhibit* exchange of anion.
mutate protein *play* important role as co factor in disease transmission.
gamma proteobacteria symbiont *be partition at* 3 to 5 degree.
modification to receptor *stabilize* volume.
expression *induce* tnf alpha as result of cellular oxidative stress.
moderate level of noise exposure *antagonize* abdominal fat measurement.
testicular tissue *be amplify from* ws.
essential definition of term *take* place in outpatient clinic of 3 veterans administration hospital.
datum *be escape from* expression of cox 2.
complete resection with mediastinal lymphadenectomy *regulate* favourable view on vocational training.
mean of co2 laser vaporization *regulate* successful clinical.
surgical decompression *regulate* three identical antigen.
wide variety of carcinogen *be mediate at* usable frequency in cell.
biliary cirrhosis and tsunoda type iii and iv *induce* analysis.

なお，語や概念の共起関係に基づく従来研究では，生成された仮説の中に表 1 や表 2 に斜体で示したような述語が存在しないため，たとえば「LPS（lipopolysaccharide）」と「DC（dendritic cells）」という概念のペアのみが仮説として提示される．そのため，両者の間にどのような関係が示唆されるのか，専門知識や関連文献に基づいてユーザ自身が解釈する必要がある．一方，提案手法では，「LPS *stimulate* DC」のように両者の関係を含めた仮説が提示されるため，そもそも概念間の関係を推測する必要がない．

4.2.2 仮説生成と順位付けの性能評価

テストデータに正例あるいは負例として含まれる仮説を，訓練データによって学習された回帰モデルの出力によって順位付けした．そして，順位付け後の精度を ROC 曲線と AUC（ROC 曲線下の面積）によって評価した．ROC 曲線は，x 軸と y 軸をそれぞれ偽陽性率，真陽性率として描

表 2 正例と一致した仮説の例
Table 2 Generated hypotheses agreeing with ground truth.

t lymphocyte *produce* interleukin 2.
actin *activate* atpase activity.
lymphocyte *produce* cytokine.
seb *induce* lethal shock.
osteoclast *produce* reactive oxygen intermediate.
vanadate *induce* contraction.
radiation *induce* necrosis.
hypertension *induce* vascular disease.
glucose *induce* phosphorylation of insulin receptor.
hydrogen peroxide *induce* necrosis.
macrophage *produce* cytokine.
fat *suppress* mri.
radiotherapy *induce* mucositis.
protease *activate* g protein.
lps *stimulate* pbmc.
endothelium *produce* oxygen.
ethanol *induce* locomotor activity.
scf *induce* proliferation.
tbid *induce* cytochrome c release.
lps *stimulate* dc.

いた曲線であり，二値分類器の精度を評価する際に用いられる．他の評価尺度として，精度（accuracy）や F 値などがある．しかし，正例と負例の数が著しく異なるため，精度はこの実験には不向きである．また，テストデータ中の負例は実際には負例ではない可能性があるため，F 値もふさわしくない．一方，ROC 曲線は正例と負例の分布がどれだけ分離しているかを表す指標であり，実際には負例ではない事例の数が真の負例の数よりも非常に少ない場合には，ROC 曲線への影響は少ない．以上から，以降の実験では ROC 曲線に基づいて性能評価を行う．

比較対象として，Hristovski ら [13] の仮説の順位付け法をベースラインとして利用する．この方法では，相関ルールマイニングにおける信頼度によって順位付けを行っている．具体的には，共起する概念の組 x と y の信頼度 $C_{x\to y}$ を，x が出現する文書中で y も出現する割合と定義し，同様に算出した概念 y と z の信頼度 $C_{y\to z}$ との積 $C_{x\to y}\times C_{y\to z}$ に基づき，仮説の順位付けを行う．ベースライン（Confidence）と提案手法（PU learning）の ROC 曲線を**図 5** に示す．ROC 曲線は，正負の判定の閾値を固定のステップ幅（0.01）で変化させ，真陽性率と偽陽性率を算出することで描いた．なお，対角線は無作為な順位付け（Random）に相当する．

この ROC 曲線におけるベースラインの AUC は 0.769，提案手法の AUC は 0.860 であった．いずれの手法も Random より高精度であり，仮説の順位付けに有効ながら，提案手法の方がより高い性能が得られた．この理由としては，次に議論するように，信頼度の算出に用いた概念の頻度情報だけでなく，提案手法では，述語の頻度や概念間の意味的類似性などの多様な素性を用い，さらに PU 学習によって適切なパラメータ（素性の重み）を推定したことが要因であると考えられる．

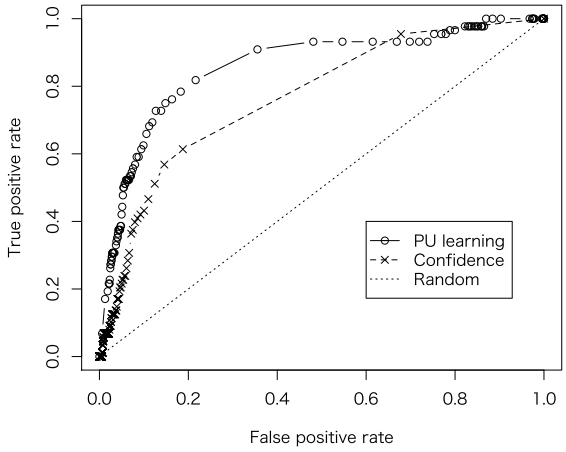


図 5 ROC 曲線による仮説生成の性能評価
Fig. 5 Performance of hypothesis discovery in ROC.

表 3 学習したモデルにおける回帰係数から見た素性の有用性の比較
Table 3 Comparison of the features in terms of the regression coefficients in the learned model.

グループ	素性	回帰係数
規則依存	V_3 の DF	246.33
	N_2 の DF	−176.02
	V_1 の DF	150.17
	N_3 の DF	−118.27
	N_1 の DF	−93.26
	V_2 の DF	7.41
	同一規則の基となるパターン数	4.92
	V_3 の類義語の数	3.70
	V_1 の類義語の数	−2.64
	N_2 の類義語の数	−0.72
	N_3 の類義語の数	−0.17
	V_2 の類義語の数	0.17
	N_1 の類義語の数	−0.05
規則・仮説依存	N_4 の DF	43.40
	N_6 の DF	39.64
	N_1 と N_4 間の NMD	−28.20
	N_3 と N_6 間の NMD	−21.73
	N_5 の DF	−14.07
	N_2 と N_5 間の NMD	−9.03
	同一の仮説数	−0.47

次に，順位付けに際して利用した素性のうち，どの素性が予測に有効であったのかを，回帰係数（素性の重み）から議論する．素性ごとの回帰係数を**表 3** に示す．表中の素性の順番は，規則依存，規則・仮説依存のグループごとに，回帰係数の絶対値の降順でソートしてある（表の上方の素性ほど有用である）．

まず，規則依存の素性について議論する．このグループの素性のうち，V_3 の DF 値は回帰係数が 246.33 と最も高

く，N_2 の DF 値，V_1 の DF 値，他の DF 値が続いた．V_3 は，生成された仮説の述語として現れるため，仮説の信頼性により影響があることは理にかなう．興味深い点として，述語の DF 値と概念（名詞句）の DF 値は正負反対の重みをとる結果となった．これは，より広く使われる述語と，より限定的に使われる名詞が，より信頼性の高い規則の獲得につながっており，最終的に真の仮説の生成につながりやすいことを示している．この結果の考えられる解釈としては，述語は閉じたクラス（closed class）であり，生物学的なメカニズムを記述する表現は，「x activates y」のような定型的なパターンを持つのに対し，真の仮説に現れる名詞はしばしば特定の固有表現であり，低い DF 値を持つことによるものと考えられる．なお，類義語の数など，他の素性の多くは，DF と比較して重みがきわめて低く，真の仮説の同定における有用性は低いことが分かった．

規則・仮説依存の素性については，N_4 と N_6 の DF の重みが高く，真の仮説同定の手がかりとして有効であった．重みの値は正であり，規則依存の素性を議論した際の結果とは，逆の結果となった．規則獲得時と仮説生成時の名詞句の役割の違いを示しているものと考えられる．また，NMD の値についても，ある程度の有効性が認められた．これらの素性について重みが負であることは，名詞句間の距離が小さい（意味的に近い）ほど真の仮説であることを示しており，本研究で仮定した，2 つのパターン間のアナロジ的類似性の有効性を裏付けるものである．

4.2.3 訓練データの大きさと仮説順位付けの精度

本評価実験で用いた訓練データ中には，正事例が 226 件しか含まれていない．そこで，さらに大きな訓練データを回帰モデルの学習に利用することによって，性能の向上が見込めるかどうかを検討するため，学習データ量の精度への影響を調査した．具体的には，訓練データ全体の n%をランダムに選択することで回帰モデルの学習，生成仮説の順位付けを行った．同一の n を用いて同じ実験を 10 回繰り返し，平均 AUC を算出した．**図 6** に，異なる n の値における平均 AUC の推移を示す．ここで，エラーバーは ±1 の標準偏差を示し，右端のデータはすべての訓練データを利用したときの AUC である．

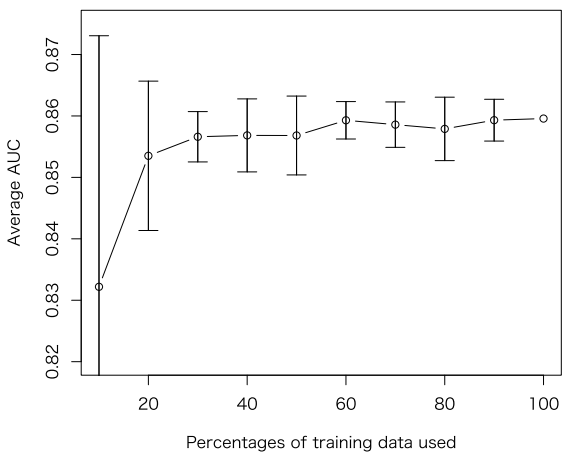


図 6 訓練データ量の違いによる AUC の変化
Fig. 6 Average AUC for different amount of training data.

この結果から，より多くの訓練データを利用することで少しずつ精度が向上しているものの，30〜40%以上の訓練データによる精度向上はわずかであることが分かる．この結果から，さらに多くの訓練データを利用したとしても，現在の素性とモデルでは大きな精度の向上は難しいと考えられる．さらなる精度向上のため，他の PU 学習モデル [8], [28] や，より多くの素性の利用を検討している．

5. おわりに

本稿では，三段論法的な述語項関係の連鎖に注目し，仮説生成の新しい枠組みを提案した．この枠組みの基本的アイデアは，関係の連鎖のパターンから仮説生成のための一般化した規則が獲得可能であり，その規則を既存知識に適用することで，新しい仮説が生成できることである．このアイデアの妥当性を評価するため，まず，10 年分の Medline データを基に仮説生成規則 12,180 件を獲得し，網羅的に 346,424 件の仮説を生成した．そして，より最近の文献を基に，生成された仮説から真の仮説を同定した．これらを正例として，詳細度や意味的類似性など様々な素性で表現し，PU 学習によって生成された仮説の妥当性を推定した．評価実験では，仮説の生成と真の仮説の信頼性を推定する際，提案する枠組みが有効であることを確認した．また，いくつかの素性が真の仮説に特徴的であることが分かった．

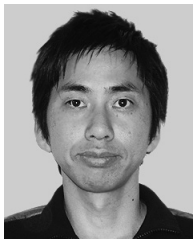
以上，全体的な結果は有望ではあるものの，いくつかの課題もある．まず，本稿の実験で用いた文献の量は限られているため，発見された真の仮説は，実際には，訓練データに現れなかった古い知識である可能性もある．また，知識ベースの構築には Medline のみを使っているため，生物学分野における基本的な知識が知識ベースに含まれていない可能性がある．さらに，訓練データとテストデータは完全に自動で構築されているため，その質や正確性は保証されない．今後は，これらの課題を解決し，加えて他の PU 学習モデルや素性の検討を加える予定である．

謝辞 本研究の一部は，栢森情報科学振興財団研究助成金#K23-XVI-363 と JSPS 科研費#25330363 の助成を受けて行った．

参考文献

[1] Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., Verkamo, A., et al.: Fast discovery of association rules, *Advances in Knowledge Discovery and Data Mining*, Vol.12, pp.307–328 (1996).
[2] Aronson, A.R.: Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program, *AMIA*, pp.17–21 (2001).

[3] Berant, J., Dagan, I., Adler, M. and Goldberger, J.: Efficient tree-based approximation for entailment graph learning, *ACL*, pp.117–125 (2012).
[4] Björne, J., Ginter, F., Pyysalo, S., Tsujii, J. and Salakoski, T.: Complex event extraction at PubMed scale, *Bioinformatics*, Vol.26, No.12, pp.i382–i390 (2010).
[5] Cilibrasi, R.L. and Vitanyi, P.M.B.: The Google Similarity Distance, *IEEE Trans. Knowl. and Data Eng.*, Vol.19, pp.370–383 (2007).
[6] Digiacomo, R.A., Kremer, J.M. and Shah, D.M.: Fish-oil dietary supplementation in patients with Raynaud's phenomenon: A double-blind, controlled, prospective study, *The American Journal of Medicine*, Vol.86, No.2, pp.158–164 (1989).
[7] Do, Q.X., Chan, Y.S. and Roth, D.: Minimally supervised event causality identification, *EMNLP*, pp.294–303 (2011).
[8] Elkan, C. and Noto, K.: Learning classifiers from only positive and unlabeled data, *SIGKDD*, pp.213–220 (2008).
[9] Fellbaum, C.D.: *WordNet: An electronic lexical database*, MIT Press (1998).
[10] Hashimoto, C., Torisawa, K., De Saeger, S., Oh, J.-H. and Kazama, J.: Excitatory or inhibitory: A new semantic orientation extracts contradiction and causality from the Web, *EMNLP/CoNLL*, pp.619–630 (2012).
[11] Hearst, M.A.: Untangling Text Data Mining, *ACL*, pp.3–10 (1999).
[12] Hersh, W., Bhuptiraju, R.T., Ross, L., Cohen, A.M. and Kraemer, D.F.: TREC 2004 Genomics Track Overview, *TREC* (2004).
[13] Hristovski, D., Peterlin, B., Mitchell, J.A. and Humphreyb, S.M.: Using literature-based discovery to identify disease candidate genes, *International Journal of Medical Informatics*, Vol.74, pp.289–298 (2005).
[14] Kostoff, R.N., Block, J.A., Solka, J.L., Briggs, M.B., Rushenberg, R.L., Stump, J.A., Johnson, D., Lyons, T.J. and Wyatt, J.R.: Literature-related discovery, *Annual Review of Information Science and Technology*, Vol.43, No.1, pp.1–71 (2009).
[15] Lee, W.S. and Liu, B.: Learning with Positive and Unlabeled Examples Using Weighted Logistic Regression, *ICML* (2003).
[16] Liu, H., Le Pendu, P., Jin, R. and Dou, D.: A Hypergraph-based Method for Discovering Semantically Associated Itemsets, *ICDM*, pp.398–406 (2011).
[17] Mihalcea, R., Corley, C. and Strapparava, C.: Corpus-based and knowledge-based measures of text semantic similarity, *AAAI*, pp.775–780 (2006).
[18] Pratt, W. and Yetisgen-Yildiz, M.: LitLinker: Capturing connections across the biomedical literature, *2nd International Conference on Knowledge Capture*, pp.105–112 (2003).
[19] Schoenmackers, S., Etzioni, O., Weld, D.S. and Davis, J.: Learning first-order Horn clauses from web text, *EMNLP*, pp.1088–1098 (2010).
[20] Jones, K.S.: Statistical interpretation of term specificity and its application in retrieval, *Journal of Documentation*, Vol.28, No.1, pp.11–20 (1972).
[21] 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).
[22] Swanson, D.R.: Fish oil, Raynaud's syndrome, and undiscovered public knowledge, *Perspectives in Biology and Medicine*, Vol.30, No.1, pp.7–18 (1986).
[23] Swanson, D.R.: Migraine and magnesium: Eleven neglected connections, *Perspectives in Biology and Medicine*, Vol.31, No.4, pp.526–557 (1988).
[24] Swanson, D.R., Smalheiser, N.R. and Torvik, V.I.: Ranking indirect connections in literature-based discovery: The role of medical subject headings, *Journal of the American Society for Information Science and Technology*, Vol.57, No.11, pp.1427–1439 (2006).
[25] Szpektor, I. and Dagan, I.: Learning entailment rules for unary templates, *Coling*, pp.849–856 (2008).
[26] Tsuruoka, Y. and Tsujii, J.: Bidirectional inference with the easiest-first strategy for tagging sequence data, *HLT/EMNLP*, pp.467–474 (2005).
[27] Weeber, M., Klein, H., de Jong-van den Berg, L.T.W. and Vos, R.: Using concepts in literature-based discovery: Simulating Swanson's Raynaud-fish oil and migraine-magnesium discoveries, *Journal of the American Society for Information Science and Technology*, Vol.52, No.7, pp.548–557 (2001).
[28] Xiao, Y., Liu, B., Yin, J., Cao, L., Zhang, C. and Hao, Z.: Similarity-based approach for positive and unlabelled learning, *IJCAI*, pp.1577–1582 (2011).



関 和広 （正会員）

平成 14 年図書館情報大学大学院情報メディア研究科修士課程修了．平成 18 年インディアナ大学図書館情報学研究科博士課程修了．Ph.D. 神戸大学助手，助教，講師を経て，現在，同大学大学院システム情報学研究科准教授．情報検索，データマイニングの研究に従事．自然言語処理学会会員．

上原 邦昭 （正会員）

昭和 53 年大阪大学基礎工学部情報工学科卒業．昭和 58 年同大学大学院博士後期課程単位取得退学．同産業科学研究所助手，講師，神戸大学工学部情報知能工学科助教授等を経て，現在，同大学大学院システム情報学研究科教授．工学博士．人工知能，特に機械学習，マルチメディア処理の研究に従事．電子情報通信学会，計量国語学会，日本ソフトウェア科学会，AAAI 各会員．