

生物医学文献からの知識抽出とイベント間のつながりを考慮した発見性を伴う仮説の提示

宮西 大樹[†] 関 和広[†] 上 原 邦昭[†]

[†] 神戸大学, 兵庫県

E-mail: [†]ai.taiki@gmail.com, ^{††}seki@cs.kobe-u.ac.jp, ^{†††}uehara@kobe-u.ac.jp

あらまし 近年, 発表される生物医学文献の数は指数的增加を見せており, 研究者個人が自身の研究分野に関するすべての文献を精査することは実質的に不可能である. そのため, 文献中に記述された知識を単純に組み合わせることで導かれる(潜在的)知識が未発見のまま存在する可能性があり, 事実, 複数の研究者によって, そのような事例が報告されている. 本研究では, 生物医学文献に記述された知識を, 語の共起を用いて自動的に抽出し, 抽出した知識を結合することにより, 概念ネットワークを構築する. この構築した概念ネットワークを探索することにより, 「潜在的知識」あるいは「仮説」を自動的に生成する枠組みを提案する. さらに, 複数の概念を構成要素とする知識間の関係を, 知識の抽出元である文献に付与された情報を用いて評価する. その評価をもとにして, それらの知識から導出された仮説に意味性や発見性を付与し, 順位付けを行う手法を提案する. 仮説手法の評価には, 既に発見された知識を, 発見される以前のデータを用いて予測できるかで評価を行った. 結果, 本手法の有効性を確認できた.

キーワード 情報抽出, 仮説生成, 生物医学文献, テキストマイニング

Knowledge Extraction from The Biomedical Literature and Hypothesis Generation Considering Event connection

Taiki MIYANISHI[†], Kazuhiro SEKI[†], and Kuniaki UEHARA[†]

[†] Kobe University

E-mail: [†]ai.taiki@gmail.com, ^{††}seki@cs.kobe-u.ac.jp, ^{†††}uehara@kobe-u.ac.jp

Abstract Recently, the number of publications of the biomedical literature is exponentially increasing. It is impossible that each researcher cannot survey all literature related their field. So, there may be implicit and undiscovered knowledge combined some simple knowledge from texts, as facts, such cases have been reported. In this paper, we introduce how to generate hypotheses automatically, which are not discovered, after structuring biomedical conceptual network consisting of knowledge from biomedical literature. Also we proposed method identify the hypotheses which will be meaningful knowledge, considering relation of knowledge which are source of them.

Key words Information Extraction, Hypothesis Generation, Biomedical Text, Text Mining

1. ま え が き

発表される生物医学文献の数は日々増大しており, 研究者個人が自身の専門分野すべてを把握することは困難になっている. そのため, 未だ発見されていない知識が大量のデータの中に埋もれていると考えられる. 事実, 生物医学文献中に記述された知識を組み合わせることで, 生物医学上の未知の知識を同定することができた事例も複数報告されている. 文献を基にした仮説生成手法の第一人者である Swanson は, 魚油とレイノー病に関係する文献を手作業で調べ上げ, 抽出した知識を組み合わせることによって両者の関係を予測した [6]. この関係

は, 後に DiGiacomo らによって実験的に証明された

このような Swanson の功績により, Swanson が手動で行った作業を自動で行うことで, 大量の文献の中から未だ知られていない発見につながるような知識を自動的に同定する試みが複数のグループによって行われている [3] [7] [5]. しかし, これらの手法には, 意味のない仮説を大量に生成したり, 仮説を生成する際に人手を要するという欠点がある. また, 概念の頻度を元に仮説の妥当性を測るため, 低頻度の概念から生成される仮説に対処できなかった.

そこで, 本研究では, 仮説の生成の際に用いる文献から抽出した知識 (イベント) 間の意味的つながりを考慮することで意

味のない仮説が明示化されることを避け、仮説の生成を完全に自動化することを目指す。さらに、仮説を生成するために必要となるイベントの意味情報に注目し、その意味情報からイベント間の関係性を定量化する。その関係性を用いて、生成した仮説の順位付けを行い、有望な仮説を同定する。

本文では以後、第2章で知識抽出と概念ネットワークの構築の仕方について説明し、第3章で仮説の生成手法を提案する。第4章では、生成した仮説の妥当性を評価するための実験と考察を行う。最後に第5章で、本文のまとめと今後の課題について述べる。

2. 知識抽出

本手法では、仮説を生成するために、生物要素間の関係を記述した概念ネットワークを必要とする。そこで、文献から抽出した知識を同義の生物要素を基に複数の生物要素間の関係を結合し、ネットワークを構築する。そのための第一段階として、生物医学文献から固有表現抽出および関係抽出を行う。

2.1 固有表現抽出

固有表現抽出とは文献中から固有表現を同定する処理のことである。今回は生物医学文献を対象とするので、たんぱく質・遺伝子・化学物質・疾患名などの生物医学要素を抽出対象とする。本研究では、辞書に基づく文字列のマッチング手法による固有表現抽出を行うツール MetaMap [1] を用いた。MetaMap は文書中の固有表現を UMLS (unified medical language system) のシソーラスと意味クラスに対応づける。これにより、語彙の曖昧性を解消することができ、意味のない関係性や概念が概念ネットワーク上に現れることを防ぐことができる。

2.2 関係抽出

関係抽出とは文献の中から抽出した固有表現間の関係 (イベント) を同定する処理のことである。本研究では、同一文中に出現する2つ以上の固有表現間には関係があるとすると、語の共起を用いた関係抽出を行う。語の共起による関係抽出を行うと、膨大な量の関係が抽出されるので、固有表現抽出の際に各語に付与した UMLS の意味クラスをもとに、特定の意味クラスに属する固有表現を含む関係だけを概念ネットワークの作成に用いる。

2.3 概念ネットワークの構築

関係抽出をすることで取得した固有表現間の関係 (イベント) を共通する固有表現を元に結合していく。この作業を複数の関係について行うことでネットワークを構築していく。図.1 は文献から抽出したイベントを結合して概念ネットワークを構築する過程を表している。上記の図から、文献1からはA,BとC,E

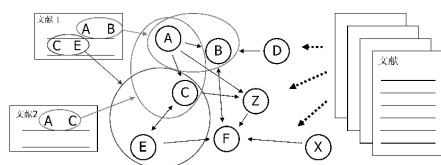


図1 Conceptual Network

のイベントが抽出され、文献2から抽出したイベントA,Cの

共通する固有表現AとCにより、文献1・文献2を個別にみているだけでは確認できないA,C,Eの関係を構築できていることが分かる。

3. 仮説生成

興味のある固有表現から他の固有表現への未だ発見されていない関係を導くために、興味のある固有表現を始点として、概念ネットワーク上のノード間のエッジ (固有表現間の関係) をたどることにより探索を行う。探索が終了したノードを終点とし、始点と終点を結ぶ関係が仮説の候補となる。この始点となるノードを A-term, 始点と終点を結ぶ中間のノードを B-term, 終点となるノードを C-term と呼ぶ。

3.1 仮説候補の検証

仮説の候補を生成した後は、その候補が概念ネットワーク上において未知の知識であるかの検証を行う。この検証により、生成された仮説が概念ネットワーク中表れればその仮説は知識として扱う。

3.2 仮説の発見性

生成した仮説が概念ネットワーク中に記述されていないだけで、現実上では当然の知識である可能性があり、そのような仮説は発見という観点からは有益でない知識であると考えられる。そこで、生成した仮説が発見につながるかどうかの尺度を設ける必要がある。この尺度を発見性とする。発見性は仮説が当たり前でない度合いを示す。

本研究では、発見性を表すために、仮説を説明するイベント (文献から抽出された知識) の分野的距離を考えた。この距離は仮説の導出の際に用いたイベント間の非類似度であるとした。

このような発見性の定義は以下の仮定に基づいている。各々のイベントには抽出元となった文献が存在し、それらの文献同士が類似していないほど、それらから抽出された知識によって新たに導かれる知識は同定されにくいはずである。なぜなら、文献が類似していないということは、文献が属する分野が類似していないことを表しているからである。逆にいえば、この仮定を基準とした尺度を用いることで、現在、細分化された分野間の知識の溝を埋めることができる。そこで本研究では、分野的に類似していない知識から導出される仮説は同時に同定 (発見) されにくい、つまり発見性を伴っているとし、これを明示する。こうすることで、ユーザが発見性の高い仮説、つまり、当たり前でない知識となる可能性の高い仮説を優先的に同定することができる。図.2 に仮説の発見性に関する概念図を示す。

発見性の具体的な定義を以下に示す。まず、イベントの抽出元である文献は必ず論文誌に掲載されている。そこで、イベントに対応する文献を掲載した論文誌間の非類似度をイベント間の距離とし、これを分野的距離とする。ある年代までに論文誌上で掲載された文献中の語の集まりを、その論文誌を特徴付ける要素であるとし、これを TF-IDF 法を用いて数値化して、論文誌を単語ベクトルで表現した。論文誌間の類似度 sim には式 (1) に示すベクトル間のコサイン類似度を用いる。

$$\text{sim}(\mathbf{j}_a, \mathbf{j}_b) = \frac{\mathbf{j}_a \cdot \mathbf{j}_b}{|\mathbf{j}_a| |\mathbf{j}_b|} \quad (1)$$

ここで, j_a, j_b は論文誌を表現した単語ベクトルである.

この論文誌間の類似度を用いて, イベント間の分野的距離を定義する. イベントには複数の論文誌 (例えば, $J_a = \{j_{a1}, j_{a2}, \dots, j_{ak}, \dots, j_{aK}\}$) が対応する場合があるので, イベント e_a, e_b 間の距離 FiM は以下の式 (2) を用いて表現する.

$$FiM(e_a, e_b) = \frac{1}{KL} \sum_k \sum_l JDist(j_{ak}, j_{bl}) \quad (2)$$

ここで, K, L は各々, 仮説の生成に用いるイベント e_a, e_b の抽出元となる論文誌の数であり, $JDist$ は論文誌を表現する単語ベクトル j_{ak}, j_{bl} 間の非類似度を表しており, 式 (3) に論文誌間の非類似度を定義する.

$$JDist(j_{ai}, j_{bj}) = 1 - \text{sim}(j_{ak}, j_{bl}) \quad (3)$$

ちなみに, 単語ベクトル j_{ak} はイベント e_a の抽出元となる, ある一つの文献の, さらにその文献を掲載した論文誌を表す単語ベクトルである.

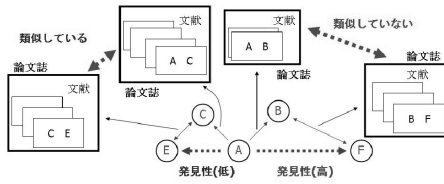


図2 Discoverability of hypothesis and fields similarity

3.3 仮説の意味性

先ほど定義した仮説の発見性は, 仮説の導出に用いるイベントの組が同時に見つからない可能性を表したものであり, たとえ発見性の高いイベントの組があったとしても, それらから導出される仮説 (仮説の説明も含む) が現実上の摂理にそぐわない, 意味のない知識である可能性がある. そこで, 生成した仮説に意味のある知識であることを示すための尺度を設ける. そのために, 仮説の導出の際に用いたイベント間の意味的距離を定義し, これを用いることで導出された仮説に意味があるかどうかを測る. このような, 仮説に意味があるかないかを表すための尺度を意味性とする.

イベントの意味的距離にはイベントの抽出元である文献間の意味的非類似度を用いて表す. 仮説の導出に使用するイベント同士が意味的に近ければ, 導出される仮説の意味性が高い, つまり, 仮説に意味があるとする. この考えは以下に示す仮定に基づいている. 複数の意味のある知識によって形成される空間を考えたとき, 類似した知識は意味的に隣接して分布する. よって, 意味のある知識は意味的に類似した知識によって導出されるはずである. 図3に仮説の意味性と仮説の元となる知識の意味的類似度の関係を表す概念図を示す.

本手法では, 文献間の意味的距離を, 文献を代表する語彙間の意味的距離の総和で表現する. 文献を代表する語彙には MeSH (Medical Subject Headings) 語彙を用いる. MeSH は NLM (National Library of Medicine) が定めた生命科学用語集であり, MEDLINE の各文献に対して, 文書分類を行う目的

で, 人手により数個から十数個程度付与される. MeSH には, 文献の内容を表すための用語として定義語・副表題・補足用語が用意され, 2008 年の時点で 24766 個の定義語が登録されている. この定義語は階層構造により意味的に細かく分類されており, 下位にいくほど厳密な意味を表す語彙となっている.

本研究では, まず, 文献間の意味的距離を測るために, 文献の MeSH 語彙間の意味的距離を計算する. 語彙間の意味的距離には Nuno らが提案した木構造を持つ分類語彙表 (シソーラス) 上での語彙の情報量に基づく語彙間の意味的類似度の計測方法 [4] を参考にした. この手法は, シソーラス中の上位の概念は厳密な語彙でないため情報量が少なく, 下位の概念は厳密な語彙であるため情報量が多いというシソーラスの特性を生かした語彙間の類似度計測法である. この特性を利用することで, 上位の概念同士が類似しているよりも, 下位の概念が類似しているときの方が類似度は大きくなる, といった類似度計算が可能となる. また, 情報量を決める際にコーパスを作成する必要がないため, 類似度の計算が容易である. 他にシソーラスを用いる利点として, シソーラスは意味に基づいて構造化されているため, 意味に関連する情報を取り出すことができるのと, コーパスと違い, データによって制限されない点が挙げられる. 後者は, コーパスはデータの性質 (年代, 流行) に影響されるが, シソーラスは語彙が表す意味に基づいて構造化されているので, データの性質に影響を受けないということを意味している.

以下にシソーラス中の語彙の情報量 式 (4) と語彙間の類似度 式 (5) および距離式 (6) を表す式を示す.

$$\begin{aligned} ic_{wn}(m) &= \frac{\log(\frac{\text{hypo}(m)+1}{\max_{wn}})}{\log(\frac{1}{\max_{wn}})} \\ &= 1 - \frac{\log(\text{hypo}(c) + 1)}{\log(\max_{wn})} \end{aligned} \quad (4)$$

$$\text{sim}(m_1, m_2) = \max_{c \in S(m_1, m_2)} ic_{wn}(c) \quad (5)$$

$$\text{dist}(m_1, m_2) = 1 - \text{sim}(m_1, m_2) \quad (6)$$

ここで, $ic_{wn}(m)$ は MeSH 語彙 m の分類語彙表上での情報量であり, $\text{hypo}(m)$ は語彙 m 以下の下位語の数である. また, $\max_{wn}$ は分類語彙表中にある語彙の数である. よって, シソーラス上で仮想的な最上位のノード (根) の情報量は 0 となる. $\text{sim}_{wn}(m_1, m_2)$ は語彙 m_1, m_2 間の類似度を表しており, $c \in S(m_1, m_2)$ が示す c は, m_1 と m_2 の共通の祖先であり, かつ, 両者から最も近い語彙を表す. $\text{dist}_{wn}(m_1, m_2)$ は語彙 m_1, m_2 間の意味的距離を表している. ちなみに, これら情報量・類似度・距離はいずれも 0 から 1 の値をとる.

本手法では, 式 (6) で定義した語彙間の意味的距離を文献間の意味的距離に拡張する. これには Corley らによる手法 [2] を参考にした. この手法は, まず, 一方の文献の中から語彙を一つ選び, もう一方の文献のすべての語彙に対して意味的類似度をもとめ, あらかじめ定めた基準に応じて, 語彙間の類似度を一つ選ぶ. この作業を一方の文献中のすべての語彙について行

選ばれた語彙間の類似度の総和をとる．これを双方の文献に対して計算し，平均をとったものを文献間の意味的類似度としている．Coley らは，ある語彙に対して一番似ている語彙の組に語彙の特性性（語が文献を特徴付ける性質）を乗することにより，特性性の大きい語彙を含む語彙間の意味的類似度の重みを大きくしている．Corley らは特性性を求めるためにコーパスを用いているが，本手法では，コーパスを作成せず，また Nuno らの手法により下位の語彙間の類似度により大きな重みが付与されているため，語彙の特性性による重み付けは行わないことにする．

提案手法を以下に示す．イベント e_a, e_b 間の意味的距離 SeM を定義する．

$$SeM(e_a, e_b) = \frac{SDist(M_a, M_b)_{M_a} + SDist(M_a, M_b)_{M_b}}{2} \quad (7)$$

ここで， M_a, M_b はイベント e_a, e_b に対応する MeSH 語彙の集合を表している．この MeSH 語彙はイベントの抽出元である文献に付与された MeSH 語彙を，イベントの抽出元である文献すべてについて，重複を許さずに集めたもの（例えば，MeSH 語彙の集合 M_a の要素は $\{m_{a1}, m_{a2}, \dots, m_{aM}\}$ となる）である． $SDist$ は一方の MeSH 語彙の集合から，もう一方の MeSH 語彙の集合への意味的距離を表したものであり，以下に定義する．

$$SDist(M_a, M_b)_{M_a} = \frac{1}{M} \sum_{m_{am} \in M_a} \min_{m_{bm} \in M_b} dist_{wn}(m_{am}, m_{bm}) \quad (8)$$

ここで， M は片方側からみた文献間の距離を測る際，基準となる MeSH 語彙の集合 M_a の大きさ（要素数） $|M_a|$ である．式 (8) は，一方の文献からもう一方への文献に対して，似ている MeSH 語彙の類似度だけを足していることに等しい．よって，この方法をとれば，イベントに対応する MeSH 語彙の数に違いが生じてても，仮説生成に関するすべての MeSH 語彙を意味的距離の対象とすることができ，また，語彙間の最小距離だけを足していくので，イベント間の類似する特性だけを考慮することができる．

MeSH 語彙に関する注意として，MeSH 語彙は重複して一つの定義語に当てはまる場合がある MEDLINE の文献には定義語の方が付与されるので，定義語に曖昧性がある場合，シソーラス中の当てはまる複数のノードを MeSH 語彙の集合に加えることにする．

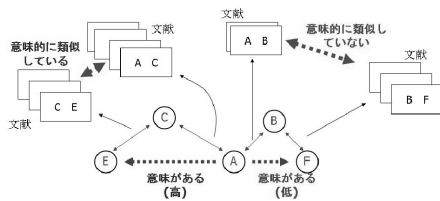


図3 Mean of hypothesis and events similarity

4. 評価・実験・考察

4.1 評価手法

評価手法としては，Swanson が 1986 年に予測し，後に Di-Giacomo らによって実験的に証明された魚油とレイノー病の関係を自動的に予測できるかで評価する．そのために，生物医学文献データベース MEDLINE の 1985 年までのデータであり，かつ，MeSH 語彙が付与されたすべてのデータを用いて概念ネットワークを構築する．そして，概念ネットワークに対して，魚油を所与として，レイノー病との関係を表す仮説を生成する．その仮説を本稿で定義した意味性・発見性をもとに順位付けを行い，魚油とレイノー病の関係を上位に順位付けできるかで，本稿で提案した仮説生成の手法を評価する．一般的に使われる評価では，レイノー病を所与として魚油との関係を予測するが，今回は魚油を所与としてレイノー病の関係を導くことを目指す．その理由は，レイノー病には対応する複数の固有表現があるため，魚油を所与とするとレイノー病に関連する多くの関係を得ることができ，そのおかげで，順位付けに使う関係が多くなり，より妥当な評価ができるからである．

仮説の評価は二段階に分けて行う．一つめは，現在，実験的に証明された知識を，その知識が予測される前までのデータを用いて予測し，その予測した仮説を他の仮説と比べて上位に順位付けできるかで評価する．今回の評価手法では，実験的に証明されている魚油とレイノー病の関係をを用い，この関係には意味がある，つまり，現実の摂理に従っているとする．この魚油とレイノー病の関係が他の仮説と比べて相対的に上位に順位付けされれば，その順位付けに用いた仮説の意味を測るための指標は，仮説に意味があるという確からしさを保証しているとする．

この評価の仕方の正当性を以下に述べる．仮説の生成のために用いるイベントは，語の共起を用いて抽出したものであるため，直接的に関係のない固有表現間の関係を多く含んでいると考えられる．ここで，直接的に関係のない固有表現間の関係を誤りと呼ぶことにする．さらに，イベントがデータベース中に出現する回数を用いてイベントの確からしさを保証し，ある一定の頻度に達していないイベントは取り除く，といったことも本手法では行っていない．そのため，文献から抽出した関係は誤りである可能性が高いと考えられる．また，誤りの多い関係の中から生成される仮説はさらに多いと考えられる．よって，このように誤りの多い中でも確実に意味のあるといえる関係を上位に順位付けできるということは，順位付けに用いた指標が妥当であるといえる．すなわち，この評価手法は正当であるといえる．

二つめは生成した仮説の説明に意味があるかの評価である．上記で示した始点と終点の関係を上位に順位付けしたことを評価するだけでは不十分だからである．なぜなら，語の共起を用いた関係抽出では，他の語彙と共起しやすい語彙により，現実の摂理にそぐわない誤った説明をもたらす仮説を生成してしまう場合があるからである．そのような誤った説明を含む仮説は偶発の産物であり，有益な知識でないと判断できる．そこで，

始点と終点の関係は摂理に従った関係でありながらも、誤った説明をもたらす仮説を下位に順位付けしなければならない。よって、誤った説明を持つ仮説を下位に順位付けするという評価も評価の基準として取り入れる。

4.2 実験

仮説生成の前段階として、まず、概念ネットワークを構築する。そのための実験条件を以下に示す。MetaMap を用いて固有表現抽出を行う際、固有表現抽出の誤りを少なくするため、マッチング精度の閾値を 800 とした。また、MetaMap は語と固有表現のマッピングに曖昧性が生じたとき、単一の語に対して、複数の固有表現を出力することがあるので、先に出力された表現を語に対応する固有表現とした。さらに、関係抽出の過程で、UMLS の意味クラスによる絞り込みを行う際、特定の固有表現の再現率を大きくするため、固有表現 Blood viscosity [Laboratory or Test Result] を Blood Viscosity [Physiologic Function] に置換した。

次に、関係抽出を行う際、特定の意味クラスに属する固有表現を含む関係だけを概念ネットワークの形成に用いる。制限に使った UMLS の意味クラスを表 1 に示す。これらの意味クラスは、Weeber [7] らが固有表現の制限のために使ったものに筆者が独自に Disease or Syndrome, Lipid を追加し、代わりに Finiding, Phenomenon or process を除去したものである。

表 1 Regulate Semantic Type

Semantic Type
Biologic Function
Cell Function
Disease or Syndrome
Lipid
Molecular Function
Organ or Tissue Function
Organism Function
Pathologic Function
Physiologic Function

上記の条件で作成した概念ネットワークを用いて仮説生成を行った。魚油を始点として合計で 15201 個の仮説が得られた。その内、レイノー病を終点とする仮説が 9 つ確認された。表 2 に、生成した魚油とレイノー病の関係を A-term, B-term, C-term を用いて掲載する。A-term は仮説の始点、B-term はイベントとイベントを結ぶ固有表現、C-term は仮説の終点を示す。

表 2 Hypotheses : Fish Oil and Raynaud Disease

A-term	B-term	C-term
Fish Oil	Atheromatosis	Raynaud Disease
Fish Oil	Blood Viscosity	Paroxysmal digital cyanosis
Fish Oil	Blood Viscosity	Primary Raynaud's
Fish Oil	Blood Viscosity	Raynaud Disease
Fish Oil	Development	Paroxysmal digital cyanosis
Fish Oil	Development	Raynaud Disease
Fish Oil	Peripheral vascular disease	Paroxysmal digital cyanosis
Fish Oil	Peripheral vascular disease	Raynaud Disease
Fish Oil	Suppression	Paroxysmal digital cyanosis

次に生成した仮説の評価を行う、3.3 節で提案したイベント間の意味的距離の指標が仮説に意味性を与えているかの評価を行う。そこで、比較の対象として、以下に追加してイベント間の意味的距離を二つ定義する。これらの比較対象となる意味的距離には定義上の問題点がある。このような明らかな問題点がある比較対象を設ける理由として、問題点のある対象と比べることで、本提案手法の妥当性を強調するためである。

一つめの比較対象を式 (9) に定義する。この意味的距離の特徴は、互いの MeSH 語彙の総数で割っていることから MeSH 語彙の数が大きくなるほど類似度が小さくなり易いことである。また、比べる必要のない類似していないもの同士も類似度を計算するので、仮説に必要なものとは関係のないものを評価していることになる。

$$SeM_{simple}(e_i, e_j) = \frac{1}{MN} \sum_m \sum_n dist_{wn}(m_{im}, m_{in}) \quad (9)$$

二つめの比較対象を式 (10) に定義する。この定義は意味的に近い語彙だけを比べることを目的としている。さらにイベントの証拠となる文献に誤りがあることを想定し、MeSH 語彙の少ない方に対応するイベントを基準とし、その語彙間の最小距離の総和を求めている。これにより、片方のイベントに都合の良い距離を測るための設定される。

$$SeM_{min}(e_i, e_j) = \frac{1}{P} \sum_{m_{pr} \in M_p} \min_{m_{qs} \in M_q} (dist_{wn}(m_{pr}, m_{qs})) \quad (10)$$

$$P = \min(M, N)$$

式 (9), 式 (10) 式 (7) の意味性の指標を用いて 表 2 の結果を順位付けした結果を、それぞれ 表 3, 表 4, 表 5 に示す。

表 3 Average of MeSH-metric

Rank	B-term	C-term
1893	Blood Viscosity	Primary Raynaud's
3879	Blood Viscosity	Paroxysmal digital cyanosis
4076	Blood Viscosity	Raynaud Disease
7219	Atheromatosis	Raynaud Disease
7657	Peripheral vascular disease	Paroxysmal digital cyanosis
7723	Suppression	Paroxysmal digital cyanosis
9587	Development	Raynaud Disease
12506	Peripheral vascular disease	Raynaud Disease
14291	Development	Paroxysmal digital cyanosis

表 4 Latest MeSH-metric (one side)

Rank	B-term	C-term
272	Blood Viscosity	Primary Raynaud's
2112	Blood Viscosity	Paroxysmal digital cyanosis
3845	Suppression	Paroxysmal digital cyanosis
5029	Development	Raynaud Disease
5778	Atheromatosis	Raynaud Disease
7918	Blood Viscosity	Raynaud Disease
10445	Peripheral vascular disease	Raynaud Disease
12068	Peripheral vascular disease	Paroxysmal digital cyanosis
14239	Development	Paroxysmal digital cyanosis

表 5 Latest MeSH-metric (both sides)

Rank	B-term	C-term
290	Blood Viscosity	Primary Raynaud's
1389	Blood Viscosity	Raynaud Disease
1963	Blood Viscosity	Paroxysmal digital cyanosis
3729	Peripheral vascular disease	Paroxysmal digital cyanosis
5533	Suppression	Paroxysmal digital cyanosis
7296	Atheromatosis	Raynaud Disease
9349	Peripheral vascular disease	Raynaud Disease
13961	Development	Raynaud Disease
14179	Development	Paroxysmal digital cyanosis

4.3 考 察

式 (10), 式 (7) は 式 (9) の結果と比べて, 魚油とレイノー病の関係を全体的に高い順位で順位付けできていることがわかる。また, 表 4 と表 5 を比べると, 式 (7) は Blood Viscosity と Primary Raynaud's の関係を 式 (10) とほぼ同順位で順位付けている。さらに, 表 5 から 式 (7) は, 中間語に Development といった他の語彙と共に起し易い語彙を含む仮説 (始点と終点はあっているが説明が間違っている仮説) を比較的低順位に順位付けできている。これは 式 (7) の定義では実現できていない。よって, 本稿で提案した手法は他の手法より意味性を反映できていると言える。

次に発見性の有効性について見ていく。発見性の低い仮説は, 裏を返せば当たり前の知識である可能性が高いことを示している。そこで, 表 5 の結果に対して, 発見性が 0 の仮説, つまり当たり前でない知識である可能性が著しく低い仮説を除去することで, さらに順位付けを行った。すると, 289 位以内の内 18 個の仮説が発見性が 0 であるため, より高い順位で魚油とレイノー病の関係を順位付けすることができた。

4.4 追加実験

補足として, 上記の結果を元に Primary Raynaud's と Raynaud Disease を始点として, 終点である C-term を意味クラス Lipid で絞り込んで検索すると, 表 6, 7 に示すように比較的高順位に Fish Oil を順位付けできている。また, 表 6 で, 2 位と 3 位に順位付けられている要素はいずれも Fish Oil に関連する生物要素である。ちなみに, 表中の Rank(num) の中の num の部分は C-term を意味クラス Lipid で絞り込んだときの C-term の総数である。

表 6 A-term:Primary Raynaud's, B-term's semantic type:Lipid

Rank(34)	B-term	C-term
1	Blood Viscosity	Fish Oil
2	Blood Viscosity	5,8,11,14,17-Eicosapentaenoic Acid
3	Blood Viscosity	Fatty Acids, Nonesterified

表 7 A-term:Raynaud Disease, C-term's semantic type:Lipid

Rank(970)	B-term	C-term
77	Blood Viscosity	Fish Oil (Fish Oils)
495	Atheromatosis	Fish Oil (Fish Oils)
624	Peripheral vascular disease	Fish Oil (Fish Oils)
785	Development	Fish Oil (Fish Oils)

5. ま と め

本論文では, 仮説の意味性と発見性を定義することにより, 有望な仮説の上位順位付けを目指した。今回もちいた発見性の指標は仮説の生成に使用したイベントの属する論文誌間の非類似度を用いて定義した。この発見性は当たり前の知識の除去に役立つことがわかった。また, 文献に付与された意味情報である MeSH 語彙を用いて, 文献から生成される新たな知識 (仮説) に意味性を付与し, この意味性をもとに順位付けを行った結果, 意味のない仮説を下位に順位付けることに成功した。また, これらの指標はいずれも, 頻度を用いず, さらに, 意味性の尺度を示す意味的距離を計算するためにコーパスを用いなかった。このことからデータベース中に現れる稀な事象 (低頻度の概念) にも対処できる。

文 献

- [1] Alan R. Aronson. Effective mapping of biomedical text to the UMLS metathesaurus: The MetaMap program. In *Proceedings of American Medical Informatics 2001 Annual Symposium*, pp. 17–21, 2001.
- [2] Courtney Corley and Rada Mihalcea. Measuring the semantic similarity of texts. In *Proceedings of the ACL 2005 Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pp. 13–18, 2005.
- [3] Michael D. Gordon and Robert K. Lindsay. Toward discovery support systems: a replication, re-examination, and extension of Swanson's work on literature-based discovery of a connection between Raynaud's and fish oil. *Journal of the American Society for Information Science*, Vol. 47, No. 2, pp. 116–128, 1996.
- [4] Nuno Seco, Tony Veale, and Jer Hayes. An intrinsic information content metric for semantic similarity in wordnet. In *Proceedings of European Conference on Artificial Intelligence 2004*, pp. 1089–1090, 2004.
- [5] Padmini Srinivasan. Text mining: generating hypotheses from medline. *J. Am. Soc. Inf. Sci. Technol.*, Vol. 55, No. 5, pp. 396–413, 2004.
- [6] Don R. Swanson. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine*, Vol. 30, No. 1, pp. 7–18, 1986.
- [7] Marc Weeber, Henry Klein, Lolkje T. W. de Jong-van den Berg, and Rein Vos. 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.