イノベーション支援に向けた知識獲得と仮説生成

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専門研究員　土田　正明

PROFILE
2005 年に日本電気株式会社に入社。
2009 年より情報通信研究機構に出向し現職。自然言語処理の研究に従事。

m-tsuchida@nict.go.jp　TEL

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専攻研究員　デ・サーガ　ステイン

PROFILE
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 年より現職。自然言語処理の研究に従事

stijn@nict.go.jp　TEL

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グループリーダー　鳥澤　健太郎

PROFILE
1995 年東京大学大学院助手、北陸先端科学技術大学院大学准教授を経て、2008 年より現職。
自然言語処理の研究に従事。

torisawa@nict.go.jp　TEL

1 はじめに

本稿では、イノベーション支援のために、インターネットから大量の知識を自動獲得し、さらに、それら知識から有望な仮説を自動生成する試みについて述べる。これは、著者の一人が昨年の Japio 年誌で述べた構想 [9] を具現化する試みで、通常多大な労力が必要な「有望そうな仮説を立てる作業」のシステマティックな支援と言える。イノベーションには、少なくとも教師や文献から知識を学び、それらの知識を元にして新しい仮説を立てるというプロセスが必要であろう。我々の方法もこれに似ている。

我々の方法では、まず、インターネット上の大量のウェブページからそこに書かれている知識を獲得し、次に、それら知識を元にアナロジーと呼ばれる推論法を機械上で行うことで仮説を生成する。アナロジーとは「似ている点をもとに他のことを推し量ること」である。

具体例として、実際に生成された仮説である「アスピリンは胃がんを予防する」を紹介しよう。本稿では、この仮説のように 2 つの単語（アスピリン、胃がん）の意味的な関係（防ぐ）を知識として扱う。この仮説は「アスピリンが大腸がんを防ぐ」という知識と「大腸がん」と「胃がん」の意味的な類似性からのアナロジーで生成された。詳細は後述するが、元となる知識と単語同士の類似性は、2007 年頃に収集された約 1 億のウェブページから自動獲得されたものである。検索エンジンでこの仮説を調べたところ、2009 年まではアスピリンを含む多くの非ステロイド抗炎症薬の使用と大腸がんとの負の相関が報告されていたものの、胃がんとの関係を検討した研究は少なかった [3] が、2009 年にアスピリンの胃がんへの有効性が報告されたようである [1][3]。これは、2007 年頃のインターネット上の知識から、未来（2009 年）に検証されるような専門家が立てる「有望な仮説」を自動生成できた例と言えるであろう。以降では、このような仮説を自動生成する方法を説明していく。

2 インターネットからの知識獲得

我々は、２つの単語が同一文中で言及される際のパタン（言語パタン）を用いて、インターネット上の大量のウェブページから知識を獲得するシステムを開発している [2][6]。インターネット上には膨大な量の知識が書かれているため、それらを網羅的に獲得し活用できるようにする価値は高い。例えば、アトピーに悩んでいる人が、その原因を断ち切りたい場合、個人の知識でいくつの原因を列挙できるであろうか？本システムで獲得した知識から検索すると、ダニ、ホルムアルデヒド、ハウスダスト、ヒスタミン、界面活性剤、リポキシゲナーゼなど多くのアトピーの原因が瞬時に分かる。また、知的なウェブサービスやソフトウェアの開発では、このような知識を低コストで収集することが肝となる。この点に関して、実際に本システムで獲得した食材とその効能（例：にんにくが冷え性に効く）などの知識を用いたレシピ検索システム [10] が開発されている。

本システムでは、獲得したい意味的関係を表す言語パタンを手がかりに、様々な知識を獲得できる。例えば、因果関係を獲得する場合には「X が Y の原因になる」「Y の原因となる X」といった言語パタンを用いてウェブページ中から X（原因）と Y（結果）の単語ペアを獲得する。実際、ウェブページ中には「カビがアトピーの原因になる」「アトピーの原因となるホルムアルデヒド」などと書かれているため、先の例のような知識を獲得できる。ただし、特定の意味的関係に絞ったとしても、その知識は様々な言語パタンで書かれている。そのため、大量の知識を獲得するには大量の言語パタンが必要で、それらを用意する作業は非常に高コストとなる。

我々は人手コストを最小限にするため、少数の言語パタン（以降シードパタンと呼ぶ）を入力するだけで稼働するようシステムを設計、開発している。その鍵は、シードパタンと同じ意味的関係を表す、一種の言い換えとなる言語パタンを自動学習する機能にある。言い換えパタンの学習では、同じ単語ペアを獲得できるパタン同士が良い言い換えであるという考えに基づいている。例えば、シードパタンとして「X が Y の原因になる」「Y の原因である X」を入力すると、これらと同じ単語ペアを獲得しやすい「X によって起こる Y」「X で Y が発生」「Y を招く X」など、多くの人がすぐには思いつかないであろう言語パタンも含め、大量の言い換えパタンを学習してくれる。最終的には、学習された全パタンを用いて大量の知識を獲得する。

さらに、本システムは、単語の意味的なカテゴリの情報を用いて、曖昧な言語パタンをうまく活用できるよう工夫している。曖昧な言語パタンとは、複数の異なる意味的関係を表せるものである。例えば「X による Y」という言語パタンは「ノロウイルスによる食中毒」の場合は因果関係、「A 社による製品 B」の場合は会社と製品の関係など、様々な関係を表すことができる。このような曖昧な言語パタンは文書中で頻繁に使われるため、うまく活用することで大量の知識を獲得できるようになる。

曖昧な言語パタンは、X、Y に当てはまる単語の意味に制限を付けることで、その曖昧性を解消できる。単語の意味カテゴリを [意味カテゴリ名] と書くことにすると、例えば、「X による Y」という言語パタンは、「[生物] による [症状]」ならば因果関係、「[組織] による [製品]」ならば会社と製品の関係となる。このように単語の意味カテゴリのペア毎に異なる言語パタンと考えることで曖昧性を解消できる。例えば「X が Y の原因になる」など因果関係を表す言語パタンの言い換えとしては、「[生物] による [症状]」など因果関係を表す意味カテゴリのペアを持つ言語パタンが学習されるようになる。単語の意味カテゴリは、我々の開発した方法 [5] で自動獲得できる [1]。詳細に興味のある読者は文献 [2][5][6] を参照されたい。

評価実験として、約 6 億のウェブページから因果関

1　実際に [5] で獲得される意味カテゴリは、[生物] や [症状] など人間に理解しやすいラベルが付くわけではなく、意味的に似た単語に同じ識別子（番号）が付いたものである。
2　３人の被験者が、評価対象の知識の２語を含むテキストを読み、２人以上が正しいと判定すれば正解とする。

係（X が Y の原因）、予防関係（X が Y を防ぐ）を獲得した結果を紹介する。それぞれシードパタンとして約 20 個を与えた。獲得された因果関係、つまり原因 (X) と結果 (Y) をそれぞれ表す単語ペアと、予防関係、つまり予防策 (X) とその対象 (Y) を表す単語ペアのランダムサンプルを評価[2]した。結果、約 3 万個の知識を、因果関係では約 80%、予防関係では約 50%の精度で獲得できていた。表 1 は実際に獲得された知識の例であるが「造血幹細胞は動脈硬化の原因である」、「ヤーコンが生活習慣病を防ぐ」など様々な知識が獲得できていることが分かる。

	関係知識(X·Y)
因果関係 [XがYの原因]	造血幹細胞・動脈硬化
	ヘリコバクター・十二指腸潰瘍
	石油ファンヒーター・一酸化炭素中毒
	有機溶剤・シックハウス
予防関係 [XがYを防ぐ]	葉酸・口内炎
	ヤーコン・生活習慣病
	中国パセリ・便秘
	シンクライアント・情報漏えい

表 1　実際に獲得された知識の例

3 アナロジーによる仮説生成

前節の方法は、インターネット上に書かれている知識を獲得していることから「機械が文献に基づき人類の知識を勉強する方法」と言える。一方、本節では、大量のウェブページにも書かれていない、すなわち、人類にとって未知の知識をも含む仮説を自動的生成する試み [8] を紹介する。この仮説生成がある程度の精度で実現できれば、人間は有望そうな仮説を立てる必要がなく、機械が生成した仮説の選別・検証が作業の中心となるため、イノベーションの加速が期待できる。

我々のアプローチは、人間がしばしば発明の際などに用いるアナロジーと呼ばれる推論法を機械上で行うことである。本稿の冒頭でも述べたが、アナロジーとは「似ている点をもとに他のことを推し量ること」で、例えば文献 [9] で紹介されている「銀イオンがカビを防ぐので、銀イオンと似ている銅イオンもカビを防ぐかもしれない」といった推論過程である。アナロジーは、既知の知識と「似ている」という基準から推論できるため、機械にもそれらを与えることで実行できる。既知の知識は前節の方法で与えることができるが、「似ている」という基準はどうすればよいであろうか？

我々は、2 つの知識の対応する単語同士、例えば、因果関係ならば原因の単語、もしくは結果の単語が意味的に似ている場合、2 つの知識が似ていると考え、元となる既知の知識のどちらかの単語をその類似語で置き換えたものを仮説として生成する。この処理では、ウェブページ中で仮説の 2 語、例えば因果関係ならば原因と結果の単語がどのように書かれているかは関係ないため、どこにも書かれていない知識も仮説として生成される。これが前節のような言語パタンに基づく方法との大きな違いである。単語の類似性には分布仮説 [4] と呼ばれる「大量の似た文脈で出現する語は意味的に似ている」という考え方を採用する。例えば、「銀イオン」と「銅イオン」は、「X が含まれる」「X を放出する」「X の殺菌効果」「X の除菌効果」など、多くの同じ文脈で出現するため似ていることになる。分布仮説に基づく類似語の言語データは、我々の開発した大規模な類似語リストの作成法 [7] によって獲得できる。

このように生成された仮説は玉石混淆であるため、より確からしい仮説を優先的に見たいという要求があるだろう。そこで「多くの既知の知識と似ている仮説がより確からしい」、「元の知識とより似ている仮説ほど確からしい」という 2 つの仮定に基づき、各仮説の確からしさのスコアを計算する。例えば、先の「銅イオンがカビを防ぐ」という仮説が、さらに「安息香酸がカビを防ぐ」という知識と「安息香酸」と「銅イオン」の類似性からも生成された場合、そのスコアが増す。また、「安息香酸」と「銅イオン」の類似性が高いほど、スコアの増す度合いが大きくなる。これらの仮定は、直感的・経験的ではあるが、後で述べるように実験で有効性が確認でき

る。詳細に興味のある読者は文献 [7][8] を参照されたい。

評価実験として、2007 年頃に収集された約 1 億の日本語ウェブページを用いた因果関係、予防関係の仮説生成の結果を紹介する。アナロジーの元となる知識には、前節の言語パタンによる方法で、それぞれ 1 万個を獲得したものを用いた[3]。また、類似語リストも 1 億ページから生成した。生成された全仮説から、既知の知識と同様のものを削除し、全てを新しい仮説とした上でランダムサンプルを評価した。具体的には、各仮説の 2 語が含まれるウェブページを Yahoo! 検索で取得し、それらを 3 人の評価者が確認し、2 人以上が正しいと判定した場合に正解とした。この評価法により、実験で用いた 1 億ページに書かれていない知識、すなわち 1 億ページからは知り得ない知識も検証可能となる。これは、Yahoo! 検索からアクセスできるウェブページが、実験に用いた 1 億ページよりも多く、新しい文書も含まれているためである。ただし、如何に膨大なウェブページを検証に用いても、全ての知識が書かれているわけではないので、実際には正しい場合でも、誤っていると判断されてしまう場合もある。また、インターネット上には間違った情報も含まれていると考えられるため、正しいと判断されたものでも、実際には間違っている知識もあるだろう。そのため、本評価法は「インターネット上で言われていること」をどの程度推論できているかを評価したものとなる。

実験の結果、上位 2 万個の仮説の精度は因果関係で 65%、予防関係で 43%であった。仮説全体はそれぞれ約 17 万個で、その全体の精度はそれぞれ 35%、21%であった。比較すると、上位の精度が高いため、スコアが有効に働いていることが分かる。

それでは、生成された仮説のうち、1 億ページからは知り得ない知識はどの程度あったのか？これを厳密に調べるには、1 億ページを読む必要があるため非常に難しいが、我々は、1 億ページ中で仮説の 2 語が近い範囲に同時に現れているか否かを通して調べた。これは、2 語が近い範囲に現れていない場合は、その 2 語の関係について何も書かれていないだろうという考えに基づいている。なぜなら、2 語の関係について何か書こうとすれば、通常は、自然と 1 つの文やせいぜい少数の文で書かれると考えられるためである。具体的には、1 億ページ内で、1 文内に同時に現れていない（1 文内言及なし）、近接 4 文内でも同時に現れていない（4 文内言及なし）の 2 段階で調査した。「1 文内言及なし」は、2 語の言語パタンが存在しないため、前節のような言語パタンによる方法では実質的に獲得できない知識と言える。「4 文内言及なし」は、1 億ページ中にその 2 語の関係が何も書かれていないと知識と考えられる。評価したランダムサンプルの結果から、上位 2 万個中で正解と判定された仮説中の「1 文内言及なし」の数を推定すると、因果関係で約 1700 個、予防関係で約 1000 個であった。同様に「4 文内言及なし」の仮説数は、因果関係で約 250 個、予防関係で約 250 と推定された。このように、本手法によって前節のような言語パタンによる方法では実質的に獲得できない知識、さらには、1 億ページ中にはおそらく書かれていないであろう知識をも仮説として獲得できたと考えられる。これは、1 億ページ中の情報のみから、そこに書かれていない知識を推論できたと言っても良いであろう。表 2 に実際に生成された仮説の例を示す。

最後に、筆者らが見つけた興味深い仮説を 2 つ紹介しよう。まずは「ネトルがアトピーを防ぐ」である。「ネトル」はハーブの一種である。図 1 は獲得した知識と生成された仮説から「ネトルが防ぐもの」を可視化した例で、濃い色の四角（花粉症、アレルギーなど）は、前節の言語パタンによる方法で獲得された「ネトルが防ぐもの」で、薄い色の四角（アトピー、水虫など）はアナロジーで生成された仮説、すなわち「ネトルが防ぐと推測したもの」である。単語同士の角度が近いほど、それらが意味的に似ていることを表している。図 1 の「アトピー」を見ると、「ネトル」が「花粉症」や「アレル

3　実際には、明らかに誤っている知識を文献 [2][6] で紹介されている簡易的なクリーニング法で除去した上で用いている。

	生成された仮説	元となった知識
因果関係	活性酸素・喘息	活性酸素・アレルギー アレルゲン・喘息
	ホルムアルデヒド・扁桃痛 （1文内言及なし）	ホルムアルデヒド・めまい ホルムアルデヒド・頭痛
	可塑剤・肌トラブル （4文内言及なし）	界面活性剤・肌トラブル 合成界面活性剤・肌トラブル
予防関係	銅イオン・カビ	銀イオン・カビ 安息香酸・カビ
	抗酸化作用・妊娠線 （1文内言及なし）	抗酸化作用・シワ 抗酸化作用・そばかす
	どくだみエキス・カミソリ負け（4文内言及なし）	グリチルリチン酸ジカリウム・カミソリ負け

表２　生成された仮説とその元の知識の例

ギー」を防ぐという知識と「アトピー」の「花粉症」や「アレルギー」との類似性から「ネトルがアトピーを防ぐ」と推測されたことが分かる。実際に検索エンジンで「ネトル」について調べると、「ネトル」には肥満細胞が蓄積できるヒスタミンの量を増やすことで、体内に放出されるヒスタミンの量を減らす効果があると言われているようである。ヒスタミンは「花粉症」や「アトピー」の原因の一つとされている。これらを考えると「ネトル」で「アトピー」を防ぐことができると考えられる。実際のところ、少し調べた範囲では、その科学的検証は見つからなかったが、これは、実験で検証する価値があ

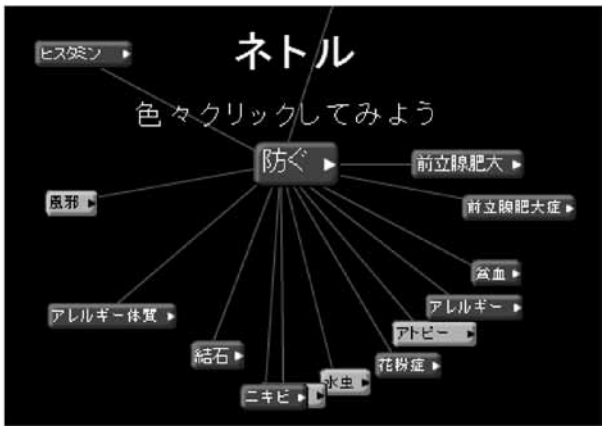

図１　ネトルが防ぐものの可視化

りそうに思えないだろうか？

次は「大豆サポニンが二日酔いを防ぐ」である。この仮説は「クルクミンが二日酔いを防ぐ」から生成されていた。「クルクミン」はウコンに含まれる黄色色素で、抗酸化作用を持ち、肝臓機能を強化することで、二日酔いの原因となるアセトアルデヒドを分解する助けとなるようである。「大豆サポニン」は、科学的な検証は見つけることができなかったものの、抗酸化作用によって肝臓に悪影響のある過酸化脂質の発生を抑え、肝機能を強化すると言われているようである。また、大豆サポニンにはブドウ糖が中性脂肪に変化することを抑える働きもあり、肥満などの生活習慣病防ぐともいわれている。「大豆サポニン」による二日酔い防止の商品は少し調べた限りでは見つからなかった。実際に検討すれば「二日酔い」と「肥満」という多くの人が宴会中に気にしているであろう２つの問題を予防してくれるヒット商品が誕生するかもしれない。

このように、気になった仮説を少し検索エンジンで調べるだけでも新しい発見ができることがわかる。これは「有望な仮説」を見つける支援と考えて良いだろう。

4 おわりに

本稿では、イノベーション支援のために、インターネット上の大量のウェブページから知識を獲得し、さらに、それら知識から有望な仮説を自動生成する試みを紹介した。日本が強い国際競争力を持ち続けるには、強い知的財産を迅速に生み出し続けることが重要であるため、このような研究テーマは重要と考えている。また、現在はインターネットという巨大で雑多な情報源を活用しているが、より専門的な知識の獲得のためにも、特許や論文など専門的な文書の活用も進めていきたいと考えている。

なお、本稿のアナロジーで用いている類似語データベースは ALAGIN フォーラムで公開されている。また、言語パタンに基づく関係獲得技術もサービスとし

て ALAGIN フォーラムで公開予定である。興味のある読者は ALAGIN フォーラムのウェブサイト (http://www.alagin.jp) を是非ご覧いただきたい。

[参考文献]

[1] Abnet, C.C., Freedman, N. D., Kamangar, F., Leitzmann, M.F., Hollenbeck, A. R., Schatzkin,A.: Non-steroidal anti-inflammatory drugs and risk of gastric and oesophageal adenocarcinomas: results from a cohort study and a meta-analysis, British Journal of Cancer, Vol. 100, Issue 3, pp. 551-557, 2009.

[2] De Saeger, S., Torisawa, K., Kazama, J., Kuroda, K. and Murata, M.: Large Scale Relation Acquisition using Class Dependent Patterns, In proc. of 9th ICDM, pp.764-769, 2009.

[3] Epplein, M., Nomura, A. M. Y., Wiklens, L. R., Henderson, B. E. and Kolonel, L. N.: Nonsteroidal Antiinflammatory Drugs and Risk of Gastric Adenocarcinoma, American Journal of Epidemiology, Vol.170, Num.4, pp.507-514, 2009

[4] Harris, Z.: Distributional structure, Word, Vol. 10, Number 23, pp. 146-162, 1954.

[5] Kazama, J. and Torisawa, K.: Inducing Gazetteers for Named Entity Recognition by Large-scale Clustering of Dependency Relations, In Proc. of 46th ACL, pp. 407-415, 2008.

[6] Stijn De Saeger, 鳥澤健太郎 , 風間淳一 , 黒田航 , 村田真樹 : 単語の意味クラスを用いたパターン学習による大規模な意味的関係獲得 , 言語処理学会第 16 回年次大会 , pp. 932-935, 2010.

[7] 風間淳一 , Stijn De Saeger, 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.

[8] 土田正明 , Stijn De Saeger, 鳥澤健太郎 , 村田真樹 , 風間淳一 , 黒田航 , 大和田勇人 : 類推による単語間の意味的関係獲得法 , 言語処理学会第 16 回年次大会 , pp. 936-939, 2010.

[9] 鳥澤健太郎 , 風間淳一 , 村田真樹 : 概念辞書によるシステマティックなイノベーション支援に向けて , Japio 2009 YEARBOOK, pp. 228-235, 2009.

[10] 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システムにおける評価 , 言語処理学会第 16 回年次大会 , pp. 138-141, 2010.