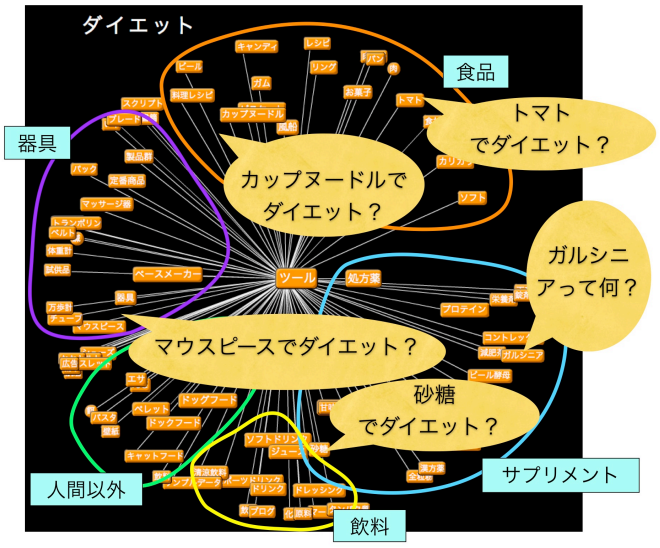
ウェブ検索ディレクトリの自動構築とその改良 –鳥式改–

鳥澤健太郎*, 隅田飛鳥‡†, 野口大輔‡, 柿澤康範‡, 風間淳一*, Stijn De Saeger*,
村田真樹*, 黒田航*, 山田一郎*, 塚脇幸代*, 太田公子*
*情報通信研究機構 MASTAR プロジェクト 言語基盤グループ
‡北陸先端科学技術大学院大学　　†奈良先端大学院大学
E-mail: torisawa@nict.go.jp

1. はじめに

適切な行動をとるための情報収集に，検索エンジンを利用するのはもはや常識である．つまりは，様々なトピックに関する問題回避，あるいは行動に関する未知のアイディア，Tips について情報を求めるため，検索エンジンを利用するということである．ところが，そうした情報を得るにはユーザから見て「意外」なキーワードを入力する必要がしばしばある．例えば，執筆者の一人が常宿としていたホテルがいわゆる建築偽装疑惑に関係した会社によって建設されていたという事実は，その執筆者にとっては全く未知であった．予約のためサーチを行なっても検索結果の上位にはそうした情報はなく，実際そのホテルに何度も宿泊した．ところが，実は通常の検索エンジンでも，ホテル名に加えて「落とし穴」という意外なキーワードを与えると，検索結果のトップに問題の事実が見つかる．重要な点は，こうした意外なキーワードはユーザの「意識に昇っていない」以上，システム側から提示する必要があることである．

我々はこうしたキーワードの想起を支援するため，「鳥式改」[1]という検索ディレクトリを開発している．これは，ユーザが最初に入力したキーワード，つまり，トピックに対して，関連語を意外なものまで含めて提示し，検索に利用できるようにする．なお，鳥式改の第一の特長は鳥式改が Web 文書に自然言語処理技術を適用することで自動生成されており，現在 180 万語という大量のトピックをカバーしていることである．第二の特長は価値ある情報を効率良く検索できるようにするため，いくつかの意味的カテゴリに属する関連語のみを提示することである．ホテルの「落とし穴」は「トラブル」というカテゴリ中の関連語として提示される[2]．現時点では，トピックを利用する行為（例：ホテルならば「宿泊」）あるいはトピックに対処する行為（例：トピック「花粉症」に対して「治療」）に関する情報収集が検索ニーズの一定部分を占めていると仮定し，それらの行為を行う上で有用なカテゴリが設定されている．具体的には，利用／対処の行為自体，利用／対処を行うための「準備の行為」（例： ホテルの「予約」），利用／対処／準備といった行為を阻害する要因としての「トラブル」，それら行為を行う際の Tips としての具体的「方法」，有用な「ツール／材料」などがある．図１は，トピック「ダイエット」の対処に利用できるツール／材料を，「トマト」「砂糖」「風船」のような意外なものも含め提示した例であるが，意味的に類似した関連語がまとまって表示され，欲しい関連語を探すのを容易にしている．



図–1：ダイエットのツールを鳥式改で提示した例

また，鳥式改の第三の特長は，広範な関連語を提示するため，トピックの上位概念の名称（例：トピック「東京大学」に対する「大学」）を自動的に獲得し，大量に保持していることである[3]．今年はじめに話題になった農薬ぎょうざ事件を例にとると，これまで開発した手法では，昨年，つまり，事件以前の Web 文書から，トピック「ぎょうざ」の関連語として「農薬」を直接認識することはできなかった．しかしながら，「ぎょうざ」の上位概念になる可能性のあるものに「冷凍

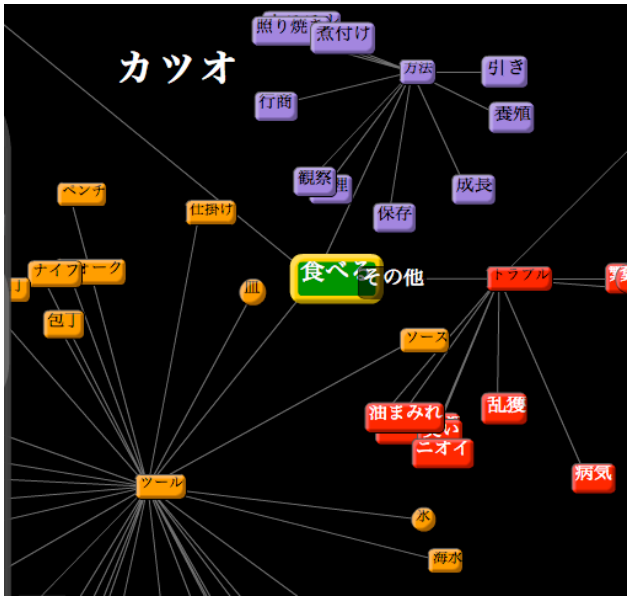
食品」があり，「冷凍食品」のトラブルとして「残留農薬」が認識できていることから，「残留農薬」を「ぎょうざ」のトラブルとして提示できる．つまり，騒ぎになる以前にぎょうざ事件をあたかも「予測」していたことになる．実際にぎょうざに付着していたものが「残留」農薬なのか意図的なものであるのかは今もって不明であるが，問題のぎょうざに関わった人々に「残留農薬」の可能性が事件の早い段階で示唆されていたとすれば，状況は改善されたかもしれない．鳥式改はトピックに対して関連語を提示するという一見単純な処理しか行なわないが，このぎょうざの例などは，そのような単純な処理ではあっても実社会でインパクトを持ち得ることを示唆しているものと考えている．

また、表１には、上述した以外の「意外でありながら有用なトラブル」の具体例をいくつか示す。

トピック	トラブル	説明
リンゴ	花粉症	花粉症の患者がリンゴを食べるとかゆみが発生する
洗濯機	アトピー性皮膚炎	古い洗濯機に発生するカビが洗濯された衣類に付着し、アトピー性皮膚炎を悪化させることがある
無洗米	鮮度低下	無洗米は普通米に比較して鮮度低下が早い
〈某ゲーム機〉	傷	ゲームのメディアに傷がつく事例が頻発していると主張されている
アガリクス	発がん性	がんに効くと一部で宣伝されているアガリクスには発がん性がある
〈某電機メーカー〉	パープルフリンジ	デジカメ等の画像劣化の一種であるパープルフリンジが特定メーカーの製品で頻発するという主張がなされている

表−1：意外でありながら有用なトラブル情報

また、鳥式改は単に関連語を提示するだけではなくて、関連語相互の関係性も提示することができる。図２は、「カツオ」の具体的な利用の行為として「食べる」を選択肢、「食べる」文脈における、方法、トラブル、ツール等を表示させた例である。現状まだ精度は低いが、このような操作により、関連語を特定の文脈内のものに限定することで有用な情報を探し出すのがより容易になっている。こうした操作をより柔軟におこなえるようにするのは、今後の拡張において重要な方向性の一つである。



図−2：「カツオ」の「食べる」文脈での関連語

2. より多様な意味的知識

以上が鳥式改の基本的な機能であるが、今後は更に人間の行う多様な推論で使われる自然言語表現間の意味的関係をシステムに導入し、人間の行う推論を補助できる一種の「考えるツール」へと拡張したいと考えている。また、現状すでに、上述してこなかった意味的関係を鳥式に導入しているが、以下ではそうしたものについて解説を行いたい。

以下では、既に鳥式に導入されている以下の意味的関係、具体的には類似の関係と因果関係、とその利用法について解説したい。

2.1. 類似表現の関係とアナロジー

まず、類似表現の関係について述べたいが、これを鳥式改で具体的に利用した状況を示したのが図３である。これは、図１にあるダイエットのツールの内、「コントレックス」（ミネラルウォーターの一種）の類似表現を提示した状況を示す。類似表現は、[4]にある単語クラスタリングを適用後、表現間の JS-divergence を計算することで得ている。

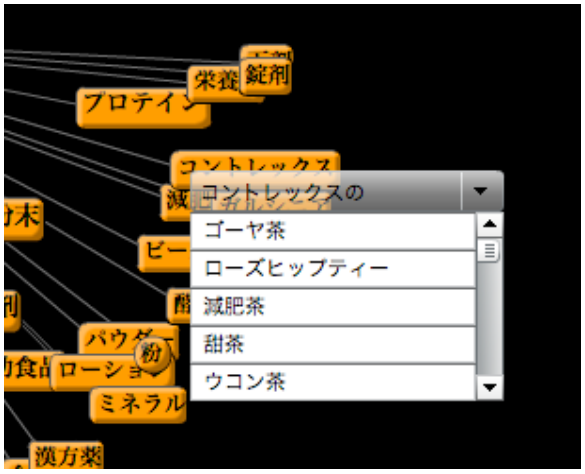


図-3：ダイエットのツールとしての「コントレックス」の類義語を提示した例

この図では、コントレックスの類義表現として「ゴーヤ茶」「ローズヒップティー」「減肥茶」等の（健康）飲料が提示されている。この内、ローズヒップティーに関しては、これを選択すると、トピック語の「ダイエット」と「ローズヒップティー」の AND 検索が商用検索エンジンで行えるようになっており、その結果、ローズヒップティーを用いたダイエット法が Web 上には書かれていることが判明する。これは、現在の鳥式改の構築手法では認識できなかったダイエットのツールが、類似表現の提示によって、新たに発見できたことを示しているが、より大きなくくりで言えば、鳥式改に格納されている類似表現を用いて、アナロジーを行ったことになる。

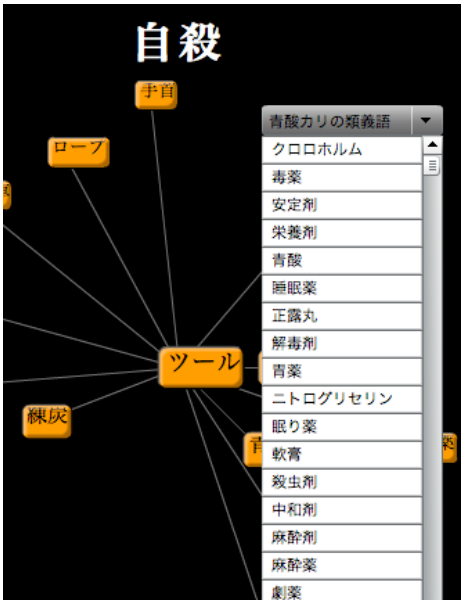


図-4：青酸カリの類似表現を鳥式改で提示した例

従来であれば、こうしたアナロジーはユーザの頭の中でやるより他になく、こうした高度な操作、推論が、システム上である程度システマティックに行えるようになったことの意義は大きい。また、こうした操作は鳥式改のデータがすでに一個人の持っている知識を越えているという事実があり、その有効性が増しているという点も指摘しておきたい。

一方で、「コントレックス」と「ローズヒップティー」の間の類似表現関係自体は比較的に自明なものである。一方で、類似表現関係自体が自明でないケースがある。図 4 は「自殺」のツールとしての「青酸カリ」を表示させ、さらにその類似表現を提示した状況を示す。この類似表現の中には薬品の「正露丸」が含まれている。類似表現自体は、トピック語である「自殺」とは無関係に計算されているため、それに起因する不適切な例かと思われたが、実際に検索を行ってみると、実は Web 上では自殺志願者とおぼしき人物が「正露丸」での自殺を検討していることが判明する。この事例などは、Web 上のデータをもとに計算された意外な類似表現が意外な発見に結びつくことを示唆している。また、鳥式改は Web 上の情報をもとに自動構築されているわけだが、その結果として、アンダーグラウンドな情報をも提示できることを示している。昨今の有害情報を巡る議論にも大きく関係して、興味深いケースであると考えている。

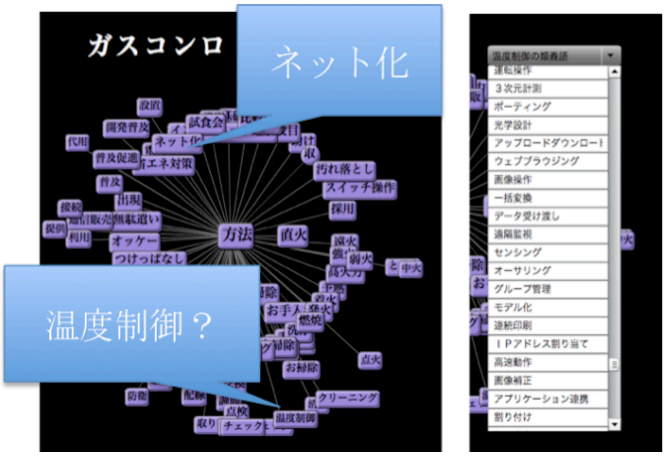


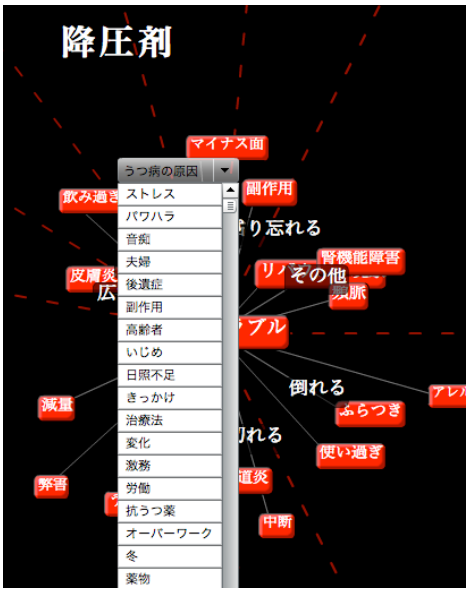
図-5：ガスコンロ利用時の具体的方法とその類似表現

また、以上のケースは、Web 上に存在する情報を鳥式改で発見したケースである。（正露丸の自殺のケースは実際にそうした自殺があったかどうかは明確には分からないが。）しかしながら、Web 上に存在しない情報を鳥式改上の膨大な知識を使って発見、あるいは産

出する場合もあり得ると考えている。例えば、図５はガスコンロに関する具体的方法を提示したものである。関連語として、「ネット化」「温度調節」の二つが出現している。ここで、ネット化はあり得る話としても、何のためのネット化ということかは明確ではない。(現在は Web 上で検索をおこなってもそれらしい情報は存在しないようである。) 一方で、「温度調整」の類義語を提示すると、「ダウンロード」「IP アドレス割当」「プログラム実行」などが表示される。例えば、こうしたキーワードを繋げていけば、「ネットからダウンロードしたプログラムの実行で温度調整をするガスコンロ」などといったものも考えることができよう。これを敷衍すれば、レシピサイトが、各料理を作る際の理想の火加減をプログラム、データの形で格納し、それにしたがって火加減を調整する「ネット化ガスコンロ」といった商品を考えることも可能となろう。こうしたものは著者がネットで調べたかぎりでは見つからず、実現の可能性はともかく、一つのアイディアとしてはありではないかと考えている。

2.2. 因果関係

鳥式では関連語に関係して、自動獲得した因果関係を表示することも可能である。例えば、図６では、「降圧剤」を利用する際のトラブルとして提示された「うつ病」の原因を提示した例を示している。こうした意味的関係は例えば、関連語の間の意味的関係を提示したり、さらには情報の深堀りを促す上で重要であり、今後さらにバリエーションを増やしていく予定である。



図−6:「うつ病」の原因の提示

また、このうつ病の例は、名詞句と名詞句の間の因果関係であるが、一方で名詞句と動詞句の間の因果関係も鳥式改では提示できる。図７では、「あじさい」のトラブルとしての「中毒」が「死亡する」という出来事を引き起こすという因果関係を示しているが、これはトラブルの深刻度を分類する上で非常に重要な手がかりになると考えている。

また、現状はトピック語の文脈に限定して原因等を提示することはできていないが、そうした文脈を適切に取り扱うことが非常に重要であり、将来的にはそうした処理を実現したいと考えている。

図−7:「中毒」が引き起こす結果としての「死亡」

3. 今後の展開

今後は前述したように、より多様な意味的関係を Web より抽出して、鳥式改を巨大な意味ネットワークとし、対話システムなどのバックエンドとして活用して、ユーザのおかれている文脈を考慮して適切なアドバイスをするといったシステムを開発することを考えている。これは膨大なネット上の情報をもとにシステムに「考えさせる」つまり、「ネットに考えさせる」ということにつながるものと考えており、一個人の持ち得る知識を一面で凌駕している鳥式改上の知識、意味的関係を新規なやり方で組み合わせることでかつてない情報、新規なアイディアが産出できるなど、今後、ある意味で現在の我々の想像の及ばないような展開があり得るものと期待している

【 参考文献 】

[1] Torishiki-kai, an Autogenerated Web Search Directory, Kentaro Torisawa, et al., in Proc. of ISUC 2008, 2008.

[2] S. De Saeger, K. Torisawa and J. Kazama, Looking for Trouble, In Proc. of Coling2008, pp. 185-192, 2008.

[3] 黒田, 李, 野澤, 村田, 鳥澤, 鳥式改の上位語データの人手クリーニング, 言語処理学会第第 15 回年次大会発表論文集, C1-3(2009)

[4] J. Kazama and K. Torisawa, “Inducing Gazetteers for Named Entity Recognition by Large-scale Clustering of Dependency Relations,” In *Proceedings of ACL-08: HLT*, pp.407-415(2008)