

# 情報信憑性判断支援のための対話型条件結論マップ 生成に向けた条件と結論の抽出

松本 拓也† 渋谷 英潔‡ 森 辰則‡

† 横浜国立大学 大学院 環境情報学府 ‡ 横浜国立大学 大学院 環境情報研究院  
E-mail: {m\_takuya,shib,mori}@forest.eis.ynu.ac.jp

## 1 はじめに

Web上に存在する情報は、ブロードバンド化の進展やブログ等の普及に伴い、爆発的に増加し続けている。これらの情報の中には、出所が不確かな情報や利用者に不利益をもたらす情報などが含まれており、信頼できる情報を利用者が容易に得るための技術に対する要望が高まっている。我々は、個々の情報利用者が情報の信憑性の最終的な判断を行うものとして、利用者による信憑性の判断を支援する技術の実現を目指して研究を行っている。

利用者の信憑性判断を支援する技術には、幾つかの種類があるが、その一つに、利用者がある言明、例えば「コラーゲンは肌に良い」の信憑性を判断したいという要求の下で支援することを目的としたものがある[1, 2]。Murakami et al.[1]は、Web上に存在するさまざまなテキスト情報について、それらの間に暗に示されている同意、対立、弱い対立、根拠などの意味的関係を解析する言論マップの生成課題を論じている。中野ら[2]は、一見対立しているようにみえるが、実際はある条件や状況の下で互いの内容が両立できるような二言明に対して、両立できる状況を短い文章で提示する調停要約の生成手法を提案している。これらの技術を用いると、利用者が信憑性を判断したい言明(着目言明)が成立する条件に関するWeb上のテキスト、例えば「コラーゲンを肌に塗っても効果はないが、サプリなどで摂取すると肌に良い」といったテキストを利用者に提示することができる。利用者はこれを読んで、「コラーゲンは肌に良い」という結論に相当する効果を得るためには、「サプリなどで摂取する」という条件を満たす必要があると理解することができる。本稿では、利用者が影響を受ける事柄を表現する語句や節を結論と定義し、結論の効果を得るために必要となる事柄を表現する語句や節を条件と定義する。なお、着目言明も結論の一つである。

こういった着目言明の条件に関する情報は、利用者の信憑性判断を支援する上で有用であるが、我々はさらに支援できる余地があると考えた。利用者が着目言明の信憑性を判断したいと考えた背景に、着目言明が真実であった場合に自分も着目言明を実践しようという意図が存在する可能性は十分に考えられる。しかしながら、ある条件から導き出される結論には着目言明以外にも複数存在する可能性が十分あり、その場合には、条件を実践することで得られる結果は着目言明の内容だけではない。「コラーゲンをサプリなどで摂取す

る」という条件の場合、「肌に良い」というメリットとなる結論の他に、例えば、「ニキビができる」というデメリットとなる結論もWeb上には存在している。したがって、着目言明が示す効果を得るために、その条件を満たそうとしたときに生じる、(着目言明を含む)メリットがある結論群に加えて、デメリットをもたらす結論群を明示することにより、着目言明を達成するための条件を実践すべきか否かが判断できる支援技術の実現を我々は目指している。

本稿では、そのための第一段階として、目的とする情報信憑性判断支援のための対話型条件結論マップについて述べた後、どのような情報を優先的にマップに提示すべきかを考察する。また、着目言明とその条件が含まれるテキストが利用者に提示されている前提で、利用者が条件となるテキストを入力(例えば、マウス操作等でマーク)した場合に、入力された条件とマッチする条件と結論のテキストの組をWeb文書から網羅的に抽出する手法を提案する。抽出される条件と結論の組は多く存在するため、利用者に見せる組を絞るためのランキング処理が必要である。さらに、利用者が理解しやすいよう最終的にはマップ化して提示したいと考えているが、本稿ではそこまで行わず、前提となる着目言明やテキストを考慮してランキングを行う手法の議論に留める。

## 2 関連研究

利用者が入力した着目言明の信憑性判断を支援する研究には、先に述べた言論マップ[1]や調停要約[2]がある。これらの研究と本研究との違いは以下の通りである。言論マップや調停要約では、着目言明そのものの信憑性判断を支援することが目的である。そのため、「コラーゲンは肌に良い」という着目言明の場合、肌に良いかどうかには焦点が当てられ、例えば、「太る」や「金がかかる」といった結論は直接の出力対象とならない。一方、本研究では、利用者がある条件を実践する上で注意を喚起するような結論を幅広く提示することを目的としている。また、1節で述べたように、本研究は言論マップや調停要約などの結果を読んだ利用者が、さらに調査するための補足的な位置づけになっている。

以下に、本研究の要素技術に関連する研究について述べる。本研究では条件と結論の組を抽出するが、これは因果関係抽出[3, 4]との親和性が高い。乾ら[3]や坂地ら[4]は手がかり表現を利用して因果関係の組を

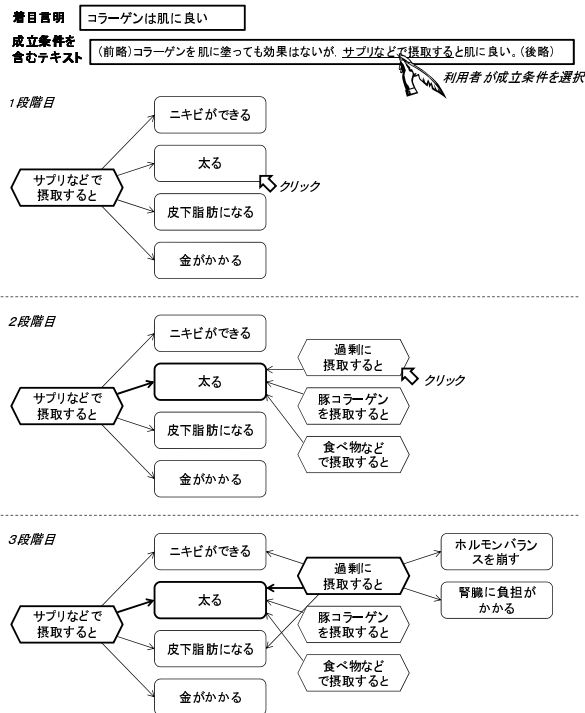


図 1: 情報信憑性判断支援のための対話型条件結論マップの例

抽出する手法を提案している。本研究では手がかり表現と係り受け構造を用いて根拠表現と結果表現の組を抽出する坂地らの手法をベースとして条件と結論の組の抽出を行う。しかしながら、坂地ら [4] は主に名詞句として表現される根拠表現を対象としているのに対し、4 節で述べるように、我々は「摂取すると」のような節を対象とするため、手がかり表現と結論となるテキストの範囲に修正を加えている。また、メリットやデメリットとなる結論を提示するという本研究の目的は、評判情報の極性判定 [5] との親和性が高い。乾ら [5] は、ブートストラップ的に拡張した評価表現辞書を用いて極性判定する手法を提案しており、本研究でも乾らの判定手法を用いてメリット・デメリットの極性判定を行う。しかしながら、本研究では、その結論が起こったときに実践者にメリットがあるのかデメリットがあるのかを判断することを目的とするため、評判情報における極性との関係は不明である。本稿では、その点に関する調査を行う。

### 3 情報信憑性判断支援のための対話型条件結論マップ

我々が目的とする情報信憑性判断支援のための対話型条件結論マップの例を図 1 に示す。前提として、利用者には着目言明とその条件を含むテキストが提示されている。利用者は条件に該当する文字列、図の例では「サプリメントで摂取する」の部分のマウス操作等により直接マークする。なお、利用者が条件に該当する文字列を厳密にマークできなくともロバストに処理できるよう、マークされた文字列を含む一文がシステムに入力されるものとする。

システムは、入力文中の条件から導かれる結論を可能な限り多く Web 文書から探して利用者に提示する (図 1 の 1 段階目)。図 1 の六角形は条件、角丸四角形は結論をそれぞれ示している。利用者は、これを見ることで、着目言明の「肌に良い」以外にも多くの結論が起こりうることを知ることができる。次に、利用者が、提示された結論から気になった結論 (図 1 の例では「太る」) を選択すると、システムは、その結論を導く他の条件を Web 文書から探し、もし存在すればそれを提示する (図 1 の 2 段階目)。利用者は、これを見ることで、気になった結論が条件から常に導かれるのか、それともある特定の状況でのみ導かれるのかを知ることができる。さらに、利用者が、提示された条件から気になった条件 (図 1 の例では「過剰に摂取すると」) があれば、システムは、その条件から導かれる他の結論を Web 文書から探して提示する (図 1 の 3 段階目)。このように、利用者が気になった条件や結論を選択することで、その条件や結論から、導かれる結論や導く条件を対話的に提示していく。

ここでシステムが提示すべき条件や結論の観点について考察する。第一に、本研究では、利用者の関心が着目言明として入力されているため、着目言明の内容と関連がある条件や結論でなくてはならない。第二に、利用者は既に着目言明の信憑性判断に関するテキストを読んでいる前提であるため、利用者がまだ目にしていない新情報を提示する必要がある。第三に、利用者は判断の手がかりを求めているため、書き手が書いている内容を断言している必要がある。第四に、利用者が実践する上で注意を必要とするような情報をより積極的に提示すべきである。例えば、着目言明「コラーゲンは肌に良い」における条件「サプリメントで摂取すると」には、「骨粗鬆症を予防する」や「毛髪を太くする」といったメリットのある結論も存在する。しかしながら、「肌に良い」というメリットに着目している利用者に対しては、「ニキビができる」や「太る」といったデメリットのある結論の方を積極的に提示した方が、実践するか否かを判断する上で有益であると考えられる。また、「体内で一旦アミノ酸に分解される」のようなメリットにもデメリットにもならない結論は、一般に注意を必要とする内容でないことが多い。したがって、提案手法では、メリットとデメリットによる分類を行った後、関連性があり、新情報を含み、断言している、条件と結論の組を上位にするようなランキングを行う。これにより、利用者はマップ化する組を可視化できる範囲に収めることができると考えられる。

### 4 提案手法

本稿で提案する抽出手法の全体の流れを図 2 に示す。システムは、入力された文に対して、坂地らの手法 [4] をベースとして条件部と結論部の組の抽出を行う。2 節で述べたように、我々は「摂取すると」のような節を条件として抽出を行うため、表 1 に示す 35 種類の手がかり表現を用いた。表 1 の手がかり表現は乾ら [6] や坂地ら [4] の調査した手がかり表現を基本にした。ただし、5 節の実験で用いた着目言明とは異なる着目言明に関する Web 文書から、第一著者が条件と結論の組を抽出するのに適切と判断した表現を加えている。文

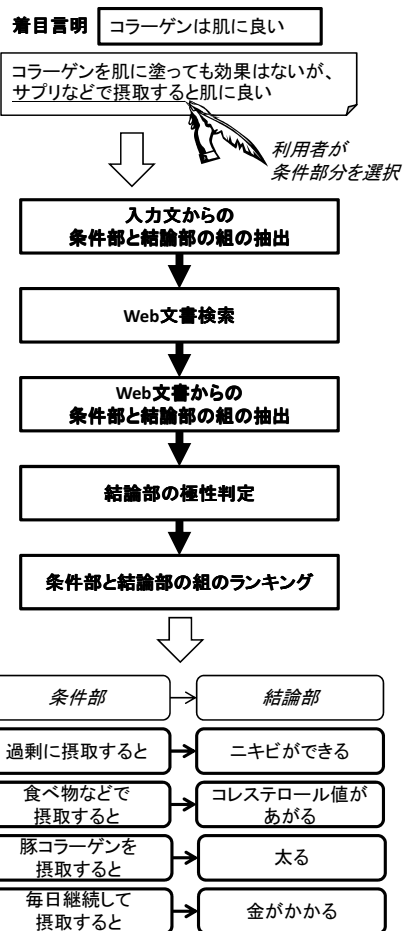


図 2: 全体の流れ

中の手がかり表現を含む文節を核文節として、核文節に係る記述を条件部、核文節の次文節から文末までを結論部として抽出する。もしも、複数の組が抽出された場合は、条件部が、マークされた文字列に最も近い位置に現れる組を採用した。

次に、抽出された条件部に含まれる助詞以外の単語列をクエリとして Web 検索を行う。検索された Web 文書集合から、入力文と同様の方法で、条件部と結論部の組を網羅的に抽出する。この抽出は網羅的に行うため、利用者の関心がある条件とは無関係の条件を持つ組も含める。そこで、抽出された組の集合を、利用者にとって有用な順序で提示する処理が必要である。提案手法では、3 節の議論にしたがって、この処理を「結論部の極性判定」と「条件部と結論部の組のランキング」の二段階で行うこととした。結論部の極性判定は、3 節で述べた第四の観点を扱うためのものである。乾らの手法 [5] を用いて、入力文と Web 文書から抽出された結論部を、ポジティブをメリット、ネガティブをデメリットと見なして処理を行う。この段階で、その他に分類された結論部をもつ組は提示すべき対象から除外される。しかしながら、評判情報抽出におけるポジティブ、ネガティブが、我々の目的とするメリット、デメリットとどのように対応するかは明らかでない。そのため、5 節において、ポジティブ、ネガティブをメリット、デメリットと見なすことに関する考察を行う。次にメリットまたはデメリットの結論部をも

表 1: 手がかり表現の一覧

直前の品詞	手がかり表現
動詞, 形容詞	と, ので, ば, たら, だら, て, で,  따라서,  だから,  たため,  だため,  た結果,  だ結果,  た場合,  だ場合,  たが,  だが,  ことで,  事,  ことにより,  事により,  こと,  ことによって,  事,  事によって,  ことが,  事が,  ことは,  事は
サ変名詞	(動詞, 形容詞の手がかり表現),  で,  により,  によって,  の結果,  の場合,  が,  は

つ組を対象に、3 節で述べた第一、第二、第三の観点に基づいて、条件部と結論部の組のランキングを行う。各組に対して以下の式 (1) に基づいてスコアを計算し、スコアの順にランキングを行う。

$$sc = r1 * r2 * rtp * rw1 * rw2 * n * e \quad (1)$$

$$r1 = \frac{match_{cd}}{N_{cdin}} \quad (2)$$

$$r2 = \frac{match_{cd}}{N_{cdex}} + 1 \quad (3)$$

$$n = 1 - \frac{match_{cc}}{N_{ccin}} \quad (4)$$

ここで、 $r1$ ,  $r2$ ,  $rtp$ ,  $rw1$ ,  $rw2$  は第一の観点のスコアであり、 $n$  は第二の観点のスコア、 $e$  は第三の観点のスコアである。また、 $N_{cdin}$  は入力された条件部の単語数、 $N_{cdex}$  は抽出した条件部の単語数、 $N_{ccin}$  は入力された結論部の単語数、 $match_{cd}$  は入力された条件部と抽出した条件部の共通する単語数、 $match_{cc}$  は入力された結論部と抽出した結論部の共通する単語数。なお、これらの単語数はすべて助詞を除いた単語数である。 $r1$  は入力された条件と比べ足りない語がある場合、 $r2$  は入力された条件と比べ余分な語がある場合にペナルティを与えるスコアであり、 $r1$  と比べ  $r2$  のほうがペナルティの比率が小さくなるように設定した。 $rtp$  はトピック語が存在する場合を 1、存在しない場合を 0 とした。トピック語とは着目言明と条件部に共通する単語である。 $rw1$  はトピック語を含む文節の係り先が手がかり表現の前の語句を含む文節の場合を優先する重みであり、 $rw2$  は手がかり表現の前の語句が入力された条件に含まれる場合を優先する重みである。これらを満たすことで入力された条件を正しく再現している可能性が高くなると考えた。それぞれ重みは 1.2 とした。 $n$  は入力された結論部と共通する単語の割合が高いほど、ペナルティを与えるスコアである。 $e$  は結論部の文末で判定を行い、「～した」や「～です」のように書き手が言い切っている文を断定しているとみなし値を 1.0、「～ようです」や「～かもしれません」のような曖昧な文は書き手が書いている内容に自信があまりないと考えられるため値を 0.7、「？」や終助詞の「か」で終わる文は書き手が書いている内容が真実であるかわからないと考えられるため値を 0 とし、それ以外の文は 0.5 とした。

表 2: 実験で用いた入力データ

着目言明	条件部	結論部
コラーゲンは肌が良い	コラーゲンを摂取すると	肌にうるおいを与え、しわやしみを防ぐ効果があります。
ヨーグルトは体に良い	ヨーグルトを毎日食べると	確実に健康効果がある。
キシリトールは虫歯の予防になる	キシリトールのガムを噛むと	唾液中の pH が上昇して歯の修復が促進されるため、虫歯予防に役立つのです。

表 3: 極性判定と正解データの対応

	システム出力		
	p	n	e
メ	281 (33%)	31 (4%)	50 (6%)
正解	26 (3%)	67 (8%)	23 (3%)
他	173 (20%)	57 (7%)	150 (17%)

また、ランキングを行った後、すでに存在している結論と似ている内容の順位を落とすため、ある組に対し、その組より上位の組の結論それぞれと、以下の式(5)を計算し、一番小さくなる値をスコアに乘算した。

$$sim = 1 - \frac{match_{cc}}{N_{cc}} \quad (5)$$

$match_{cc}$  は結論部同士の共通する単語数、 $N_{cc}$  はある組の結論部の単語数である。

## 5 実験

本稿では、極性判定とランキングの精度に関する実験を行う。実験に用いた着目言明、入力文の条件部、入力文の結論部を表 2 に示す。正解データは以下のようにして作成した。着目言明ごとに、網羅的に抽出した条件部と結論部の組からランダムで 300 組選択し、この 300 組に対して、「メリット」「デメリット」「その他」の 3 値の極性と、利用者に提示すべきか否かの 2 値の情報を付与した。この作業は第 1 節で述べた、着目言明を達成するための条件を実践すべきか否かを判断する上で、参考になる結論を得られるかという基準で第一著者が行った。

システムが判断した評判情報としての極性と正解データ(メリット, デメリット, その他)の対応の結果を表 3 に示す。システム全体として対応が正しく行われたのは 58% (33%+8%+17%) である。ポジティブをメリット、ネガティブをデメリットとして分類を行うことに大きな問題はないと考えている。しかし、その他に分類すべき結論がメリットやデメリットに分類されることが 27% (20%+7%) 存在した。これは「ヨーグルトと食べると」と「よく合います。」の組のように、抽出した組の結論部が本研究で結論と定義している利用者が影響を受ける事柄になっていない場合が多かった。そのため結論部が利用者が影響を受ける事柄であるのかという判断をする必要があると考えている。

次に、表 4 に着目言明ごとの、メリット、デメリットに分類された組の集合に対するランキングの R 精度を示す。R 精度とは、正解データ数が R 個の場合、結果

表 4: ランキングの R 精度

着目言明	メリット	デメリット
コラーゲンは肌が良い	0.47 (8/17)	0.33 (1/3)
ヨーグルトは体に良い	0.19 (3/16)	0.40 (2/5)
キシリトールは虫歯の予防になる	0.38 (8/21)	0.25 (2/8)

の上位 R 個にある正解データの割合である。R 精度を見ることで実験データ中における正解データをどれだけ上位にすることができているかを評価することができる。結果より、ある程度の正解を上位にすることはできているが、本研究の網羅的に提示するという目的のためには、改善の必要があると考えている。誤り分析をした結果、「ヨーグルトは毎日取れば」と「胃痛予防に最適です」の組は提示すべきであるが、システムではポジティブの 27 位と下位に出力した。これより、「食べる」と「取る」のような同義語句を柔軟に扱える枠組みを導入する必要があると考えている。また、「キシリトールを 100% 使用したガムで 1 日 3 回食後、歯磨き前に 1 粒噛む」と「効果があります。」の組は提示すべきでないのが正解だが、システムはポジティブの 3 位と上位に出力した。このようなメリットであるが、どのようなメリットなのかかわからない結論は、ユーザにとってあまり有用でないと考えられるため、順位を下げる必要があると考えている。

## 6 まとめ

本稿では、情報信憑性判断支援のための対話型条件結論マップを目的として、どのような情報を優先的にマップに提示すべきかを考察した。また、利用者が入力した条件とマッチする条件と結論のテキストの組を Web 文書から網羅的に抽出する手法を提案し、極性判定とランキングの精度に関する実験を行った。今後、マップによる可視化を行いたいと考えている。

## 参考文献

- [1] K. Murakami, E. Nichols, J. Mizuno, Y. Watanabe, S. Masuda, H. Goto, M. Ohki, C. Sao, S. Matsuyoshi, K. Inui and Y. Matsumoto, "Statement Map: Reducing Web Information Credibility Noise through Opinion Classification", In Proc. of the Fourth Workshop on Analytics for Noisy Unstructured Text Data (AND 2010), 2010.
- [2] 中野正寛, 洪木英潔, 宮崎林太郎, 石下円香, 金子浩一, 永井隆広, 森辰則, "情報信憑性判断支援のための直接調停要約生成手法", 電子情報通信学会論文誌 (D), vol.J94-D, no.11, pp.1019-1030, 2011.
- [3] 乾孝司, 乾健太郎, 松本裕治, "接続標識「ため」に基づく文章集合からの因果関係知識の自動獲得", 情報処理学会論文誌, Vol. 45, No. 3, pp. 919-933, 2004.
- [4] 坂地泰紀, 竹内康介, 関根聡, 増山繁, "構文パターンを用いた因果関係の抽出", 言語処理学会第 14 回年次大会, pp.1144-1147, 2008.
- [5] 乾孝司, 乾健太郎, 松本裕治, "出来事の望ましさ判定を目的とした語彙知識獲得", 言語処理学会第 10 回年次大会, pp.91-94, 2004.
- [6] 乾孝司, 奥村学, "文章内に現れる因果関係の出現特性調査", 情報処理学会研究報告, SLP-056, pp. 81-86, 2005.