文間意味的関係認識による言論マップ生成

村上浩司† 水野淳太† 後藤隼人† 大木環美† 松吉俊† 乾健太郎‡† 松本裕治†
奈良先端科学技術大学院大学† 東北大学‡
{kmurakami, junta-m, hayato-g, megumi-o, matuyosi, inui, matsu}@is.naist.jp

1 はじめに

ウェブ上には大量のテキスト情報が存在し，そこでは様々なトピックに関して多角的な意見が述べられている．情報検索技術の発展により，あるトピックに関連する文書集合を容易に入手できるようになった．しかしながら，これらの文書に記述されている情報は，そのすべてが真実というわけではなく，不正確な記述，偏りのある意見などが混在している可能性が高い．そのため，あるトピックに対する情報の集合を俯瞰するには，ユーザは個々の情報の信憑性を判断する作業を繰り返すことを強いられる．しかし，限られた時間で各意見にそのような作業を行うことは容易ではない．これらの作業に関してユーザを支援する技術が必要である．

我々は現在，こうしたユーザによる Web 情報の信憑性分析を支援するために，言論マップ生成課題 [17, 8] に取り組んでいる．これは，例えばユーザが「イソフラボンは健康維持に効果がある」と思っていた場合，それをクエリとして入力すると，図 1 のような入力クエリに関連する情報を提示する言論マップを出力するものである．我々は，ユーザに提示すべき情報を 6 種類のカテゴリに分類した．こうした情報を中心として言明を整理することで，ユーザによる言明の信憑性判断の支援情報とする．

同意 クエリとおおよそ同じ意味を持つ言明
同意言明の根拠 同意言明を支持する根拠を述べる言明
矛盾 クエリと同時に成り立たない言明
矛盾言明の根拠 矛盾言明を支持する根拠を述べる言明
限定 クエリの範囲や程度を制限する，弱い対立を表す言明
限定言明の根拠 限定言明を支持する根拠を述べる言明

言論マップの特長は，個々の言明が関連する言明の意味的関係性の中に相対的に位置づけられる点であり，それぞれの言明に対する根拠，同意，矛盾などの意味的関係が周辺情報となり，ユーザに対し，各言明の立場や信憑性の判断を支援することができると考えられる．これにより，情報の偏りや思いこみによる誤信の可能性を抑えることができるようになる．

本稿では，言論マップ生成の課題設計，生成のための構成要素に対するこれまで取り組みと，それらを統合した言論マップ生成システムの現状について述べる．2 節でまず文間の意味的関係に関する関連研究について述べる．3 節で言論マップ生成のための課題を整理し，4 節では言論マップ生成のための個別タスクの現状について述べる．5 節でプロトタイプシステムによる意味的関係認識例を示し今後の課題を考察し，6 節でまとめを述べる．

2 関連研究

文間の「含意」や「矛盾」等の意味的関係の自動認識は，近年 NLP が実現すべき課題として精力的に研究され，いくつかのタスクが提案されてきた．その 1 つに与えられた文対 (t,h) が含意関係，矛盾関係もしくは不明であるかを判定する課題，RTE Challenge (Recognizing Textual Entailment Challenge)[1] がある．また，複数文書中の文間の関係解析には，Radev らの CST (Cross-Document Structure Theory)[9] がある．RST[13] に基づく談話構造解析が単一文書内の構造を解析するのに対し，CST はこれを文書横断構造解析に拡張するものであり，18 種類の意味的関係が定義された．

しかしながら，RTE では「含意」「矛盾」のみが認識の対象であり，また CST においても，2 文間の情報の差に着目しており，本課題で認識すべき「限定」などの意味的関係は対象外である．RTE と CST のタスク仕様では，以下の文対に対する意味的関係を取り扱うことが出来ない．

(1) A 副腎皮質ホルモンには副作用がある
　　B ステロイド剤は<u>ある程度の期間なら</u>副作用はほとんど出ません （限定）

(2) A イソフラボンは健康維持に効果がある
　　B <u>僅かだがエストロゲン様の作用により</u>大豆イソフラボンは人体に悪影響が出る（矛盾の根拠）

CST では，Zhang らが比較する 2 文からのみ素性を抽出して素性空間に表現し，主要な関係を 1 つの分類器により分類 [14] を行った．また，正解ラベルが付与されたデータが少ないため，ラベルなしデータも用いた Boosting を利用した手法を提案 [15] したがどちらにおいても精度は高くなく，類義語や反義語等の語彙知識の適用，素性の洗練，各意味的関係認識のための個別処理の検討など，多くの課題を残した．このことから，複数の意味的関係クラスへの分類手法そのものの確立も重要な課題となる．

3 言論マップ生成のための基本方針

3.1 全体の位置づけ

1 節に示した通り，ユーザの思っているクエリに関連する様々な Web 文書に記述される文（以下，検索対象文と呼ぶ）を，クエリとの間に持つ意味的関係により整理する事で言論マップを実現するため，言論マップ

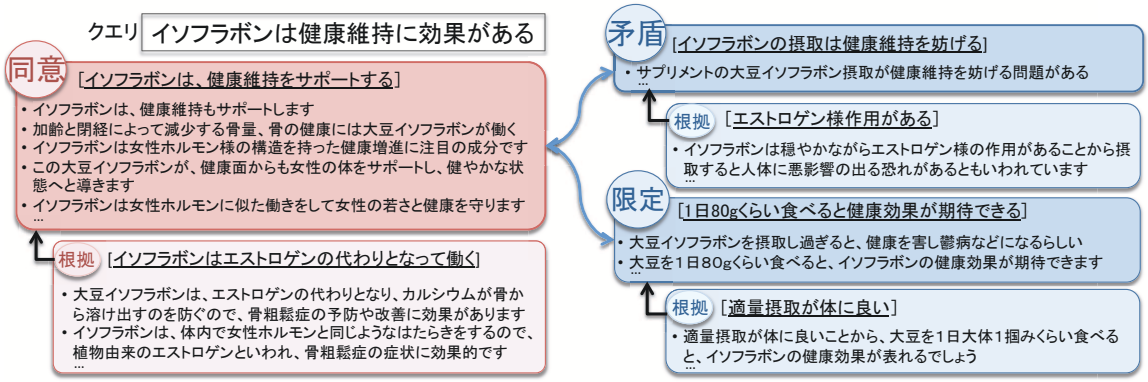

図 1: 言論マップ（クエリ：イソフラボンは健康維持に効果がある）

生成課題は，(i) クエリに関連する文集合を得る情報検索，(ii) それぞれの検索対象文とクエリを対とし，1 節で説明しその間の意味的関係を文間関係認識，の 2 つの課題に帰着することができる．

3.2 文間意味的関係認識

文間の関係認識である RTE では，含意関係は文対 (h,t) のうち，h の情報を t が適切に含んでいるかを判定する課題であることから，語彙的な言い換え知識を用いて含意を判断する手法（例えば [3]）が中心的だが，認識のために語彙的な情報だけでなく構造的な情報に着目する必要がある矛盾や限定などを同時に扱うことは難しい．一方，文の統語・意味構造を考慮して文対中の単語アライメントを求め，その結果から関係認識を行う，より柔軟なアプローチが（例えば [6]）ある．語彙的な知識に加え文構造も考慮した単語間の対応付けは，3 クラス分類タスクの RTE においても高精度の認識を実現している．

本課題の文間関係認識はこうした理由から，(i) 各文の多様な表現を捉える係り受け，述語項構造解析等の言語解析，(ii) 文間の依存関係も考慮する局所構造アライメント，(iii) 大域的構造やその他の意味情報などを利用した関係分類，の流れで実現する．対象とする意味的関係は，同意／矛盾／限定／同意の根拠／矛盾の根拠／限定の根拠，と更に，どれにも当てはまらない「負例」の合計 7 種類とする．このうち根拠関係の認識は 2 文間の関係分類の対象としない．これは [18] で述べたように同一文内で根拠関係を判断すべきと考えているためであり，文内で根拠関関係が認識されたのちに，主節とクエリとを比較して同意または矛盾のとき，それぞれ「同意の根拠」,「矛盾の根拠」と認識する．

関係分類では，対象の文対間の構造アライメントで得られた意味情報付きのグラフを対象として，アライメント結果を解釈して意味的関係の分類を行う．文対の正しい構造アライメントは，精度の高い関係分類のための重要な手がかりとなる．しかしながら構造アライメントによる文節の対応づけ，各文内と文間の局所構造の対応づけにより，もし文対が語彙的，構造的に類似していても次の例のように，矛盾や限定を認識するには更に他の情報も必要となる．構造アライメントでは添字のついた下線部分の各文節が依存関係とともに対応づけられる．

(3) A ステロイド剤は$_0$副作用が$_1$ある$_2$
B ステロイド剤は$_0$ **短期間の使用では重大な副作用は$_1$ある$_2$** わけでは**ない** （限定）

正しくこの文対の関係分類を行うためには，太字で示される様な，文間における情報の差異に着目する必要がある．これまでの矛盾認識 [2, 4] でも肯否極性，語彙的反義，事実性，モダリティなどが重要な意味的情報であるとしている．これらの情報は，負例の判定にも有効であることが示されている [12, 10]．限定認識においては，まずクエリと検索対象文の間に同意もしくは矛盾が分類される必要がある．更に，検索対象文中に存在するクエリの意味内容の範囲や程度などを制限する表現を捉えるために，程度副詞や条件節，部分否定といった特徴的な語彙や表現を同定する必要がある．本課題における関係分類においても，こうした意味的情報および文間の大域的構造を考慮して，総合的に判断して関係分類を行う必要がある．

こうした文間の局所構造アライメントや関係分類をより汎用的な手法で実現するためには，統計的手法の導入が考えられる．しかしながら機械翻訳の分野と異なり，アライメント，関係分類ともに単純に統計的手法を試すことができるほど十分な訓練事例が用意できない問題がある．そこでまず，個々の事例の分析からルールを作成し，ルールベースのシステムを構築して関係分類の評価を行う事を考える．これにより最適化が必要な部分をある程度特定した後に改めて事例収集を行い，統計的手法の適用を検討する．

4 言論マップ生成システム

前節で述べた言論マップ生成のための技術的課題に対する我々の現在の取り組みについて述べる．

4.1 情報検索

言論マップを生成するためには，ユーザのクエリに関連する文集合が必要となる．その文集合は，例えば図 1 中のクエリに対して,「効果がある」という表現を数多く検索するのではなく，様々な種類の情報を網羅的に検索する必要がある．我々は永井らのパッセージ抽出法 [21] を用いて文集合を得る．

4.2 言語解析

言論マップ生成のためには，クエリと検索対象文の語彙的情報や統語的な構造だけではなく，更に意味的な構造も言語解析により得る必要がある．そこでまず，パッセージ中の各文に対して CaboCha[11] による係り受け解析を行い，さらに SynCha[5] による述語項構造解析を行う．文内根拠関係解析については飯田らの研究成果 [22] 等を適用する．文中の肯否や時制，態度等を表す拡張モダリティの解析については次節で述べる．

4.3 拡張モダリティ解析

拡張モダリティ解析は，係り受け・述語項構造解析結果を入力とし，文中に存在する各々の事象に対して，モダリティや肯否極性などに関する 7 つの項目のタグ (拡張モダリティタグ) を付与する [20].

拡張モダリティ解析結果の例を以下に示す（下線は事象の核となる述語）.

- キシリトールを 噛め$_{\mathrm{A}}$ ば，虫歯を 予防できる$_{\mathrm{B}}$ といわれている.
- サプリメントによる大豆イソフラボンの過剰 摂取$_{\mathrm{C}}$ が健康 維持$_{\mathrm{D}}$ を 妨げ$_{\mathrm{E}}$ たのです.

	態度表明者	時制	仮想	態度	真偽判断	価値判断	焦点
A	wr_arb	非未来	条件	叙述	成立	0	0
B	wr_arb	非未来	帰結	叙述	成立	0	0
C	wr	非未来	0	叙述	成立	0	0
D	wr	非未来	0	叙述	不成立	0	0
E	wr	非未来	0	叙述	成立	0	0

4.4 局所構造アライメント

局所構造アライメントは，前節で述べた言語解析により統語的な依存構造と意味的な依存構造が付与された 2 文の解析結果を入力として，文節の対応と，クエリ中で対応付く 2 文節間の関係と検索対象文側での関係の対応づけ，の 2 処理を行う．図 2 に局所構造アライメントの例を示す．文節の対応付けは，語彙の意味的類似度に基づき行う．構造の対応付けは，クエリ側の対応づいた 2 文節間の語彙的・意味的な依存関係が，検索対象文側の対応する 2 文節間にも含まれるかにより判定する．図中の”○”印は文節の対応を示す．具体的な局所構造アライメントのアルゴリズムについては，[19] にて説明を行っている.

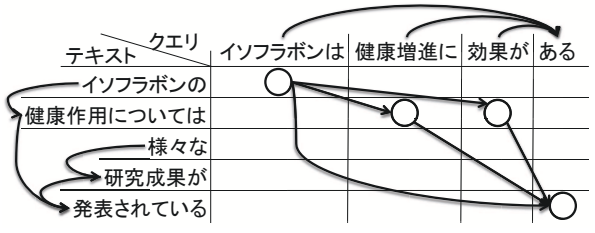

図 2: 局所構造アライメントの例

4.5 関係分類

関係分類は，局所構造アライメントの結果を入力として，その構造を解釈し，さらに各文中の意味的情報を考慮して，ただ 1 つの意味的関係に分類する．具体的には，(i) 局所構造と大域構造を考慮してクエリ側の文節間依存関係がもれなく検索対象文に含まれるかを判定，(ii) 肯否極性や語彙的反義，事実性，モダリティ情報等を考慮して意味的関係に分類，の 2 処理から構成される．限定関係の認識に関しては，[16] にて詳細にタスク設定と事例分析を行った.

5 意味的関係認識例と考察

これまで述べたそれぞれの処理を統合し，プロトタイプ言論マップ生成システムを構築した．このシステムに対して，実際にいくつかのクエリを与え意味的関係認識分類を行った．システムは基本的にドメイン非依存であり，かつ任意のクエリの入力に対して動作することができる．処理に必要な時間は，クエリとその関連文 500 文を入力としてアライメントと関係分類の処理時間はシングルプロセスで 1 文対当たり 0.02 秒程度である．図 3 にクエリに対してそれぞれの意味的関係分類のシステムのスナップショットおよび，表 1 にその他のクエリの場合に得られたシステムの出力例と正解の関係ラベルを示す.

キシリトールの限定の例では，クエリと検索対象文中のクエリ相当部分の間に「効果がある：効果がない」の否定を捉えて矛盾関係に分類した上で，かつクエリが成立するための「キシリトール 100%でなければならない」「正しい使い方であれば」という制限部分を適切に解析できたことにより正しく関係分類されている．また還元水の例では，クエリ中の文節の単語がおおよそ実文側とアライメントされるだけではなく，「健康に良い」に対して「健康を支える」「健康を維持してくれる」のような事象間関係知識 [7] ではカバーしていない，ドメインや文脈に依存した事象の同義関係を適切に対応付けできたことで正しく関係分類された．これは係り受け構造や意味構造の類似性を利用する柔軟な構造アライメントによるものである.

しかしながら適切に関係分類できない例も多く，解決すべき課題が残されている．例えば虫歯に関して，「予防に効果がある」と「予防ができる」や「予防に効く」，「虫歯予防」と「虫歯菌抑制」等の間のアライメントが適切にできないことがある．これには関係知識ベースの規模の増強，効率的な利用法を検討する必要がある.

また基本方針でも述べたが関係分類における最大の課題は，統計的手法を適用するために局所構造アラインメントやその他の意味的な情報，解析の結果をどのような素性空間に写像すれば分類に有効であるかを検討し，そのモデルを定量的に評価することである.

6 おわりに

本稿では，Web 情報信憑性評価のための言論マップ生成について，課題設定，現在の取り組みと生成システムの現状を述べた．また，いくつかのクエリに関して意味的関係の分類を行い，現状のシステムで捉えられた例とこれから取り組むべき課題について考察した.

今後はそれぞれのタスクに対してのシステムの精度を高めるとともに，多くのクエリで言論マップ生成システムを適用して定量的な評価を行い，システムの有効性を示すことが課題である.

謝辞

本研究は，(独）情報通信研究機構の委託研究「電気通信サービスにおける情報信憑性検証技術に関する研究開発」の一環として実施した.

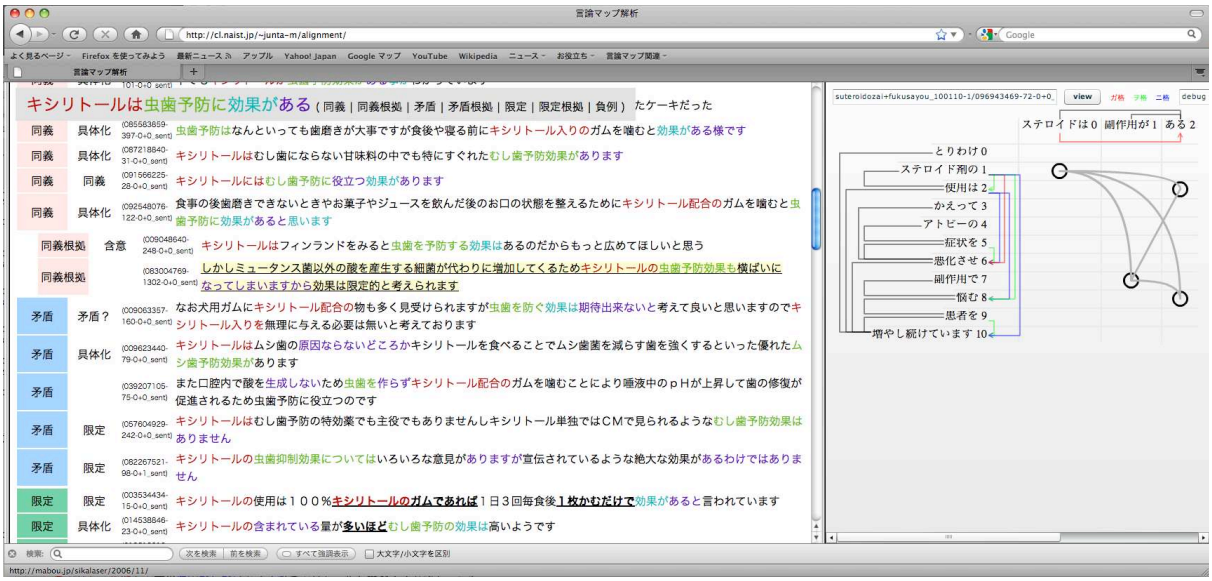

図 3: 局所構造アライメントと関係分類の動作例（クエリ：キシリトールは虫歯予防に効果がある）

表 1: 任意のクエリに対する言論マップ生成システムの動作例

クエリ	検索対象文	出力（正解）
還元水は健康に良い	弱アルカリ性の アルカリイオン還元水が あなたと家族の 健康を支えます	同意（同意）
	還元水は 活性酸素を除去すると言われ 健康を維持し てくれる働きをもたらす	同意（同意）
	美味しくても体を 酸化させる水は健康には役立ちません	矛盾（矛盾）
ステロイドは副作用がある	しかし 経口ステロイド剤を 服用すると 副作用が出ます	同意（同意）
	ステロイド剤は **医師が行う治療であれば** 副作用の心配はない ようです	限定（限定）
	ステロイド剤は **局所治療では**副作用の心配は**ほとんど** ありません	限定（限定）
バイオエタノールは環境に良い	バイオエタノールは 高い可能性を秘めた高品質の燃料であり，現在我々が直面する 環境問題に対処し得る 潜在能力を持っている	同意（同意）
イソフラボンは健康維持に効果がある	大豆イソフラボンを サプリメントで**過剰摂取すると** 健康維持には 負の影響を与える 結果となります	限定（限定）

参考文献

[1] Ido Dagan, Oren Glickman, and Bernardo Magnini. The pascal recognising textual entailment challenge. In *Proc. of the PASCAL Challenges Workshop on RTE*, 2005.

[2] Marie-Catherine de Marneffe, Anna N. Rafferty, and Christopher D. Manning. Finding contradictions in text. In *Proc. of ACL 2008*, pp. 1039–1047, 2008.

[3] Oren Glickman, Ido Dagan, and Moshe Koppel. Web based textual entailment. In *Proc. of the First PASCAL Recognizing Textual Entailment Workshop*, 2005.

[4] Sanda Harabagiu, Andrew Hickl, and Finley Lacatusu. Negation, contrast and contradiction in text processing. In *Proc. of AAAI-06*, pp. 755–762, 2006.

[5] Ryu Iida, Kentaro Inui, and Yuji Matsumoto. Capturing salience with a trainable cache model for zero-anaphora resolution. In *Proc. of ACL-IJCNLP 2009*, pp. 647–655, 2009.

[6] Bill MacCartney, Michel Galley, and Christopher D. Manning. A phrase-based alignment model for natural language inference. In *Proc. of EMNLP 2008*, pp. 802–811, 2008.

[7] Suguru Matsuyoshi, Koji Murakami, Yuji Matsumoto, , and Kentaro Inui. A database of relations between predicate argument structures for recognizing textual entailment and contradiction. In *Proc. of the 2nd International Symposium on Universal Communication (ISUC2008)*, pp. 366–373, 2008.

[8] Koji Murakami, Eric Nichols, Suguru Matsuyoshi, Asuka Sumida, Shouko Masuda, Kentaro Inui, and Yuji Matsumoto. Statement map: Assisting information credibility analysis by visualizing arguments. In *Proc. of WICOW*, pp. 43–50, 2009.

[9] Dragomir R. Radev. Common theory of information fusion from multiple text sources step one: Cross-document structure. In *Proc. of the 1st SIGdial workshop on Discourse and dialogue*, pp. 74–83, 2000.

[10] Rion Snow, Lucy Vanderwende, and Arul Menezes. Effectively using syntax for recognizing false entailment. In *Proc. of NAACL 2006*, pp. 33–40, 2006.

[11] Yuji Matsumoto Taku Kudo. Japanese dependency analysis using cascaded chunking. In *Proc of CoNLL 2002*, pp. 63–69, 2002.

[12] Rui Wang, Yi Zhang, and Guenter Neumann. A joint syntactic-semantic representation for recognizing textual relatedness. In *Proc. of Recognizing Textual Entailment*, 2009.

[13] Mann William and Sandra Thompson. Rhetorical structure theory: towards a functional theory of text organization. *Text*, Vol. 8, No. 3, pp. 243–281, 1988.

[14] Zhu Zhang, Jahna Otterbacher, and Dragomir Radev. Learning cross-document structural relationships using boosting. In *CIKM '03*, pp. 124–130, 2003.

[15] Zhu Zhang and Dragomir Radev. Combining labeled and unlabeled data for learning cross-document structural relationships. In *IJCNLP '05*, pp. 32–41, 2005.

[16] 大木環美, 村上浩司, 水野淳太, 増田祥子, 乾健太郎, 松本裕治. 文間の限定関係認識：課題設計および分析と予備実験. 言語処理学会第 16 回年次大会発表論文集 D3-1, 2010.

[17] 村上浩司, 松吉俊, 隅田飛鳥, 森田啓, 佐尾ちとせ, 増田祥子, 松本裕治, 乾健太郎. 言論マップ生成課題：言説間の類似・対立の構造を捉えるために. 情報処理学会研究報告 2008-NL-186, pp. 55–60, 2008.

[18] 村上浩司, 増田祥子, 松吉俊, Eric Nichols, 乾健太郎, 松本裕治. 言明間の意味的関係の体系化とコーパス構築. 言語処理学会 第 15 回年次大会, 2009.

[19] 後藤隼人, 水野淳太, 村上浩司, 乾健太郎, 松本裕治. 文間関係認識のための構造的アライメント. 言語処理学会第 16 回年次大会発表論文集 E3-7, 2010.

[20] 江口萌, 松吉俊, 佐尾ちとせ, 乾健太郎, 松本裕治. モダリティ、真偽情報、価値情報を統合した拡張モダリティ解析. 言語処理学会第 16 回年次大会発表論文集 E3-8, 2010.

[21] 永井隆広, 金子浩一, 渋木英潔, 中野正寛, 宮崎林太郎, 石下円香, 森辰則. 多層ネットワーク型 textrank による根拠関係を考慮した重要パッセージ抽出. 言語処理学会第 16 回年次大会発表論文集 PA-1, 2010.

[22] 飯田龍, 乾健太郎, 松本裕治. 根拠情報抽出の課題設計と予備実験. 言語処理学会 第 15 回年次大会, pp. 817–820, 2009.