言明間の意味的関係の体系化とコーパス構築

村上浩司 †　増田祥子‡†　松吉俊†　乾健太郎†　松本裕治†
奈良先端科学技術大学院大学 †　大阪府立大学‡
{kmurakami, shouko, matuyosi, inui, matsu}@is.naist.jp

1 はじめに

情報検索，質問応答などの情報アクセス技術の更なる発展のために NLP が実現すべき課題に，文書中の文間の「類似」「矛盾」「根拠」などの意味的な関係の自動認識がある．こうした文間の関係認識は，これまでいくつかのタスクが提案されてきた．その 1 つに，一対の文について一方の文が他方から含意されるかを判定する含意関係認識がある．英語を対象とし，含意関係だけでなく 2 文が矛盾関係にあるかを判定する課題，RTE Challenge (Recognizing Textual Entailment Challenge)[1] がある．ここで用いられているコーパスは，情報検索，質問応答，情報抽出，自動要約の 4 つの情報アクセスが対象であり，それぞれの情報アクセスで必要な情報を得るための処理を想定し，独立して文ペアが作成されている．日本語においての含意関係認識は，梅基ら [12]，小谷ら [16] が研究を進めている．

複数文書中の文間の関係解析には，Radev らの CST (Cross-Document Structure Theory)[6] がある．RST[9] 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コーパス中で定義された．衛藤らは，CST を元に日本語に適応した 14 種類の関係を再定義して，文書横断文間関係コーパス [14] を作成した．また，このコーパスを用いて宮部らは＜同等＞関係の認識 [15] を行った．

しかしながら，RTE は一対の文における事実関係に認識対象を限定しており，CST は横断する関連文書は同じイベントに関する複数の新聞記事を利用することから，RTE と同様に認識対象は事実関係のみである．

本稿では，これまでの文間関係認識に関する研究を吟味し，Web 文書中の多様な言明を扱うための仕様と Web 文書中の言明間の意味的関係を体系化するための仕様を議論し，コーパス構築の方法論について述べる．2 節では想定するタスクと必要な意味的関係について述べ，3 節で先行研究を踏まえ，具体的な意味的関係を体系化する．4 節では，現在構築中の言明間意味的関係コーパスついて議論し，5 節で，まとめを述べる．

2 Web 情報の信憑性分析

我々は現在，Web 情報の信憑性を分析するため，言論マップ生成課題 [13] に取り組んでいる．これは，ユーザが着目したある言明に関するトピックの文書集合から，そのトピックに対する多種多様な言明を抽出し，それらの間の意味的関係を解析して，俯瞰図となる言論マップをユーザに提供することを目的とした課題である．ここで我々は，(i)“〜である/〜でない，〜した/〜してない” など，真偽が問える陳述，(ii)“すべきだ/すべきでない，望ましい/望ましくない” などの，何らかの価値判断，(iii)“〜できる/できない，可能性がある/可能性がない” といった，ものごとの可能/不可能，などの文を言論として扱う．

ユーザが種々の言明を俯瞰的に概観するためには，文書集合から抽出された言明を単に羅列するのではなく，それぞれの言明を区別，整理して提示することが必要である．本研究で着目するのは，ユーザが着目した言明の信憑性判断のために必要となる情報である．

Metzger らは Web 情報の信憑性評価に使われる要因として，情報を記述するときの品質だけでなく，科学的データへの引用などを挙げた [2]．また，Moela ら [3] は，着目する Web 情報を他の情報と比較することでトピックの議論が明らかにできること，多くの情報源からの事実や意見を集約させる実証作業により情報の正しさや信頼性が評価できると示した．これらの研究を踏まえて，本研究ではユーザに提示すべき情報を大きく 5 種類に分類した．ここで，ユーザの着目した言明が「バナナダイエットは痩せるのか？」であると仮定した場合に基づいて例を挙げる．

着目言明と同等の言明 朝ご飯をバナナと水にすると痩せる

着目言明の根拠・実例・詳細化 バナナの持つ酵素の働きにより優れたダイエット効果がある

着目言明に対する対立言明 朝バナナと水を飲んだだけで，痩せるわけがありません！

対立言明の根拠・実例・詳細化 バナナに痩せる成分があるわけではない

着目言論の根拠と対立言明の根拠の対立 ダイエット効果がある⇔痩せる成分があるわけではない

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．

3 言明とその意味的関係

文書間の関係認識のため先行研究を踏まえて，言論マップ生成課題において (1) 対象とする言明，(2) 対象とする意味的関係，(3) 事例収集の戦略，の 3 つに関して議論を行う．

3.1 対象とする言明

Web 上の多様な言明を大きく分けると，1. 意見などの主観的要素を含むもの，2. 評価極性を暗に持つ事実，3. 客観的事実，に分類することができる．

表 1: 言明間関係認識における対象関係の比較

関係カテゴリ	具体的な関係	RST	CST	RTE	言論マップ
態度関係	同意，反対，評価など	○			○
論理関係	含意，矛盾	○	○	○	○
比較関係	類似，対比など	○			○
論争的関係	根拠，証拠，正当化など	○	○		○
ソースを考慮した関係	Attribution，参照など		○		○
時間を考慮した関係	追加，実現，観点の交替など		○		○

1. 〜で嬉しい，〜は値段が高い，〜かどうか疑わしい
2. 〜は虫歯予防に効果がある，〜を使い続けている
3. 〜の本社は東京にある

客観的言明は新聞記事などにおいての中心的なものであり，Factoid 型の質問応答や情報抽出など，これまで多くの研究が対象としてきた．主観的言明はここ近年研究が進められており，意見や評価表現に焦点をあてた代表的な研究に Wiebe らによる MPQA (Multi-Perspective Question Answering)[8] がある．主観的/客観的言明はそれぞれ示す情報が異なるために同時に扱われることは少なかったが，近年，上記の 1 と 2 までを対象とした研究として，イベントの事実性やモダリティ解析，言明の極性判定を行う乾らによる経験マイニング [11]，Nakagawa らによる評価表現抽出 [4] がある．これに対し言論マップ生成課題では，前節で示したように，Web 上の任意のトピックに関する様々な言明を取り扱うため，1〜3 の全言明を対象とする．

3.2 対象とする意味的関係

これまでの文間の意味的関係を取り扱ってきた中心的なタスクである RST，CST，RTE と言論マップについて，対象とする言明と関係の種類の観点で整理する．表 1 に，それぞれのタスクで取り扱われる意味的関係を示す．RST (Rhetorical Structure Theory) では，同一文書中の文を対象とし，関係解析が行われてきた．単一文書の一貫性に着目するという目的から，文書中の主観的，客観的表現にかかわらず態度関係，論理関係，比較関係，論争的関係などの関係を認識する．これに対し CST は，複数の新聞記事要約という目的から，関連する複数新聞記事を横断して意味的関係を発見する．新聞記事は基本的に陳述などの客観的言明によって構成され，CST が扱う関係は論理関係，Attribution などのソースを考慮した関係，時間を考慮した関係などを対象としている点が RST とは異なる．RTE は関係認識を行う対象は任意の 2 文であり，文書から抽出した文を対象とする RST，CST とは異なるが，基本的に文は異なる文書から抽出されることから，文章中の談話標識などの手がかりに頼らない点で CST と同様であると考えられる．また認識する関係は含意関係及び矛盾関係という論理関係のみを目的としている．

これに対して言論マップ生成課題は，Web 上の任意の 2 文書中の言明を意味的関係認識の単位とするため，CST や RTE と同様，文書間を横断する関係を認識する．Web 文書は新聞記事とは違い主観的言明も多く，書き手の意見や評価表現などからなる言明に対しても関係認識を行うため，図 1 が示すように，すべての関係を言論マップ生成課題では対象とする．

3.3 事例収集の戦略

表 1 中の各意味的関係の事例収集を考える．時間を考慮した関係は，Web 文書内の時間情報の記述だけでは時間の推定が難しいことから，ここでは事例収集を行わないこととする．ソースを考慮した関係は，佐尾ら [10] の事実性解析用のコーパスから事例収集を行う．文内の論争的関係や比較関係は，修辞構造解析により得られる．文書間の論争的関係や比較関係は，RTE などでの提案方法を用いて直接得るべきではないと我々は考える．なぜなら例えば，論争的関係中の＜根拠＞に関して，任意の事象をある事象の根拠とみなせ情報抽出においては，閉じた文書内で，明確にそれらの間に＜根拠＞があると判断できる文対を抽出すべきであると考えるからである．文書間の論争的関係や比較関係は＜同義＞を通しての認識が妥当であると思われる．例えば，文 (1) が記述された文書と文 (2a) と文 (2b) が記述された文書において＜根拠＞の認識を考える．

(1) クローン技術を法律で規制する必要がある．
(2) (a) 法律でクローン技術の応用の制限をすることが必要だ．(b) なぜならクローン技術は悪用される懸念があるからだ．

文 (2a) と文 (2b) の間には，根拠を示す明確な談話標識があり，修辞構造解析により，それらの間に＜根拠＞を認めることができる．一方，RTE などの技術を用いると，(1) と (2a) の間に＜同義＞を認めることができる．これらをあわせると，(1) と (2b) の間に根拠と呼べるべき関係を推測することができる．我々は，このようにして，文書間の論争的関係や比較関係を認識する．これらの関係の事例は，いわゆる，談話構造のアノテーションにより得られるので，根拠情報を対象にした飯田らの研究 [17] を参考に，文書内の言明に着目し，別コーパスを構築する予定である．

構築する言明間意味的関係コーパスにおける関係の体系を表 2 に示す．我々は，RTE で扱われている論理関係に加えて，価値判断などの態度関係を対象としてコーパスを構築する．各行は，関係の種類，意味的関係ラベル，定義，及び例文で構成される．より詳しい定義や情報はコーパスのサイト[1]にて公開の準備中である．

4 言明間意味的関係コーパス構築

これまで述べてきた仕様を元に現在，構築を進めている言明間意味的関係コーパスにおける，コーパスの構造，および構築手順について述べる．

4.1 言論間意味的関係コーパス

本コーパスは，各意味的関係の例文ペア “文 1-文 2-意味的関係” および付加情報から構成されるレコードの

[1] http://cl.naist.jp/stmap/corpus

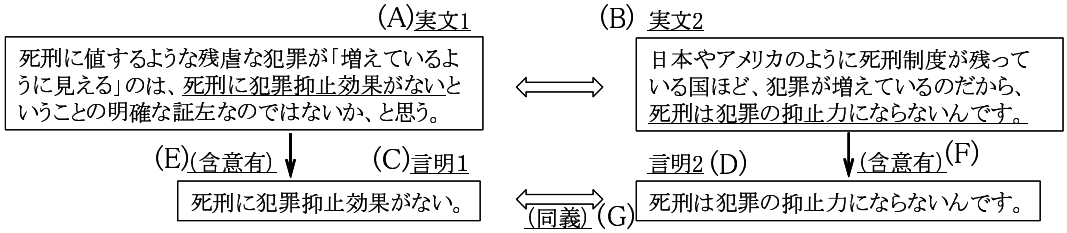

図 1: コーパスを構成する情報

集合である．言論マップ生成課題における解析対象はWeb 上の言明であることから，コーパスも同様に Web 文書中の実際の文 (以下，実文と呼ぶ) を用いることが望ましい．実文は複数の言明から構成されることが多いため，2 つの実文が全体として同義や矛盾関係を満たすことは実際には稀である．そこで，実文から適当な構成要素を言明として取り出し，得られた言明同士の間に表 2 で示す意味的関係のうちの 1 つを付与することとした．コーパスの 1 レコードは，図 1 で示す (A)〜(G) の 7 つの情報から構成される．

実文 (A) および (B)．実文集合から次節の処理により選別されペアとなる．言明のペアを作成するための元情報である．

言明 (C) および (D)．実文 1，実文 2 からそれぞれ抽出される．1 実文から抽出されるのは 1 言明とは限らず，実文の内容などにより複数の言明が抽出される．抽出される言明は基本的に実文の一部分を元に作成されるが，抜粋では情報が不足する場合，述語，代名詞などを補い，文を完結する形に修正する．

実文-言明間含意・引用関係フラグ (E) および (F)．抽出された言明と，抽出元の実文との間に含意・引用関係があるかを示すフラグ．引用関係とは，実文中の判断主体以外の価値判断などの部分を言明とし抽出する場合に付与する関係で，実文「政府は死刑制度を支持するが，私は断固反対だ.」から言明「政府は死刑制度を支持している.」を作成した際，言明が判断主体の意志ではないことを示すために用いる．

言明間の意味的関係のラベル 図中の (G)．3.3 で述べた言明間の意味的関係を示すラベル．表 2 で示される 9 種類の関係ラベルのうちの 1 つが付与される．

4.2 コーパス構築の手順

我々は，以下の手順で前処理を行い Web 文書からペア候補となる実文集合を得る．まず，(i)TSUBAKI[7] により，特定のトピックをクエリとして関連文書を検索し，(ii) 中心的に議論されていると思われるサブトピックを半自動で選別した．例えば，トピック「クローン技術」から選別されたサブトピックは “規制”，“研究”，“可能性” などであった．最後に (iii) サブトピック毎の文書集合から広告，極めて類似した実文をヒューリスティクスによりフィルタリング，を行う．

言明ペアの作成には，得られた実文集合中の実文から言明抽出を予め行い，任意の 2 言明でペアを作成する方法も考えられるが，正例となりえる言明ペアを含む 2 つの実文を予めペアにし，それぞれの実文から言明を抽出してペアとする方法を採用した．実文からの言明抽出の人的コストを抑え，ペアとならない言明をできるだけ作らずに言明ペアを作成できるためである．

正例となりえる言明ペアを含む実文ペアには，同一単語，言い換え可能な語や表現などが含まれていると考えられる．そこで，2 つの実文間が一定以上に類似した表現を持つ実文からペアを作成することとした．予備実験から，文字ユニグラム（漢字，カタカナ），単名詞，複合名詞，動詞，形容詞を素性にしてベクトルとして，コサイン距離により類似性を計算した．

4.3 検討中の課題

現在検討中の課題に，言明ペアが条件を伴って意味的関係が成立する場合に対する対応がある．

(3) a. キシリトールは虫歯予防に効果がある
　　b. 普段の食生活，適切な歯磨きがあれば，キシリトールは虫歯予防に役に立つ
(4) a. キシリトールは虫歯予防になり，歯石などをふせぐことができます
　　b. キシリトールを食べれば必ず虫歯予防ができるわけではありません

これらの例では，条件部分が真のときに同義関係が成り立つため，このままでは表 2 のどの関係も付与出来ない．(4) では否定を伴った「必ず」が予防可能性を限定しており，必ずしも条件節を伴わないことが分かる．現在，こうした例を収集しており，今後このような条件の取り扱いを検討する予定である．

5 まとめと今後の予定

本稿では，言論マップ生成課題においての対象とすべき言明間の意味的関係を体系化するための仕様について，関連研究を踏まえて検討を行った．また定義した体系をもとに，言論マップ生成システムを評価するための言明間意味的関係コーパスを構築するための準備とコーパスの構造について説明した．言明間意味的関係コーパスは現在構築中であり，マイナスイオン，キシリトールなど科学的な証拠が根拠となり，議論が収束できるトピックを中心に，2009 年 3 月中に，約 3,000 の言明ペアに対して意味的関係を付与する予定である．

現在このコーパスの他に，同じトピックについて議論している，リンク構造を持つ複数のブログを対象に，文書間の 2 文に意味的関係のアノテーションを行う英語のコーパスについて検討を進めており [5]，日本語においても同様のコーパスを構築する予定である．

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

表 2: 言明間意味的関係コーパスにおける関係の体系

関係	関係ラベル	定義	例文
論理的関係	同義	A が成り立てば B も成り立つ．また B が成り立てば A も成り立つ（双方向に含意）	A: 死刑は犯罪を予防する力がある B: 死刑制度には犯罪抑止力がある
	矛盾	A が成り立つとき B は成り立たない．また B が成り立つとき A は成り立たない	A: 水産庁は，調査捕鯨を中止すると発表した B: 日本は調査捕鯨は強行すると発表した
	含意	A が成り立つとき B が成り立つ	A: 日本のメタボの 診断基準は女性で胴囲 90cm です B: メタボリックシンドロームの 日本の診断基準は腹囲である
態度関係	意見一致	異なる判断主体が同義，もしくは含意関係の言明を表する	A: 私は捕鯨が正しいとは思えない B: 俺の彼女は捕鯨に反対している
	意見対立	異なる判断主体が矛盾関係の言明を表する	A: 私はマイナスイオンに効果があるとは思えない B: 両親はマイナスイオンは体に良いと考えている
	極性一致	ある側面に対する何らかの評価を A，B が行っており，それらの極性が一致する	A: 私は iPod は持ち運びに便利だと思う B: 自分では iPod の値段は安いと感じる
	極性対立	ある側面に対する何らかの評価を A，B が行っており，それらの極性が一致する	A: 私は iPod は持ち運びに便利だと思う B: 僕は iPod は使い勝手が悪いと感じる
	評価 (P/N/O)	A に対して（P）ポジティブ,（N）ネガティブ,（O）その他の評価，を行う	A: 日本では少子化が進行している B: 少子化は非常に問題であると思う （N 評価)
負例	負例	関係がない，または論理的関係，論争的関係，態度関係に当てはまらないもの	A: 遺族の気持ちを考えれば，死刑制度廃止には反対だ B: 死刑制度廃止を目指す弁護士の活動は怪しく思える

究開発」の一環として実施した．また，ジー・サーチ株式会社の岡田真穂氏，宗意幸子氏，六条範俊氏には意味的関係の定義やコーパス構築方法について多くの示唆を頂いた．記して深く感謝する．

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 Recognising Textual Entailment*, 2005.

[2] Miriam J. Metzger. Makig sense of credibility on the web: Models for evaluating online information and recommendations for future research. *Journal of The America Society for Information Science and Technology*, Vol. 58, No. 13, pp. 2078–2091, 2007.

[3] Marc Moela. Chucking the checklist: A contextual approach to teaching undergraduates web-site evaluation. *Libraries and the Academy*, Vol. 4, No. 3, pp. 331–344, 2004.

[4] Tetsuji Nakagawa1, Takuya Kawada, Kentaro Inui1, and Sadao Kurohashi. Extracting subjective and objective evaluative expressions from the web. In *Proc. of the 2nd International Symposium on Universal Communication (ISUC2008)*, pp. 251–258, 2008.

[5] Eric Nichols, Koji Murakami, Kentaro Inui, and Yuji Matsumoto. Constructing a scientific blog corpus for information credibility analysis. 言語処理学会 第 14 回年次大会, 2009.

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

[7] Keiji Shinzato, Tomohide Shibata, Daisuke Kawahara, Chikara Hashimoto, and Sadao Kurohashi. Tsubaki: An open search engine infrastructure for developing new information access methodology. In *Proc. the 3rd International Joint Conference on Natural Language Processing (IJCNLP2008)*, pp. 189–196, 2008.

[8] Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, Vol. 39, No. 2-3, pp. 165–210, 2005.

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

[10] 佐尾ちとせ, 乾健太郎江口萌. 日本語文のモダリティ・極性情報を捉えるために. 言語処理学会 第 14 回年次大会, 2009.

[11] 乾健太郎, 原一夫. 経験マイニング：web テキストからの個人の経験の抽出と分類. 言語処理学会 第 14 回年次大会, pp. 1077–1080, 2008.

[12] 梅基宏, 杉原大悟, 大熊智子, 増市博. LFG 解析と語彙資源を利用した日本語含意関係判定. 情報処理学会研究報告 2008-NL-188, pp. 57–64, 2008.

[13] 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.

[14] 衛藤純司, 奥村学. 文書横断文間関係タグ付コーパスの構築. 言語処理学会第 14 回年次大会, 2005.

[15] 宮部泰成, 高村大也, 奥村学. 異なる文書中の文間関係の特定. 情報処理学会研究報告 NL-168, pp. 35–42, 2005.

[16] 小谷通隆, 柴田知秀, 中田貴之, 黒橋禎夫. 日本語 textual entailment のデータ構築と自動獲得した類義表現に基づく推論関係の認識. 言語処理学会 第 14 回年次大会, pp. 1140–1143, 2008.

[17] 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 14 回年次大会, 2009.