社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

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

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††

† 奈良先端科学技術大学院大学情報科学研究科
〒 630–0192 奈良県生駒市高山町 8916-5
†† 情報通信研究機構 知識処理グループ
〒 619–0289 京都府相楽群精華町光台 3-5

あらまし　Web 文書には様々な情報が存在し、あるトピックについて多角的な言論などが述べられていることが多い。こうしたトピックに関わる種々の言論は、単純なクエリ検索だけでは広く網羅することができず、偏在する特定の立場の言論を中心的に捉えてしまう危惧がある．本稿では Web 情報中の数的に優勢な立場の言論だけでなく、与えられたトピックに関して存在する多様な言論を抽出し、それらの言論間の類似、対立、含意等の論理的関係を解析してマップ化する言論マップ生成課題について論じる。また、述語項構造レベルの言論間の関係解析について、既存の事態関係知識を利用した予備実験について報告し、言論マップ生成のために必要な個々の技術課題について述べる。
キーワード　言論マップ、含意関係認識、述語項構造、事実性解析

Generating Statement Maps for Capturing Supportive and Contrastive Relations between Statements

MURAKAMI KOJI†, MATSUYOSHI SUGURU†, SUMIDA ASUKA†, MORITA HIRAKU†, SAO CHITOSE†, MASUDA SHOKO†, MATSUMOTO YUJI†, and INUI KENTARO††

† Nara Institute of Science and Technology
8916-5 Takayama-cho, Ikoma city, Nara, 630-0192 JAPAN
†† National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Sagara-gun, Kyoto, 619-0289 JAPAN

Abstract　This paper proposes a novel NLP task, **statement map generation**, which is the task of analyzing logical relations, such as equivalence, entailment and contradiction, between statements involved in a set of documents relevant to a given topic. The aim of the task is to provide Web users with multiple viewpoints on a variety of statements so that they can avoid believing wrong information or disinformation. This paper argues the technical issues to address for modeling statement map generation and reports on the current results of a research project we just launched for this goal.
Key words　Statement Map, recognizing textual entailment, predicate-argument structure, factuality

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、陳腐化した情報などが混在している可能性が非常に高い。そのため、あるトピックに対する言論の集合を俯瞰するためには、ユーザは、個々の言論の**信憑性**や**有効性**を適切に判断する作業を繰り返すことを強いられる。しかし、限られた時間で各言論の信憑性を判断し、言論間の構造を把握することは容易ではない。これらの作業の実行に関してユーザを支援するシステムが必要である。

このような背景により我々は、言論間に存在する類似や対立、正当化などの論理的関係を解析する基盤技術を開発している。以下に「アトピー性皮膚炎にはステロイド剤が日常的に使われている」という言論に対する論理的関係の例を示す。

類似　皮膚科専門医たちは通常アトピー性皮膚炎にはステロイド剤を処方している

含意　アトピー性皮膚炎には免疫抑制剤を使用する
対立　ステロイド剤によるアトピー性皮膚炎の治療はお薦めできません
正当化　皮膚科専門医はステロイド剤の副作用をほとんど経験していない
これらの論理的関係を、**言論マップ**という構造で示すことにより、ユーザーが各言論の信憑性を判断する作業を支援し、情報の偏りや思いこみによる誤信の可能性を抑え、あるトピックに対する言論の俯瞰図を提供することを目指している。

論理的関係によって関係付けられた言論の集合は、それぞれの言論をノード、論理的関係をエッジとしたときに 1 つのグラフを形成する。本稿ではこれを言論マップと呼ぶ。この言論マップの特長は、個々の言論が他の言論との論理的関係性の中に相対的に位置付けられる点であり、これにより各言論の立場や信憑性が判断しやすくなると思われる。河原らの WISDOM [1] は情報内容の信頼性判断を支援するため、あるトピックに関する主要・対立表現を俯瞰的に提示する機構を持っているが、扱っている論理的関係の種類が言論マップより少ない。

本稿では、上記の言論マップを生成するために達成すべき課題について論じ、それぞれの課題に対して現在我々が行なっている取り組みについて説明する。その中でも、特に、単純命題レベルの言論間の関係解析とその予備調査について詳しく報告する。

本稿は次のように構成される。まず 2. 章で関連研究について述べる。次に、3. 章において、言論マップ生成課題について論じ、続く 4. 章で、個々の課題に対する我々の取り組みについて述べる。5. 章において、単純命題レベルの言論間の関係解析とその予備実験について報告し、6. 章で全体をまとめる。

2. 関連研究

2.1 文書横断文間関係解析

複数の文書間の論理的関係解析には、Radev らの Cross Document Theory（CST）[2], [3] がある。この理論は基本的に談話構造解析は単一文書内の構造を解析するものであったが、複数文書を対象とした異なる文書間の構造解析に拡張したものであり、この中で 24 種類の論理的関係が定義された。日本語の文書においては、CST をベースに衛藤ら [4] が日本語に適応した 14 種類の論理的関係を再定義して、文書横断文間関係コーパスを作成した。これらの研究はいずれも複数文要約を目的としており、関係付けの対象は新聞記事である。文書横断文間関係解析モデルに関しては、Zhang ら [5], [6]、宮部ら [7], [8] の研究があるが単語ベースの素性空間における分類器学習に留まっており、語の係り受けや文構造などの情報は用いていない。言論マップ課題においては、対象は新聞記事ではなく更に多様な Web 文書が対象であるため、CST などで規定される論理的関係の他にも特有な関係の認識が必要になることも考えられる。また分類モデルに関しても単語ベースの素性ではなく、文の構造を表現する述語項構造なども用いることで、より多くの情報を利用する。また関係解析には機械学習だけではなく、事象間関係知識や実体間関係知識などの知識を整備して適用する。

2.2 含意関係認識

2 つの言論間の類似関係、あるいは対立関係の判定にはまず、両言論間の共通部分と差異を認識する必要がある。含意関係認識 (Recognizing Textual Entailment:RTE) は、一対のテキストが与えられたときに一方の記述が他方から含意もしくは推論することが出来るかを判別する課題は、Pascal RTE Challenge [9] を契機に注目を集めている研究分野である。RTE では、様々なアプローチを用いて研究が行われており、表層的な情報のみならず、述語項構造解析、関係解析などの深い解析に基づく手法が研究されている。しかしながら言論マップ課題においては、RTE と同様に一対の記述から含意関係の識別が必要不可欠であるが、認識する関係は含意に限らず他の関係も同時に認識する必要がある。また、RTE で利用されるテキスト t は新聞や Web などの実例文を用いているが、仮説 h に関しては必ずしもそうではなく人工的に作られた、ある程度単純な記述であることが多い。それに対して言論マップ課題では、比較する対象の文はすべて Web 上に存在する文であるため、より深い解析などを考慮する必要がある。

2.3 大規模語彙知識獲得

事象間関係知識獲得（Inui [10], Torisawa [11] など）や実体間関係知識（Shinzato [12],Sumida [13],）などの、知識獲得の研究が近年急速に発展している。しかしながら、獲得された知識の精度が実問題に対して、どれくらい適用可能かという実証はまだ殆ど行われていない。こうした知識を言論マップ生成課題に適用することで、さらに利用可能な知識を構築できると考えられる。

3. 言論マップ生成課題

言論マップとは，論理的関係で関係づけられた言論の集合である。ここでは、言論マップ生成のための言論そのものと識別すべき関係、及びマップ生成手順について述べる.

3.1 言論の単位と構成

本研究で対象とする言論とは文を基本単位とし、出来事や行為を参照する命題部分があることが第一の条件である。さらに、文書の書き手もしくは第三者の、真偽や善悪などの心的態度が付加されたものと定義する。本研究課題で取り扱う言論の最小単位は、事態を表す述語とそれに関係のある項からなる述語項構造を基本とし、述語の極性、時間情報（テンス、アスペクト）話者態度（モダリティ）、否定を加えた形で表現する。以下に例を示す。

(1) a. 鯨の数が十分に回復している。: 鯨の数ガ 十分に 回復する（現在・ている）

しなしながら下のような文の場合、 1 つの述語とそれに係る項の情報のみから構成されるため、条件節と帰結節が別々の命題として認識されてしまい、この文の構造を適切に捉えたとは言えない。そこで、条件節や仮定節などが文中に存在する場合は、帰結節の述語と周辺の項情報が単独で命題を構成せずに条件・仮定節を考慮した形式で文の意味構造を反映させる必要がある。これを複合命題と呼ぶ。

(2) a. ステロイド剤を長期に使用すると，皮膚表面の免疫を失調させる：条件［ステロイド剤ヲ 長期に 使用する（現在・)］　皮膚表面の免疫ヲ 失調する（現在・させる）

3.2 言論間の論理的関係

言論マップ生成のために、言論間の論理的関係を認識する必要がある。我々は命題間の論理的関係の定義するために、実際に 10 個のトピックに関する Web 文書集合を用意し、先行研究 [4] を参考に単純命題間の関係を調査した。その結果、以下に示す関係を言論マップ生成課題用の静的な論理的関係として採用した。

(3) a. 同義・類似：ステロイド剤が皮膚炎を抑制する ＝ ステロイド剤で皮膚炎の症状を抑えられる
b. 矛盾・対立・反論：鯨が絶滅の危機にさらされている ⇔ 鯨の頭数は十分である
c. 視点の交替：タバコの煙による健康障害が出る ⊃ タバコは人体に障害を与える
d. 認識：喫煙は健康を損ねる ⊃ タバコの煙による健康被害が知られる
e. 言明：裁判員制度は多くの問題点が指摘される ⊃ 裁判員制度には問題がある
f. 結果：ニコチンが体内に吸収される ⊃ ニコチンを体内に取り込む
g. 例示：裁判員制度の一番の問題は守秘義務です ⊃ 裁判員制度には問題がある
h. 情報付加：裁判員制度が刑事裁判に導入される ⊃ 2009 年 5 月までには刑事裁判に裁判員制度が導入される
i. 狭義の含意：ステロイド剤が皮膚表面の免疫を失調させる ⊃ ステロイド剤には副作用がある
j. 根拠/正当化：皮膚科専門医はステロイド剤の副作用を経験していない → 皮膚科専門医はステロイド剤を日常的

表 1: 事実性タグ付きコーパスの一部

対象文 (下線は対象述語)	時間情報	モダリティー	モダリティー時間
ヘルシア緑茶とヘルシアウォーター、私はウォーターのほうが 続きそうです。	<・, ・, 状態 >	推量	<・, 状態, ・>
かずさん、またレガシィのこといろいろ 教えてください。	<・, ・, 瞬間 >	依頼	<・, 状態, ・>
レガシィに 乗れるのなら それくらいの我慢はへっちゃら！	<・, ・, 瞬間 >	仮定	<・, 状態, ・>

に使っている

3.3 技術的課題

言論マップを生成するためには、以下のような研究に取り組む必要がある。

a）論点抽出 あるトピックで検索された文書集合内において、どのような論点で議論されているかを認識する必要がある。現在、文書集合での頻度ベースで論点を抽出している。

b）意味的な言論抽出 文書から意味的な言論の情報を抽出するためには、文の構造を捉える述語項構造解析の他に、照応・省略解析 [14] が必要となる。書き手の態度を表すムード情報を適切に捉える事実性解析が重要である。事実性解析に詳しくは 4.1 で述べる。

c）知識の整備

述語項構造辞書 事象間の関係認識を目的として、事象間関係知識をもつ述語項構造辞書を作成する。4.3 節で詳しく説明する。

事象に関する知識 言論が示す意味を適切に捉えるには、言論内の事象を示す格要素を伴う述語もしくは事態性名詞を正しく解析する必要がある。原らは [15] はこうした事象の事実性解析システムを開発した。我々は 4.1 節で詳しく述べるが、このシステムを改良することで、高精度な事実性解析を行う研究を進めている。また、人手による知識の整備だけではなく、大規模なコーパスからの事象間関係知識 [16] の獲得も利用する。

実体に関する知識 言論内の実体を示す名詞についても、関係認識を行う必要がある。

(4) a. 副流煙は健康に甚大な被害を与えると発表された
b. タバコの煙による健康侵害が報告される

この例では、「副流煙」と「タバコの煙」が同一の実体を示すことを認識しなければこの 2 つの言論間の類似関係を認識することが出来ない。我々は、Torisawa ら [11] の知識獲得を元に更に高精度の上位下位関係知識を作成する。詳細は 4.2 節で述べる。

d）論理的関係解析 2. で述べたように、論理的関係の解析については RTE などの研究がある。外部知識と学習ベースの統計的手法を組み合わせた柔軟なアライメントなど、興味深い研究課題が多数ある。

しかしながら実文の構造は複雑なため、論理的関係の解析は条件・仮定節を省略することで単純な述語項構造に変換しその中で論理的関係解析を行うことから始める。具体的には 4 節で述べる。

4. 個々の課題に対する取り組み

前章で言及した技術的課題のうち、現在、我々が取り組んでいる課題について報告する。

4.1 事実性解析器

我々の最終目標は、3. で定義した複合命題レベルの言論マップを作成することである。これを達成するためには、少なくとも次の 2 つの処理が必要である。

（1） 述語の後に続く文末表現を解析する
（2） 従属文が仮定節であるかどうか判定する

我々は、原らが定義した事実性に関するタグ体系 [15] と GRMM ツール [17] を用いることにより、与えられた述語に対してその事実性を解析するシステムを構築している。このシステムは、上記の 2 つの処理を同時に行なうことができるという特長を持っている。現在、次の 2 つの方向から、高精度で頑健な事実性解析器の実現に向けた研究を進めている。

表 2: 隅田らの実体間上位下位関係知識の内訳

抽出元データ	知識数
HTML 文書	627,791
日本語 WordNet v0.6	876,861
Wikipedia のカテゴリ体系	642,695
Wikipedia の階層構造など	1,885,502
異なり数	3,534,357

表 3: 隅田らの実体間上位下位関係知識の例

上位語	下位語
そば焼酎	雲海
医科大学	千葉医科大学
蒸留酒	ラム酒
米加工品	日本酒

事実性タグ付きコーパスの開発 原らのタグ体系を見直すとともに、ブログから抽出した文に存在する約 7,000 の述語に対してタグ付け作業を行なっている。このタグ付きコーパスの一部を表 1 に示す

有効な素性の発見 GRMM を用いた学習において様々な素性の組み合わせを試行し、解析システムの精度向上を目指している

これらの成果に基づく、我々の事実性解析システムとその評価実験の詳細については、別稿 [18] で報告する予定である。

4.2 実体間上位下位関係知識

3.2 節で述べたように、言論間の関係を捉えるためには、実体と事象に関する大規模な知識を利用する必要がある。例えば、実体に関するシソーラスを用いれば、そこに記述される親子のつながりにより実体間含意関係を、姉妹の関係により実体間対立関係を認識することができる。

隅田らは、次の 4 つのデータを合わせることにより、3,534,357 件の実体間上位下位関係知識を作成した。

（1） 括弧を用いた統語パターンや名詞連続を利用して HTML 文書から抽出した上位下位関係 [13]
（2） 日本語 WordNet v0.6 [19] に記述されている上位下位関係
（3） Wikipedia 日本語版(注1)のカテゴリ体系による上位下位関係
（4） Wikipedia 日本語版に存在する階層構造・定義文・カテゴリから抽出した上位下位関係 [20], [21]

関係知識の内訳を表 2 に、関係知識の例を表 3 に示す。

隅田らの実体間関係知識を利用することにより、名詞の観点から言論間の類義・含意関係を判定することができる。例えば、表 3 の 1 つめの関係知識を用いると、次の 2 つの言論が類義関係にあることが分かる。

- 彼は そば焼酎 が大好きだ
- 彼は 雲海 が大好きだ

「バイオエタノール」と「少子化問題」という 2 つのトピックに関して、上記の実体間関係知識のみを用いて言論マップを作成する予備調査を行なった結果、入力された言論の約 6 割に対して類義関係を認定することができた。後の 5 章の予備調査においては、まだこの実体間関係知識を利用していないが、これらを利用することにより、クラスタリングの再現率が向上することが期待される。

(注1)：http://ja.wikipedia.org/wiki/

4.3 事象間関係知識を持つ述語項構造辞書

ある事象と別のある事象の間に存在する関係を認識するためには、それらの関係に関する大規模な言語資源が必要である。これまでに我々のグループでは、次の 2 つの言語資源を作成してきた。

事象間関係知識　岩波国語辞典 [22] の述語に対する語釈文を利用して、人手により作成した事象間関係知識 [23], [24]

シソーラス　語彙概念構造による動詞意味分析の枠組み [25] に基づいて、Lexeed [26] に存在する高頻度動詞約 4,000 語、約 7,000 語義に対して、項構造と意味クラスを記述した述語項構造辞書 [27]

本研究では、これら 2 つの言語資源を計算機で利用しやすい形に変換して統合し、事象間関係知識を持つ述語項構造辞書を編纂した。

この述語項構造辞書は、1 つの XML 形式のファイルであり、ある述語に対して、語義ごとに、以下に示す情報が記述されている。

ID　この辞書における ID

見出し語　ひらがなで記述された見出し語

表記　送り仮名のゆれなどを含む漢字表記やカタカナ表記

意味クラス　5 階層からなる述語の意味クラス。最下層には約 1,000 のクラスが存在する

出現頻度　Senseval-2 コーパス [28] における出現頻度

岩波国語辞典 ID　岩波国語辞典 [22] における ID

LexeedID　Lexeed [26] における ID

項構造　述語がとる必須項のリスト。それぞれの項に対して、次のような情報が記述されている

- **変項**　関係知識において、後件の項と対応づけるために用いる
- **定項**　格の前に現れる語が確定している場合に、その語を記述する
- **深層格**　表層格に対する深層格
- **項番号**　半統制意味構造記述 [27] における項番号
- **格の交替**　「に」→「へ」のように、代わりに用いることができる格
- **例**　格の前に現れる語の例

事象間関係知識　事象間関係知識を表す、関係と後件の述語項構造の対のリスト

この辞書におけるエントリーの例を表 4 に示す。現在、編纂した辞書には、29,555 エントリーが登録されており、計 45,907 個の事象間関係知識が含まれている。それぞれの関係に対する知識数を表 5 に示す。

表 4: 編纂した辞書におけるエントリーの例

ID	04992
見出し語	きゅうしゅうする
表記	吸収する
意味クラス	状態変化あり. 位置変化. 位置関係の変化（物理）. 吸入／排出. 吸入
出現頻度	0
岩波 ID	0011538-0-0-0-0-x0
lexeedID	06027950-4
が格	変項: 何か A; 深層格: causer; 項番号: 2; 例: 水が
を格	変項: 何か B; 深層格: 対象; 項番号: 1; 例: 二酸化炭素を
関係知識	同義 (言い換え): ＜何か A＞が＜何か B ＞を＜何か A＞に吸い込む
関係知識	同義 (言い換え): ＜何か A＞が＜何か B ＞を吸い取る
関係知識	付帯状況: ＜何か A＞が＜何か B ＞を取り入れる
関係知識	同義・上位: ＜何か A＞が＜何か B ＞を自分のものとする

表 5: 事象間関係知識の数

関係名	英訳	知識数
同義 (言い換え)	near synonym	17,816
同義・上位	hypernym	11,487
反義語	antonym	540
前提条件	presupposition	3,037
結果 (状態)	effect	2,163
付帯状況	cooccur	4,274
不可分	inseparable	174
原因・理由	cause	2
目的	goal	882
手段	means	5,532
計		45,907

4.4 言論マップ評価コーパス

言論マップ生成システムの出力を評価するためには、その正解出力例となる評価コーパスが必要である。現在、我々は、**単純命題**レベルの言論に対するこのような評価コーパスを構築する作業を進めている。本稿では、単純命題を次のように定義する。

単純命題 = 項構造 + 述語 + 助動詞

ここで、助動詞としては、次の機能を持つもののみを認める。

否定: ない、ぬ、受け身: (ら) れる、使役: (さ) せる

言論マップ評価コーパスは、少なくとも次の 2 つの要請を満たす必要がある。

要請 1　コーパスに含まれるすべての言論に対して、それが属する言論クラスターが明示してある

要請 2　言論クラスター間に存在するすべての関係が、その関係名とともに明示してある

我々は、このようなコーパスを構築するために、1 つのトピックに対して、論点ごとに、表 6 に示すような形式の、1 行が 1 言論であるデータを用意した。このデータに存在するすべての言論に対して、「クラスター番号」を付与することにより、上記の要請 1 を満たすコーパスを作成することができる。また、同時に、この作業により、言論クラスター集合が確定するので、任意の 2 つの言論クラスターの間に関係が存在するかどうか検証し、それを「クラスター番号 × 関係名」の表に記述することにより、このコーパスは上記の要請 2 を満たすようにすることができる。

言論マップ評価コーパス構築作業手順を以下に示す。

（1）　表 6 の形式のデータに存在する言論のうち、項が不自然に省略されているものに対して、「不完全」フラグを立て、データから除去する

（2）　個々の言論に対して、その項構造、述語、原文を観察し、「クラスター番号」を付与する。このとき、類義の言論に対して同じクラスター番号を付与する

（3）　個々の言論クラスターの中から、そのクラスターを代表する言論を 1 つ選び、その言論に「代表」フラグを立てる

（4）　言論クラスターの間に関係が存在するかどうか検証し、それを「クラスター番号 × 関係名」の表に記述する

現在、上記のコーパス構築作業を、言語学を専攻している大学院学生 1 名が行なっており、次の 0 つのトピックに対する計 1000 個の論点に対して作業が終了した。これに対する平均作業時間は、1 トピックあたり約 2000 時間であった。

温暖化、裁判員制度、喫煙、ステロイド、還元水

この作業においては、元のブログ記事中の出現頻度が 2 以上の言論のみを対象とし、言論クラスター間の関係としては、次の 10 個の関係を定義して用いた。

対立、視点の交代、認識、言明、確信度の強弱、未実現、結果、例示、情報付加、狭義の含意

表 6: 評価コーパス構築のためのデータの一部 (トピック: ステロイド、論点: 医師)

言論 ID	項構造	述語	クラスタ番号	代表	不完全	原文 ID	原文
108	医師から ステロイドを	処方する れる				100381	医師からステロイドを処方され、 ···
198	ないと 医師は	信じる				569102	力はあまりないと医師は信じて ···

5. 単純命題レベルの言論間の関係解析

3. 章で述べたように、我々の最終目標は、複合命題レベルでの言論マップを作成することである。これを実現するための第一歩として、我々は、単純命題レベルの言論マップ生成システムを試作した。この章では、我々が作成した言論マップ生成システムについて説明し、その予備実験について述べる。

5.1 言論マップ生成システム

我々が試作した言論マップ生成システムは、入力されたクエリーから、各論点ごとに単純命題レベルの言論マップを生成し、それらを出力する。このとき、主に類義関係知識を用いることにより言論をクラスタリングし、反義関係知識を用いることにより言論クラスター間に反義という論理的関係を導入する。言論マップ生成アルゴリズムを以下に示す。

（1） クエリーから関連文書集合を得る

ウェブ検索エンジン TSUBAKI [29] を用いて、クエリーが少なくとも 1 度以上出現する文書を収集する

（2） 関連文書集合から論点のリストを獲得する

名詞句を対象とした用語抽出アルゴリズム [30] を利用して用語集合を獲得し、そこからサ変名詞を除去した後、出現頻度が高い上位 10 語ほどを人手で選択し、論点として用いる

（3） 関連文書集合から単純命題を抽出する

Cabocha(注2) を用いて文書を解析し、その解析結果から、動詞もしくは形容詞を中心とする述語項構造を抽出する。このとき、用言はすべて基本形に変換する。否定、受け身、使役を表す助動詞が述語句内に存在した場合、それらも合わせて抽出する。例えば、「喫煙がやめられない」から、「喫煙が やめる られる ない」を抽出する

（4） 否定表現にタグを付ける

単純命題内に存在する否定を表す助動詞「ない」、「ん」、「ぬ」に対して、< 否定 > というタグを付与する

（5） とりたて助詞を標準化する

単純命題の項にとりたて助詞「は」、「も」が存在する場合、それらを取り除く。必要ならば、関連文書集合における「直前名詞+格助詞+述語」の出現頻度が一番高い格助詞を選択することにより、とりたて助詞があった位置に「が」や「を」を補う。例えば、「ニコチン は 脳に及ぼす」を「ニコチン が 脳に及ぼす」に標準化する

（6） 項の順番を無視し、表記が全く同じ単純命題をまとめる

（7） 項構造を包含する単純命題をまとめる

実際の文書集合においては、述語の必須格が省略されていることが多々あるので、不完全な述語項構造の削減を目的としてこの処理を行なう。例えば、「ニコチンを 取り入れる」を「ニコチンを 体内に 取り入れる」とまとめる。包含する単純命題が複数存在する場合は、そのうち、最も頻度が高い単純命題とまとめる

（8） 格の交替関係にある項を持つ単純命題をまとめる

例えば、「二酸化炭素が 大気中 で 増える」と「二酸化炭素が 大気中 に 増える」をまとめる

（9） 関係知識を用いて単純命題をクラスタリングする

4.3 節で述べた述語項構造辞書に含まれる類義関係知識(注3) に基づいて、単純命題をクラスタリングする。ここでは、同じ意味クラスに属するという知識と表 5 における次の 8 種類の知識を、類義関係知識と定義して用いる。

同義 (言い換え)、同義・上位、結果 (状態)、付帯状況、不可分、原因・理由、目的、手段

この処理により、例えば、次の 3 つの単純命題が 1 つの言論クラスターにまとめられる。

- ニコチンを 体内に 入れる
- ニコチンを 体内に 吸収する
- ニコチンを 体内に 取り入れる

（10） 態の変換関係にある言論クラスターをまとめる

単純命題の述語句内に、受け身もしくは使役を表す助動詞が存在する場合、項構造を検証することにより、2 つの言論クラスターをまとめる。例えば、「ニコチン が 含む れる」を含む言論クラスターを、「ニコチン を 含む」を含む言論クラスターとまとめる

（11） 言論クラスター間に反義関係を認定する

単純命題に含まれる否定のタグと 4.3 節で述べた述語項構造辞書が持つ反義関係知識に基づいて、2 つの言論クラスターの間に反義関係があるかどうか検証する。ここでは、反義の意味クラスに属するという知識と、表 5 における反義語関係知識を、反義関係知識と定義して用いる。この処理により、例えば、「京都議定書に 賛成する ない < 否定 >」を含む言論クラスターと「京都議定書に 賛成する」を含む言論クラスターの間に反義関係を認定する。また、別の例として、「ニコチンを 含む」を含む言論クラスターと「ニコチンを 放出する」を含む言論クラスターの間に反義関係を認定する。

（12） 言論クラスターのフィルタリングを行なう

その中に 1 つの単純命題しか含まず、かつ、それと反義関係にあるクラスターが存在しない場合、その言論クラスターを除去する。ただし、その中の単純命題の出現頻度が閾値以上の場合、除去しない

5.2 予備調査

正解データ中で反義、同義・類義関係が付与された命題を、事象間関係知識によってその関係を再現できるかを予備的に調査した。

5.2.1 調査設定と結果

実験に用いたデータは、「喫煙」(論点：肺がん、害、ニコチン、健康、危険性、リスク、マナー、女性、問題、受動喫煙、禁煙)、「ステロイド」(論点：医師、皮膚、かゆみ、ステロイド剤、ステロイド軟膏) の 2 トピック、16 論点である。

正解データは 4.4 で説明したコーパスである。この中には、10 種類の関係が定義されているが、その中で反義関係と、同義・類義関係のみを用いた。言論マップ生成には関係を有する命題だけを利用するため、正解データのうち関係が付与されている命題だけを対象にした。これにより、「喫煙」では 839、「ステロイド」では 777 命題が正解の命題数となる。類似関係の認識は、事象間関係知識を用いて正解命題を再現できた割合を求める。反義関係の認識は、正解データ中で反義関係が付与されているクラスタを知識により再現できた割合を求める。16 論点での単純命題の類似・反義関係の認識精度を表 7 に示す。表の項目はそれぞれ、各トピックについて論点数、類義関係の再現率、および反義関係の再現率である。

類義、反義関係共に再現率に低い値となった。この要因の 1 つは、利用した事象間関係知識ではまだ知識が記述されていない、動詞「なる」が広範囲に高頻度で出現したことである。以下の例は動詞「なる」を含む、類似関係の認識誤りである。

(5) a. ニコチンは 血流を 悪化させる ⇔ ニコチンで 体内血流が 悪く なる

b. 喫煙で ニコチン濃度が 高まる ⇔ 喫煙で 体内ニコチン濃度が 高く なる

(注2)：http://chasen.org/ taku/software/cabocha/

(注3)：4.2 節で述べた実体間関係知識はまだ利用していない

表 7: 単純命題の類似・反義関係の認識精度

トピック	論点数	Recall(sym)	Recall(ant)
喫煙	11	0.319 (268/839)	0.266 (25/95)
ステロイド	5	0.422 (328/777)	0.114 (4/35)

こうした関係認識誤りは、事象間関係知識を増やすことで対応できると考えられる。

現在は 1 つ以上の項が同一で、かつ述語中の動詞間に類義関係がある場合のみ類義関係と認識している。しかしながら、正解データは述語の類似関係だけではなく、項構造にも着目して総合的に命題の類似関係を付与しており、項の上位下位関係や類義語により類似関係となっている命題が多く見られた。

(6) a. ニコチン依存症に 喫煙者が 陥る ⇔ ニコチン依存に 陥る

b. ニコチン依存に なる ⇔ ニコチン中毒に なる

これらの命題は、名詞の類義語や実体間関係知識などを用いて名詞の観点から類義判定を行うことで、命題全体でも類義関係を認識することができると考えられる。

6. おわりに

本稿では、言論マップ生成課題について我々の方針と、現在取り組んでいる個々の課題について述べた。また、単純命題の述語を対象として事象間関係知識を用いて命題の類似・反義関係認識の予備実験を行い、取り組むべき個々の課題について簡単に考察した。今後は事実性解析、実体上位下位関係知識を利用して言論マップ作成を行う予定である。

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究開発」の支援の下に実施した。

文　　献

[1] 河原, 黒橋, 乾：“主要・対立表現の俯瞰的把握ーウェブの情報信頼性分析に向けて”, 情報処理学会研究報告 2008-NL-186, 2008-NL-186 (2008).

[2] D. R. Radev: “Common theory of information fusion from multiple text sources step one: Cross-document structure”, Proc. the 1st SIGdial workshop on Discourse and dialogue, pp. 74–83 (2000).

[3] D. R. Radev, J. Otterbacher and Z. Zhang: “Cst bank: A corpus for the study of cross-document structural relationships”, Proc. the 4th International Language Resources and Evaluation (LREC'04) (2004).

[4] 衛藤, 奥村：“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).

[5] Z. Zhang, J. Otterbacher and D. R. Radev: “Learning cross-document structural relationships using boosting”, Proc. the 12th International Conference on Information and Knowledge Management, pp. 124–130 (2003).

[6] Z. Zhang and D. Radev: “Combining labeled and unlabeled data for learning cross-document structural relationships”, Proc. the Proceedings of IJC-NLP (2004).

[7] 宮部, 高村, 奥村：“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).

[8] 宮部, 高村, 奥村：“文書横断文間関係の特定”, 言語処理学会 第 12 回年次大会, pp. 496–499 (2006).

[9] I. Dagan, O. Glickman and B. Magnini: “The pascal recognising textual entailment challenge”, Proc. of the PASCAL Challenges Workshop on Recognising Textual Entailment (2005).

[10] T. Inui, K. Inui and Y. Matsumoto: “Acquiring causal knowledge from text using the connective marker tame”, ACM Transactions on Asian Language Information Processing, **4**, 4, pp. 435–474 (2005).

[11] K. Torisawa: “Acquiring inference rules with temporal constraints by using japanese coordinated sentences and noun-verb co-occurrences”, Proc. the NAACL (2006).

[12] K. Shinzato, S. Sekine, N. Yoshinaga and K. Torisawa: “Constructing dictionaries for named entity recognition on specific domains from the web”, Proc. the Web Content Mining with Human Language Technologies workshop on the 5th International Semantic Web (2006).

[13] A. Sumida, K. Torisawa and K. Shinzato: “Concept-instance relation extraction from simple noun sequences using a search engine on a web repository”, Proc. the Web Content Mining with Human Language Technologies workshop on the 5th International Semantic Web (2006).

[14] R. Iida, K. Inui and Y. Matsumoto: “Exploiting syntactic patterns as clues in zero-anaphora resolution”, Proc. of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL), pp. 625–632 (2006).

[15] 原, 乾：“事態抽出のための事実性解析”, 情報処理学会研究報告 2008-FI-89, 2008-NL-183, pp. 75–80 (2008).

[16] S. Abe, K. Inui and Y. Matsumoto: “Two-phased event relation acquisition: Coupling the relation-oriented and argument-oriented approaches”, Proc. of the 23rd International Conference on Computational Linguistics (COLING) (2008). (to appear).

[17] C. Sutton: “GRMM: A Graphical Models Toolkit” (2006). http://mallet.cs.umass.edu.

[18] 森田, 松吉, 佐尾, 松本, 乾：“テキスト情報の事実性解析”, 第 7 回情報科学技術フォーラム (FIT2008) 発表論文集 (2008).

[19] F. Bond, H. Isahara, K. Kanzaki and K. Uchimoto: “Boot-strapping a wordnet using multiple existing wordnets”, Proc. the 6th International Language Resources and Evaluation (LREC'08) (2008).

[20] A. Sumida, N. Yoshinaga and K. Torisawa: “Boosting precision and recall of hyponymy relation acquisition from hierarchical layouts in wikipedia”, Proc. the 6th International Language Resources and Evaluation (LREC'08) (2008).

[21] 隅田, 吉永, 鳥澤, 萬成：“Wikipedia からの大規模な上位下位関係の獲得”, 言語処理学会第 14 回年次大会, pp. 769–772 (2008).

[22] 西尾, 岩淵, 水谷：“岩波国語辞典第五版”, 岩波書店 (1994).

[23] 青山, 阿部, 大西, 乾, 松本：“事態間関係の獲得のための動詞語釈文の構造化”, 言語処理学会 第 13 回年次大会発表論文集, pp. 286–289 (2007).

[24] 大西, 乾, 松本：“事態間関係知識の整備と含意文生成への応用”, 言語処理学会第 14 回年次大会発表論文集, pp. 1152–1155 (2008).

[25] 影山：“動詞の意味と構文”, 大修館書店 (2001).

[26] 笠原, 佐藤, 田中, 藤田, 金杉, 天野：“「基本語意味データベース:lexeed」の構築”, 情報処理学会研究報告 2004-NL-159, pp. 75–82 (2004).

[27] 竹内, 乾, 竹内, 藤田：“意味の包含関係に基づく動詞項構造の細分類”, 言語処理学会 第 14 回年次大会発表論文集, pp. 1037–1040 (2008).

[28] 黒橋, 白井：“Senseval-2 日本語タスク”, 電子情報通信学会技術研究報告 NLC2001-36, pp. 1–8 (2001).

[29] K. Shinzato, T. Shibata, D. Kawahara, C. Hashimoto and S. Kurohashi: “Tsubaki: An open search engine infrastructure for developing new information access methodology”, Proc. the 3rd International Joint Conference on Natural Language Processing (IJCNLP2008), pp. 189–196 (2008).

[30] 村上, 乾, 橋本, 内海, 石川：“専門用抽出における助詞情報の利用に関する一考察”, 情報処理学会 研究報告 NL-182, pp. – (2007).