E-053

# 言論マップ生成のための事象間類似・対立関係の認識
# Recognizing Synonymous and Antonymous Relations between Events for Generating Statement Map

村上 浩司[†] 松吉 俊[†] 増田 祥子[†§] 松本 裕治[†] 乾 健太郎[‡]
Murakami Koji Matsuyoshi Suguru Masuda Shoko Matsumoto Yuji Inui Kentaro

## 1. はじめに

情報検索技術の発展により、あるトピックに関連するウェブ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、これらの文書には不正確な記述、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、あるトピックに関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、可視化を行う**言論マップ**を開発している [1]。これにより、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。

本稿では、上記の言論マップを生成するためにまず必要となる、単純命題レベルの言論間の類似、対立関係の認識と、その予備調査について報告する。

## 2. 関連研究

複数文書間の論理的関係解析には、Radev らの Cross Document Structure Theory (CST) [2, 3] がある。日本語の文書においては、CST をベースに衛藤ら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 [4] を作成した。これらの研究はいずれも関係づけの対象は新聞記事である。

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[5] 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。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。言論マップ生成では、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。また RTE において認識すべき関係は現状においては t が h を含意、矛盾するかの判定である。河原らの WISDOM[6] においても、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ものである。これらに対し言論マップ生成においては、RTE や WISDOM と同様に一対の記述から含意、対立関係の識別が必要不可欠であるが、認識する関係はそれらに限らず他の関係も同時に識別する必要がある。現在我々は、Web 上の実文を対象とし、[4] で定義されている関係を元に 10 種類の言論間の論理的関係を定義した言論マップ評価コーパス [1] を作成中である。

[†]奈良先端科学技術大学院大学
[‡]独立行政法人 情報通信研究機構
[§]大阪府立大学大学院

## 3. 単純命題レベルの言論間の関係解析

### 3.1 単純命題

Web 文書中の実文の構造は複雑なため単純な構造に変換し、その中で類義・反義関係解析を行う。我々は述語項構造に否定、受け身、使役の助動詞を加えたものを「単純命題」と定義し、本稿で扱う言論の単位とする。

### 3.2 言論マップ生成システム

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。このシステムは、入力されたクエリから各論点ごとに単純命題レベルの言論マップを生成、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間に反義という論理的関係を導入する。言論マップ生成過程を簡単に示す。生成過程の詳細は我々の報告 [1] を参照されたい。

1. クエリから関連文書集合を取得
2. 関連文書集合から論点のリストを獲得
3. 関連文書集合から単純命題を抽出
4. 否定表現にタグを付与
5. とりたて助詞を標準化
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
8. 格の交替関係にある項を持つ単純命題をまとめる
9. 述語項構造辞書 [7] の類義関係知識を用いて単純命題をクラスタリング
10. 態の変換関係にある言論クラスタの結合
11. 言論クラスタ間に反義関係を認定
12. 言論クラスタのフィルタリング

### 3.3 予備調査

正解データの反義・類義関係が付与された命題を、述語項構造辞書によってその関係を再現できるかを予備実験により検証した。

#### 3.3.1 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、受動喫煙、禁煙)、「ステロイド」(論点：医師、皮膚、かゆ

表 1: 単純命題の類義・反義関係の認識精度

| トピック | 論点数 | Recall(sym) | Precision(sym) | Recall(ant) |
|---|---|---|---|---|
| 喫煙 | 10 | 0.192 (192/999) | 0.627 (192/306) | 0.266 (25/95) |
| ステロイド | 5 | 0.342 (287/839) | 0.634 (287/452) | 0.114 (4/35) |

み、ステロイド剤、ステロイド軟膏）の 2 トピック、15 論点である。

正解データは予め作成した言論マップ評価コーパス [1] 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 999、「ステロイド」では 839 命題が正解の命題数となる。類義関係の認識は、述語項構造辞書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述語項構造辞書により再現できた割合を求める。

15 論点での単純命題の類義・反義関係の認識精度を表 1 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、精度及び反義関係の再現率である。類義、反義関係共に再現率は低い値となった。再現率が低い原因の 1 つは、正解データは作業者が単純命題の項と述語のすべてに着目して総合的に類義・反義関係を記述しているのに対し、システムは述語項構造辞書のみを利用していることから、命題間の項構造の上位下位関係、類似関係などを考慮していないことである。システムが類義関係を認識できなかった例を示す。

(1) 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

ここでは、「ニコチン依存症⇔ニコチン依存」、「ニコチン依存⇔ニコチン中毒」の関係を認識する必要がある。名詞の類義語や実体間関係知識 [8] などを用いて名詞の観点から論理的関係の判定を行うことで、命題全体の類義関係を認識することができると考えられる。

また他の原因として、利用した述語項構造辞書ではまだ記述されていない、動詞「なる」が広範囲に出現したことである。以下の例は動詞「なる」を含む、類義関係の認識誤りである。

(2) 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

こうした関係認識誤りは、述語項構造辞書を増強することで対応できると考えられる。

この実験で用いた、「喫煙」トピック中の「ニコチン」を論点としたときの言論マップイメージの一部を図 1 に示す。図中のクラスタは、類義関係と認識された単純命題により構成される。また、異なった 2 つのクラスタにそれぞれ属する命題の間に反義関係がある場合、それらの命題が属するクラスタ間に反義関係を認定した。この例では、「ニコチンを 放出する」と「ニコチンを 体内に 入れる」の間に反義関係が認識されたため、それらを含むクラスタを反義関係とした。

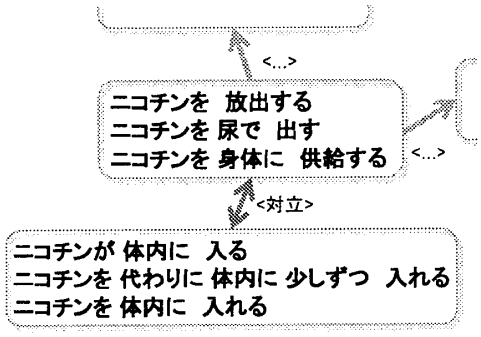


図 1: 言論マップイメージ

## 4. おわりに

本稿では言論マップ生成にむけた、単純命題の述語を対象とした述語項構造辞書を用いて単純命題の類似・反義関係認識について述べた。予備実験を行い、60%程度の精度でこれらの論理的関係の認識結果を得た。今後は実体上位下位関係知識 [8] を利用した単純命題の論理的関係認識、事実性解析 [9] を導入した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。

## 参考文献

[1] 村上, 松吉, 隅田, 森田, 佐尾, 増田, 松本, 乾：“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, 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] I. Dagan, O. Glickman and B. Magnini: “The pascal recognising textual entailment challenge”, Proc. of the PASCAL Challenges Workshop on Recognising Textual Entailment (2005).

[6] 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).

[7] 松吉, 村上, 松本, 乾：“含意・矛盾認識のための事象間関係知識の整備”, 第 7 回情報科学技術フォーラム (FIT2008) 発表論文集 (2008).

[8] 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).

[9] 森田, 佐尾, 松吉, 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).