# 主題連鎖と文モダリティの分類に基づく文章の論理展開

Logic Structure of Text Based on Thematic Chain and Categorization of Modality

藤田 彬†　鈴木春菜‡　楊 華†‡　前川眞一‡‡　田村直良†‡‡

横浜国立大学大学院環境情報学府†
横浜国立大学教育人間科学部‡
東京工業大学大学院社会理工学研究科†‡, ‡‡
横浜国立大学大学院環境情報研究院†‡‡
E-mail: {fujita† hsuzuki‡}@tamlab.ynu.ac.jp　you.k.aa@m.titech.ac.jp†‡
mayekawa@hum.titech.ac.jp‡‡　tam@ynu.ac.jp†‡‡

## 1　はじめに

本稿では、文間の主題連鎖と文のモダリティの分類に基づいて、論説文における論理展開を把握する手法について述べる。

近年、e-learning システムのように無人でシステムとのインターラクションを進めていく形態が求められているが、その中で自由記述式問題の自動評価に対するニーズが高まっている。与えられた問題文に対して回答者が意見を述べる論説文を小論文と呼ぶが、小論文の自動評価システムは自由記述式問題の自動評価の中でも特にニーズが高く、活発に研究が進められている。

小論文の自動評価に関する先行研究は[1]のサーベイが詳しい。代表的なシステムとして、英文を対象にした ETS の e-rater などが挙げられる[2]。また、和文を対象にしたシステムも、石岡らが e-rater を参考に Jess というシステムを開発している[3]。しかし、Jess では、接続表現のみを手掛かりにして文脈を把握するなど、筆者の主張についての論理の展開を評価の観点に加えることに対して十分に配慮されていないのが現状である。我々はこの状況に対し、主題連鎖を捉えることで筆者の主張についての論理の展開を把握して評価する手法を検討した[5]が、論理の展開の把握に関して精度に不十分な点がある。

そこで本研究では、論説文の文脈理解の精度向上に向けて、文章の結束性を支える主題の連鎖（**主題連鎖**）を捉えた上で、主題連鎖に含まれる各文における筆者の陳述態度（**文モダリティ**）を考慮する手法を用いて、「文章中である話題を提起し、論理的な誘導を行い、結論づける」という一連の論理の展開（**論理展開**）を捉える手法を検討する。

文のモダリティのうち意見モダリティをもつ文は、先行研究において主に「意見文」と呼ばれる。ある文が意見文であるか否かを自動判定する研究は Web からの評判情報抽出など様々な目的で行われており、サーベイがある[4]。これらの研究を参考に、我々は高い精度で文のモダリティを判定する手法を示している[6]。本研究では、まず[5]で示した手法で文間の主題連鎖を把握することで、文の連鎖構造を捉える。次に[6]で示した手法で連鎖構造中に含まれる文のモダリティを同定する。このようにして、文章中にどのような文モダリティの連鎖が含まれているかを考察して、論理展開のパターンを把握する。

以下、本稿では第 2 章で文モダリティの分類手法について述べる。第 3 章では、文間の主題連鎖について述べる。第 4 章では、文モダリティを含む文間の連鎖構造について述べる。第 5 章では、専門家による評価が高い文章と低い文章の間での、文モダリティを含む文間の連鎖構造の違いについての考察し、本研究で提案する手法の有用性を示す。

## 2　文モダリティの分類

### 2.1　意見文と叙述文

本研究における意見文と叙述文の定義について述べる。意見文は、文脈上で筆者の意見が述べられている文と定義する。また叙述文は、意見文ではない文と定義する。論説文では、意見性を持つ表現が使用された文であっても、意見文であるか叙述文であるかは文章の主題に依存する文がある。

例えば、以下に示す文は単独で意見文であるか叙述文であるかを判断することが不可能である。

(1) *リンゴは赤い。*

(1)が含まれる文章の主題がリンゴの色と直接関連のないものである限り、(1)は文脈上で筆者の意見が述べられている文ではなく、叙述文と捉えられる。ところが、(1)が属する文章の主題が「リンゴの色をどう表現するか」に関して論ずるものであれば、(1)は文脈上でリンゴの色の表現に関して筆者の意見を陳述する意見文である。

このように、論説文に含まれる文の意見性について判断を行う際は、文脈を考慮する必要がある。

## 2.2 文モダリティ・カテゴリ

本研究では、文脈を詳細に捉えるために、意見文を以下の 2 つの観点から 4 種類に分類する。

表 1：文モダリティ・カテゴリの分類

| | | 特定陳述相手 | |
|---|---|---|---|
| | | 有 | 無 |
| 特定対象 | 有 | **カテゴリ A** | **カテゴリ B** |
| | 無 | **カテゴリ C** | **カテゴリ D** |

1） 特定陳述相手の有無

筆者の意見陳述が成立するために意見を訴えかける相手を**特定陳述相手**と呼ぶことにする。

(2)*無意識に英語を学ぶのではなく自分の将来を考えてほしい。*

（特定陳述相手、有）

(3)*小学校における英語の早期教育は必要である。*

（特定陳述相手、無）

(2)では、「考えてほしい」と要求しているため（「要求」）、要求する相手（特定陳述相手）が存在しなければ意見陳述が成立しない。この他にも、「問い掛け」、「疑問」、「提案」、「要望」等のモダリティを持つ意見文は、意見を陳述する相手である特定陳述相手が存在しない限り、意見陳述が成立しないという共通点を持つ。一方(3)では、特定陳述相手が存在しなくても意見陳述が成立する。

2） 特定対象の有無

筆者が陳述する意見が言及している対象を**特定対象**と呼ぶことにする。

(4)*初等教育は人間の基礎をつくる重要な過程である。*

（特定対象、有）

(5)*国際人を育てるのであれば、まず日本語を身につけさせるべきだ。*

（特定対象、無）

(4)では、「初等教育」という特定対象に関して筆者の印象を述べている（「印象」）。一方(5)では、特定対象が示されていない。

この他にも、「評価」、「感想」、「賛否の表明」、「問い掛け」、「疑問」等のモダリティを持つ意見文は、意見を陳述する対象である特定対象が存在しない限り、意見陳述が成立しないという共通点を持つ。

以上の 2 つの観点で意見文を表 1 のように 4 種類に分類する。また、叙述文を**カテゴリ E** とする。

# 3 文間の主題連鎖

## 3.1 文の構造を捉えるモデル

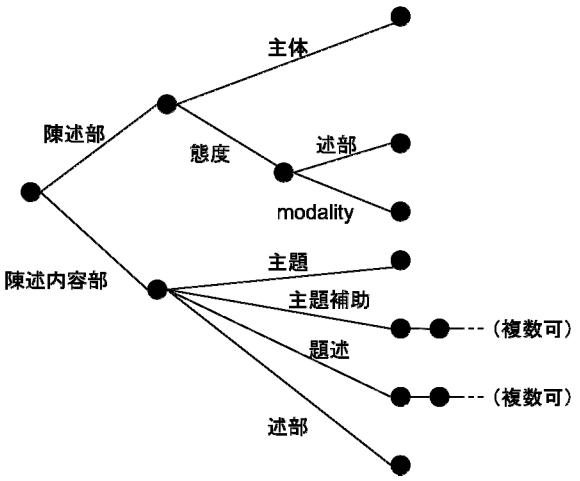


図 1：文のモデル

図 1 に示したグラフ構造で、文の構造をモデル化する。以下でこのモデルについて述べる。

**陳述部**　陳述内容に対する発話者についての記述

- **主体**　発話者
- **態度/述部**　主体の陳述態度に関する表現
- **態度/modality**　主体の陳述態度（文モダリティ）

**陳述内容部**　陳述内容についての記述

- **主題**　文の中で筆者が話題の中心として取り上げている対象
- **主題補助**　主題を修飾する語
- **題述**　主題と主題補助以外の語
- **述部**　主題が係り、様相を含む述部

意見文では、陳述部に陳述内容に対する発話者について、陳述内容部に陳述内容について記述し、叙述文では、陳述内容部に陳述内容について記述する（陳述部は未定義）。

## 3.2 主題連鎖

本研究では以下の 3 種の主題連鎖の関係に基づいて文間の連鎖構造を捉える。

- **主題維持**　ある文の主題またはそれに類似する語が直後の文の主題になっているか、または隣接する 2 文の主題のうち後文の主題が省略されている 2 文間の関係
- **主題変化**　ある文の題述またはそれに類似する語が直後の文の主題になっている 2 文間の関係

- **主題回復**　ある文の主題またはそれに類似する語が後の文(直後の文を除く)の主題になっている 2 文間の関係

図 2 に以下の文章を例とって、主題連鎖の例を示す。

*「果たしてαはγなのだろうか。まず、αはβと言われている。そして、βはγであると言われている。よってαはγである。」*

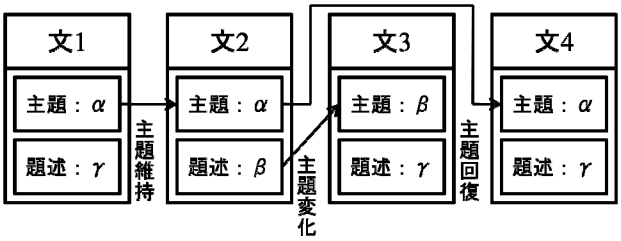


図 2：主題連鎖の例

## 4　文モダリティを含む文間の連鎖構造

本研究では、主題連鎖に基づいて捉える文間の連鎖構造と各文の文モダリティを複合したモデルを用いて、文章中の論理展開を把握する。

文間の連鎖構造は、論理が展開される流れ(**パス**と呼ぶ)を示すものである。また文モダリティは、文が陳述される時点で論理がどこまで誘導されているかを示すものである。例えば、論理展開が開始される文、すなわち話題が提起される文は、文モダリティが読み手に話題を投げかける種類の文モダリティ(カテゴリ A、カテゴリ C)で、パス上で始点に位置する文である。

このように、各文についてパス上の位置とモダリティを捉えることで文章の論理展開を把握することができる。そこで本研究では、以下のように文間の主題連鎖と連鎖する文のモダリティを同時に捉えるモデルを提案する。ただし、「$m(\ )m'(\ )\cdots$」は文の並びについてモダリティ、主題、題述を示すものである。

主題維持 $m(t,r)m'(t,r')$
主題変化 $m(t,\{\cdots,t',\cdots\})m'(t',r')$
主題回復 $m(t,r)\cdots m'(t,r')$

$$\begin{cases} t,t' : \text{主題} \\ r,r' : \text{題述} \\ m,m' : \text{文モダリティ・カテゴリ} \\ \quad \text{ただし} m,m' \in \{A,B,C,D,E\} \end{cases}$$

このモデルを用いて図 2 の例を表すと、以下のように表される。

$$A(\alpha,\gamma)E(\alpha,\beta)E(\beta,\gamma)B(\alpha,\gamma)$$

## 5　実験・考察

### 5.1　実験と結果

提案モデルで実際の論説文の論理展開を捉える実験を以下の手順で行った。

≪1≫ 被験者（高校生）が書いた小論文 605 編を用い、それぞれを専門家 2 名と国語教師 2 名が採点した。これらの文章を以下のようにクラス分けした:①得点順に 5 段階にクラス分けをする、②クラスの文章数の分布が正規分布になるようにクラスの境界を調整する。最も高い評価を受けた文章を含むクラスを最高クラス、最も低い評価を受けた文章を含むクラスを最低クラスと呼ぶことにする。

≪2≫ 最高クラスと最低クラスに属する文章について採点者 4 名の間での採点結果の標準偏差を測定し、標準偏差が小さい文章を各クラス 20 編ずつ抽出した。最高クラスには 286 文、最低クラスには 181 文がそれぞれ含まれている。ただし、最低クラスには完結していない文章が含まれており、このような文章は論理展開が含まれていないと判断し、省いた。

≪3≫ 次に、抽出した 40 編の文章(計 467 文)に含まれる主題連鎖と文モダリティを手動で判定した結果から、提案モデルで捉えられる文間の主題連鎖と$\{m,m'\}$の組合せパターン(以下、展開パターン)について、各クラスでの出現頻度を求めた。

以上の結果、表 2 に最高クラスでの展開パターンの出現頻度、表 3 に最低クラスでの展開パターンの出現頻度を示す。両クラスの平均値の差の検定(t 検定)を行なったところ、有意水準 5%において有意差があると判定された（P(T≦t) 両側の値が 0.0049)。

### 5.2　考察

「$A(t,r)\cdots B(t,r')$」は、読み手に投げかける形で問題提起を行い、提起した問題を確認しながら結論を述べる展開パターンである。この「$A(t,r)\cdots B(t,r')$」は、最高クラスに出現する反面、最低クラスには出現しない。このことから、明確な問題提起をおこなった上で何に対する結論であるかを提起段階に遡って結論を示す論理展開になっている場合、読み手に好印象を与える要因となりうることがわかる。

「$B(t,r)\cdots B(t,r')$」は、論理展開中の離れた文間で意見陳述する特定対象が一致していることを示す展開パターンである。これは、全体的に話題が一貫していることが表れる展開パターンと解釈することができる。この「$B(t,r)\cdots B(t,r')$」が各クラス内で出現する割合は、最高クラスで 12.07%である事に比べ、最低クラスでは 4.24%と少ない。このことから、読み手に与える印象が良い文章では、より意見陳述の特定対象が一定した一貫性のある論理展開になっていることがわかる。

「$B(t,\{\cdots,t',\cdots\})E(t',r')$」は、ある特定対象に対する意見陳述の後に、題述の内容に関する事実確認を行う展開パターンである。これは、意見陳述の後に判断の根拠を示す

展開パターンと解釈することができる。この「$B(t,\{\cdots,t',\cdots\})E(t',r')$」が各クラス内で出現する割合は、最高クラスで 7.33%である事に比べ、最低クラスでは 0.85%と少ない。このことから、読み手に与える印象が良くない文章では、事実確認による判断の根拠を示した意見陳述が少ないことがわかる。

「$D(\cdots)D(\cdots)$」の主題連鎖は全て、主題が明示されていない文間に連鎖がある展開パターンである。この展開パターンは、表層のみから連鎖を読み取ることができず、読み手に主題を推測させることになる。場合によっては、読み手にこの部分で論理展開が遮断されているような印象を与えることもある。この「$D(\cdots)D(\cdots)$」の主題連鎖は、最高クラスでは出現しないが最低クラスでは出現する。このことから、論理の流れを明確に示さない論理展開は読み手に良くない印象を与える要因となりうることがわかる。

「$E(\cdots)E(\cdots)$」の主題連鎖は全て、事実確認をする文間に連鎖がある展開パターンである。この展開パターンは、事実陳述が連続していることを示すものである。意見陳述を目的とする論説文における事実陳述は意見陳述の判断根拠の役割を果たすものであるため、冗長な事実陳述は望ましくない。例えば、参考資料の内容の羅列が大部分を占める文章にこの展開パターンが多く見受けられる。この「$E(\cdots)E(\cdots)$」の主題連鎖の各クラス内で出現する割合は、最高クラスで 10.34%であることに比べ、最低クラスでは 18.64%と多い。このことから、連続した事実陳述は度合いによって読み手に良くない印象を与える要因の一つとなりうることがわかる。

## 6 おわりに

本稿では、文モダリティと主題連鎖を分類し、「主題連鎖に基づいて捉えた文間連鎖」と「各文の文モダリティ」を複合したモデルを用いて論説文の論理展開を把握する手法について述べた。

提案手法は、主題連鎖のみで論理展開を把握する手法に比べて、連鎖している文が論理展開上で果たす役割を捉えることができるため、詳細な論理展開の把握を可能にすることがわかった。

今後の課題として、論理展開の部分的な把握にとどまらずに、提案したモデルで論理展開全体を表現して、文章全体の論理展開のパターンを把握する試みが挙げられる。また、考察対象とする論説文文章を増やすことが挙げられる。また、カテゴリ C の出現頻度が少ないことから、他のカテゴリと合併するなど文モダリティの分類を最適化することが挙げられる。

## 参考文献


[1] 石岡恒憲. 小論文およびエッセイの自動評価採点における研究動向. 人工知能学会誌, Vol.23, No.1, pp.17-24(2008)

[2] Yigal Attali, Jill Burstein. Automated essay scoring with e-rater v.2. *Journal of Technology, Learning and Assessment*, 4. Retrieved June 22(2007)

[3] 石岡恒憲, 亀田雅之. コンピュータによる小論文の自動採点システム jess の試作. 計算機統計学, Vol.16, No.1, pp.3-18(2003)

[4] 乾孝司, 奥村学. テキストを対象にした評価情報の分析に関する研究動向. 自然言語処理, Vol.13, No.3, pp.201-241,(2006).

[5] 藤田 彬, 田村直良. 文章構造解析に基づく小論文の自動評価. 第 7 回情報科学技術フォーラム FIT2008 講演論文集第 2 分冊, pp.285-288(2008)

[6] 藤田 彬, 鈴木春菜, 田村直良. 文脈理解のための文モダリティの分類と自動判定. 情報処理学会研究報告, 2009-NL-189(2009)


表 2：最高クラスでの展開パターンの出現頻度

| | $\{m,m'\}$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | AA | AB | AC | AD | AE | BA | BB | BC | BD | BE |
| 維持 | 0 | 2 | 0 | 0 | 1 | 3 | 31 | 0 | 4 | 5 |
| 変化 | 0 | 1 | 0 | 1 | 0 | 2 | 23 | 0 | 4 | 17 |
| 回復 | 0 | 4 | 0 | 0 | 0 | 3 | 28 | 0 | 2 | 5 |
| | CA | CB | CC | CD | CE | DA | DB | DC | DD | DE |
| 維持 | 0 | 0 | 0 | 0 | 2 | 0 | 1 | 2 | 0 | 1 |
| 変化 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 3 |
| 回復 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| | EA | EB | EC | ED | EE | | | | | |
| 維持 | 0 | 16 | 1 | 2 | 13 | | | | | |
| 変化 | 3 | 18 | 0 | 2 | 8 | | | | | |
| 回復 | 0 | 13 | 0 | 2 | 3 | | | | | |

表 3：最低クラスでの展開パターンの出現頻度

| | $\{m,m'\}$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | AA | AB | AC | AD | AE | BA | BB | BC | BD | BE |
| 維持 | 0 | 1 | 0 | 0 | 1 | 2 | 16 | 0 | 1 | 8 |
| 変化 | 0 | 0 | 0 | 1 | 0 | 0 | 5 | 1 | 3 | 1 |
| 回復 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 | 2 | 3 |
| | CA | CB | CC | CD | CE | DA | DB | DC | DD | DE |
| 維持 | 0 | 0 | 1 | 1 | 1 | 0 | 2 | 0 | 2 | 0 |
| 変化 | 0 | 0 | 1 | 0 | 0 | 0 | 2 | 0 | 3 | 1 |
| 回復 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 1 | 0 |
| | EA | EB | EC | ED | EE | | | | | |
| 維持 | 2 | 8 | 0 | 1 | 10 | | | | | |
| 変化 | 0 | 8 | 0 | 4 | 7 | | | | | |
| 回復 | 0 | 5 | 0 | 1 | 5 | | | | | |