手がかり表現に基づく評論文への内容・関係属性の自動付与

佐野　正裕 †　　佐藤　理史 ‡　　松崎　拓也 ‡
† 名古屋大学工学部電気電子・情報工学科　‡ 名古屋大学大学院工学研究科
m_sano@nuee.nagoya-u.ac.jp

1　はじめに

大学入試の問題は、大きく、センター試験に代表される選択式の問題と、東京大学の二次試験に代表される記述式の問題がある。記述式問題は、選択式問題よりも難易度が高いため、これまで「ロボットは東大に入れるか」プロジェクトで対象としてきた記述式問題は、数学と社会 (世界史) の二科目であり、国語現代文の記述式問題は対象としてこなかった。

国語現代文の特徴は、数学や社会とは異なり、かなり長い文章が「本文」として与えられ、その本文に即して解答することが求められる点にある。このため、文章の読解力が直接問われることになる。さらに、記述式の問題の場合は、文章の生成力も必要となる。つまり、国語現代文の記述式問題をコンピュータで解く試みは、コンピュータで言語の読解・生成力をどこまで実現できるか・できているかを測る試金石となる。

このような考えのもと、我々は、今年度から、大学入試「国語」評論文の記述式読解問題の解法に関する研究を本格的にスタートさせた。選択式の問題とは異なり、記述式の問題は、簡単には分類問題 (選択問題) に落とし込むことはできない。そのため、本文がどのような構造となっているかを把握し、それに基づいて問われている部分を同定して、それをコンパクトに言い換えることが必要となる。

本研究では、その最初のステップとして、文章を構成するユニット (おおよそ文または節) に対する属性付与と、ユニット間に対する関係付与に取り組んだ。文献 [1] では、英語文章を、(1) 文章中の主張や根拠を示す部分 (component)、および、(2) その component 同士を結ぶ関係 (relation) から構成されるものと捉え、component や relation の特定に試みている。本研究では、日本語で書かれた評論文を対象に、(1) のような「文章における各文・節の役割 (内容属性)」と、(2) のような「文や節間の関係 (関係属性)」の 2 つの情報の特定を試みる。具体的には、評論文に特徴的な 26 種類の内容・関係属性を、それらを示唆する言語表現と対応づけて整理し、各ユニットに属性を自動付与するシステムを作成した。

> まず自然は、近代の自然科学的な見方からいえば、それ自体としては価値や目的を含まず、因果的・機械論的に把握される世界である。人間ももちろん自然の一部分であるから、人為と自然の対立はない。人間が自然にどのような人為を加えても、それは自然に反するものではなく、人間による自然破壊というようなことはありえないであろう。自然のある状態とかある段階に特に価値があるとする理由もない。ァすべての事象は等しく自然的である。
>
> **(設問 1)** (傍線部ア) とはどういうことか、説明せよ。
>
> **(解答例)** 人間を含めたすべての自然界の現象は、それ自体としての価値や目的を含まず機械論的に把握されるものであること。

図 1: 評論記述問題の例 (東大 2000 年第 1 問設問 1)[2]

2　記述式読解問題へのアプローチ

本研究では、大学入試の国語現代文の評論文に対する記述式読解問題を対象とする。具体例を図 1 に示す。典型的な記述式読解問題のひとつは、この図のような、本文中の傍線部に対して「どういうことか説明せよ」という設問である。

このタイプの設問に対しては、傍線部をいくつかの部分に分割し、それらを本文に基づいてパラフレーズすることによって解答を構成できる [3] ことが多い。図 1 の場合、表 1 のような対応関係となる。

このような方法に従って解答を作成するためには、次のような 2 種類の情報が不可欠である。

1. **文章における各文・節の役割**　たとえば、第 1 文「まず自然...」が、この文章における「自然」という用語を**定義**している文であること。
2. **文や節間の関係**　たとえば、第 2 文の従属節「人間も... であるから」が、主節「人為と自然の対立はない」の**理由**となっていること。

表 1: 傍線部のパラフレーズによる解答作成

	傍線部	解答例	根拠となる本文
(1)	すべての事象は	人間を含めたすべての自然界の現象は、	人為と自然の対立はない
(2)	等しく自然的である	それ自体としての価値や目的を含まず機械論的に把握されるものである (こと。)	自然は、近代の自然科学的な見方からいえば、それ自体としては価値や目的を含まず、因果的・機械論的に把握される世界である

本研究では、前者を**内容属性**、後者を**関係属性**と呼ぶ。本稿では、与えられた評論文に対して、これらの属性を自動的に付与する方法を検討する。

3 設定する属性と手がかり表現

評論文 (論説文) を、「論じる・主張する」ために「説明する」文章と捉えれば、それらは「主張部分」と「説明部分」の 2 つに大別できる [4]。主張部分は、筆者の考えを直接述べる部分である。ここでは、主張であることを際立たせる強調表現や、説明からの帰結であることを示唆する接続表現などが用いられることが多い。一方、説明部分では、主張の根拠の説明や、具体例の提示、対立する主張との比較などが行われる。

これらの点を考慮し、表 2 に示す 26 種類の属性を設定した。これらの属性のうち、No.1 から No.8 までの 8 属性は、主に、接続表現によって導かれる属性である。No.9 から No.11 までの 3 属性は、主張部分を示唆する属性である。No.12 と No.13 の 2 属性は、議論の前提の提示に関係する属性であり、No.14 から No.22 までの 9 属性は、説明部分に典型的に現れる属性である。

設定した 26 種類の属性は、以下の 3 つのタイプに分類される。

- 内容属性のみの属性
- 関係属性のみの属性
- 内容属性と関係属性の両方を併せ持つ属性

たとえば、図 1 の第 2 文の従属節「人間も... であるから」に付与される**理由**という属性は、その従属節が**理由**を述べていること (内容属性) と同時に、その節が主節に対する**理由**という関係 (関係属性) を持つことの両方を意味する。

これらの属性の多くは、特定の言語表現によって示唆される。このような表現を本研究では**手がかり表現**と呼ぶ。東京大学の入試問題 (国語第 1 問) の本文 (2000〜2009 年)[2] から、合計 270 個の手がかり表現を収集した。これらの手がかり表現には、接続表現、接続助詞、副詞的表現、文末表現、こそあど、に加え、「理由」や「問題」といった内容語も含まれる。表 2 には、手がかり表現の例も示した。

本研究では、手がかり表現を形態素列パターンとして定義する。それらには、次の 2 種類がある。

- **連続パターン** 連続する形態素列として定義されるもの
  (例) 考えられる = [考え, られる]
- **呼応パターン** パターンの内部に、任意の形態素列とマッチする要素 (*) を含むもの
  (例) 〜は〜である = [*, は,*, で, ある]

以下では、(*) の部分を含まない形態素列を、**手がかり表現の構成要素**と呼ぶ。上記の 2 種類のパターンのうち、連続パターンは一つの構成要素となる。一方、呼応パターンは複数の構成要素をもつ。たとえば、呼応パターン「〜は〜である」では、「は」と「である」が構成要素となる。

4 属性の自動付与システム

4.1 付与手順

表 2 の整理に基づき、文に属性を付与するシステムを作成した。システムの付与手順を以下に示す。

1. **形態素解析** MeCab (IPADIC) を用いて、文を形態素列に変換する。
2. **構成要素の発見** 形態素列パターンマッチングにより、手がかり表現の構成要素を見つける。
3. **属性の付与** ステップ 2. で発見した構成要素に基づき、属性を付与する。手がかり表現が連続パターンの場合は、無条件で対応する属性を付与する。手がかり表現が呼応パターンの場合は、構成要素がすべて存在し、かつ、それらが強い節区切りをまたがない場合にかぎり属性を付与する。

ここで、強い節区切りとは、以下の節末を意味する。

- 並列節の節末を表す接続助詞「が」「けれど」「けれども」+ 句点
- 連用節の節末を表す「動詞・形容詞・助動詞の連用形」+ 句点 + （以降に格助詞「が」または係助詞「は」が存在）
- テ節の節末を表す接続助詞「て」+ 句点

表 2: 設定した属性の一覧と手がかり表現の例

No.	属性名 *	上段：属性の説明、下段：手がかり表現 (代表的なもの)
1	＜逆接＞	前の文とは異なる立場で内容を述べる しかし / だが / ところが / にもかかわらず / けれども / しかしながら
2	＜転換＞	異なった話題に移行する では、 / 他方、 / ところで、 / それでは / そもそも / 一方
3	＜順接＞	前のユニットから導かれる主張や結論を述べる だから / したがって / ということは、 / しかるに
4	{ 追記 }	前のユニットに付随もしくは関連する内容を、付け加えて述べる も / そして / その上 / もっと / さらに / さらにいえば、 / それのみでなく / しかも
5	{ 並列 }	他のユニットと対になって文章を構成する まず / また / 第 N に / まずは / あるいは / 同じように
6	＜まとめ＞	これまでの議論をまとめて、比較的簡潔に言い換える 以上に見てきたように / 以上の / つまり / 言い換えると / すなわち / このように、
7	[提起]	これから議論する内容を提示する 問題 / う。 / か。 / よう。 / みたい。 / 問う
8	{not-onlyA, but-alsoB}	比較対象を掲げながら、but-also 以降を強調する A だけではなく B / A もとより B / A のみでなく B / いや B にも
9	[筆者主張]	筆者が (事実や理由ではなく) 明示的に独自の主張を述べる 考えられる / のである / 考察 / 思われる / 私には / ちがいない / いわざるをえない
10	[強調]	(より簡易な表現などがあるが) 強調する言葉を含んでいる 重要 / 必要 / べき / なければならない / 大事 / はず / こそ / にほかならない / すら / さえ / 実は / だけ / 少しも / しか / まちがいなく / 何よりも / まさに /
11	[可能性]	断定的ではないが、(弱い) 主張を伝える うる / かもしれない / できる / できない / えた / えなかった / ありえ
12	[定義]	それぞれの言葉に (確認のため/独自の) 定義を与える A は A’ である / A は A’ と定義される / A とは A’ である / A は A’ といわれる
13	[一般]	一般的な考え方・発想について述べる よく / もちろん / いうまでもなく / 一般に / 私たち / だれもが知っている / といわれる / 今日 / 普通 / 当然 / 常識 / しばしばあげられる / あきらか
14	{ 理由 }	主張や事実に対して、その理由や根拠を述べる から / 理由 / のため / からである / ため / ゆえ / 根拠 / ので、 / ゆえん
15	{ 例示 }	読者に理解を促すための具体例を挙げる A とか A’ といった / いわゆる / A や A’ や / 例えば / など / の場合 / のように、
16	{ 譲歩 }	一般的な事実や想定される反論に対し、譲歩の構造を用いて主張を補強する にしても / としても / なるほど / たしかに
17	{ 仮定 }	仮定の構造を用いて、別の観点から主張を補強する ば、 / かぎりは / ならば / 必ずしも / ても / もし / とすると、 / さもなくば / たとえ A とも / かぎりは / すれば / もしも
18	{notAbutB}	比較対象を掲げて、別の観点から主張を補強する A ではなく B である / A であって B ではない / A とは異なり B / A よりは B
19	[否定]	筆者の主張と、異なる主張もしくは事実を述べる ない / まい / ぬ / 違う / 異なる / なく / ず
20	[引用]	一般的な事柄を述べたり、書籍・伝聞などから内容を引用したりする とされる / によれば / が言った / ことを聞かされた / らしい / と述べている
21	[固有名詞]	固有名詞を含む (MeCab が固有名詞と判断したもの)
22	{ 推論 }	論を展開してきた結果として導き出される主張・事実を述べる であろう / なろう / ことになる / というのである / と言ってよかろう / といえる
23	[過去]	主節の時制が過去である た / かつて
24	[「」]	かぎ括弧に囲まれたフレーズを含む 「 / 」
25	＜指示＞	指示語や指示表現を含む これ / この / これら / それ / 以上に / その / あの / そこ / そう / かかる / ここ / 次 / こう / かよう / それら / そんな / こんな / そのような
26	{ 手段 }	主張を実現するための、手段や目的について述べる ために / によって / をもって

* 属性名の列の凡例：[内容属性], ＜関係属性＞, { 内容・関係両方を併せ持つ属性 }

4.2 評価と考察

東京大学の入試問題 (国語第 1 問) の本文 (評論文)[2] を使用して、システムの評価を行った。評価に使用した年度を以下に示す。

(1) 2000〜2009 年。(手がかり表現の収集に利用)
(2) 2010〜2011 年。(システムにとって未知のテキスト)

評価では、人手で作成した正解データとシステムの出力を比較し、表 3 の TP, TN, FN, FP の数を数えた。その結果を表 4 に示す。なお、recall と precision は、以下で定義される。

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}},\ \text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

表 4 において、recall は、(2) が (1) よりも約 5%低い。一方、precision は 70%強で、(1) と (2) でほぼ同じである。前者は、現在使用している手がかり表現の網羅性に若干問題があることを示唆する。後者は、手がかり表現に基づく属性の付与が可能であることを意味すると同時に、精度を上げるためには、他の情報を併用する必要があることを示唆する。

正しく属性を付与できなかったケースを、どういった問題点があるか、正解を導くためにはどういった情報がさらに必要か、という観点から分析した。以下では、(·) に、(1) 内の誤り数 (FP + FN) に対する割合を、また【·】に、各例文の【誤って付与された属性/理想的な属性】を示す。

(1) 表現の影響範囲の特定が必要 (23%)
(例) 極端な立場によれば、改変をまったく受けない自然が最善であろう。【否定/なし】

(2) 文における語の役割や、文の構造の把握が必要 (17%)
(例) 自然は、元来は没価値的であり、人間によって守るべき価値を付与される。【定義/なし】

(3) 複数の意味・用法を持つ表現の区別が必要 (17%)
(例) だから、環境保護は第一義的に人間のためのものである。【理由/手段】

(4) 前後の部分との関係を加味した判断が必要 (13%)
(例) 生物共同体を構成する他の生物たちには権利や義務の意識はない。【なし/強調】

(5) 表現と文の伝達内容との対応の特定が必要 (11%)
(例) 自然のある状態とかある段階に特に価値があるとする理由もない。【理由/なし】

表 3: システム出力の分類

		システム出力	
		属性付与	付与せず
正解データ	属性付与	TP	FN
	付与せず	FP	TN

表 4: 評価結果

対象	TP	TN	FN	FP	recall	precision
(1)	1645	11831	121	665	93.1 %	71.2 %
(2)	266	2549	36	98	88 %	73 %

その他、さらなる (何らかの) 情報が必要 (9%)、形態素解析結果の影響 (4%)、プログラム仕様上の課題 (3%)、(人間でも) 微妙な判断 (2%)、本文の問題 (1%)、などがあった。

5 まとめ

評論文において特徴的な内容および関係の 26 種類の属性を、手がかり表現と対応づけ、それらの属性を文章の構成ユニットに自動付与するシステムを作成した。

人間が文章を読解するときも、本文にある手がかりを用いて、文章における重要な部分を同定していると考えられる。しかし、そのときに使用している手がかりは、本稿で整理したような、語やフレーズに基づく手がかり表現だけに限定されるわけではない。たとえば、文の構造が類似していることや、段落内での位置など、他の手がかりも利用していると思われる。

今後は、手がかり表現をさらに充実させるとともに、上述したような、語やフレーズ以外の他の手がかりの利用も検討していく必要がある。

謝辞 本研究は JSPS 科学研究費基盤研究 (B)「文章の読解と産出のための言語処理技術」(課題番号 15H02748) の助成を受けている。

参考文献

[1] Christian Stab and Iryna Gurevych. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 46–56, 2014.

[2] 桑原聡. 東大の現代文 25 カ年 [第 8 版]. 教学社, 2016.

[3] 今井健仁. 現代文の解法 第 3 版 (東京大学への道). データハウス, 2015.

[4] 神田邦彦. 現代文 標準問題精講. 旺文社, 2015.