

採点ルール学習とその説明機能をもつ 小論文の採点支援システム

泉谷 達庸^{†1} 片上 大輔^{†1} 新田 克己^{†1}

採点基準が定められた論述式小論文試験を対象とした、採点結果への説明機能を持つ採点支援システムの開発が本研究の目的である。各採点項目を予測した結果をユーザに提示するだけでなく、評価に対する根拠も示すことで、より説得的な支援を可能とする。本研究では「論点の言及」、「論理性」の評価に的を絞り、決定木によって支援情報提示の観点から評価ルールを生成するシステムを構築する。決定木の学習には J48 を採用し、事前に収集した採点済み答案を基に評価ルールを学習する。評価実験の結果、「論点の言及」については人間とほぼ同等に評価でき、その評価の根拠として「論点の言及箇所」を精度よく特定できることが分かった。

Essay Grading Support System with Grading Rules Learning and Explanation Function

MICHINOBU IZUMIYA,^{†1} DAISUKE KATAGAMI^{†1}
and KATSUMI NITTA^{†1}

In this paper, our goal is to develop a system which can support essay (statement-type) examination rating by the fixed rating criteria. Not only the predicted point of rating are provided to users, but the basis of explanation is also given, which makes the support more convincing. We focus on the evaluation of "mentioned issues" and "logicality", and develop a system which can generate assessment rules in perspective of supporting information presentation by decision tree. We adopt J48 as the decision tree learning, and also learn the assessment rules based on the marked paper we collected preliminarily. As a result of the evaluation experiment, the system has the same evaluation capability in rating "the mentioned issues" as human. To be more precisely, we found that "the mentioned issues" can be accurately identifiable.

1. はじめに

近年、授業の課題や入試科目として小論文問題を取り扱うケースが増えてきた。小論文問題は論理的思考力や表現力を量ることができ、有用性が高い問題といえる。しかし、暗記力や読解力を問う問題のように選択肢から選ばせる形式ではなく、一般的に自由記述形式を取るため、採点時に以下の問題を生じる。

- 採点者にかかる時間的な負担
- 採点する小論文の順序などが要因で生じる評価の偏り
- 採点者の価値観の違いなどが要因で生じる採点者間での評価の違い

このような背景から、コストの削減や採点の安定性向上を目的とした、小論文の自動採点に関する研究が国内外で精力的に行われてきた¹⁾⁻⁶⁾。そのうちの多くの研究は論理展開の自由度の高い小論文(以下、エッセイ)を対象としており、問題文と答案の間の関連度や文章の読みやすさを定量化し、それを基に評価している。

それに対し、詳細な問題設定の下、より細かい論理性が問われる小論文(以下、論述式小論文)を対象とした研究も報告されている⁷⁾⁸⁾。我々の研究室では、過去に日本語に対応した論述式小論文の採点システムとして2手法が提案された。それらはアプローチが対象的で、藤田ら⁹⁾は内面的情報(論点の関係性)を手動で定義し、ルールベースのシステムを考案したのに対し、村田ら¹⁰⁾は表面的情報(単語の出現傾向)から SVM によって評価を識別するシステムを考案した。両者は論述式小論文の一例として、法科大学院統一適性試験¹¹⁾の表現力問題を対象とし、人間と同程度の精度で評価できることを示した。

しかし、既存の採点システムは採点結果に対する説明機能が十分とはいえず、評価の根拠に説得力がないという批判¹²⁾がある。自然言語処理の技術進歩に伴い、形態素解析や構文解析といった統語論的な解析は高精度化されたが、意味的な解析は未だ困難とされており、答案に記述された内容を根拠に説明を提示することは難しい。

本研究では、論述式小論文を対象とした採点支援システムを提案する。上記の通り、意味に踏み込んだ説明機能の実現は困難だが、論述式小論文は問題設定が詳細なので記述される論点の種類が限られてくる。そこで、その論点同士の関係性などを機械的に抽出し、評価ルールを学習することによって、間接的ではあるが意味的な評価を実現し、その評価への説明機能を持つ採点システムの開発を目的とする。

^{†1} 東京工業大学 大学院 総合理工学研究科

2. 対象とする論述式小論文

論述式小論文問題の一例として、本研究では法科大学院統一適性試験の表現力問題（以下、表現力問題）を採点対象とする。本章では、表現力問題の問題形式、および採点基準について述べる。

2.1 表現力問題の問題形式

表現力問題は年度を問わず、問題文及びアンケートなどの参考資料が A4 用紙 2 枚分に記載されており、受検者は用意された相反する 2 種類の立場のうち一方を選択し、その理由について意見書を作成する。問題文中には、それぞれの立場を主張する根拠となる論点が挙げられており、自分が主張する立場の方が相手方よりも優れている理由を論理的且つ説得的に説明する能力が問われる。

例えば 2006 年度問題 1 の問題設定では、小規模でありながら専門性の高い書物を扱う老舗出版社の社員として受検者は意見書を作成する。この会社は近年の出版不況に伴って事業方針を転換するかどうかの決断を迫られており、受検者は営業部長が主張する「路線拡大」、もしくは編集部長が主張する「従来路線の継続」のどちらかの立場を選択しなければならない。問題文中には以下に挙げる路線拡大、従来路線の継続の根拠を説明するための論点を推察できる内容が記述されている。

- 路線拡大
 - － 倒産の危機
 - － より幅広い読者の開拓
 - － 自費出版事業による収益安定
 - － 良書の刊行継続は可能
- 従来路線の継続
 - － 従来刊行物への高い評価・固定ファンが存在
 - － 老舗ブランド力に傷がつくおそれ
 - － 取引先や書店が離れていくおそれ
 - － リストラによる編集部員の士気下降、離職のおそれ

このように、両主張について論点が存在し、受検者は相手方の主張を考慮した文章作成能力が問われる。問題設定は違いますが形式は毎年度同じであり、必ず 2 種類の立場が設けられている。

2.2 採点基準

論述式小論文問題はエッセイとは異なり、一般的に問題設定が詳細なため、答案に出現する論点は似通う傾向にある。よって、読みやすさやストーリー性だけでなく、挙げられた論点の数や種類、そして結論に対する論拠説明の結び付きなどが採点基準として公開されている。以下では、表現力問題の採点項目を紹介する。

C1 結論の明示（1 項目）

表現力問題には必ず主張すべき立場が 2 種類用意されており、受検者は片方の立場を選択した上で小論文を作成する。そこで、受検者が明確に立場を明記しているかどうかを評価する項目が設けられている。それが結論の明示の採点項目である。

C2 論点の言及（10 項目前後）

各年度とも問題設定が非常に詳細なため、おのずと受検者が触れるべき論点は絞られてくる。そのため、論点項目があらかじめ設けられており、各々の論点が言及されているかどうかをみる。

C3 論理性（3 項目）

挙げるべき論点は問題の内容によって当然異なるが、論理性は特に問題の内容に依存しないため、例年以下に示す 3 項目が評価項目として設けられている。

- 提起
 - 問題状況を冒頭で説明しているかどうかをみる。
- 結び付き
 - 結論支持のための論拠説明が、結論と整合性のある記述となっているかどうかをみる。
- 反対論拠
 - 予想される反対論拠を挙げて反駁することにより、自らの選択を説得的に提示しているかどうかをみる。

C4 説得性（1 項目）

文章構成、ストーリー性、分かりやすさを総合的に判断して評価される。また、論点を 5 つ以上挙げ、論拠を補強するなどの効果をあげていれば加点される。

C5 資料の適切な活用（1 項目）

資料を正確に読み取り、説得力を高める要因として活用できているかどうかをみる。

この他、文章の冗長性や誤字脱字などの減点項目などはあるが、基本的には以上の C1～C5 の合計が答案の総合得点となる。

3. システム概要

本研究では、論述式小論文の採点支援に重点を置いたシステムを提案する。研究対象の表現力問題は 2.2 節の通り 5 つの採点項目が存在するが、そのうち統語論的なアプローチで評価の根拠を提示することが可能で、支援情報があれば役に立つと思われる C2 論点の言及、C3 論理性について評価するシステムを開発した。ただし、C3 論理性のうち提起については対象外とし、論理的構造を評価する点に的を絞った。なぜなら、提起は答案の最初で述べられるため見落とすことはまず考えられず、特に支援する必要はないと考えたからである。

なお、C2 論点の言及は言及されたか否かについて $\{0,1\}$ の 2 段階評価、C3 論理性中の項目は $\{0,1,2\}$ の 3 段階評価とした。

本章では、まずシステムの機能を示し、次にシステムの全体像について述べる。続いてシステム内の各モジュールの役割について、そこに用いた技術や実装方法を含めて説明する。

3.1 システムの機能

本システムは採点対象となる電子化された答案を入力とし、C2 論点の言及、C3 論理性の評価値およびその根拠となる説明を出力する。本システムは採点者を支援することを目的としており、採点者は一般的に紙媒体での採点を好むと考えられるので出力形式は PDF とし、印刷して活用することを想定している。

一方で、評価ルールの学習するためのデータを入力するインターフェースを別に用意している。こちらは紙媒体での収集ではなく、学習データ収集用の GUI を介して評価を入力してもらう。

このように、システムへの入力には採点対象の答案、および学習用データであり、出力は採点結果の PDF ファイルという構成になっている。

3.2 システムの全体像

本システムは、ルールの生成時に必要となるサンプル答案を採点するための学習データ収集モジュール、ルール生成モジュール、採点モジュールを備えた、論述式小論文の採点支援システムである（図 1）。以下では、採点結果出力までの流れを説明する。

まず最初に、採点対象の小論文をサンプルデータと採点対象データに分類する。前者は採点ルール生成に活用され、後者は文字通り本システムが採点対象とする小論文データとなる。ちなみに、これらのデータはすべて電子化されたテキストデータである。

システムに入力されたサンプルデータはルール生成モジュールの前処理部および学習データ収集モジュールに送られる。学習データ生成モジュールでは採点の専門家により評価が与

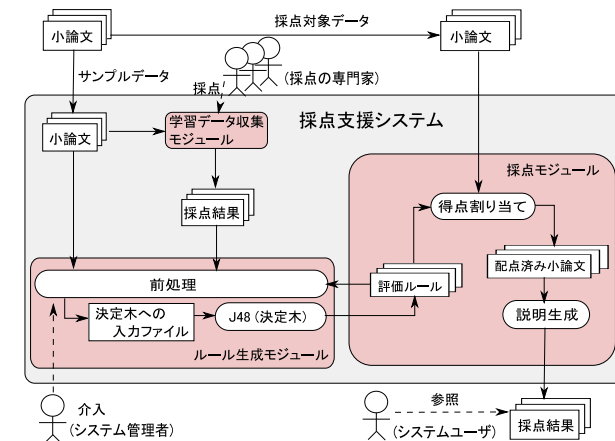


図 1 採点支援システムの全体像
Fig. 1 Outline of essay grading support system

えられ、その結果はルール生成モジュールの前処理部に渡される。

本システムではルール生成に J48¹³⁾ を採用する。ルール生成モジュールの前処理部は主に統語論的な解析を行い、決定木に与える属性情報を生成する役割を担っている。決定木によって生成されたルールは採点モジュールに送られるが、C3 論理性のルール生成には統語論的な解析結果のみならず、言及された論点の種類や書き手が主張する立場も活用すべきである。そこで、C2 論点の言及、主張する立場に関するルールを生成したのち、そのルールを C3 論理性のルール生成に再利用する。図 1 において採点モジュールとルール生成モジュール間でデータの流れがループしているのはこのためである。

採点モジュールでは、採点対象データに評価ルールを適用し、得られた採点結果に対して根拠となる説明を付加したのち、システムユーザにそれを提示する。

次節から、各モジュールについて詳述する。

3.3 学習データ収集モジュール

採点ルールの生成に用いる学習データを収集するために、採点用インターフェースを開発した（図 2）。採点の専門家はこのインターフェースを介して答案を採点する。採点された結果は全て XML ファイルとして保存され、ルール生成モジュールに渡される。

このインターフェースは 2.2 節で挙げた採点項目を評価するための機能を備えている。中でも結論の明示、C2 論点の言及に関しては言及された箇所を特定するための文字列選択機

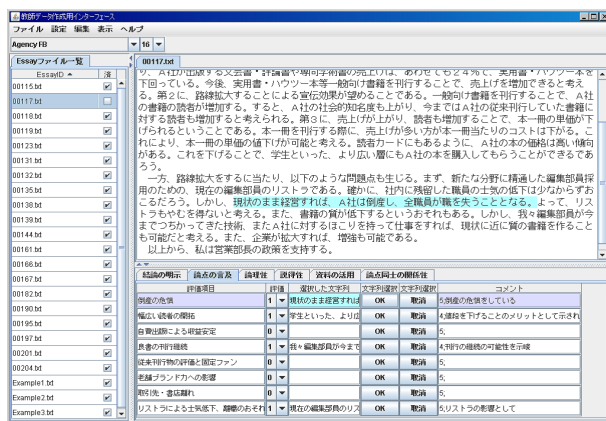


図 2 採点用インターフェース
Fig.2 Evaluation interface

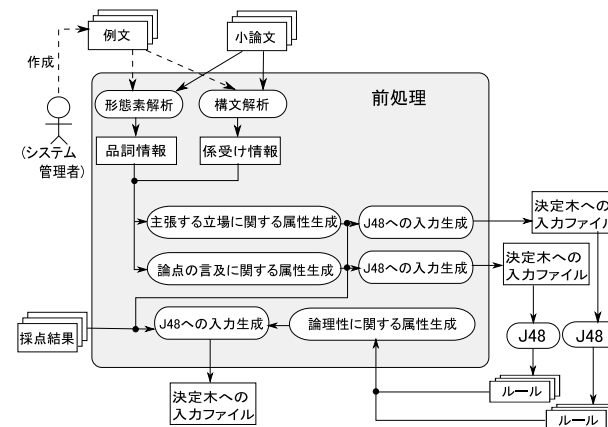


図 3 ルール生成モジュールの詳細
Fig.3 Detail of rule generation module

能を有しており、この選択機能がルール生成および説明生成に大きく貢献する。なぜなら、C2 論点の言及もしくは結論の明示がなされているかどうかを判別するために、このインターフェースを介して選択された文字列を用いることで、ただ単に 0 もしくは 1 の評価を下すだけでなく、言及されている箇所を特定し提示することが可能となるからである。

採点者には答案中で結論および各論点が記述されたと思われる、尤もらしい部分をそれぞれ 1 箇所だけ選択させるようにしている。1 答案中で同じ論点は何箇所も出現する可能性は否定できないが、その全てを採点者に選択させるには負担が大きく、間接的に述べられ選択するのに迷うような箇所も多く存在するので、このような措置を取った。また、選択する箇所について句単位、文節単位、文単位といった制限は設けず、自由に選択させるようにした。

3.4 ルール生成モジュール

ルール生成モジュールは前処理部とルール生成部 (J48) から成る。本節では前処理部を中心に説明し、ルール生成までの流れを追う。

図 1 のルール生成モジュールを詳細化したものを図 3 に示す。先に説明した通り、C3 論理性のルール生成において C2 論点の言及および主張する立場のルールを利用するため、当該モジュールはループ構造を有する。C3 論理性の評価には、一般的に書き手が主張する立場や言及された論点の繋がり等が密接に関係するので、このようなデータの流れは自然だと考えられる。

本節では、C2 論点の言及、主張する立場、C3 論理性について、ルール生成までの流れを説明する。なお、以後「論点の言及に関するルール」、「主張する立場に関するルール」、「論理性に関するルール」はそれぞれ「論点ルール」、「主張ルール」、「論理性ルール」と記す。ルール生成について説明する前に、事例データに与えるクラス値を一意的に決める仕組みを 3.4.1 節で説明し、続いて論点ルールの生成 (3.4.2 節)、主張ルール (3.4.3 節)、論理性ルールの生成 (3.4.4 節) の順に見ていく。

3.4.1 多数決採点方式

ルール生成には、学習データ収集モジュールで事前採点されたサンプル答案が教師データとして用いられる。しかし、1 答案当たりの採点者が複数人いるため、サンプル答案の採点結果を評価ルールの学習に用いるには、評価の個人差を修正し、各評価項目に対して一意的な評価を与える仕組みが必要となる。そこで、本研究では 10) と同様に、次のような方式でサンプル答案に対する評価を決定することにした。

- 採点者がつけた各評価値のうち、最も支持人数の多い値を採用 (多数決)
- 支持人数が等しくなった場合は、全員の平均値を整数に丸めた値を採用

以下、本研究ではこの方式を多数決採点方式と呼ぶ。

3.4.2 論点ルールの生成

決定木に対する属性の生成方法および、事例データの与え方について説明する。

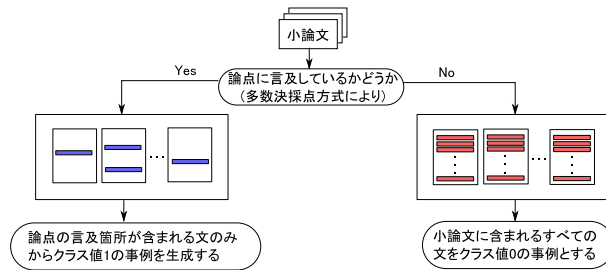


図 4 クラス値の割り当て方法
Fig. 4 Class value assignment method

§1 属性の生成方法

3.3 節で述べたとおり、サンプル答案の採点時、各答案毎に論点の言及箇所が与えられるため、属性生成にはこの文字列（以下、論点文字列）を利用する。ただし、属性生成に用いる論点文字列はサンプル答案のうち多数決採点方式で評価 1 を受けたものに限り、この文字列から共通する言い回しを抽出することとした。つまり、1 答案あたり 5 人が採点し、その内 2 人がある論点について言及していると判断した場合は、その 2 人が選択した論点文字列は属性生成時には破棄される。

具体的には、論点文字列から形態素、文節、係り受け関係、文節同士の位置関係、答案における論点文字列の位置情報を抽出し、J48 に与える属性を生成をする。なお、これらの属性生成には形態素解析器 chasen¹⁴⁾、構文解析器 cabocha¹⁵⁾ から得られた情報を利用した。

§2 事例データの与え方

事例データは多数決採点方式によって決定された評価値と属性情報のペアで登録される。その際、図 4 のように、多数決採点方式より言及されていないと判断された答案中の文は全てクラス値 0 の事例として、言及されていると判断された答案については論点言及文字列がクラス値 1 の事例として登録される。

3.4.3 主張ルールの生成

主張ルールは結論を述べた文字列（以下、結論文字列）に基づいて生成される。そして、ルール生成に用いる属性情報、事例データは論点ルールと同様の手順で登録される。そのため、説明は省略する。

3.4.4 論理性ルールの生成

C3 論理性を評価するには、文同士の繋がりを見たり、根拠や反対論拠の扱い方を考慮し

たりなど、意味的な関係に踏み込んだ評価が必須であり、統語論的なアプローチのみで評価することは困難である。この点、本研究で対象とする小論文は論述式であるため、論点の言及箇所が特定できれば、論点の出現傾向を C3 論理性の評価に活用することができる。論理的な矛盾や論理の飛躍などは検知不可能であるが、高い評価あるいは低い評価を受ける答案に見られる特徴的な論点同士の繋がりなどが抽出できれば、システムユーザに対する支援が可能だと考えた。

C3 論理性の評価項目として結び付きと反対論拠があるが、まずは両評価項目で共通して用いる属性情報について述べ、その次に各固有の属性について説明する。そして最後に決定木に対する事例データの与え方についても言及する。

§1 共通属性

以下に、結び付き、反対論拠のルール生成に共通して用いられる属性を挙げる。

- 主張する立場
2 種類の立場のうちどちらを主張するかによって論理展開が変わることが予想されるため、この属性を用意した。この属性は 3.4.3 節で生成した主張ルールを適用した結果から値が決定される。主張ルールの適用方法は 3.5.1 節を参照。
- 言及された論点の数・種類
結論への結び付きや、反対論拠に対する説明をするにあたり複数の論点を挙げる必要があり、また、挙げる論点の種類も重要となるため、この属性を用意した。この属性は 3.4.2 節で生成した論点ルールを適用した結果から値が決定される。論点ルールの適用方法は 3.5.1 節を参照。

§2 結び付き特有の属性

結び付きは 2.2 節で述べたとおり、結論に対し整合性を保ち、矛盾なく論拠説明ができていないかどうかを評価する項目である。そこで、論証関係にある文字列同士の関係を属性にする。答案から論証関係にある文字列を自動的に抽出するために、テンプレートを用意する。具体的には以下の 3 種類を用意した。

- 段落間の論証関係
段落間を結ぶ論証関係を表す接続詞によってテンプレートを作成する。例えば「根拠となる段落 + すなわち + 結論を表す段落」、「結論を表す段落 + なぜなら + 根拠となる段落」など。
- 文間の論証関係
文間を結ぶ論証関係のテンプレートを作成する。用いられる接続詞は段落間の論証関係

と全く同じである．

- 句間の論証関係

句間を結ぶ論証関係のテンプレートを作成する．例えば「根拠となる句 + ので、+ 結論を表す句」、「根拠となる句 + であるため + 結論を表す句」など．

以上の 3 種のテンプレートを答案とマッチングさせ抽出した文字列中の根拠部分と結論部分には、図 3 の通りそれぞれ主張ルールと論点ルールが適用される．その際、段落間の論証関係として生成された属性は $\text{cntPa}(A,B)$ 、文間、句間の場合はそれぞれ $\text{cntSe}(A,B)$ 、 $\text{cntPh}(A,B)$ という形式を取る．このとき、A は根拠となる論点もしくは主張、B は結論となる論点もしくは主張を表す．

これに加えて $\text{cnt}(A,B)$ という属性を設けた．この属性は $\text{cntPa}(A,B)$ 、 $\text{cntSe}(A,B)$ 、 $\text{cntPh}(A,B)$ のうちどれかに「yes」が割り当てられたとき、自動的に $\text{cnt}(A,B)$ も属性値「yes」を取る．つまり、段落間、文間、句間にかかわらず、論証関係にある論点、主張が存在すれば「yes」が割り当てられる属性である．

§3 反対論拠特有の属性

反対論拠は 2.2 節で述べたとおり、反対論拠に対する反駁が適切になされているかどうかを評価する項目である．そこで、対立関係にある文字列同士の関係を属性にする．

論証関係と同様に 3 種類のテンプレートを用意した．

- 段落間の対立関係
- 文間の対立関係
- 句間の対立関係

これらのテンプレートを答案とマッチングさせ抽出した文字列は、論証関係の属性 cnt 、 cntPa 、 cntSe 、 cntPh と同様に obj 、 objPa 、 objSe 、 objPh の属性生成のために用いられる．

§4 事例データの与え方

多数決採点方式から得られた評価値をクラス値として、属性情報と共に登録される．C3 論理性は 3 段階評価なので $\{0,1,2\}$ のうちのどれかがクラス値として割り当てられる．

3.5 採点モジュール

採点モジュールは、ルール生成モジュールから得た評価ルールを基に、未知答案に評価値を割り当てると共に、その評価値を与えた根拠を提示する役割を担っている．本節では、評価ルールの適用方法および、根拠の説明機能について述べる．

3.5.1 評価ルールの適用

各論点ルールは、未知答案中に存在する全ての文に適用される．その際、ある論点ルール

が 1 文でも評価 1 となるルールが適用された場合、答案においてその論点と言及されたものとして扱う．よって、全ての文において評価 0 となった答案のみが C2 論点の言及なしと判断される．

主張ルールも論点ルールと同様に、未知答案中の全ての文に適用される．ただし、2 種類の主張する立場についてルールを生成するため、ルールの適用時に以下の 4case が生じる．

case1 $\text{claim}(\text{主張 } 1)=\text{yes}$, $\text{claim}(\text{主張 } 2)=\text{yes}$

case2 $\text{claim}(\text{主張 } 1)=\text{yes}$, $\text{claim}(\text{主張 } 2)=\text{no}$

case3 $\text{claim}(\text{主張 } 1)=\text{no}$, $\text{claim}(\text{主張 } 2)=\text{yes}$

case4 $\text{claim}(\text{主張 } 1)=\text{no}$, $\text{claim}(\text{主張 } 2)=\text{no}$

本システムでは case2 および case3 のとき、それぞれ主張 1、主張 2 を取っていると判断する．case1、case4 のときは識別不能とし、システムユーザに正しい主張を問い合わせる．システムユーザに対し高確率で問い合わせが起こらなければ、識別精度を高めるために止むを得ないと考え、このようなルール適用形式を採用した．

論理性ルールは最も単純であり、答案単位でルールが適用され、評価が割り当てられる．

3.5.2 説明生成

C2 論点の言及、C3 論理性の評価に対する説明は次のようにして生成される．

§1 論点の言及

論点ルールは単語レベルの属性から生成されるため、そのルールを根拠として提示しても分かりやすい情報とは言えず、役には立たない．そこで、論点ルールを適用したときに評価 1 となった文を提示することにした．つまり、論点と言及されていると思われる位置を特定する形でユーザを支援する．その位置特定は文単位でなされ、1 論点につき複数の文が提示される場合もある．

§2 論理性

論理性ルールは論点の種類・数、論証関係、対立関係といった属性から生成されるため、単語レベルで生成される論点ルールに比べて抽象度の高い．そこで、適用されたルールを基に説明を生成する．

具体的な説明生成例を図 5 に示す．このように、C2 論点の言及では位置情報が示され、C3 論理性については箇条書きで説明がなされる．

答案ID:○○○○

段落	文
1	1 営業部長の主張を支持する。
2	2 確かにブランドにこだわり、実績のある分野で販売を続けていくことは大事である。 3 しかし、今は経営を安定させることが重要であり、・・・、倒産してしまいかねない。 4 そうなれば、これまで実績・定評のあった分野での出版もできなくなってしまう。
3	5 実用書・ハウツー本は資料にある通り、・・・、収益を高めることができると考える。 6 これまでの販売実績の無さはその分野に精通した編集部員の雇用と研究を重ねることで克服できると考える。 7 またブランド力に傷がつき取引先や書店が離れていく恐れに対しては、・・・この説明を入れ、 経営が安定すれば評論書や文芸書の分野も拡大すると約束することで大丈夫と考える。
4	8 一番危険なのは職人の士気の低下による離職であるが、・・・により防ぐことができると考える。 9 ...にまで社名が知れわたれば、これまで評価の高かった学術書や評論書にも・・・期待できる。

言及された論点

倒産の危機
文3:しかし、今は経営を安定させることが重要であり、・・・、倒産してしまいかねない。
老舗ブランド力への影響
文2:またブランド力に傷がつき取引先や書店が離れていく恐れに対しては、・・・この説明を入れ、
経営が安定すれば評論書や文芸書の分野も拡大すると約束することで大丈夫と考える。
取引先・書店離れ
文2:またブランド力に傷がつき取引先や書店が離れていく恐れに対しては、・・・この説明を入れ、
経営が安定すれば評論書や文芸書の分野も拡大すると約束することで大丈夫と考える。

論理性(結び付き):評価1

- ・ 路線拡大を主張している。
- ・ 論点が3つ以上挙げられている。
- ・ 論点「倒産の危機」が言及されている。
文3:しかし、今は経営を安定させることが重要であり、・・・、倒産してしまいかねない。

論理性(反対論拠):評価1

- ・ 論点が3つ以上挙げられている。
- ・ 文2、3間において「老舗ブランドへの影響」と「倒産の危機」を対立させている。
文2:またブランド力に傷がつき取引先や書店が離れていく恐れに対しては、・・・この説明を入れ、
経営が安定すれば評論書や文芸書の分野も拡大すると約束することで大丈夫と考える
文3:しかし、今は経営を安定させることが重要であり、・・・、倒産してしまいかねない。

図 5 ユーザに示す説明情報

Fig. 5 Explanation information presented to system user

4. 評価実験

本節では C2 論点の言及、C3 論理性の評価で用いる論点ルール、論理性ルールについて適用精度を検証する。紙面の関係上、主張ルールの評価は省略する。

4.1 実験設定

評価実験には表現力問題の答案を用いる。本実験では 2006 年度問題 1 の答案を 80 件用意した。そして、学生に採点の協力を仰ぎ、次のようにして採点データを収集した。

- 学習データ収集モジュールを介して 1 答案あたり 5 人が採点
- 学生 5 人に採点を依頼し、1 人あたり 80 件の答案を採点

表 1 論点の言及における採点者との精度比較 (2006 年度問題 1)

Table 1 Accuracy comparison between our system and scorers in perspective of issue point (2006 problem 1)

論点	本システム		採点者				
	言及箇所特定あり	なし	A	B	C	D	E
1. 倒産の危機	0.75	0.78	0.85	0.76	0.94	0.90	0.71
2. 幅広い読者の開拓	0.80	0.80	0.91	0.86	0.96	0.93	0.81
3. 自費出版による収益安定	0.94	0.94	0.93	0.86	0.96	0.96	0.93
4. 良書の刊行継続	0.58	0.65	0.85	0.76	0.85	0.71	0.76
5. 従来刊行物の評価と固定ファン	0.73	0.73	0.88	0.76	0.84	0.71	0.76
6. 老舗ブランド力への影響	0.80	0.83	0.89	0.84	0.96	0.98	0.95
7. 取引先・書店離れ	0.86	0.86	0.99	0.94	0.98	0.93	0.98
8. リストラによる士気下降	0.58	0.61	0.66	0.84	0.93	0.76	0.93

表 2 論理性における採点者との精度比較 (2006 年度問題 1)

Table 2 Accuracy comparison between our system and scorers in perspective of logicity (2006 problem 1)

論点	本システム	採点者				
		A	B	C	D	E
結び付き	0.38	0.59	0.70	0.72	0.55	0.76
反対論拠	0.41	0.69	0.59	0.72	0.53	0.66

また、評価実験は全て 10-fold cross validation により実施した。

4.2 論点ルールの評価実験の結果・考察

論点ルールの適用時、多数決採点方式による評価と一致した精度を表 1 に示す。表中の採点者の列はサンプル答案の採点時に人間が与えた評価と多数決採点方式との一致率を表している。また、本システムの列における言及箇所特定の「あり」「なし」とは事前採点時に採点者が選択した文字列を特定できたかどうかを、条件に含める場合とそうでない場合を示している。この結果をみると、システムと多数決採点方式で得られた評価が一致したとき、論点の言及箇所を精度よく特定できていることが分かる。また、システムと採点者の精度を比較すると、論点 4、論点 8 は若干劣るものの、ほぼ同程度の精度が得られている。よって C2 論点の言及については十分支援可能であると考えられる。

4.3 論理性ルールの評価実験の結果・考察

論理性ルールの適用時、多数決採点方式による評価と一致した精度を表 2 に示す。この結果から、結び付き、反対論拠ともに人間よりも精度が低いことが読み取れ、現状では支

援情報として活用できるレベルではない。このような結果が得られた要因は2点あると考える。まず1つ目は学習データの不足である。論点ルールは、決定木の生成時に与える事例データが文単位であり、数百の事例からルールが生成されるのに対し、論理性は答案単位で事例データを与えるため80の事例しかない。更に論点は2値分類なのに対し論理性は3値分類なので、過学習が起こってしまったと考えられる。2つ目は統語論的アプローチによる論理性評価の困難さである。論証関係や対立関係など、論理性に関係する属性を含めてルールを生成しているが、文脈の矛盾や論理の飛躍といった意味情報を検知できないことが、ルール生成時にノイズとなって精度が落ちたと考えられる。

今後、学習データを増やした検証実験が必要であり、他の属性を含めるかどうかについても検討していく必要があると思われる。

5. おわりに

本研究では「論点の言及」と「論理性」を対象として、事前採点された答案から評価ルールを学習し、採点時に割り当てた評価への説明機能を持つ採点支援システムを開発した。評価実験の結果「論点の言及」については人間とほぼ同等に評価でき、その評価の根拠として論点の言及箇所を精度よく特定できることが分かった。一方、「論理性」については論理性ルールの適用精度が悪く、現状では支援情報として役立つ可能性は低いことが示唆された。今後、論理性ルールの生成手法に問題があるのかどうかをより多くの実験データを用意した上で検証・改善し、また、支援情報がシステムユーザにとって本当に役に立つ情報なのかどうかについても確かめる必要がある。

参 考 文 献

- 1) Rudner, L.M & Liang, L.: Automated essay scoring using Bayes' Theorem, National Council on Measurement in Education, New Orleans, LA (2002).
- 2) Foltz, P.W., Laham, D., & Landauer, T.K.: Automated Essay Scoring, Applications to Educational Technology. In Proceedings of EdMedia (1999).
- 3) Burstein, J., Kukich, K., Wolff, S., Lu, C., Chodorow, M., Braden-Harder, L., & Harris, M.D.: Automated Scoring Using A Hybrid Feature Identification Technique., In Proceedings of the Annual Meeting of the Association of Computational Linguistics, Montreal, Canada (1998).
- 4) Attali, Y. & Burstein, J.: Automated Essay Scoring with e-rater V.2., Journal of Technology, Learning, and assessment, Vol.4, No.3 (2006).
- 5) Jess : <http://coca.rd.dnc.ac.jp/jess/>

- 6) 石岡恒憲, 亀田雅之: コンピュータによる日本語小論文の自動採点システム, 電子情報通信学会, 信学技報 Vol.102, No.491, pp.43-48 (2002).
- 7) Elliot, S.: Construct validity of IntelliMetric with international assessment, Yardley, PA., Vantage Technologies(RB-323) (1999).
- 8) Rudner, L.M., Garcia, M. & Welch, C.: An Evaluation of the IntelliMetric Essay Scoring System, The Journal of Technology, Learning, and Assessment, Vol.4, No.4 (2006).
- 9) 適性試験委員会, JLF 業書 Vol.12, 法科大学院統一適性試験テクニカル・レポート2006, 商事法務, pp.132-146 (2007).
- 10) 村田淳哉, 片上大輔, 新田克己, SVM を利用した小論文の採点支援システム, 電子情報通信学会技術研究報告, Vol.107, No.428, pp.7-12 (2008).
- 11) 日弁連法務研究財団, 法科大学院統一適性試験, <http://www.jlf.or.jp/tekisei/index.shtml>
- 12) Shermis, M.D., Koch, C.M., Page, E., Keith, T.Z., & Harrington, S.: Trait Rating for Automated Essay Grading, Educational and Psychological Measurement, 62, [1], pp.5-18 (2002).
- 13) weka: <http://www.weka-jp.info/>
- 14) 形態素解析器 chasen: <http://chasen-legacy.sourceforge.jp/>
- 15) 構文解析器 cabocha: <http://www.chasen.org/taku/software/cabocha/>