答案診断グラフを用いた国語記述式答案へのフィードバックの生成

古橋萌々香 [1,2]　舟山弘晃 [1,2]　岩瀬裕哉 [1,2]　松林優一郎 [1,2]
磯部順子 [2]　菅原朔 [4]　乾健太郎 [3,1,2]
[1] 東北大学　[2] 理化学研究所　[3]MBZUAI　[4] 国立情報学研究所
{furuhashi.momoka.p4,h.funa,yuya.iwase.t8}@dc.tohoku.ac.jp
y.m@tohoku.ac.jp　yoriko.isobe@riken.jp
saku@nii.ac.jp　kentaro.inui@mbzuai.ac.ae

概要

学校教育現場では記述式問題が盛んに使われている．しかし，人手による記述式の答案の採点では，学習者の答案に対して個別最適なアドバイスが難しいという問題がある．本研究では，国語の記述式問題を対象に，学習者の答案の誤りに応じた個別のフィードバックを生成するシステムの構築を目指す．その手法として，実際の教育現場で扱われている本文の論理構造関係に着目し，本文の論理構造，談話関係，フィードバックのテンプレートを統合した**答案診断グラフ**と呼ぶ構造を構築し，答案に記述されている内容と模範解答の内容の対応から，適切なフィードバック文を生成する枠組みを提案する．

1　はじめに

国内の国語教育では，与えられた文章を読んで，その内容に関する設問について数十字程度の答案で記述する**記述式問題**が盛んに使われている．こうした記述式問題は，文章を適切に理解し，論理的な思考力や表現力を育むことができる一方で，実践的な課題も生じている．第一に，選択式問題と比べ，記述式の答案に対する採点や教育的なコメントの付与は教師に多大な負担を強いる．第二に，記述式問題の答案は多様な誤りを含み，改善の要素として注目すべき点も個々の答案によって異なるため，適切なアドバイスのためには，それぞれの学習者の誤り方に応じた対応が必要である．しかし，通常の採点結果では正否や得点のみが返却されるため，誤答の理由が不明瞭になりがちであり，学習者は画一的な解答・解説文を読んで自ら振り返るほか術がない．

本稿では，こうした記述式問題を教育実践に適応

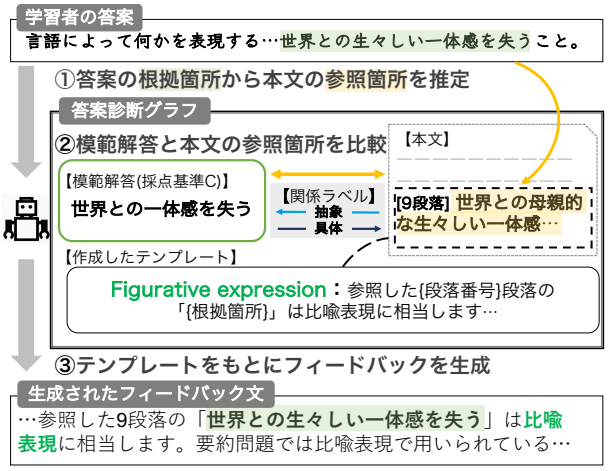

図 1　本研究のフィードバック生成の概要図．**答案診断グラフ**は採点基準と関係ラベル，フィードバックのテンプレートの要素が含まれている．答案の根拠箇所から類似度を用いて参照箇所を推定し，模範解答と答案の参照箇所の比較を行う．そして，参照箇所に紐づいているテンプレートをもとにフィードバックを生成する．

する上での課題を改善するため，学習者の答案の誤りに応じた個別のフィードバックを生成するシステムの構築を目指す．第一の課題については，近年，記述式答案の自動採点システムの研究が盛んに行われている．Mizumoto ら [1] や Sato ら [2] の研究では，答案に対して，項目点の採点根拠となる箇所が明瞭となるように，複数個ある採点項目ごとに採点の根拠箇所と点数を出力している．読解問題の記述式答案に対するフィードバック生成の先行研究として岩瀬ら [3] の研究が挙げられる．岩瀬らは，既存の談話構造ラベルを用いて本文構造グラフとテンプレートを作成し，フィードバックを生成した．本研究では，この手法を拡張し，教育現場で扱われている関係論理構造ラベルを調査し再定義するとともに，問題横断的な答案の分析をもとにフィードバッ

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

クテンプレートの種類を拡充した（4 節）．図 1 に拡張したフィードバック生成の概要を示す．この方法では，自動採点システムが各答案に対して出力する項目点の根拠箇所を利用し，答案内の該当フレーズが参照している本文の箇所を推定し，この参照箇所と模範解答との論理的関係に基づいて予め定めておいたテンプレートを選択することで，学習者の答案の誤りに応じた個別のフィードバックを生成する．システムは，各採点項目に対し 50%以上の精度で適切なフィードバックを生成することが確認できた．

2　データセット

本研究では，理研記述問題採点データセット [1, 4] を題材としてフィードバックの生成を行う．このデータセットには問題ごとに答案と採点者によってアノテーションされた点数のペアが含まれている．採点基準は複数の独立した採点項目に分かれており，採点項目ごとの得点（項目点）とその項目点に関連付けられる答案中の部分文字列（根拠箇所）がアノテーションされている (付録 A)．いくつかの採点基準はさらに細分化された小項目に分類されているが，データセットでは小項目ごとのアノテーションはなされていない．本研究では，小項目に対しても詳細なフィードバックを返却するために，小項目ごとに根拠箇所を追加でアノテーションした．

本研究では，本文中のある文（傍線部）に対して，その理由を説明させる問題を傍線部説明問題と呼ぶ．この問題タイプは，利用するデータセット内で最も数が多い問題であったため，これらをフィードバック生成の対象として取り上げた．また，データセットに含まれる全 13 問のうち，5 問に対して 4 節で説明する答案診断グラフを作成し，この中からさらに 2 問を対象として，フィードバック生成の評価実験を行った．

3　フィードバック生成タスク

本研究では，学習者が記述した答案の誤りに対して，学習者の読解力向上に寄与することを目指しフィードバックを生成する．効果的なフィードバックを生成するためには，学習者の答案を細分化して，誤りを特定する必要がある．そこで，我々は採点基準に含まれる独立した採点項目ごとに，フィードバックの生成を目指す．

Mizumoto ら [1] は，採点項目ごとに項目点を予測し，さらに，その項目点に関連付けられる根拠箇所

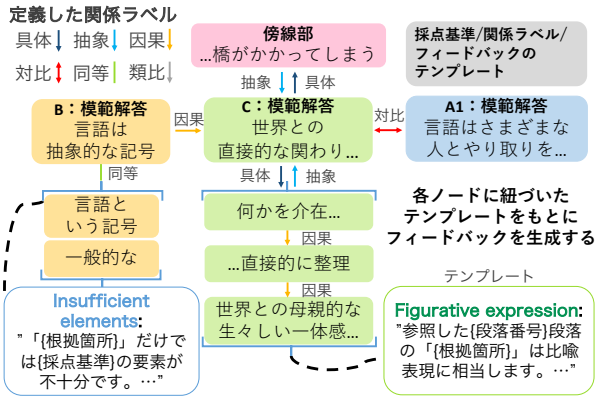

図 2　作成した答案診断グラフの例．ノードには問題本文を適切な単位で分割した文と模範解答を番号で置き換えたものが格納されている．エッジは実際の教育現場で扱われている教科書や参考書を参考にして我々が設計したラベルセットを使用している．ノードに我々が作成したフィードバックのテンプレートが 1 対 1 で紐づいている．

を特定する自動採点タスクを提案した．この自動採点タスクに従い，本研究におけるフィードバック生成は，自動採点モデルから学習者の答案と出力される項目点，およびその根拠箇所を入力として受け取り，適切なフィードバックを出力するタスクとして定義する．

4　フィードバックの設計

本研究では，岩瀬ら [3] の手法を参考にし，より教育現場に適応したフィードバックを生成する．その概要を図 1 に示した．このフィードバック生成において中核的な役割を果たすのは**答案診断グラフ**である．答案診断グラフは，問題文本文に含まれる各文と模範解答をノードとして持ち，それらの論理関係を表す関係ラベルをエッジとして持つグラフ構造である．フィードバック生成の過程では，まず入力された答案の根拠箇所と各ノードの類似度を計算することで，その答案が参照しているノード（参照箇所）を推定する．次に，答案の参照箇所と模範解答との間に張られた関係ラベルより，その差分を明らかにすることでフィードバックに使用するテンプレートを決定する．本節では，このようなフィードバックの設計について詳しく述べる．

4.1　答案診断グラフ

本研究で対象とする傍線部理由説明問題は，本文中のある文について，その理由や根拠を本文に則して説明する問題である．したがって，答案は本文中のある特定箇所の書き抜きや，言い換えや要約表現

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

表 1 採点項目ごとの参照箇所推定精度 (%).Y14_2-2_1-4 と Y15_1-1_1-4 はそれぞれ採点項目 A と C に小項目を含む.

問題	A1	A2	B	C1	C2
Y14_2-2_1-4	67.3	100	96.1	72.4	-
Y15_1-1_1-4	95.6	-	51.1	81.1	50.0

を含む．誤答は誤った箇所の書き抜きや言い換えで構成されるため，その箇所と模範解答との論理関係を明らかにすることで，その誤答と模範解答との間のギャップを分析することができる．そこで，本文中の文と模範解答をノードとして持ち，それらの論理関係をエッジとして持つ**答案診断グラフ**を作成した．答案診断グラフの例を図 2 に示す．

グラフを作成するためには，本文中の文章間の関係をラベル付けする必要がある．このラベルは，後のフィードバック生成時に学習者にとって理解理解しやすい説明を与える関係として設計する必要がある．そこで，Rhetorical Structure Theory (RST) [5] の関係ラベルを出発点として，これらのラベルと教育現場で使用されている複数の教科書や問題集に記載のある関係ラベルとの対応を整理し，最終的なラベルセットを設計した．

答案診断グラフの作成に当たっては，まず，問題本文を一文単位に分割し，設問の参照箇所はさらに適切な単位に人手で細分化し，これらをノードとした．さらに，採点基準をもとに模範解答の該当する部分文字列をノードに追加した．次に，各ノード間にその文間の関係を表すラベルを人手で付与し，それをエッジとした．最後に，各エッジに対して後述するテンプレートを紐づけて，さらに各ノードに段落番号や問題を解くためのヒントなどの，テンプレートに適用するための情報を付与して答案診断グラフを構築した．

4.2 テンプレートの構築

テンプレートの構築にあたって，我々は分析対象の 5 問について，開発セットに含まれる答案の誤答のタイプを問題横断的にいくつかの共通したパターンに分類することを試みた．この手続きは，はじめに各答案に対して理想的と思われるフィードバックを人手で作成し，この内容を類型化して整理し，最終的に 10 種類の問題横断的なフィードバックテンプレートを構築した（付録 B）．例として，模範解答と比べて参照した本文の要素が不十分であるときに用いられるテンプレートを以下に示す:

{ 根拠箇所 } だけでは { 採点基準 } の要素が不十分です．参照した { 段落番号 } と関係のある段落から, { ヒント } に着目して, もう一度確認してみましょう．

このテンプレートには，答案の根拠箇所が参照している答案診断グラフのノードに応じて，参照箇所の段落番号や該当する採点基準の抜粋など，付加的な情報を挿入することができる．

一方で，問題や採点項目ごとに特有の性質が存在するため，問題横断的な枠組みで対応することが困難な誤りタイプが存在することもわかった．そこで，我々は採点基準に応じて，採点項目固有のテンプレートを追加した．採点項目固有のテンプレートについては，以降の生成実験で用いる 2 問に対してのみ作成した．この 2 問に含まれる計 8 個の採点項目のうち，3 つの採点項目については固有のテンプレートが用いられた．

4.3 フィードバックの生成

フィードバックの生成は，答案の根拠箇所を入力として採点項目ごとに行う．採点根拠箇所のトークン列に対し，訓練済み Sentence-BERT [6][1] を用いて埋め込み表現を得る．同様に，答案診断グラフの各ノードに対応するトークン列に対しても予め埋め込み表現を得ておき，これらと答案の根拠箇所のコサイン類似度を計算し，最も類似度の高いノードを答案の参照箇所と推定する．この結果，答案診断グラフにおいて，対象の部分採点項目に対して，推定された参照箇所が答案に記載された場合に選択されるテンプレートが決定される．このテンプレートは，模範解答に対応するノードと答案診断グラフ内の各ノードとの論理関係から予め各ノードに対して紐付けられている．最終的に，参照箇所のノードに含まれた段落番号などの付加情報をテンプレートに埋め込むことでフィードバックを生成する．

5 実験

実験では 4 節で提示したフィードバック設計に従ってフィードバックの自動生成を行い，その有効性について確かめる．特に本研究で提示した設計では，各答案の参照箇所を正しく推定することが，適切なテンプレートを選ぶために重要である．そこで答案の参照箇所の推定精度に注目してフィードバッ

1) 計算速度と性能を考慮して https://huggingface.co/microsoft/MiniLM-L12-H384-uncased を用いた

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

表 2 生成されたフィードバックの例．一つの部分採点項目に対応するフィードバックのみを記載している．根拠箇所の表現はほとんど同じであるにも関わらず，異なるフィードバック文が出力されている．上の例では正しく参照箇所を推定できているが，下の例では，参照箇所の推定を誤っているため不適切なテンプレートが用いられている．

答案	A1 - フィードバック	テンプレート名
言語によって何かを表現することは、言葉という記号では生のままで…	「言語によって何かを表現する」だけでは要素が不十分です．参照した 9 段落と関係のある段落から…	Insufficient elements
言葉によって何かを表現しようとすると言葉では全てを伝えられない可能性が…	言語は様々な人とやり取りを行うという要素は適切に読み取れています．	Rubric criteria available

ク生成システムの評価を行う．

5.1 実験設定

岩瀬ら [3] 同様，本研究でも学習者にとって有益なフィードバックを生成することが目的であるため，採点項目について満点と 0 点の答案を扱わない．また，フィードバックの生成について集中的に分析を行うために，各採点項目の根拠箇所は，データに付与されている教師信号を利用する．4.2 節で述べたように，実験では Y14_2-2_1-4 と Y15_1-1_1-4 の 2 問を用いて評価した．それぞれ評価データとして 205 件（採点項目 C のみ 203 件）と 90 件の答案を用いた．

5.2 結果

データセットには参照箇所に関する情報は含まれていないため，参照箇所の推定精度を人手で確認した．参照箇所の推定精度を表 1 に示す．Y14_2-2_1-4 の採点項目 A2,B や Y15_1-1_1-4 の A のように 95% を超える高い精度を示した採点項目があることが分かる．これらの採点項目では，根拠箇所が比較的短く，本文中の特定単語の抜き書きのみが根拠箇所として答案に含まれているため根拠箇所の多様性に乏しいことが，その理由として考えられる．一方で，Y15_1-1_1-4 の採点項目 B や C2 のように参照箇所を適切に推定することが困難な採点項目も見られた．これは本文の内容を要約し，学習者自身の言葉で表現している答案が多いことや根拠箇所の字数が多いことなどが要因として考えられる．

表 2 に，フィードバックの生成例を示す．この例から「表現すること」と「表現」のように，動詞の活用形の差のような小さな要素が参照箇所の推定に影響を及ぼすことが確認できる．この結果は適切なフィードバックテンプレートを選択するためのさらに優れた方法をを考える必要があることを示唆している．

6 関連研究

教育工学におけるフィードバック 教育分野において，フィードバックは学習者が現在の理解と「望ましい理解」のギャップを縮小するために有効である [7]．一方で，学習者にとって「良いフィードバック」を一意に定めることは難しいことが知られている [8, 9, 10]．その要因として，フィードバックの受容には学習者の好みや習熟状況，そして内容の間接性など様々な要因が関与していることが挙げられる．しかし，いずれの場合でも，テストの点数に加えて，短いコメント文の形でフィードバックすることは有効であることが示されている [11, 12]．本研究では，問題の解き方の過程を提示する間接的なフィードバックを選択した．

フィードバックの自動生成 本研究で扱う読解問題の記述式答案に対するフィードバック生成の先行研究として岩瀬ら [3] の研究が挙げられる．岩瀬らは英語の談話関係コーパスの代表例である Penn Discourse Treebank (PDTB) [13] のタグセットを用いて，問題本文に含まれる文に対して関係ラベルを付与し本文構造グラフを作成した．関係ラベルに応じたテンプレートを作成することでフィードバック生成を試みた．

7 おわりに

本研究では，記述式問題の本文と模範解答との論理的関係性を表す答案診断グラフを用いてフィードバックを生成する枠組みを提案した．また，答案の誤答タイプを分析し，フィードバックのテンプレートを構築した．実験により，我々のシステムは各採点項目に対し 50%以上の精度で適切に参照箇所を推定し，フィードバックを生成できることが確認できた．今後は，生成方式のより詳細な検討によって精度の改善をはかるとともに，教育現場での実証実験を行い，本研究で設計したフィードバックの有効性について調査することを予定している．

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

謝辞

本研究は JSPS 科研費 JP22H00524, JST 次世代研究者挑戦的研究プログラム JPMJSP2114 の助成を受けたものです．実際の模試データを提供していただいた学校法人高宮学園代々木ゼミナールに感謝します．また，フィードバックのテンプレート作成にあたり，東北大学大学院情報科学研究科 長濱准教授より助言を賜りました．ここに深謝の意を表します．

参考文献

[1] Tomoya Mizumoto, Hiroki Ouchi, Yoriko Isobe, Paul Reisert, Ryo Nagata, Satoshi Sekine, and Kentaro Inui. Analytic score prediction and justification identification in automated short answer scoring. In **Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications**, August 2019.

[2] Tasuku Sato, Hiroaki Funayama, Kazuaki Hanawa, and Kentaro Inui. Plausibility and faithfulness of feature attribution-based explanations in automated short answer scoring. In Maria Mercedes Rodrigo, Noburu Matsuda, Alexandra I. Cristea, and Vania Dimitrova, editors, **Artificial Intelligence in Education**, pp. 231–242, Cham, 2022. Springer International Publishing.

[3] 岩瀬裕哉, 舟山弘晃, 松林優一郎, 乾健太郎. 文章構造グラフを用いた 国語記述式答案への自動フィードバック生成. 言語処理学会 第 29 回年次大会, pp. 1333–1338, 2023.

[4] Hiroaki Funayama, Yuya Asazuma, Yuichiroh Matsubayashi, Tomoya Mizumoto, and Kentaro Inui. Reducing the cost: Cross-Prompt pre-finetuning for short answer scoring. In **Artificial Intelligence in Education**, pp. 78–89. Springer Nature Switzerland, 2023.

[5] William C. Mann and Sandra A. Thompson. Rhetorical structure theory: A theory of text organization. Technical Report ISI/RS-87-190, Information Sciences Institute, June 1987 1987.

[6] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing**. Association for Computational Linguistics, 11 2019.

[7] John Hattie and Helen Timperley. The power of feedback. **Review of Educational Research**, Vol. 77, No. 1, pp. 81–112, 2007.

[8] 村山航. テスト形式が学習方略に与える影響. 教育心理学研究, Vol. 51, No. 1, pp. 1–12, 2003.

[9] Valerie J. Shute. Focus on formative feedback. **Review of Educational Research**, Vol. 78, No. 1, pp. 153–189, 2008.

[10] 鈴木雅之. ルーブリックの提示による評価基準・評価目的の教示が学習者に及ぼす影響. 教育心理学研究, Vol. 59, No. 2, pp. 131–143, 2011.

[11] Hedy McGarrell and Jeff Verbeem. Motivating revision of drafts through formative feedback. **ELT Journal**, Vol. 61, No. 3, pp. 228–236, 07 2007.

[12] Zichao Li, Prakhar Sharma, Xing Han Lu, Jackie Cheung, and Siva Reddy. Using interactive feedback to improve the accuracy and explainability of question answering systems post-deployment. In **Findings of the Association for Computational Linguistics: ACL 2022**. Association for Computational Linguistics, 2022.

[13] Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. The Penn Discourse TreeBank 2.0. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, editors, **Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)**, Marrakech, Morocco, May 2008. European Language Resources Association (ELRA).

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

A　理研記述式問題データセット

図 3 に理研記述式問題データセットより抜粋した問題，答案およびその項目点と根拠箇所，そして採点基準の例を示す．この問題の採点項目 A における採点基準はさらに A①，A②という小項目に分類されている．本研究では, これらの小項目に対して新たに根拠箇所と項目点のアノテーションを行った．

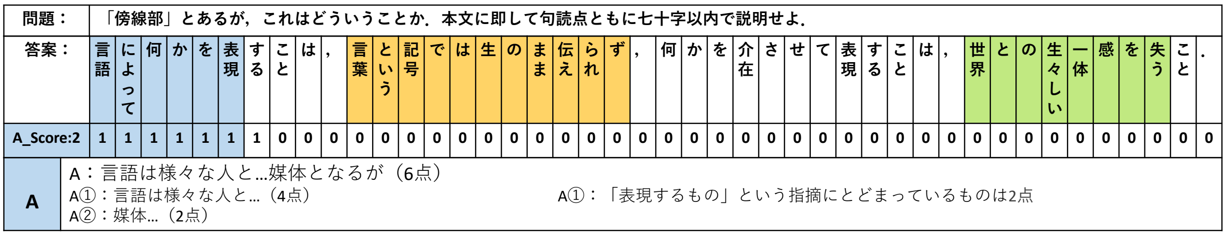

図 3　理研記述問題採点データセットの答案（上）と採点基準の抜粋（下）．（上）：A_score, B_score,⋯は項目点, 0/1 は根拠箇所の有無を表している．（下）：左側：A は採点基準, 右側：A は採点項目，A①, A②⋯は細分化した小項目．

B　フィードバックテンプレート

表 3 に作成したテンプレートを示す．本研究では，10 個の問題横断的なテンプレートを作成し，さらに，問題や採点基準特有の誤り方に対応するために，問題ごとのフィードバックテンプレートを作成した．フィードバックの生成では，{} で囲まれた部分を，4.3 節で説明した方法で対応する表現に置き換える．

表 3　作成したテンプレートの一覧．{} で囲まれた部分をそれぞれ対応する表現や情報に置き換えることでフィードバックを生成する．

テンプレート名	テンプレートの種類	フィードバック文
Rubric criteria available	採点基準 (各基準全体) が読み取れている場合	{ 採点基準 } という要素は適切に読み取れています．
Includes underlined words	傍線部の言葉を含んでいる場合	今回は傍線部は { 傍線部の内容 } の説明を求められているため, 答案に { 答案 } と用いることはできません．{ 答案 } を適切な形で言い換えてみましょう．
Insufficient elements	要素が不十分である場合	{ 答案 } だけでは { 採点基準 } の要素が不十分です．参照した { 段落番号 } と関係のある段落から, { ヒント } に着目して, もう一度確認してみましょう．
Figurative expression	比喩表現が含まれている場合	参照した { 段落番号 } の { 答案 } は比喩表現に相当します．要約問題では比喩表現で用いられている言葉の性質の説明を記述する必要があります．{ 答案 } と関係のある段落から { ヒント } 踏まえた上で参照箇所を広げて，自分の言葉で一般化して考えてみましょう．
Specific examples	具体例が含まれている場合	参照した { 段落番号 } の { 答案 } は具体例に相当します．要約問題では具体例ではなく，一般化した表現を記述する必要があります．{ 答案 } と関係のある段落から同等の意味で言い換えられた記述を探してみましょう．
Deduction due to direct quotations	書き抜きによる減点をされている場合	参照した { 段落番号 } の { 答案 } の着眼点は適しています．一方で，今回はより適した表現をするために参照した段落と同等の記述がされた部分を探してみましょう．
Vague expressions	漠然とした表現が含まれている場合	{ 答案 } は漠然としているため，違う表現で言い換えましょう．
Demonstrative pronouns	指示語が含んでいる場合	要約問題では指示語はしっかりと明示する必要があります．今回の答案は何に当たるのか参照した段落番号と関係のある段落からもう一度考えてみましょう．
Punctuation errors	句読点ミスをしている場合	また、記述式では句点（“.”）で終わらせるようにしましょう．
Demonstrative pronouns	指示語が含んでいる場合	要約問題では指示語はしっかりと明示する必要があります．今回の答案は何に当たるのか参照した段落番号と関係のある段落からもう一度考えてみましょう．
Default	設問の参照ノード以外を参照している場合	問題をもう一度よく読んで回答しましょう．
{ 問題特有のテンプレート }	各ルーブリックに記載された表現のうち答案に記述されていない表現	e.g. 傍線部の中の「橋」という言葉は筆者がわかりやすく伝えるための比喩表現です．よって，まず「橋」とはどのような意味で用いられているのか考えた上で明記する必要があります．

This work is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).