TYPIC: A Corpus of Template-Based Diagnostic Comments on Argumentation

Shoichi Naito$^{1,2,5}$, Shintaro Sawada$^{2}$, Chihiro Nakagawa$^{3,2}$, Naoya Inoue$^{4,2,*}$, Kenshi Yamaguchi$^1$, Iori Shimizu$^1$, Farjana Sultana Mim$^1$, Keshav Singh$^1$, Kentaro Inui$^{1,2}$

$^1$Tohoku University, $^2$RIKEN, $^3$Osaka Prefecture University, $^4$Stony Brook University, $^5$Ricoh Company, Ltd.

Abstract

Providing feedback on the argumentation of the learner is essential for developing critical thinking skills, however, it requires a lot of time and effort. To mitigate the overload on teachers, we aim to automate a process of providing feedback, especially giving diagnostic comments which point out the weaknesses inherent in the argumentation. It is recommended to give specific diagnostic comments so that learners can recognize the diagnosis without misinterpretation. However, it is not obvious how the task of providing specific diagnostic comments should be formulated. We present a formulation of the task as template selection and slot filling to make an automatic evaluation easier and the behavior of the model more tractable. The key to the formulation is the possibility of creating a template set that is sufficient for practical use. In this paper, we define three criteria that a template set should satisfy: expressiveness, informativeness, and uniqueness, and verify the feasibility of creating a template set that satisfies these criteria as a first trial. We will show that it is feasible through an annotation study that converts diagnostic comments given in a text to a template format. The corpus used in the annotation study is publicly available.

Keywords: argument, argumentation, debate, formative feedback, diagnostic comment

1. Introduction

Argumentation and debate are known to be effective tools for developing critical thinking skills (Roy and Macchiette, 2005). In such argumentation-based education, teachers’ feedback helps learners efficiently develop their skills (Durón et al., 2006; Tsui, 1999; Paulus, 1999). Learning becomes more efficient when specific feedback is given to the learners (Shute, 2008). However, it is impractical to ask all teachers to do this because they will need a substantial amount of time to provide specific feedback to each student, and not all teachers have been trained to teach argumentation (Driver et al., 2000). It would be highly beneficial to develop a technology that automatically gives specific feedback to students.

Essay Scoring aims for assessing students’ arguments and giving feedback as a score. Some studies give a single holistic score for the entire essay (Burstein and Chodorow, 1999; Dong and Zhang, 2016), while others give multi-dimensional scores such as organization, clarity, and justification (Persing et al., 2010; Persing and Ng, 2013; Persing and Ng, 2014; Persing and Ng, 2015; Wachsmuth et al., 2017b; Carlile et al., 2018). However, we argue that these score-based feedback is not specific enough for students to develop their skills. Consider an example argument in Fig. 1 where we imagine students receive feedback such as Partially justified: The thesis justifies some of the author’s opinions (Carlile et al., 2018), these students may not know how to revise their arguments because they are not told how weak their justification is.

Here, we aim to support learners by automatically giving more specific diagnostic comments highlighting the weaknesses of their argumentation. One challenge with this approach is that it is not clear what task setting should be designed. One possible approach is to formulate it as a pure generation task, but there is an evaluation issue. Automatic evaluation metrics such as BLEU (Papineni et al., 2002) are controversial in generation tasks such as machine translation and dialogue (Liu et al., 2016; Reiter, 2018).
However, it is costly and time-consuming to evaluate generated comments manually. To address this issue, we propose formulating the task of giving diagnostic comments by \textit{template selection} and \textit{slot filling} as shown in Fig. \ref{fig:overview}. The task is to select the most likely template from a predefined set of templates and to extract or generate phrases for slots in the selected template. We assume that diagnostic comments that occur frequently are limited, and having them covered with a predefined set of templates is sufficient for practical use. Compared with the generation approach, this formulation enables us to use more interpretable evaluation measures such as accuracy, precision, and recall of template selection. This also helps with error analysis, such as determining which diagnoses the model fails to recognize.

Our research question is the following. Can we create a set of templates that satisfies the following properties?: (i) \textit{expressive}: represent a reasonable amount of common diagnostic comments, (ii) \textit{informative}: preserve the meaning of original comments, and (iii) \textit{unique}: ensure one-to-one mapping between comments and templates. To investigate this, we collect 1,000 counterarguments and ask people who are experienced in debate education to provide feedback on some of them (§3). Further, we identify common patterns of feedback to induce a predefined set of templates (§4) and evaluate the quality of the induced set of templates (§5). Our main contributions can be summarized as follows:

- We propose expressive, informative, and unique templates for argumentative diagnostic comments. Our study shows that 92.2\% of unseen diagnostic comments can be represented using our templates with moderate inter-annotator agreement (Cohen’s Kappa of 0.517), 78\% of which are judged as informative.
- We publicly release TYPIC corpus\footnote{TYPIC corpus will be publicly available at: \url{https://github.com/cl-tohoku/TYPIC}} consisting of 1,000 counterarguments, 197 of which are annotated with 1,082 diagnostic comments in both natural language and template formats.

\section{Related Work}

\textbf{Essay scoring } Essay scoring has been studied as a diagnostic tool for argumentation \cite{Dikli2006}. Some studies give a single holistic score for an entire essay \cite{Burstein1999}, \cite{Dong2016}, while others score for multiple dimensions such as organization \cite{Persing2010}, clarity \cite{Persing2013}, prompt adherence \cite{Persing2014}, and argument strength \cite{Persing2015}. \cite{Wachsmuth2017} conducted a comprehensive survey on the dimensions of assessing argumentation and created a dataset with scores from 15 dimensions. Furthermore, \cite{Carlile2018} annotated the scores of multiple dimensions on Persuasive Essay Corpus, and \cite{Ke2019} on the ICLE. In particular, \cite{Carlile2018} annotated the scores to the fine-grained target of the argumentative discourse unit. However, these approaches did not consider giving specific diagnostic comments. Even if a scoring rubric is given as a diagnostic comment, it will produce an abstract comment such as “Unjustified” and “Partially justified.” Because the reasons for the scores are not entirely clear, it is challenging for learners to understand why and what improvements are required.

\textbf{Missing premise detection } Several studies give feedback to indicate the lack of premises for the results of analyzing argumentation structure. Some studies have proposed an annotation scheme to detect the absence of the premise, considering that a proposition must be supported by an appropriate premise to be evaluable \cite{Park2015}, \cite{Park2018}. \cite{Morio2018} and \cite{Egawa2020} extended Park’s annotation scheme to capture interactions between users for online discussion forums and \cite{Iida2019} proposed an agent system that prompts users to provide additional premises when few premises support a claim. \cite{Wambgsanns2020} proposed a support system that analyzes an argumentation made by learners, displays the argumentation structure, and highlights unsupported claims. However, these approaches cannot provide an in-depth diagnosis of what premise is lacking.

\textbf{Revision support for argumentative writing } There exists some research that supports the revision process of argumentative writing. \cite{ArgRewrite} automatically classifies the objective of the revision based on the student’s draft and revision from eight categories, such as reasoning, rebuttal, evidence, fluency, etc. \cite{Zhang2014}, \cite{Zhang2015}, \cite{Zhang2016}, \cite{Zhang2017}, \cite{Afrin2021}. By showing the classification results, learners can check whether a revision is consistent with their intentions. These studies focus on feedback on the revisions made by learners, and not on what revisions should be made.

\textbf{Recognition of weakness in argumentation } Some studies recognize the types of weakness in argumentation and give them to learners as feedback. \cite{Stab2017} created a dataset that annotated whether an argument satisfied sufficiency, a criterion evaluated whether premises provided sufficient evidence to accept or reject a claim. \cite{Persing2017} defined five error types: grammar error, lack of objectivity, inadequate support, unclear assertion, and unclear justification, as factors that make an argument unpersuasive and annotated the presence of these errors in addition to the holistic score. However, these ap-
Original argument
Hello everyone. Today’s topic is “Homework should be abolished”. We have two points: The first point is “free time” and the second point is “decrease burden on teachers”. I will explain the first point of “free time”. We believe that if homework were to be abolished, we could have more free time. As a result, we could do more of what we really wanted like club activities, hobbies, or playing with friends. In my case, I go to tennis club after class until 5:00 pm and then I go to cram school until 8:00 pm. After this full day, I arrive at my home around 8:40 pm to eat dinner and take a shower. At nearly 10:00 pm I start my homework. I have a lot of homework. As a result, I go to bed late at night at nearly 1:00 am in the morning and I don’t have the opportunity to sleep for a long period of time. It is not healthy. Therefore, homework should be abolished. Thank you.

Counterargument
They said that present situation are lack of free time for students and students cannot do what really they want, so homework should be abolished. However, not all students are lack of time in present situation. For example, I belonged to Judo club in junior high and high school and I could join extracurricular activity that I wanted to join in, but I could implement homework and I could guarantee enough time for sleeping. Even if homework is accumulated, some people can control the tasks. It is also training to manage having time in school. Therefore, lack of time is not necessarily caused by a lot of homework, but lack of management skills, therefore, the government opinion is rejected.

Diagnostic comments for Counterargument
The specific example of being able to balance homework and club activities is no more than a personal story and is not generalized.

Figure 2: An example of parliamentary debate style argumentation and diagnostic comment on counterargument to be collected.

Proaches can provide the presence of errors, but they do not provide a specific diagnosis that is specific enough for the learner to recognize how to correct them. The most related work to ours is Song et al. (2014). They proposed a modified Critical Question (CQ) based on Argumentation Scheme (Walton et al., 2008), which provides specific feedback on the weaknesses inherent in the argumentation. However, datasets annotated with CQ are not publicly available, and feasibility in terms of expressiveness and semantic change cannot be assessed. Besides, their annotation scheme of selecting a single segment, such as a sentence or clause, and assigning a static CQ is insufficient to provide specific feedback. For example, if a CQ is about cause and effect, at least two segments must be selected, cause and effect, otherwise ambiguity remains. Additionally, diagnosis using phrases not mentioned in the argumentation cannot be expressed. Our annotation scheme in the form of a template with slots can naturally express both.

3. Dataset

3.1. Counterarguments
We collect counterarguments in the form of parliamentary debate as the target for giving diagnostic comments, as seen in Fig. 2. A parliamentary debate is an impromptu debate in which two groups, the government and the opposition, argue about a given topic. The government takes a position in favor of the topic, while the opposition takes a position against the topic and they give a speech in turn. This study focuses on the first two speakers, the Prime Minister (PM) on the government side and the Leader of the Opposition (LO) on the opposition side (i.e., original arguments and counterarguments, respectively).

We prepared 10 PM speeches as the original argument on the topics “Homework should be abolished” and “Death penalty should be abolished,” as shown in Table 1. For each of the 10 PM speeches, 100 LO speeches are collected. Out of 1,000 LO speeches, 250 speeches are written by experienced debaters affiliated with the Parliamentary Debate Personnel Development Association (PDA). The remaining 750 speeches are written by Amazon Mechanical Turk workers with Master Qualification. We pay $1.60 as a reward per speech.

3.2. Diagnostic Comments
We randomly selected 200 LO speeches (100 for each topic) from the collected counterarguments and asked four assessors to give diagnostic comments on these speeches. The assessors have at least 4 years of debating and judging experience in high school debate competitions held by PDA. The assessors read an original argument and counterargument, select the target sentences to be diagnosed, and give a diagnostic comment in natural language sentences. The instruction is designed to give diagnostic comments on content rather than grammatical errors or expressions to focus on developing thinking skills. In particular, diagnostic comments are given in terms of relevance to the original argument, justification, and appropriateness of the examples.

All counterarguments are evaluated by two assessors each. The annotation is conducted under the descrip-

https://pdpda.org/
Homework should be abolished

HW1 Abolishing homework gives students more free time
HW2 Forcing students to do homework makes them passive in character
HW3 It is not good for students to be obliged to study by their teachers or parents
HW4 Students have memorized the incorrect way to study with homework
HW5 Schools should take responsibility for the academic skills of children, not parents at home

Death penalty should be abolished

DP1 Death penalty is an inhumane punishment
DP2 Abolishing death penalty will prevent the ending the life of innocent people
DP3 Because of the high stress on the executioner, death penalty should be abolished
DP4 Death penalty deprives criminals of the opportunity for rehabilitation
DP5 The society is brutalized by the use of death penalty

Table 1: Points of the original argument.

| # Counterargument | 1,000 |
|--------------------|-------|
| Avg. tokens per argument | 124.0 |
| Avg. sentences per argument | 7.1 |
| # Diagnostic Comments | 1,082 |
| Avg. # comments per argument | 5.5 |

Table 2: Statistics of TYPIC Corpus.

The challenge is the difficulty in creating a template set that meets all these criteria. To see the trade-off between these properties, assume two extreme cases: (i) a template set consisting of only one versatile template (e.g., *This argument is unpersuasive.*), and (ii) a template set consisting of very specific templates designed for every single diagnostic comment. Case (i) satisfies the expressiveness property because the abstract template can represent almost all types of diagnostic comments. It can also satisfy the uniqueness property since there is only one template. However, the informativeness is not satisfied because the template-based representation is significantly different from the original diagnosis comment. In Case (ii), on the other hand, the informativeness is satisfied since the template is identical to the original diagnostic comment. However, the expressiveness cannot be satisfied because these specific templates are rarely applied to other diagnostic comments.

4. Inducing Template Set

As discussed in §1, we formulate the task of argumentative diagnosis as a template-based task. We assume that frequently occurring diagnostic comments are limited, and having them covered with a predefined set of templates is sufficient. Toward this end, we manually induce a predefined set of templates from the collected arguments.

4.1. Design Choice

We assume that an ideal set of templates should satisfy the properties listed below:

**Expressive** It should be able to cover most of the diagnostic comments. This ensures that learners receive a wide variety of diagnoses.

**Informative** It should preserve the meaning of the original diagnostic comment and maintain the same level of specificity. This is important for our goal of giving specific diagnostic comments.

**Unique** Only one unique template must be identified for one diagnostic comment. It is essential to ensure the reliability of the annotations.

The challenge is the difficulty in creating a template set that meets all these criteria. To see the trade-off between these properties, assume two extreme cases: (i) a template set consisting of only one versatile template (e.g., *This argument is unpersuasive.*), and (ii) a template set consisting of very specific templates designed for every single diagnostic comment. Case (i) satisfies the expressiveness property because the abstract template can represent almost all types of diagnostic comments. It can also satisfy the uniqueness property since there is only one template. However, the informativeness is not satisfied because the template-based representation is significantly different from the original diagnosis comment. In Case (ii), on the other hand, the informativeness is satisfied since the template is identical to the original diagnostic comment. However, the expressiveness cannot be satisfied because these specific templates are rarely applied to other diagnostic comments.

4.2. Templates

Having the design choice in our mind, we manually designed a template set by analyzing common patterns of the diagnostic comments in DEVSET. Table 3 summarizes our template set. Each template consists of a natural language comment and placeholders (henceforth, slots). These slots are intended to be filled with phrases extracted from input arguments or newly generated. Our templates can be categorized into a standard dimension used in argumentation quality assessment; see Wachsmuth et al. (2017b) and Wachsmuth et al. (2017a) for more information.

4.3. Task Setting

Given the predefined templates, we formulate the task of argumentative diagnosis as two subtasks: **template selection** and **slot filling**, as shown in Fig. 1.

**Template Selection** Given (i) an argument, (ii) its counterargument, and (iii) the target argument (indicated by sentences in the counterargument), the task is to identify the target argument’s flaw and to choose a template from a list of a predefined set of templates.
| Quality Dimension | Category | Template |
|-------------------|----------|----------|
| Local Acceptability | CA2 | なぜ \( x \) によって \( y \) という悪い結果が起こるのかが不明確 \( It \) is unclear why \( x \) causes a bad result of \( y \) |
| VAL1 | なぜ \( y \) にとって \( x \) が良いことなのかが不明確 \( It \) is unclear why \( x \) is good for \( y \) |
| CLS1 | なぜ \( x \) は \( y \) という特性を持つと言えるのかが不明確 \( It \) is unclear why \( x \) has the property of \( y \) |
| CLS2 | なぜ \( z \) という点において \( x \) と \( y \) は同じ/似ているのかが不明確 \( It \) is unclear why \( x \) and \( y \) are same/similar in terms of \( z \) |
| Local Sufficiency | EX3 | \( x \) というのは限定的な状況である \( It \) is a limiting situation that \( x \) |
| CMP2 | \( y \) を実現するのに、なぜ \( z \) ではなく \( x \) という方法が良いのかの説明が不足している \( It \) lacks an explanation of why \( x \) is a better method to realize \( y \) instead of \( z \) |
| Local Relevance | LR1 | なぜ \( x \) という理由が \( y \) という結論を支持するのかが不明確 \( It \) is unclear why a premise \( x \) supports a claim \( y \) |
| Global Relevance | GR2 | 肯定側の \( x \) という主張に \( y \) というのは直接的な反論になっていない \( It \) is not a direct objection to the government’s claim \( x \) to say \( y \) |
| Global Sufficiency | GS1 | なぜ否定側の \( y \) という主張よりも否定側の \( x \) という主張が優位だと言えるのかが不明確 \( It \) is unclear why the opposition’s claim \( x \) is superior to the government’s claim \( y \) |
| GS2 | 肯定側からの \( x \) という反論が想定される \( It \) is expected that the government side will object that \( x \) |

Table 3: An excerpt of the template set used in the annotation study (See Table 8 in the Appendix for the full version).

that best reflects the flaw. We assume that there can be multiple weaknesses for one target argument and consider template selection as a multi-label classification task. That is, the output is formally defined as a label vector \( l = (l_1, l_2, \ldots, l_n) \), where \( l_i \in \{0, 1\} \) indicates whether \( i \)-th template is an answer or not, and \( n \) is the size of the template set. One can evaluate the models’ prediction with evaluation metrics for multi-label classification tasks, such as F1 and accuracy.

**Slot Filling** Given the selected template in the template selection, the task is to fill the slots of the template. We assume some slot fillers can be extracted from an original argument or counterargument, and some of them must be generated. One can evaluate the models’ prediction with the similarity between the predicted fillers and gold-standard fillers such as n-gram overlaps.

### 5. Annotation Study

We verify the template set described in §4.2 satisfies the three criteria through an annotation study. In the annotation study, we annotate the diagnostic comments in EVAL.SET with a template using the predefined template set (§5.1). We then use corresponding evaluation metrics to see if the three criteria have been met (§5.2).

#### 5.1. Annotation Procedure

We conduct an annotation study to convert diagnostic comments in natural language text into template form using a template set.

The annotation of the template application involves template selection and slot filling. In the template selection, the annotator selects a template from the template set that expresses the same point as the diagnostic comment. In slot filling, the annotator fills the slots of the selected template with a phrase.

1. **Diagnostic comment (natural language):**
   
   倫理を吹くすれば、生徒は性格的に受動的になるという主張の根拠となる理由や例がなく、PMの主張を否定しきれていない。
   
   No reasons or examples for the argument that students will become more passive in character if homework is abolished, failing to completely refute the PM’s argument.

   For the example above, the diagnosis questions the causality between abolishing homework and the passive personality of students. In template selection, the appropriate template is CA2, which asks for causality between \( x \) and a bad consequence \( y \). The two slots \( x \) and \( y \) in template CA2 are filled with \( x = \text{abolishing homework} \) and \( y = \text{students becoming passive in character} \), respectively. The final annotation result is as follows.

2. **Diagnostic comment (templated):**
   
   (CA2) なぜ 倫理吹くによって 生徒は性格的に受動的になるかが不明確
   
   (CA2) It is unclear why abolishing homework
| Score | Description                                                   |
|-------|---------------------------------------------------------------|
| 3     | It gives the same diagnosis as the original, without lacking specificity. |
| 2     | It gives the same diagnosis, but is less specific.            |
| 1     | It gives different diagnosis than the original.               |

Table 4: Description of the informativeness scores.

If there is no applicable template in the template set, the annotator selects “Not Applicable.”

Two annotators who are not authors, annotate the template application. To calculate inter-annotator agreement (IAA), 74 diagnostic annotations are annotated with overlap between two annotators. One round of calibration is conducted before the main annotation.

5.2. Evaluation Metrics

We evaluate how well the template set satisfies the three criteria described in §4.1.

**Expressiveness** is evaluated using the coverage of the template set for unseen diagnostic comments. The template set is built based on the observation of 25% of the diagnostic comments. We evaluate the coverage of the remaining 75% of diagnostic comments as a metric.

**Informativeness** is evaluated to determine whether it can express the same points as the diagnostic comments in the text without lacking specificity. We use crowdsourcing to evaluate the template’s informativeness manually. We compare the diagnostic comments in the text with those after applying the template and evaluate them using a numerical score from 1 to 3. Table 4 shows the rubric. Each diagnostic comment is judged by five workers, and the results are aggregated by majority voting.

**Uniqueness** is evaluated by IAA of template selection. This is because if the appropriate template can be uniquely identified, the IAA for template selection should inevitably be high. We use the agreement rate for the 74 cases where two annotators annotate overlapping as a metric.

5.3. Results

**Expressiveness** The coverage of the template set for unseen comments is 92.2% (757/821). Fig. 3 shows the distribution of the templates. Some examples of diagnostic comments in natural language and those after applying the template are shown in Table 5. The result indicates that the types of frequently occurring diagnostic comments are limited and that typical templates can cover numerous diagnostic comments.

**Informativeness** Fig. 4 shows the distribution of the informativeness scores. The result shows that 78.6% (857/1090) have the same specificity as a diagnostic comment in the text even after applying a template. This indicates that template-based diagnostic comments can adequately express the intent of the original diagnosis.

**Uniqueness** The IAA is 0.517 for Cohen’s Kappa (Cohen, 1960), which corresponds to the moderate agreement (Landis and Koch, 1977). Additionally, we evaluate whether the contents of the slots are the same for cases where the template selection is agreed. The results of the manual evaluation showed that 89% (65/73) of the slots were substantially consistent in content. These results indicate that annotators can select the appropriate template and fill in the slots with a certain degree of reliability.

Overall, we conclude that it is feasible to create a template set that satisfies three criteria of expressiveness, informativeness, and uniqueness.

---

3 We used Yahoo! Crowdsourcing [https://crowdsourcing.yahoo.co.jp/](https://crowdsourcing.yahoo.co.jp/) as a platform.

4 The inter-annotator agreement between workers is 0.265 in Krippendorff’s α with ordinal distance function (Krippendorff, 1980).

5 The agreement of the slots is evaluated based on lenient matches. If the fillers of a slot have the same meaning, it is considered as agreed even if the phrases are not exactly the same.
### 5.4. Analysis

#### 5.4.1. Characteristics of Fillers

To analyze the difficulty of slot filling, we randomly select some diagnostic comments and analyze whether or not fillers can be extracted from the original argument or counterargument. Table 5 shows the proportion of fillers that can be extracted from an original argument or counterargument.

For 75.9% (126/166) of the fillers, they can be extracted from the arguments with minor modifications such as changing the part of speech and paraphrasing. For 8.4% of the fillers, the basic phrases can be extracted, but require substantial changes. The following is a case where the extracted concepts need to be combined.

(3) **Counterargument (excerpt):**

That is to say even if abolishing homework, students become passive in character. This is because students are instructed by teachers in club activity or cram school in many situation.

**Diagnostic comment (templated):**

(CLS2) It is unclear why **passivity due to homework** and **passivity due to club activities** are same/similar.

Consider the second filler, “passivity due to club activity”. The phrases “passive in character” and “club activity” can be extracted from the counterargument, but they need to be combined to make the filler.

For 15.7% of the fillers, it cannot be extracted from an original argument or counterargument. Some fillers belong to this case for diagnoses that provide the learner a new perspective.

(4) **Diagnostic comment (templated):**

(CMP2) It lacks an explanation of why **homework** is a better method to realize **students’ mastery of basic skills** instead of **independent study**.

In the above case, the third filler, “independent study,” which is taken up as a comparison, is not mentioned in the original argument or counterargument.

(5) **Diagnostic comment (templated):**

(GS2) It is expected that the government will object that **the death penalty is still one of the factors in the brutalization of modern society.**

Similarly, the filler of the template for the anticipated objection (GS2) is not extracted, because it highlights aspects that have not been considered in the counterargument (see Example (3)).

This result suggests that the approach to extracting fillers from the arguments can cover most fillers. We think this characteristic of the filler will increase the feasibility of automated models. Despite recent breakthroughs in pre-trained language model, it is still challenging to generate argumentative knowledge with reasoning (Saha et al., 2021). The approach to extracting fillers can alleviate the difficulty of the problem.

#### 5.4.2. Different Diagnoses for the Same Target

To analyze whether the template selection should be a single-label classification or a multi-label classification, we examine how many cases in which the same
home is a better method to realize curiosities. These diagnostic comments differ in how the relationship between homework and students’ curiosities is viewed, but both are reasonable. The results indicate that there are multiple reasonable diagnoses for the same target sentence. Therefore, we adopt multi-label classification as a task setting of template selection.

6. Discussion

Limitations We did not verify whether the template set satisfies the three criteria for diagnostic comments on different topics and given by different assessors. Whether it can be generalized to other conditions should be explored in the future. Furthermore, there are still some issues in the current template set. Template CLS1 is applicable in many cases, and there is room for further subdivisions. After analyzing the disagreement cases, templates VAL1 and VAL4 are often confused, and they may need to be consolidated into one. Although some issues still need to be addressed, we believe they can be effectively resolved by repeating the error analysis.

Data Collection of Counterargument This study collects counterarguments intensively by focusing on a few topics. We think it would be difficult to adequately address this task with a broad and shallow approach to collecting counterarguments. To present appropriate diagnostic comments, it is necessary to analyze the argumentation structure of an original argument and counterargument accurately. Although pre-trained language models have improved the performance, it is still challenging to analyze argumentation structures with implicit relations (Atwell et al., 2021). We aim to alleviate the difficulty by covering the counterarguments that typically appear in a topic.

Of course, the cost of expanding to other topics will be higher than a broad and shallow approach to collecting argumentation. However, if we think about actual usage scenarios, there will be many situations where only a few topics will be immediately useful. For example, in a high school classroom, five topics may be sufficient. It is not necessary to have different topics for each class, nor is it necessary to change them yearly.

Data Collection of Diagnostic Comments To collect various diagnostic comments and analyze the task’s characteristics, we collected diagnostic comments based on the descriptive paradigm, which does not adjust for agreement among assessors. It is not suitable for training a model to make consistent predictions. Based on the results of this trial, we plan to refine the corpus to make it more suitable for model training and evaluation.

7. Conclusion

We proposed a template-based formulation to make the task of giving specific diagnostic comments on argumentation more tractable. As a first attempt, we verified the feasibility of creating a template set that satisfies three criteria of expressiveness, informativeness, and uniqueness. We showed it is feasible through an annotation study that converts diagnostic comments given in text into a template format. We publish the corpus used for the annotation study. The corpus consists of 1,000 counterarguments, 197 of which are annotated with 1,082 diagnostic comments in both natural language and template formats. Our future work is to refine the corpus to be more suitable for model training and evaluation based on this annotation study. For modeling, we plan to analyze what kind of information needs to be captured by the model to give a correct diagnostic comment.

| # Different diagnoses | Percentage |
|-----------------------|------------|
| 1                     | 71.1%      |
| 2                     | 18.9%      |
| 3                     | 6.8%       |
| 4                     | 2.5%       |
| 5                     | 0.7%       |

Table 7: Percentage of filler that can be extracted from original argument or counterargument.

| Percentage | Extractable 75.9% (126 / 166) |
|------------|--------------------------------|
|            | Extractable (some essential changes are required) 8.4% (14 / 166) |
|            | Not extractable 15.7% (26 / 166) |

Table 6: Percentage of different diagnoses given to target sentences.

target sentences have been given different templates. Table illustrates the percentage of different templates that were given to the target sentences. For 28.9% of the cases, two to five different diagnoses (templates) are given for the same target. The following is an example of such a case.

(6) Counterargument (excerpt):

Training through homework can cultivate students’ perspectives, values, and curiosities.

Diagnostic comment 1 (templated):

(CA1) It is unclear why homework causes a good result of cultivating students’ curiosity.

Diagnostic comment 2 (templated):

(CMP2) It lacks an explanation of why homework is a better method to realize cultivating perspectives, values, and curiosity instead of other methods.

Diagnostic comment 1 questions the belief that homework cultivates students’ curiosities. Alternatively, diagnostic comment 2 questions whether homework is more suitable than other methods to cultivate students’ curiosities. These diagnostic comments differ in how the relationship between homework and students’ curiosities is viewed, but both are reasonable. The results indicate that there are multiple reasonable diagnoses for the same target sentence. Therefore, we adopt multi-label classification as a task setting of template selection.
8. Acknowledgments

This work was partly supported by JSPS KAKENHI Grant Number 22H00524 and NEDO JP1234567. The authors would like to thank Paul Reisert, members of the Tohoku NLP Lab, and the anonymous reviewers for their helpful feedback. We also would like to thank the assessors and the annotators for their time and effort.

9. Bibliographical References

Afrin, T., Kashefi, O., Olshefski, C., Litman, D., Hwa, R., and Godley, A., (2021). Effective Interfaces for Student-Driven Revision Sessions for Argumentative Writing. Association for Computing Machinery, New York, NY, USA.

Atwell, K., Li, J. J., and Alikhani, M. (2021). Where are we in discourse relation recognition? In Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 314–325, Singapore and Online, July. Association for Computational Linguistics.

Burstein, J. and Chodorow, M. (1999). Automated essay scoring for nonnative English speakers. In Computer Mediated Language Assessment and Evaluation in Natural Language Processing.

Carlile, W., Gurrapadi, N., Ke, Z., and Ng, V. (2018). Give me more feedback: Annotating argument persuasiveness and related attributes in student essays. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 621–631, Melbourne, Australia, July. Association for Computational Linguistics.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1):37–46.

Dikli, S. (2006). An overview of automated scoring of essays. The Journal of Technology, Learning and Assessment, 5(1), Aug.

Dong, F. and Zhang, Y. (2016). Automatic features for essay scoring – an empirical study. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1072–1077, Austin, Texas, November. Association for Computational Linguistics.

Driver, R., Newton, P., and Osborne, J. (2000). Establishing the norms of scientific argumentation in classrooms. Science Education, 84:1–312, 05.

Durón, R. R., Limbach, B., and Waugh, W. S. (2006). Critical thinking framework for any discipline.

Egawa, R., Morio, G., and Fujita, K. (2020). Corpus for modeling user interactions in online persuasive discussions. In Proceedings of the 12th Language Resources and Evaluation Conference, pages 1135–1141, Marseille, France, May. European Language Resources Association.

Hovy, D., Berg-Kirkpatrick, T., Vaswani, A., and Hovy, E. (2013). Learning whom to trust with MACE. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1120–1130, Atlanta, Georgia, June. Association for Computational Linguistics.

Ida, M., Morio, G., Iwasa, K., Tatsumi, T., Yasui, T., and Fujita, K. (2019). Can you give me a reason?: Argument-inducing online forum by argument mining. In The World Wide Web Conference, WWW ’19, page 3545–3549, New York, NY, USA. Association for Computing Machinery.

Ke, Z., Inamdar, H., Lin, H., and Ng, V. (2019). Give me more feedback II: Annotating thesis strength and related attributes in student essays. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3994–4004, Florence, Italy, July. Association for Computational Linguistics.

Kocmi, T., Federmann, C., Grundkiewicz, R., Junczys-Dowmunt, M., Matsushita, H., and Menezes, A. (2021). To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. arXiv preprint arXiv:2107.10821.

Krippendorff, K. (1980). Content Analysis: An Introduction to Methodology. Sage Publications, Inc., Beverly Hills, CA.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1):159–174.

Liu, C.-W., Lowe, R., Serban, I., Noseworthy, M., Charlin, L., and Pineau, J. (2016). How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2122–2132, Austin, Texas, November. Association for Computational Linguistics.

Mathur, N., Baldwin, T., and Cohn, T. (2020). Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4984–4997, Online, July. Association for Computational Linguistics.

Morio, G. and Fujita, K. (2018). Annotating online civic discussion threads for argument mining. In 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI), pages 546–553.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Park, J. and Cardie, C. (2018). A corpus of eRule-making user comments for measuring evaulability of arguments. In Proceedings of the Eleventh Interna-
Zhang, F., Hashemi, H. B., Hwa, R., and Litman, D. (2017). A corpus of annotated revisions for studying argumentative writing. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1568–1578, Vancouver, Canada, July. Association for Computational Linguistics.

Appendix:

Table 8 shows the template set of the full version used in the annotation study. Table 9 shows the examples of annotation for template application.
| Quality Dimension | Category | Template |
|-------------------|----------|----------|
| Local Acceptability | CA1 | なぜ x によって y という良い結果が起こるのかが不明確  
It is unclear why x causes a good result of y |
|                   | CA2 | なぜ x によって y という悪い結果が起こるのかが不明確  
It is unclear why x causes a bad result of y |
|                   | CA3 | なぜ x によって y という良い結果が抑制されるのかが不明確  
It is unclear why x suppresses a good result of y |
|                   | CA4 | なぜ x によって y という悪い結果が抑制されるのかが不明確  
It is unclear why x suppresses a bad result of y |
|                   | VAL1 | なぜ y にとって x が良いことなのかが不明確  
It is unclear why x is good for y |
|                   | VAL2 | なぜ y にとって x が悪いことなのかが不明確  
It is unclear why x is bad for y |
|                   | VAL3 | なぜ x は y とすべきと考えているのかが不明確  
It is unclear why x should be y |
|                   | VAL4 | なぜ x は y とすべきでないと考えているのかが不明確  
It is unclear why x should not be y |
|                   | CLS1 | なぜ x は y という特性を持つと言えるのかが不明確  
It is unclear why x has the property of y |
|                   | CLS2 | なぜ z という点において x と y は同じ似ているのかが不明確  
It is unclear why x and y are same/similar in terms of z |
|                   | PR1 | なぜ x を実現可能なのかが不明確  
It is unclear why x can be feasible |
| Local Sufficiency | EX1 | x の例として具体的には何があるか  
It lacks the specificity of what exactly is an example of x |
|                   | EX2 | x はどの程度 y かの具体性に欠ける  
It lacks the specificity regarding the extent to which x y |
|                   | EX3 | x というのは限定的な状況である  
It is a limiting situation that x |
|                   | CMP1 | なぜ y よりも x を優先すべきかの説明が不足している  
It lacks an explanation of why x should be preferred over y |
|                   | CMP2 | y を実現するのに、なぜ z ではなく x という方法が良いのかの説明  
It lacks an explanation of why x is a better method to realize y instead of z |
| Local Relevance   | LR1 | なぜ x という理由が y という結論を支持するのかが不明確  
It is unclear why a premise x supports a claim y |
| Clarity           | CLR1 | x という表現が何を意味しているのか分からない  
It is hard to understand what the statement x is means |
|                   | CLR2 | x については具体例はあるが一般化した説明がない  
There is a specific example for x, but it lacks a generalized justification |
| Global Relevance  | GR1 | x という主張/理由が論題とどのように関係するのかが不明確  
It is unclear how the statement x relates to the topic |
|                   | GR2 | 肯定側の x という主張に y というのは直接的な反論になっていない  
It is not a direct objection to the government’s claim x to say y |
|                   | GR3 | x というのは肯定側が定義している y を考慮できていない  
The statement x fails to consider y, which is the definition of the government side |
| Global Sufficiency| GS1 | なぜ肯定側の y という主張よりも否定側の x という主張が優位だと言えるのかが不明確  
It is unclear why the opposition’s claim x is superior to the government’s claim of y |
|                   | GS2 | 肯定側からの x という反論が想定される  
It is expected that the government will object that x |

Table 8: The template set used in the annotation study (Full version).
| Category | Diagnostic comment in natural language text | Diagnostic comment after applying template |
|----------|---------------------------------------------|-------------------------------------------|
| CA4      | 死刑が十分な抑止力になるという根拠が示されていない。No evidence is shown that the death penalty is a sufficient deterrent. | なぜ死刑によって犯罪という悪い結果が抑制されるのかが不明確。It is unclear why the death penalty suppresses a bad result of crime. |
| VAL3     | なぜ、学校でたくさんのスキルを学べるよう担保するべきなのかが述べられていない。No discussion of why it should be guaranteed that children can learn many kinds of skills at school. | なぜ学校は生徒がたくさんのスキルを学べるよう担保すべきと考えているのかが不明確。It is unclear why schools should be responsible for guaranteeing that students learn many skills. |
| PR1      | 死刑執行ボランティアは本当に自由意志で募ることができるのかという点が疑問として残る。Doubts remain over whether volunteer executioners would really apply out of their own free will. | なぜ死刑執行のボランティアを本当に自由意志で募ることを実現可能なのかが不明確。It is unclear why recruiting volunteer executioners who actually do it out of their own free will can be feasible. |
| EX1      | 「生徒のスキル」や「生徒の習得状況に合わせた授業の内容」という部分が抽象的。Abstract use of phrases like “students’ skills” and “class content suited to students’ acquisition of learning.” | 「生徒のスキル」や「生徒の習得状況に合わせた授業」の具体例には何があるか。It lacks the specificity of what exactly is an example of “students’ skills” and “class content suited to students’ acquisition”. |
| EX2      | 犯罪者が刑務所に入れておくにはどれくらいコストがかかるのかという説明がない。No explanation exists of the costs required to keep criminals in prison. | 犯罪者が刑務所に入れておくに は ど の 程度 セットがかかるかの具体性に欠ける。It lacks the specificity regarding the extent to which keeping criminals in jail costs money. |
| EX3      | 宿題と部活動の両立ができたという具体例は、体験談にとどまっており、一般化されていない。The specific example of being able to balance homework and club activities is no more than a personal story and is not generalized. | 宿題と部活動を両立できるというのは限定的な状況である。It is a limiting situation that being able to balance homework and club practice. |
| CMP1     | 勉強が必要な生徒のために勉強が嫌いな生徒のことを無視しても良いのかという理由づけがない。No reasoning for why students who do not like studying can be ignored for the sake of those who do. | なぜ 勉強が嫌いな生徒も 勉強が好きな 生徒を優先すべきかの説明が不足している。It lacks an explanation of why students who like studying should be preferred over students who dislike studying. |
| GR1      | 犯罪を犯した者が無効化する方法としても刑務所は有効であるという主張は、論題に関係ない説明に見える。The argument that incarceration is effective to incapacitate those who commit crimes does not seem relevant to the topic. | 犯罪を犯した者が無効化する方法としても刑務所は有効であるという主張が論題とどのように関係するのかが不明确。It is unclear how the statement “incarceration is a way to incapacitate those who commit crimes” relates to the topic. |
| GS1      | リスクがゼロでないことを問題として挙げているのに対し、なぜ最小限にするので十分なのかの説明が不足している。Insufficient explanation exists regarding why minimizing risk is sufficient in response to the point that the risk exists. | なぜ 勘的側の リスクがゼロでないことが問題という主張よりも勘的側の リスクを最小限にするので十分という主張が優位だと言えるのかが不明确。It is unclear why the opposition’s claim “minimizing the risk is enough” is superior to the government’s claim of “the problem raised is that the risk exists”. |
| GS2      | 授業中に宿題をするということは人によっては悪いことだと思ってしまう可能性がある。Some people might consider it a bad thing to do homework during class. | 認定側から 授業中に宿題をするのは悪いことだという反論が想定される。It is expected that the government will object that doing homework during class is bad. |

Table 9: The examples of annotation for template application.