An Argument Annotation Scheme for the Repair of Student Essays by Sentence Reordering

Jan Wira Gotama Putra[†] Simone Teufel[‡†] Takenobu Tokunaga[†]
[†]Tokyo Institute of Technology [‡]University of Cambridge
gotama.w.aa@m.titech.ac.jp simone.teufel@cl.cam.ac.uk take@c.titech.ac.jp

1 Introduction

Writing good texts is a skill that needs to be learned like many other skills. Our long-term goal is to automatically improve human texts of insufficient quality and to provide an explanation for the changes made. Such a tool would be invaluable for the teaching of writing skills and for language education.

It is well known that the highest quality texts are often the result of many revision cycles, starting from an initial draft and moving towards a highly polished result. To achieve this goal, a writer has to consider grammar, writing style, argumentation and coherence during the revision process [6, 11].

Out of the many ways how texts can be improved, we focus on the task of reordering sentences according to the role they play in the argumentation. Our working hypothesis is that it is foremost the argument structure that should be instrumental for the reordering. For example, should we find a *major claim* (*main stance*) in the middle of an essay, it is to be moved to the beginning or the end, while respecting all the other aspects of how it is connected to the argument. The argument structure provides not only a means to reorder sentences automatically but also explains the ways which the improved texts are better than the original texts. Therefore, in our approach, the annotation of argument structure is the first step towards automatic reordering.

In this study, we analyse student essays written in various Asian countries, which are expert-scored for writing quality [4]. We concentrate particularly on *medium*-quality essays. They are an ideal target for our study: if they were any better, they would not need improvement; if they were any worse, the intention of the writer with respect to the argumentation would no longer be clear. In fact, we can think of them as being like a first draft that requires improvement.

This paper discusses an annotation scheme that is suitable for our task. We performed three pilot studies with different definitions of argument structure, incrementally improving our annotation guideline. This process leads to a reasonable inter-annotator agreement among small samples in the final scheme. This paper discusses the stepwise improvements, since they may be useful to others during the creation of similar schemes.

2 Text Collection

In this study, we target student essays from ICNALE[1] [4], a corpus of 5,600 persuasive essays. ICNALE essays contain 200–300 words and are written in response to a prompt. All essays are scored in five aspects: content, organisation, vocabulary, language use and mechanics (capitalisation, punctuation and spelling), which are combined into a total score in the range of $[0, 100]$. The ICNALE dataset, in this respect, is highly controlled, and enables us to study the different aspects of text improvement (e.g., sentence order) separately. 640 of the essays are corrected concerning grammar and mechanics, and we take these as our starting point.

We divided the revised essays into 10 percentile groups by total score, and sampled one from each group randomly to understand the characteristics of the essays. Our impression is that the essays in the 0 to 40 percentile are poor in quality and need rewriting; it is typically hard to understand what the authors want to convey in these essays. These are 4.1% of all essays and we refer to this group as *low*. The essays scored in the range of 40 to 80 (80.8%; *medium*) are understandable and fairly good, and we believe it is possible to improve them by sentence reordering. The essays with scores of 80 or more (15.2%; *high*) are well-written. In the context of this paper, we assume that they don't require any improvement (while it might of course be possible to improve their persuasiveness even further). While the *medium*-quality are the main focus of our study, we also experimented with the *low* and *high* quality essays.

3 Annotation Design

3.1 Annotation Scheme

Since the essays we aim to annotate are persuasive, we can use the existing techniques in Argument Mining (AM) to analyse their argument structures. AM aims to produce a structured output (tree or graph) for a given text [7]. The structured output explains

[1] http://language.sakura.ne.jp/icnale/

the role that units (sentences, clauses or clause-like segments) play in the discourse, and/or how they relate to each other with respect to their argumentative role. AM involves three main tasks: (a) argument component identification, (b) argument component classification and (c) argument structure prediction [7, 11].

Argument component identification determines the exact boundaries of units and differentiates them as either argumentative and non-argumentative [7, 11]. Furthermore, argumentative components (ACs) can be classified according to their rhetorical function in the discourse, such as *claim* or *premise* [9]. The ACs are then connected to each other in the argument structure prediction task, commonly forming a tree-like structure [10, 11]. Existing studies have proposed several labels to describe the relations between ACs, e.g., *support*, *attack*, *detail* and *sequence* [5, 11].

We aim to use as few labels and perform as few annotation tasks as possible while keeping the scheme expressive enough for our goal. We decided to annotate at the sentence-granularity level. Our annotation differentiates argumentative and non-argumentative components, in a sense, because we introduce the notion of an *omittable* sentence – one that is hardly connected to the argument and could be ignored without affecting the argument structure. Examples include meta-information, redundant material (repeated facts) and disconnected sentences with no clear connection to the argument.

Rather than annotating the status of the argumentative components themselves, we label relations between them with the following labels: *support* (`sup`), *attack* (`att`), *detail* (`det`) and *restatement* ("`=`"). `Sup`, `att` and `det` are directed, whereas "`=`" is an undirected relation. In a `sup` relation, the source AC asserts or justifies reasons and ideas for supporting the target AC. When the source AC considers counter-arguments that argue for the opposite opinion, they are in an `att` relation. Our `det` relation is used when the source AC further explains, describes, elaborates or provides background for the concept(s) mentioned in the target AC. It roughly corresponds to a combination of *elaboration* and *background* in Rhetorical Structure Theory (RST) [5,8]. One of the pilot studies reported in this paper will try to establish whether we can replace `det` with a simpler subset (Section 4). Inspired by the study of Skeppstedt et al. [10], we also use a "`=`" (*restatement*) relation to express a situation where important parts of an argument are summarised for the second time. Unlike mere repetitions of facts (that can easily be omitted), restatements happen at a higher level of argumentation (*claims*) and often have a perfectly good rhetorical function, so we decided to mark them separately. With respect to their connection to the overall argument, we treat the two argumentative components connected by a restatement relation as equivalent.

3.2 Annotation Procedure

This paper focuses on the annotation of argument structure, since it is a prerequisite for the subsequent analyses. On the one hand, the analysis of sentence reordering can only be done if the argument structure has been annotated reliably; on the other hand, given a good argument structure, the best reordering often follows automatically. Our annotation scheme consists of only two tasks: (1) argument component identification and (2) argument structure prediction.

4 Pilot Annotation

We performed three pilot annotations in an attempt to find the best annotation setting for our task, explore pitfalls in the annotation and improve the annotation guideline. We deliberately used different settings in the pilot studies to detect problems that might not be apparent if that setting remained fixed. This is because we incrementally improve our annotation guideline and setting over time, which finally resulted in the final annotation scheme (Section 3.1). The pilot studies are thus not directly comparable, and we rely on qualitative analysis.

We measure the inter-annotator agreement (IAA) of argument component identification and argument structure prediction task. For the first task, we measure to what extent all annotators agree on the binary classification of each sentence as argumentative/non-argumentative component, calculated using Fleiss' Kappa [3]. We split the IAA of the second task into linking and labelling scores. As linking agreement, we measure the extend to which the annotators consider each possible pair of ACs as being connected or not (binary), calculated using Fleiss Kappa [3]. For all AC pairs which have been confirmed as being connected, we measure in a second step whether the annotators agree on the relation label that connects them (calculated using Cohen's Kappa [1]). We believe the separation of the linking and labelling agreement gives better insight into the quality of our guideline, instead of collapsing the two metrics.

In reality, there might be multiple acceptable structures for the same essay. This is an inherent problem in high-level interpretative tasks such as discourse annotation [2]. Nevertheless, it is desirable to reach the highest possible IAA, especially when the resulting dataset is to be used for machine learning purposes. Hence, we use the IAA scores as an indicator of improvement in our annotation.

4.1 Pilot 1: Simplification of label "detail"

At first, we wanted to confirm whether the *elaboration* relation on its own is sufficient to explain

	A	B	C
WG	0.45	0.37	0.49
A	-	0.36	0.13
B	-	-	0.57

Table 1. Pairwise labelling agreement (computer scientist) of scheme with *elaboration* (Pilot 1)

the existing relations in essays, or whether we need a more general category *detail*. In this pilot study, we employed four labels: *support*, *elaboration*, *attack* and *restatement*. Annotators build a tree structure of each essay in which the prompt acted as the *root* and the essay's stance towards the prompt was explicitly annotated.

Four in-house annotators annotated three essays of *medium*-quality using Microsoft Excel. The set of annotators is composed of the first author (WG) and three laboratory members (A, B, C), all having a computer science background. An annotation guideline of six pages was used.

The argument component identification and linking agreement scores are 0.33 and 0.61, respectively. Table 1 shows the pairwise labelling agreement between annotators. We also asked the annotators about their experience with the task. They mentioned that they could often identify clusters of sentences concerning the same sub-argument, but with our initial guideline, this information could not be expressed. We decided that a more structured annotation process is needed in reaction to this observation. Furthermore, they also commented that the set of available relation labels felt insufficient to them in order to express the relations that were there according to their intuition; in particular, they requested an additional *background* relation. This lead to our definition of *detail* as the union of *background* and *elaboration* in the subsequent pilot studies.

4.2 Pilot 2: Linguistic experts

We suspected that the difference in argument structure and labelling among annotators in the previous pilot study was due to a difference in linguistic knowledge. In this pilot study, we used three annotators, the first author (WG) and two paid annotators (D, E). The new annotators graduated from the department of linguistics, are fluent (non-native) English speakers and have experience in teaching English. Another change in this pilot study is the use of the newly introduced *detail* relation, leading to the final relation set as described in Section 3.1. The new annotation guideline remained six pages in length. Also, as a quick confirmation whether the annotation quality is affected by essays' quality, we analysed the correlation between essay's score and linking agreement (per essay). If it is true that *high*-quality essays are easier to interpret (therefore, annotate) and *low*-quality essays are harder, this should be visible in the correlation between the essay scores and linking.

This time, eight essays were annotated: one *low*, five *medium* and one *high*-quality. The annotation tool was Microsoft Excel as before.

	D	E
WG	0.47	0.54
D	-	0.63

Table 2. Pairwise labelling agreement (Linguistic experts); final annotation scheme (Pilot 2)

The argument component identification and linking agreement scores are 0.47 and 0.48, respectively. Table 2 shows the pairwise labelling agreement between annotators. The argument component identification and labelling agreement scores were improved, which we attributed to the background of the new annotators in linguistics. However, they still produced different structures and this signals that the problem of distinctive structures cannot be fully eliminated by simply selecting annotators with more expertise. To address this problem, we introduced a sub-argument annotation procedure in the next pilot study. Also, to our surprise, there was no correlation between the score and the linking agreement per essay (Pearson coefficient= 0.09). The low correlation score is possibly attributed to the small sample size. In another viewpoint, this might mean that argumentation skills are developed separately from general L2 skills such as vocabulary and language use. Since, at this point, improvement of the annotation guideline and setting is more important for us, we leave the investigation of this issue with more data in the future.

Furthermore, some errors were introduced by the fact that Microsoft Excel is not a specialised annotation tool. For this reason, we decided to build our own annotation tool which was to implement the minimum necessary features to perform the tasks in our annotation. This included rejecting illogical annotation (e.g., connecting a sentence to itself) during the annotation process. The tool is called TIARA. It also provides a visualisation function, which we hope is helpful during annotation.

4.3 Pilot 3: Sub-argument Instructions

In this pilot study, where we used TIARA for the first time, we also introduced several other changes to resolve the problems in previous pilot studies. Firstly, we introduced a sub-argument annotation procedure to guide annotators to produce similar structures. Text is broken recursively into sub-arguments ("clusters"). Annotators then annotate intra-cluster relations. For each cluster, they se-

	B	F
WG	0.65	0.79
F	-	0.62

Table 3. Pairwise labelling agreement (computer scientists); Sub-argument instructions (Pilot 3)

lect one sentence as the cluster's representative and connect it to another cluster's representative or the *major claim*, forming a global structure. This results in a tree covering the entirety of the text. The guideline grew to 10 pages. The annotation guideline also includes a whole-text annotation example (this was a recommendation from pilot study 1), in addition to the sentence-pair examples we used up to this point. These were presented without context, and while being sufficient to illustrate each relation in isolation, they could not explain what the most-connected sentence-pairs and their connection labels in a given context were supposed to be, when there is competition between sentences.

Secondly, we asked the annotators to use the essay's *major claim* as the *root* of the tree (previously, the *root* was the prompt, and the essay's stance toward the prompt was explicitly annotated). Since the past pilot studies allowed ACs to be connected to the prompt, it might wrongly make annotators think that the prompt is to be treated as a part of the essay and thus the argument structure, which it is not. Additionally, the essay's stance towards the prompt is not of any interest for our task.

Due to limited resources, we used in-house annotators and fewer essay samples than in the previous pilot study. Three annotators participated in this pilot study: the first author (WG) and two laboratory members (B who participated in pilot 1 and F). They annotated one *low*, one *medium* and one *high*-quality essay (three in total). The essays were mixed in quality for a comparable setting to the previous pilot study. The same four relation labels as in the pilot study 2 were used.

The argument component identification and linking agreement scores are 0.44 and 0.55, respectively. Table 3 shows the pairwise labelling agreement between annotators. The linking agreement is higher than in the previous pilot study. According to the annotators, the sub-argument annotation procedure matched their intuitions about the structure well. This new rule, possibly, allowed annotators to build more similar structures since it provides a more detailed description of how to handle inter- and intra-cluster relations, rather than forcing the annotator to connect argument parts arbitrarily. The labelling agreement scores were also reasonably improved compared to the previous pilot study. We also believe introducing a whole-text example has made it easier for annotators to distinguish the labels.

The use of the TIARA, compared to using Microsoft Excel, decreased the average annotation time of annotator WG and B from around 40 minutes to 25 minutes. However, the argument component identification agreement has not improved. This may be due to the expertise level of the annotators. In the future, we plan to administer the latest annotation guideline to annotators with a background in linguistics.

5 Conclusion

Designing an annotation scheme is always an iterative process since it is impossible to recognise all parameters and problems in the beginning. But even in the light of these difficulties, our pilot annotation provides evidence that even imperfect texts can be annotated with a reasonable inter-annotator agreement. The current paper details several aspects of the annotation procedure and the guideline which helped us in reaching this goal. Employing a more intuitive label, annotators with more expertise and a more detailed sub-argument annotation procedure have all contributed to improving the inter-annotator agreement. We also found that developing our own annotation tool is worth the investment, considering it decreased annotation time and can prevent careless mistakes.

Acknowledgement

This work was partly supported by Tokyo Tech World Research Hub Initiative (WRHI) Program of Institute of Innovative Research, Tokyo Institute of Technology. We also thank the 7 annotators for their work.

References

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

[2] Debopam Das and Maite Taboada. Signalling of coherence relations in discourse, beyond discourse markers. *Discourse Processes*, 55(8):743–770, 2018.

[3] Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971.

[4] Shinichiro Ishikawa. The icnale edited essays: A dataset for analysis of l2 english learner essays based on a new integrative viewpoint. *English Corpus Linguistics*, 25:1–14, 2018.

[5] Christian Kirschner, Judith Eckle-Kohler, and Iryna Gurevych. Linking the thoughts: Analysis of argumentation structures in scientific publications. In *Proceedings of the ARGMINING workshop*, pages 1–11, 2015.

[6] John Lee and Jonathan Webster. A corpus of textual revisions in second language writing. In *Proceedings of ACL*, pages 248–252, 2012.

[7] Marco Lippi and Paolo Torroni. Argumentation mining: State of the art and emerging trends. *ACM Trans. Internet Technol.*, 16(2):10:1–10:25, 2016.

[8] William C. Mann and Sandra A. Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281, 1988.

[9] Andreas Peldszus and Manfred Stede. From argument diagrams to argumentation mining in texts: A survey. *Int. J. Cogn. Inform. Nat. Intell.*, 7(1):1–31, 2013.

[10] Maria Skeppstedt, Andreas Peldszus, and Manfred Stede. More or less controlled elicitation of argumentative text: Enlarging a microtext corpus via crowdsourcing. In *Proceedings of ARGMINING workshop*, pages 155–163, 2018.

[11] Christian Stab and Iryna Gurevych. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING*, pages 1501–1510, 2014.