Paibi Student Essay Dataset - Abbreviated Annotation Guide

Stephen Bothwell

July 27th, 2023*

1 Summary

The document below describes our annotation task. It provides a general description of our research in Section 1, a set of annotation guidelines in Section 3, and a guide to using the Brat Rapid Annotation Tool in Section Section 4 (abbreviated here).

2 Premise

In our current research, we are studying ***rhetorical parallelism***—the intentional use of many similar linguistic features in close proximity to achieve an effect. Frequently, we use rhetorical parallelism to strengthen our arguments, producing a stronger relationship between distinct chunks of those arguments. For instance, the argumentative phrases "on the one hand" and "on the other hand" connect themselves and the text following them through their lexical similarity (*i.e.*, by using similar words). Other instances rely on a variety of linguistic relationships in terms of sound, grammar, and meaning. For instance, English uses the following as a proverb:

> Give a man a fish and you feed him for a day;
> teach a man to fish and you feed him for a lifetime.

As these examples imply, parallelisms can take various lengths; they can be a few words, or they can be their own sentences in and of themselves.[1]

Our work on rhetorical parallelism attempts to use computational methods to detect and delimit rhetorical parallelisms automatically. Because the study of rhetorical parallelism via computational methods is relatively novel, we need data to train and test our systems. We have already collected one such dataset in Latin. However, to show the ability of our systems to generalize, we also wanted to perform our task in another language. That is where the Paibi Student Essay (PSE) dataset comes in.

The PSE dataset is a collection of mock exam essays written by senior high school students in China. This data was gathered by Song *et al.* in their attempts to study the automatic evaluation of essays [1]. They had annotators examine these essays and mark up sentences as to whether they were parallel. Annotations were done at the granularity of a sentence; thus, the dataset was oriented toward linking similar sentences together. However, Song *et al.* recognized that some sentences may contain a parallelism on their own. So, they allowed annotators to tag these sentences as such.

The problem with those annotations is that they only indicate that a parallelism exists; they do not indicate where the actual parallelism is. Moreover, these in-sentence parallelism annotations cover roughly a fourth (247 out of 907) of the annotated data. Because we define a parallelism as an innately

*The date presented here is the "last updated" date.

[1]For more examples, the Wikipedia page on this topic provides quite a few solid ones: `https://en.wikipedia.org/wiki/Parallelism_(rhetoric)`.

interconnected object—as a connection between two or more spans of text which work together to create an effect—these annotations are insufficient. As such, we are asking you to add this specificity: to examine the annotated sentences provided and to determine what spans are parallel with one another.

3 Annotation Guidelines

In general, the detection of a span of parallel text, which we call a *branch*, should be on the basis of exhibiting at least *two* of these criteria:

- **Length**: They contain a nearly identical (*e.g.*, ±2) number of words.
- **Distance**: They are a short distance from one another (*i.e.*, they had few [roughly ±2] intervening words).
- **Phonology**: They have at least two words that sound similar in some manner (*e.g.*, that rhyme).
- **Morphology**: They have two or more pairs of words that are the same or are derived from the same base form (*e.g.*, as "sink", "sank", and "sunk" are all derived from the same verb "sink").
- **Syntax**: They have identical syntactic structure, or they consist of two or more pairs of words of an identical grammatical form in identical order.
- **Semantics**: They have two or more pairs of words that are semantically related (*e.g.*, are synonyms, cognates, or antonyms).

By definition, parallelisms come together on the basis of multiple branches. Because the smallest unit of any text span is a single word, there are many ways in which one could divide up and connect spans into branches of a parallelism. Because of this, we provide further principles on how to select branches from multiple possibilities.

- **Maximality**: In general, the *largest possible branches* should be selected for a parallelism. Parallelisms should only be kept separate if the pertinent spans concern distinct linguistic relationships.
  - *Interlocking*: Branches should **not** interlock—that is, maintain an A-B-A-B structure on their own if each "A" and "B" pair can be merged to form a larger "C" branch containing each of the smaller branches and pairing those branches in effectively the same manner.
  - *Nesting*: Branches are permitted to *nest* in one another; that is, a span can be labeled more than once so long as one of the branches entirely contains the other. Each branch should also be a part of a *distinct* parallelism, meaning that the linguistic features emphasized by the parallelism should be different.
- **Intent**: One critical fixture of *rhetorical* as opposed to merely *syntactic* parallelism is its intentionality. Syntactic parallelism can be required by the grammatical constructions of a given language; meanwhile, rhetorical parallelism is an effort by a speaker or writer to create an effect. We elaborate upon a few items to aid in determining what is intentional.
  - *Conjunctions*: In many cases, it is unclear whether conjunctions are simply fulfilling a syntactic role or are involved in a wider juxtaposition. We rule that conjunctions should be included in parallelisms for a series of branches only if *every branch* contains a conjunction. Otherwise, conjunctions should be left out.
  - *Laundry Lists*: As is implied by the description above, syntactic parallelism is required but is not necessarily intentional, whereas rhetorical parallelism is not necessarily required but is intended. By this logic, a series of items piled up in a "laundry list" is not innately a rhetorical parallelism unless some other feature-based connection ties the items together.

4 Guide to the Brat Rapid Annotation Tool

4.1 Omitted Content

In this section, we gave a visual guide to using the BRAT annotation tool. To focus on the exact procedure of collecting annotations (regardless of the tool used to do so), we have omitted that information here.

4.2 Brat Annotation Procedure

To annotate the relevant data for parallelism, please adopt the following work process (repeating these steps as necessary):

1. Using the `log.txt` file provided to you, locate a sentence for annotation. The file, paragraph number, and sentence number should allow the line to be located exactly. Moreover, the line will be prefixed with three "$>$" symbols (as in "$>>>$").

2. Having logged in, use the `Branch` tag (alongside the annotation guidelines given in the previous section) to annotate the sentence for its parallel structure. Link related branches in the order that they appear and from left to right. Moreover, only link adjacent branches. Finally, while having annotations nest within one another is permitted, do not allow text span annotations to overlap (*i.e.*, have both content shared between them and content unique to each). BRAT should highlight the annotations in red if this occurs.

    - Note that, although only the lines designated with "$>>>$" are meant to be annotated, you are allowed to add annotations to other branches as you see fit. For example, if a nearby sentence or phrase is also parallel to the designated sentence, you may tag it.
    - If the sentence seems to be falsely annotated for rhetorical parallelism, annotate it with the "Finished" tag, and use the "Notes" of the annotation to indicate that it has no parallel structure.

3. Once the designated lines in the `log.txt` file exhausted for a given essay, scroll to the top of the essay and annotate the first line with the `Finished` tag to let others know that it has been completed. Also, in the "Notes" section of that annotation, please write your name so that we know who annotated the file.

References

[1] Wei Song et al. “Learning to Identify Sentence Parallelism in Student Essays”. In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. Osaka, Japan: The COLING 2016 Organizing Committee, Dec. 2016, pp. 794–803.