Annotation Guidelines for Research Workflow

This document outlines the annotation guidelines employed to identify research workflow phrases and paragraphs within the analyzed texts.

**(1) Annotation Principles**

In this study, the annotation of research workflow was simplified into a two-step, binary classification procedure. The first step involves determining whether a paragraph heading can potentially function as a “research workflow phrase”. If so, the second step confirms whether the corresponding paragraph content actually describes a research workflow. Pre-annotation revealed that the vast majority of paragraph headings identified as potential research workflow phrases did indeed correspond to content describing research workflows. Consequently, the majority of the annotation effort was focused on the initial step of identifying potential research workflow phrases within paragraph headings. To this end, two primary principles were established for the first step:

*Principle 1: Phrasal Constraint*

This study introduces the concept of “research workflow phrase” to delineate specific and granular research steps. A research workflow phrase should be concise, clear, and, crucially, must constitute a phrase. During the initial data preparation stage, bold text from the beginning of each paragraph was extracted to serve as paragraph headings. However, bold text at the beginning of a paragraph may represent phrases, sentences, or other textual forms. Furthermore, rule-based extraction may occasionally yield incomplete textual units. Therefore, during annotation, non-phrasal paragraph headings were initially filtered out before assessing their suitability as research workflow phrases.

*Principle 2: Specificity Requirement*

This study employs research workflow phrases to denote specific and detailed research steps. These phrases should not encompass overly broad meanings; for example, “Results and Analysis” would not be considered a research workflow phrase. Rather, research workflow phrases should be focused and targeted, highlighting specific elements of the research methodology.

**(2) Action-Oriented Research Workflow Phrases**

In academic papers, descriptions of research steps often involve specific operations or manipulations of research elements, such as data. These operations are inherently action-oriented. Consequently, we consider “actionality” as a crucial criterion for identifying research workflow phrases, meaning that a phrase should contain or imply action-oriented words. During the pre-annotation process, we identified a range of action-oriented research workflow phrases, such as “data cleaning”, “variable selection” and “model training”. The following summarizes common types of action-oriented research workflow phrases:

* Data Collection: This category encompasses phrases like “Data collection”, “Dataset Statistics”, and “Constructing the STC-Auto-SeFun dataset”.
* Data Preprocessing: This includes phrases such as “Sentence Splitting”, “Data preprocessing”, and “Data Filtering”.
* Data Operation: This category involves phrases like “Template extraction”, “Passage generation”, and “Computing XX”.
* Model Evaluation: This comprises phrases such as “Measuring XX” and “XX evaluation”.
* Results Analysis: This includes phrases like “XX analysis”, “analysis of/on XX”, and “Analyzing XX”.

**(3) Nominal Research Workflow Phrases**

During pre-annotation, we observed that relying solely on the criterion of actionality predominantly yielded research workflow phrases related to data preparation or data processing, with fewer phrases pertaining to data analysis. This observation is also evident in the previously summarized action-oriented research workflow phrases. This situation hinders the extraction of complete research workflows from academic papers. Therefore, it is necessary to expand the scope of research workflow phrases from action-oriented to nominal phrases.

Nominal research workflow phrases frequently appear in the “Results and Analysis” sections, implicitly representing research steps related to data analysis. For example, “Effect of Pointer Generation” is a nominal phrase that implicitly signifies an analysis of the effect of “Pointer Generation”. Adding the word “Analyzing” transforms it into the action-oriented phrase “Analyzing the Effect of Pointer Generation”, which explicitly expresses a data analysis process. During the annotation of nominal research workflow phrases, the following rule can be applied: If a paragraph heading contains a preposition such as “of”, “on”, “for”, “by” or “over” and expresses an analytical perspective (e.g., influence, effect, performance, limitations), it can be annotated as a research workflow phrase. The following summarizes common types of nominal research workflow phrases:

* Indicating analysis of the influence of XX: Examples include “Influence of XX” and “Impact of XX”.
* Indicating analysis of the effect of XX: An example is “Effect of XX”.
* Indicating evaluation of the performance on XX: An example is “Performance on XX”.
* Indicating assessment of the quality of XX: An example is “Quality of XX”.
* Indicating analysis of the limitations of XX: An example is “Limitations of XX”.

**(4) Paragraph Headings that Should Not be Annotated as Research Workflow Phrases**

It is crucial to differentiate descriptions and introductions of concepts, models, etc., which should not be considered research workflow phrases. For instance, in the examples below, “String Similarity Features” and “Keyword Intensity” are not process phrases. The former’s paragraph content introduces string similarity features, while the latter’s introduces the concept of keyword intensity.

*Example 1*

**Paragraph Heading:** String Similarity Features

**Paragraph Content:** String Similarity Features include the following similarity measures: Jaro Winkler, Dice Sorensen, Hamming, Jaccard, Levenshtein, NGram overlapping and common prefix matching. Also, two binary features are added for identical match and identical match ignoring case.

*Example 2*

**Paragraph Heading:** Keyword Intensity

**Paragraph Content:** To compute n, we introduce another parameter r that controls the ratio of keyword tokens to original reference summary tokens. Higher values of r lead to extracting sentences in a manner more closely approximating keyword matching, but yielding poor standalone summaries. On the other hand, lower values of r may lead to generic summaries insensitive to the keywords. In practice, the number of times a keyword w is concatenated to the original summary S is defined as n = r × len (S) #(keywords) where len (S) is the number of tokens in the original summaries and #(keywords) is the total number of keywords available. When r = 1, the concatenated keywords have the same length of the original summary.

**(5) Further Confirmation of Paragraph Content’s Relevance to the Research Workflow**

To alleviate the annotation burden and enhance efficiency, we optimized the data annotation process. The initial step involves determining whether a paragraph heading can be classified as a research workflow phrase. Since paragraph headings are significantly shorter than paragraph content, annotators can quickly read them and make a judgment. For headings identified as potential research workflow phrases, the corresponding paragraph content is then reviewed to confirm whether it genuinely describes a research workflow. The criteria for this confirmation are: Does the paragraph content relate to data preparation (e.g., describing data collection or dataset construction), data processing (e.g., describing data transformation, model building, or model training), or data analysis (e.g., describing results evaluation, error analysis, or ablation studies)?