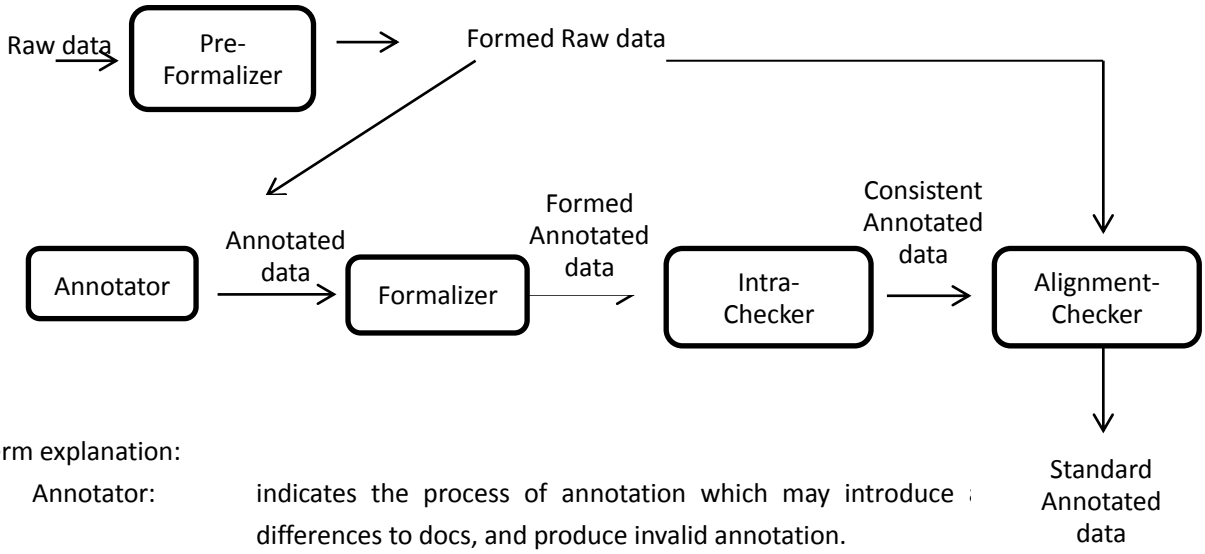
Schema for annotated dataset management

1 Structure

2 Usage

1 Structure



Term explanation:

Annotator: indicates the process of annotation which may introduce differences to docs, and produce invalid annotation.

Formalizer: Uniform EOL from different system, delete unnecessary blank

Pre-formalizer: Conduct the work of Formalizer and delete '<' and '>' other than '<doc>'.

Intra-Checker: Check balance of '<','>' and annotation, also ensure that all annotation content inside '<' and '>' follow the guideline of annotation.

Alignment-Checker: Ensure the annotation doesn't introduce accidental differences to docs.

2 Usage for annotator

Formalizer and Pre-formalizer doesn't need annotators to conduct further modification to the docs.

Intra-Checker and Alignment-Checker need annotators to correct detected errors or inconsistency in the docs.

3 Notes:

3.1 Currently, I haven't make an easy used formalizer, so basically we will omit the step of pre-formalizer and formalizer except that we should definitely delete '<' and '>' in raw data.

3.2 Intra-Checker is the "checker.py" I sent through email. Once you guys decide the final annotation rules, I can improve "checker.py" to be able to check annotation content.

3.3 Alignment-Checker is the "alignmenter" attached in this email.

All these tools might not be so matured, so if you encounter anything you think might be the problem of the tools , don't hesitate to tell me :)