HerBERT - Reference Graph Builder

Group 11

November 2025

Members

- Dominik Haring — Dataset / Reference Extractor
- Julian J. Kravanja — Dataset / Usage Validator
- Monika Windhager — Reference Extractor / Usage Validator
- Konrad Zalar — Reference Tree Builder / Usage Validator

Abstract

In this project we work on HerBERT - a system for automatically detecting weak and / or circular citation practices in scientific writing by constructing a weighted reference graph from research papers. We plan to combine automated citation extraction with context-aware assessment of how each referenced source is used within the citing document. With using a pretrained distilBERT model, we extract citation relations from full text to reconstruct reference edges, while a fitted BART model evaluates citation context to assign weights that reflect argumentative relevance and evidential strength. The then resulting graph allows us to identify claims or topics that rely on poorly supported references or citation chains that are prone to amplification effects.

With our project we therefore want to offer a proof-of-concept approach for uncovering "bad practice" referencing patterns and enhancing the transparency in knowledge distribution and generation in scientific papers.

Idea (Goal / Research Question)

The goal of this project is to develop a method for extracting citation dependencies from the full text of research papers and constructing a weighted reference graph in which edge weights represent how critically a cited source contributes to the citing document's arguments, methodology, or findings. Using this we want to detect bad citations and non-proofed claims which have been build on. Which brings us to our research question:

”Can an automatically constructed, weighted citation graph reveal how well scientific claims are supported and where citation weaknesses and / or circular referencing occurs?”

Visual Depiction

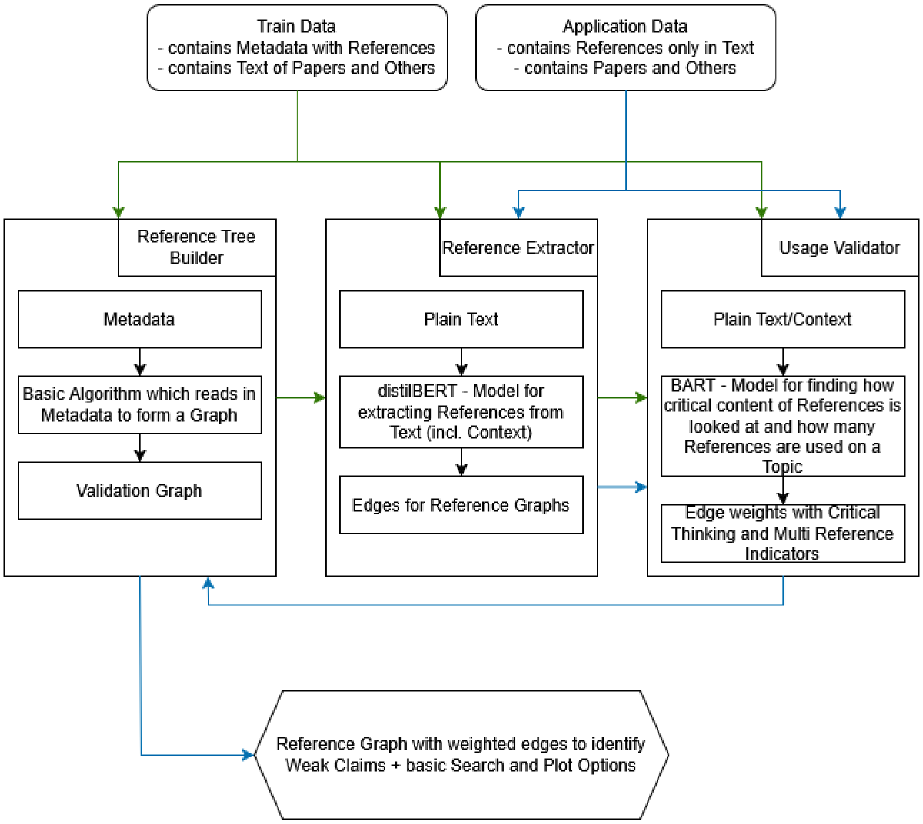
Train Data
- contains Metadata with References
- contains Text of Papers and Others
Application Data
- contains References only in Text
- contains Papers and Others
Reference Tree Builder
Metadata
Basic Algorithm which reads in Metadata to form a Graph
Validation Graph
Reference Extractor
Plain Text
distilBERT - Model for extracting References from Text (incl. Context)
Edges for Reference Graphs
Usage Validator
Plain Text/Context
BART - Model for finding how critical content of References is looked at and how many References are used on a Topic
Edge weights with Critical Thinking and Multi Reference Indicators
Reference Graph with weighted edges to identify Weak Claims + basic Search and Plot Options

Figure 1: Test and Application Flow

Main Task

The Tasks within the work are split into different sub-tasks:

- Reference Extraction: From the plain text of scientific works extracting the references to different work and continuing this path building a graph of references.
- Usage Validation: Out of the Text (or some snippets, surrounding references and given by the reference extractor) create edge weights for the

network which tell how critical a work uses some source.

- Graph: A full Graph over the given Database which makes it possible to identify topics/contents which are badly cited or not proofed.

Dataset and Processing

For this project we will have two different type of data sets, the training data and application data.
As training data, we will be using a data set that includes papers and other scientific works with already extracted citations as meta data. For one part, it will be used to train the model to extract the citations out of the application data and test the resulting model.
Additionally, by applying a simple algorithm, which systematically parses the citation meta data, we will build a graph – the reference tree – out of the data set to improve the accuracy of the model. The training data in combination with the reference tree will also be used to test the context extraction capabilities of the second model in the project.
The application dataset consists of papers and other scientific writings where the citations have not been extracted. This dataset will be significantly smaller in comparison to the training data and, for the purpose of this project, will contain examples of cross referencing.
Data sets we are currently looking into using as training data are a subset of the S2ORC data set[1] and the physics scholarly article data set[2].

Methods / Models

To enable our graph creation we will start with a basic reference tree builder which builds a graph from a dataset which already includes metadata including the edges. With that graph we will fit a pretrained distilBERT such that it will be able to, from text only, create the needed metadata for creating a graph. Furthermore we will fit a BART such that it finds out context for how references are used and score them accordingly to how critically they are viewed and how strongly claims are backed up.

Evaluation

We will try to measure our system by making a visualization function for the graph which then should show some specific works as roots of ”bad practice” referencing which we hid within the database. The overall structure should not be taken as a fully fetched version but more as a proof of concept.

[1] https://github.com/allenai/s2orc?tab=readme-ov-file

[2] www.kaggle.com/datasets/shivd24coder/physics-scholarly-articles