Automatic Detection of Core Sections in Scientific Papers

Monica Emilia Văruicu, Paul Stefan Popescu, Marian Cristian Mihăescu

University of Craiova, Romania

varuicu.monica.i6x@student.ucv.ro, stefan.popescu,cristian.mihaescu@edu.ucv.ro

Academic research tasks can be substantially optimised in terms of time efficiency by employing techniques that provide targeted access to the most relevant sections of research papers, eliminating the need for exhaustive reading of the entire document. In the context of scientific research, a commonly encountered challenge is the necessity to review and analyse a large number of articles, which often proves to be a demanding and time-consuming task. Typically, the interest in a particular paper focuses on its key points, such as the research issue, the proposed approach or the main results. In this context, tools that automatically detect the structure of research papers can facilitate access to relevant content and contribute to a more rapid research process. Most of the scientific articles are organised into standardised sections, which provide a clear structure for presenting research. In practice, however, quickly locating and understanding these sections can be challenging, especially when dealing with extensive collections of documents. Enabling automatic identification of these sections could significantly improve both navigation and content retrieval. We tackle the task of training a classifier that labels each sentence in a research paper with its corresponding section category, such as Introduction, Related Work, Proposed Approach, Body, Results, and Conclusion. We elaborate on this premise by introducing a sentence classification model trained on a high-quality dataset obtained through a dedicated data filtering and selection pipeline. The model is designed to assign each sentence from a research paper to its corresponding structural section, enabling accurate and fine-grained discourse segmentation. The pipeline leverages two pre-trained models to preprocess the data, which was initially split into smaller subsets serving as training and testing sets. Each sentence in these subsets was labelled according to its source section in raw form, based on the predictions of the two models. The subsets were later merged to reconstruct the original dataset, retaining only the sentences for which both models predicted the same label, thus ensuring high annotation consistency. Although several architectures were explored during experimentation, the final and most effective version of the classifier is based on the BERT-Base architecture. This model was fine-tuned on the refined dataset, achieving notable results in sentence-level section classification and demonstrating the effectiveness of both the dataset construction process and the underlying model. The classifier architecture, based on BERT-Base and trained on the filtered dataset, demonstrates stable and well-balanced performance, with an average precision of 84.67%, a recall of 83.78%, and an F1 score of 84.86%. The close alignment of these metrics indicates no significant trade-offs between accuracy and coverage. Performance variation is moderate, precision ranging from 78.66% to 90.93%, recall from 77.51% to 91.94%, and F1 from 81.27% to 91.43%, reflecting a robust and consistent classification across all classes. These results are further supported by a solid overall accuracy of 86.49%, highlighting the strong impact of data quality on model reliability. To enable reproducibility, the dataset [1], the model [2], and the code [3] implementing the classifier are publicly available at this GitHub repository. We have shown that it is possible to achieve a reliable classifier capable of accurately labelling sentences from research papers, which can significantly support research workflows by improving both processing time and task efficiency. The classifier would enhance the accessibility of research articles by consistently and precisely labelling sentences, facilitating the identification of the sections within the paper. This research opens the way for the future development of architectures that support summarisation, information retrieval, and semantic search, facilitating the efficient processing of large volumes of scientific documents regardless of the specific research task. By enabling structured understanding at the sentence level, it lays the groundwork for more advanced and task-agnostic tools that can adapt to diverse academic needs.

References

[1] Data available at: https://www.kaggle.com/datasets/monicavaruicu/sentence-split-arxiv-papers-with-section-tags.

[2] Model available at: https://www.kaggle.com/models/monicavaruicu/bert-section-labeler.

[3] Code available at: https://github.com/monicavaruicu/document-sections-detection.