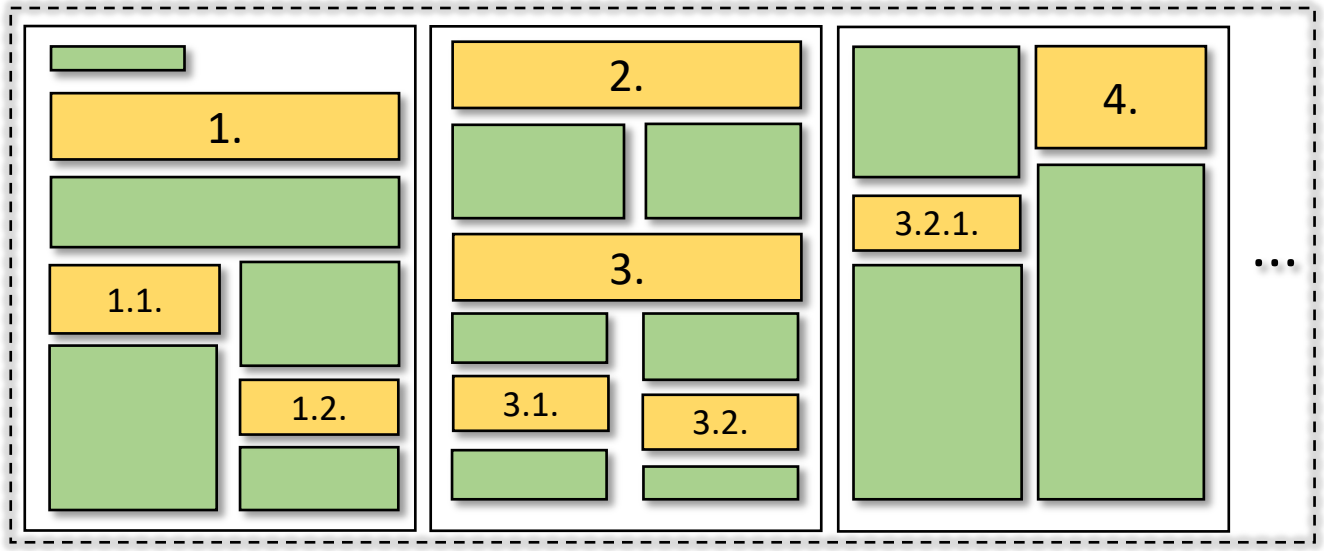
A Scalable Framework for Table of Contents Extraction from Complex ESG Annual Reports

Scan to check the paper!

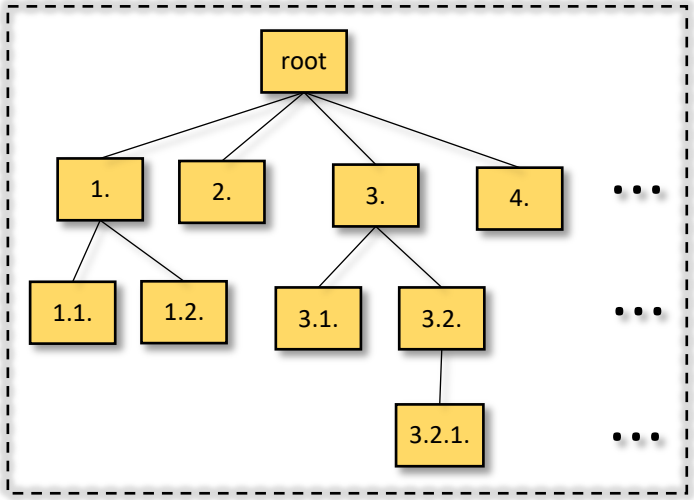
Xinyu Wang, Lin Gui, Yulan He

Table of Contents Extraction

Document



Tree of Table of Contents



Previous Dataset

- HierDoc: focus on scientific papers; well-structured and short.

A Lie conformal algebra of Block type

Lamei Yuan
Academy of Fundamental and Interdisciplinary Science,
Harbin Institute of Technology, Harbin 150080, China
lmyuan@hit.edu.cn

Abstract: The aim of this paper is to study a Lie conformal algebra of Block type. In this paper, conformal derivation, conformal module of rank 1 and low-dimensional comohology of the Lie conformal algebra of Block type are studied. Also, the vertex Poisson algebra structure associated with the Lie conformal algebra of Block type is constructed.

Keywords: Lie conformal algebra, vertex Lie algebra, cohomology, vertex Poisson algebra

MR(2000) Subject Classification: 17B65, 17B69

1 Introduction

The notion of Lie conformal algebra, introduced by Kac [9], encode an axiomatic description of the operator product expansions of chiral fields in conformal field theory. It is a powerful tool for the study of infinite-dimensional Lie (super)algebras, associative algebras and their representations. Lie conformal algebras have been extensively studied, including the classification problem [5, 6], cohomology theory [2, 12] and representation theory [3].

The Lie conformal algebras are closely related to vertex algebras. Primc [11] introduced and studied a notion of vertex Lie algebra, which is a special case of a more general notion of local vertex Lie algebra [4]. As it was explained in [10], the notion of Lie conformal algebra and the notion of vertex Lie algebra are equivalent. In this paper, we shall use Lie conformal algebra and vertex Lie algebra synonymously.

With the notion of vertex Lie algebra, one arrives at the notion of vertex Poisson algebra, which is a combination of a differential algebra structure and a vertex Lie algebra structure, satisfying a natural compatibility condition. The symmetric algebra of a vertex Lie algebra is naturally a vertex Poisson algebra [7]. A general construction theorem of vertex Poisson algebras was given in [10]. Applications of vertex Poisson algebras to the theory of integrable systems were studied in [1].

In the present paper, we study a nonsimple Lie conformal algebra of infinite rank, which is endowed with a $\mathbb{C}[\partial]$-basis $\{J_i | i \in \mathbb{Z}^+\}$, such that

$$[J_{i\,\lambda}J_j] = ((i+1)\partial + (i+j+2)\lambda)J_{i+j}, \text{ for } i,j \in \mathbb{Z}^+. \tag{1.1}$$

The corresponding formal distribution Lie algebra is a Block type Lie algebra, which is the associated graded Lie algebra of the filtered Lie algebra $W_{1+\infty}$ [13, 14, 15, 16, 18]. Thus we call this Lie conformal algebra a *Lie conformal algebra of Block type* and denote it by $\mathscr{B}$ in this paper. It is a conformal subalgebra of $\mathrm{gr}\, gc_1$ studied in [17]. In addition, it contains the Virasoro conformal algebra $\mathrm{Vir} = \mathbb{C}[\partial]\mathrm{J}_0$ with $[J_{0\,\lambda}J_0] = (\partial + 2\lambda)J_0$ as a subalgebra.

Provably Convergent Algorithms for Solving Inverse Problems Using Generative Models

Viraj Shah, Rakib Hyder, M. Salman Asif, and Chinmay Hegde[¶]

Abstract

The traditional approach of hand-crafting priors (such as sparsity) for solving inverse problems is slowly being replaced by the use of richer learned priors (such as those modeled by deep generative networks). In this work, we study the algorithmic aspects of such a learning-based approach from a theoretical perspective. For certain generative network architectures, we establish a simple non-convex algorithmic approach that (a) theoretically enjoys linear convergence guarantees for certain linear and nonlinear inverse problems, and (b) empirically improves upon conventional techniques such as back-propagation. We support our claims with the experimental results for solving various inverse problems. We also propose an extension of our approach that can handle model mismatch (i.e., situations where the generative network prior is not exactly applicable). Together, our contributions serve as building blocks towards a principled use of generative models in inverse problems with more complete algorithmic understanding.

1 Introduction

1.1 Motivation

Inverse problems arise in a diverse range of application domains including computational imaging, optics, astrophysics, and seismic geo-exploration. In each of these applications, there is a target signal or image (or some other quantity of interest) to be obtained; a device (or some other physical process) records measurements of the target; and the goal is to reconstruct an estimate of the signal from the observations.

Let us suppose that $x^* \in \mathbb{R}^n$ denotes the signal of interest and $y = \mathcal{A}(x^*) \in \mathbb{R}^m$ denotes the observed measurements. The aim is to recover (an estimate of) the unknown signal x^* given y and $\mathcal{A}$. Based on the forward measurement operator $\mathcal{A}$, the inverse problem can be defined in two broad categories of linear and nonlinear problems. Many important problems in signal and image processing can be modeled with a *linear* measurement operator $\mathcal{A}$; examples include *compressive sensing*, the classical problem of *super-resolution* or the problem of *image inpainting*. In case of nonlinear inverse problems, the operator $\mathcal{A}$ exhibits a nonlinearity; examples include *phase retrieval*, *blind deconvolution*, and *de-quantization*.

When $m < n$, the inverse problem is ill-posed, and some kind of prior (or regularizer) is necessary to obtain a meaningful solution. A common technique used to solve ill-posed inverse problems is to seek the minimum of a

[*] V. Shah was with Iowa State University. He is now with the ECE Department at the University of Illinois, Urbana-Champaign. (e-mail: vjshah3@illinois.edu).
[†] C. Hegde was with Iowa State University. He is now with the Tandon School of Engineering at New York University. (e-mail: chinmay.h@nyu.edu).
[‡] R. Hyder and M. Asif are with the ECE Department at the University of California Riverside. (e-mail: sasif@ece.ucr.edu).
[§] This work was completed when VS and CH were at Iowa State University, and were supported in part by grants CAREER CCF-1750920/2005804 and CCF-1815101, a faculty fellowship from the Black and Veatch Foundation, and an equipment donation from the NVIDIA Corporation. RH and MA were supported in part by grants CAREER CCF-2046293 and ONR N00014-19-1-2264, and equipment donation from NVIDIA. Parts of this manuscript appeared in short conference papers [1], [2], and [3].

achieve good cross-lingual performance. We take on a different approach by using the monolingual model itself instead of extracting knowledge from it.

Rust et al. (2020) compared multilingual and monolingual models on monolingual tasks (i.e., the tasks whose language is the same as the monolingual model). They found that both the size of pretraining data in the target language and vocabulary have a positive correlation with monolingual models' performance. Based on our results, we hypothesize that a model pretrained with MLM using a large monolingual corpus develops both language-specific and language-agnostic properties, being the latter predominant over the former.

3 Methodology

Our method consists of a pretrain-finetune approach that uses different languages for both. We call **source language** as the language used for pretraining our models. We refer to **target language** as a second language, different from the one used for pretraining our model. We apply the following steps:

1. Pretrain a monolingual model on the **source language** with masked language modeling (MLM) objective using a large, unlabeled dataset.
2. Finetune and evaluate the model on a downstream task with a labeled dataset in the **target language**.

The novelty of our approach is to perform a cross-lingual evaluation using *monolingual models* instead of bi-lingual or multi-lingual ones. We aim to assess if the model is able to rely on its masked language pretraining to achieve good representations for a task even when finetuned on a different language. If successful, this would suggest that MLM pretraining provides the model with representations for more abstract concepts rather than learning a specific language.

Pretraining data. Our monolingual models are pretrained on a large unlabeled corpus, using a source language's vocabulary. Some high-resource languages, such as English, have a high presence in many datasets from other languages, often created from crawling web resources. This may influence the model's transfer ability because it has seen some examples from the foreign language during pretraining. However, the corpora used to pretrain our models have a very small amount of sentences in other languages. For instance, Portuguese pretraining corpus has only 14,928 sentences (0.01%) in Vietnamese.

Control experiment. To discard the hypothesis that the monolingual model can learn patterns from the finetuning dataset, instead of relying on more general concepts from both finetuning and pretraining, we perform a control experiment. We train the models on the target language tasks without any pretraining. If models with monolingual pretraining have significantly better results, we may conclude that it uses knowledge from its pretraining instead of only learning patterns from finetuning data.

Evaluation tasks. We follow a similar selection as in Artetxe et al. (2020) and use two downstream types of tasks for evaluation: natural language inference (NLI) and question answering (QA). Even though a classification task highlights the model's ability to understand the relationship between sentences, it has been shown that the model may learn some superficial cues to perform well (Gururangan et al., 2018). Because of that, we also select question answering, which requires natural language understanding as well.

4 Experiments

In this section, we outline the models, datasets and tasks we use in our experiments.

4.1 Models

We perform experiments with four base models, highlighted in Table 1. The experiments run on the Base versions of the BERT model (Devlin et al., 2019), with 12 layers, 768 hidden dimensions and 12 attention heads. We use models initialized with random weights. We also report the number of parameters (in millions) and the pretraining dataset sizes in Table 1.

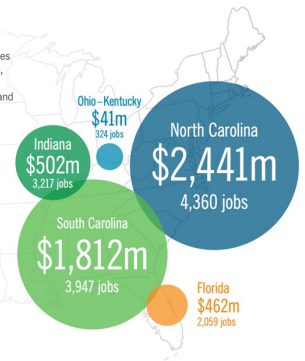
4.2 Pretraining procedure

The selected models are pretrained from random weights using a monolingual tokenizer, created from data in their native language. We also select models that have been pretrained using the Masked Language Modeling (MLM) objective as described by Devlin et al. (2019).

Our Collected ESGDoc dataset

ECONOMIC DEVELOPMENT

Duke Energy works with state and local authorities to promote economic growth in our communities, helping attract business investment and jobs. Duke Energy helped attract nearly 14,000 jobs and $5.3 billion of investments in 2018.



$5,300m
Total Capital Investment
$ (million)

14,000
Total Jobs

- **Transforming the customer experience.** Duke Energy is working hard to further improve the customer experience. New technology is shortening and sometimes eliminating power outages. Smart meters are giving customers new ways to manage and reduce electricity usage, saving them money. Electric vehicle charging stations are giving customers new transportation fuel options.
- **Engaging stakeholders.** Fortune magazine named Duke Energy to its 2019 "World's Most Admired Companies" list – an indication that Duke Energy's many diverse stakeholders recognize and value the company's significant progress on its future-focused journey. The company continues to work collaboratively with regulators, legislators, environmentalists, consumer advocates and many others on its multiple sustainability and modernization initiatives.

Economic Development: Jobs and Major Investment

Duke Energy's economic development team in 2018 helped bring nearly 14,000 new jobs and $5.3 billion in private-sector investment – through 94 projects – to the six states served by the company's electric utilities.

Site Selection magazine named Duke Energy to its "Top Utilities in Economic Development" list for the 14th consecutive year.

Duke Energy's economic development specialists work to attract new industry to North Carolina, South Carolina, Florida, Indiana, Ohio and Kentucky. The team also encourages existing companies in those states to expand at home, rather than look elsewhere.

In 2018, the team evaluated 26 properties for potential business and industrial development through Duke Energy's Site Readiness Program. The program identifies potential business and industrial sites, then

Diverse, Engaged and Capable Workforce

We invest in people to strengthen organizational capability and develop a talented global workforce that gets results the right way. Our success in attracting, developing and retaining a diverse workforce comes from strategies, programs and processes based on The Chevron Way.

We are committed to building a workforce that represents the many countries where we operate. We believe that sustainable high performance is achieved by creating a culture that encourages and values people with a wide range of experiences and knowledge.

21,000
EMPLOYEES

In 2014, an estimated 21,000 employees–about one-third of our regular workforce–participated in Chevron's many employee networks that celebrate cultural and lifestyle differences.

CATALYST
2015 AWARD WINNER

Chevron was honored to receive the prestigious 2015 Catalyst Award, the premier honor for companies committed to expanding opportunities for women.

Our application for the Catalyst Award–titled "The Chevron Way: Engineering Opportunities for Women"–details how The Chevron Way's focus on people over the past two decades established a culture that attracted and retained more diverse talent in our workforce.

The application also shows how The Chevron Way served as a primary driver for our diversity strategy that includes programs, processes and tools to facilitate gender diversity and help increase the number of women in leadership roles.

To be considered for the Catalyst Award, we participated in a rigorous, yearlong application and review process that included extensive documentation and interviews with more than 60 employees and leaders. We shared information and perspectives with Catalyst regarding our business rationale, senior leadership support, accountability, communication, employee engagement, innovation and measurable results.

OCCUPATIONAL HEALTH

As part of the Company's efforts to ensure a healthy workforce, Ferrexpo's medical department at FPM conducted 8,792 medical examinations in 2016, the equivalent of 98% of Ferrexpo's employees in Ukraine (2015: 9,222[1], 97%[1] of employees).

98%
EQUIVALENT PROPORTION OF EMPLOYEES AT OPERATIONS IN UKRAINE THAT RECEIVED A HEALTH CHECK.

1 2015 figure restated to included Ferrotrans employees.

OCCUPATIONAL HEALTH GOALS AND PERFORMANCE IN 2016

GOAL	PERFORMANCE
Improve awareness of occupational health risks and increased reporting of instances of illness caused by occupational risk	Efforts to increase awareness of occupational health issues and reporting of new conditions continue to be developed.

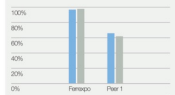
In 2016, the statutory annual review was carried out by representatives of the Scientific and Research Institute of Preventative Medicine, assessing individuals for occupational illnesses and injuries. If an occupational illness is identified or suspected, employees are referred to the Department of Occupational Health and Illnesses at the Kharkov National Medical University. There were no occupational illnesses or diseases identified during the period.

Employees who have worked for over ten years in conditions that exceed the exposure limits recommended are on prophylactic monitoring and undergo sanatorium-resort treatment in the specialised resorts of Ukraine. In 2016, a total of 450 Ferrexpo employees attended these sanatorium resorts (2015: 750).

None of the Company's operations are located in a geographical location where occupational activities would expose workers to a high risk of a location-specific disease, such as malaria.

BENCHMARKING PERFORMANCE
In 2016, Ferrexpo's medical centre in Ukraine performed health checks on the equivalent of 98% of Ferrexpo's employees located in Ukraine, which compares favourably with the Company's peers.

EMPLOYEE HEALTH MONITORING
Medical checks (% of total employees)



Ethics and Compliance

We are committed to upholding the highest ethical standards and ensuring compliance in the workplace and our business activities. Our core values include responsibility and integrity and we understand that these values are foundational to our success. They are vital to creating a culture of mutual trust and respect in our interactions with stakeholders, including workforce, customers, suppliers and the public.

STANDARDS OF BUSINESS CONDUCT

To promote a strong and consistent culture of ethics, we have uniform Standards of Business Conduct (SBC) that apply to our global workforce around the world. These standards guide personnel in making the best possible decisions. The SBC is available to our workforce in 11 languages and to the public online.

Upon joining Applied, all personnel must review and certify adherence to our SBC and must refresh their knowledge of it every two years by completing an online training course. They are also required to verify annually that they have read, understand and agree to comply with the SBC by completing a recertification process. In 2017, we achieved a 97% SBC-recertification rate.

TRAINING

Each member of our workforce undergoes comprehensive web-based and/or classroom training in our Anticorruption, Conflict of Interest, Global Gift, Expense Reporting, Global Travel and Insider Trading policies. Personnel must refresh their knowledge of these policies at regular intervals.

Through training on privacy issues and IP (intellectual property) protection, everyone working at Applied learns the importance of keeping customers' and suppliers' data and information secure. We provide an ethics training to suppliers to ensure that they understand our requirements and how we operate.

2017 WORLD'S MOST ETHICAL COMPANIES®
WWW.ETHISPHERE.COM

In 2017, for the sixth year running, we were recognized by the Ethisphere Institute as one of the World's Most Ethical Companies®

Our Winnebago Industries Ethics Hotline provides an avenue for employees and suppliers to report instances of misconduct anonymously and without fear of retaliation. All reports are investigated fully and are reviewed by our CEO and Audit Committee of the Board and handled in accordance with our Whistleblower Policy.

View the Winnebago Industries Code of Conduct and the Winnebago Industries Supplier Code of Conduct at winnebagoind.com.

Board Diversity

We recognize the importance of having diverse perspectives on our Winnebago Industries Board of Directors and aspire to promote diversity as we build and refresh our Board. Our developing diversity, equity and inclusion framework will inform board and leadership development. Since 2015, we have more than doubled board gender and racial diversity, with three women and three racially and ethnically diverse directors on our board of eleven members.

Governing Corporate Responsibility

Winnebago Industries SVP, General Counsel, Corporate Secretary and Corporate Responsibility Stacy Bogart is responsible for the oversight and governance of Winnebago Industries Corporate Responsibility efforts, in partnership with the Board's Nominating and Governance Committee, chaired by William Fisher, an independent director. The Winnebago Industries Corporate Responsibility Advisory Team, comprising business and enterprise functional leaders, provides strategic guidance to the company's ESG priorities. ESG goal owners, a subset of the advisory team, develop and drive specific strategies and goals for the priorities we identified through our ESG materiality assessment. Visit our investor relations site at winnebagoind.com to view the corporate responsibility advisory team charter.

BOARD TENURE AND REFRESHMENT

7 of 11 Directors have joined since 2015

Average Age: **59**

BOARD INDEPENDENCE

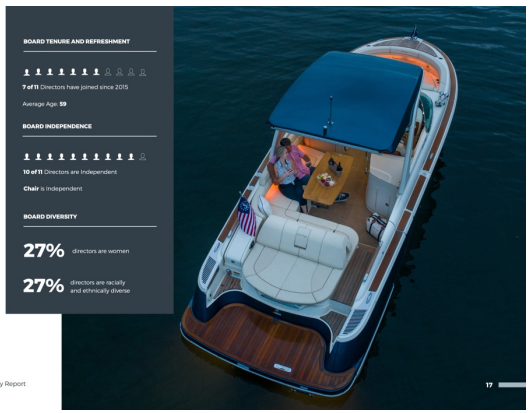
10 of 11 Directors are independent

Chair is independent

BOARD DIVERSITY

27% directors are women

27% directors are racially and ethnically diverse



ESGDoc Dataset

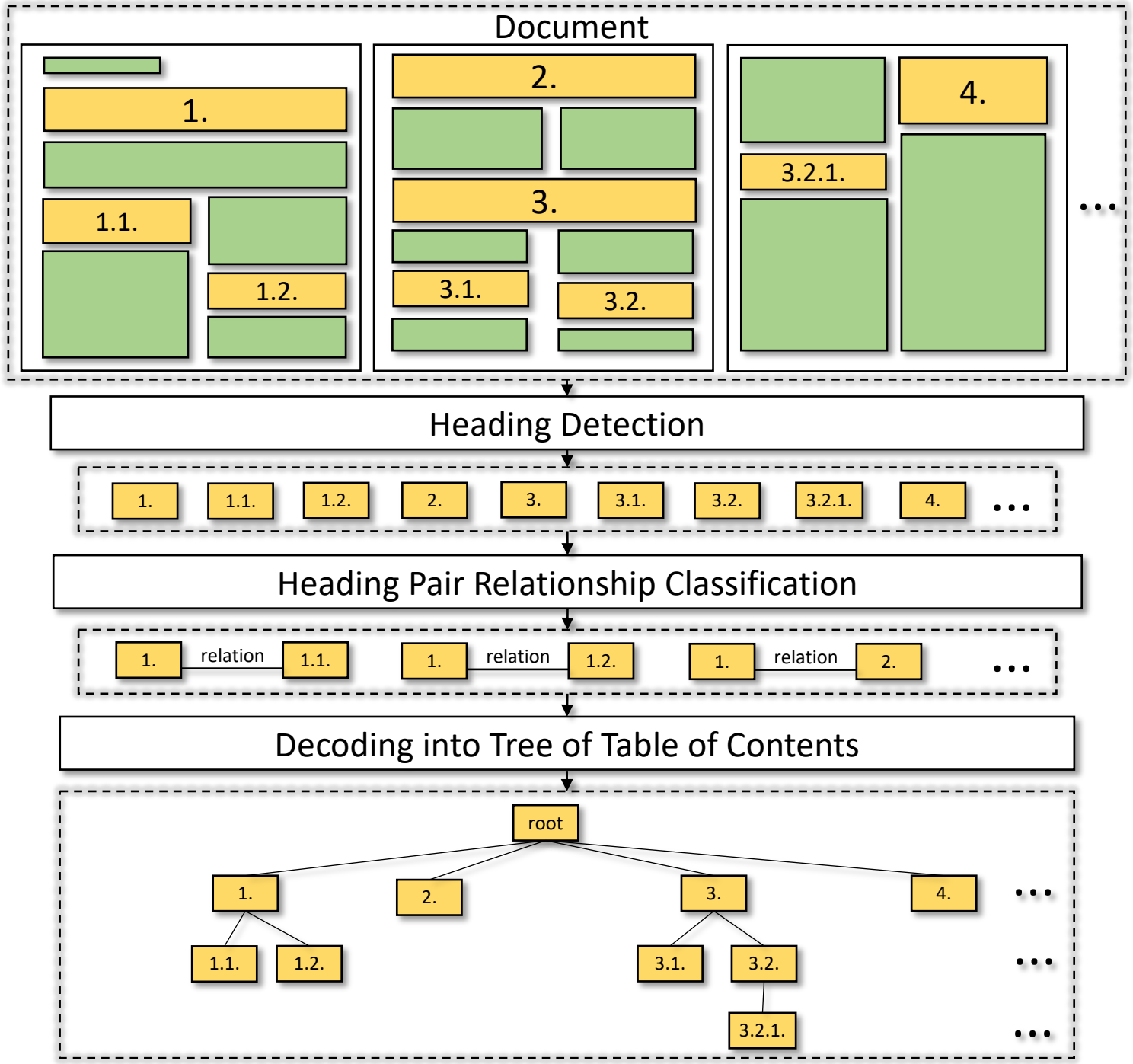
- ESGDoc comprises 1,093 publicly available ESG annual reports, sourced from 563 distinct companies, and spans the period from 2001 to 2022.

	HierDoc	ESGDoc
Total Document	650	1,093
Average pages per document	19	72

- Documents in ESGDoc are extensive, lengthy, diverse, and complex.

Previous Method

- Trained from scratch.
- Modelling relationships of heading pairs, consuming more GPU memory as document size increases, which is impractical for lengthy documents.



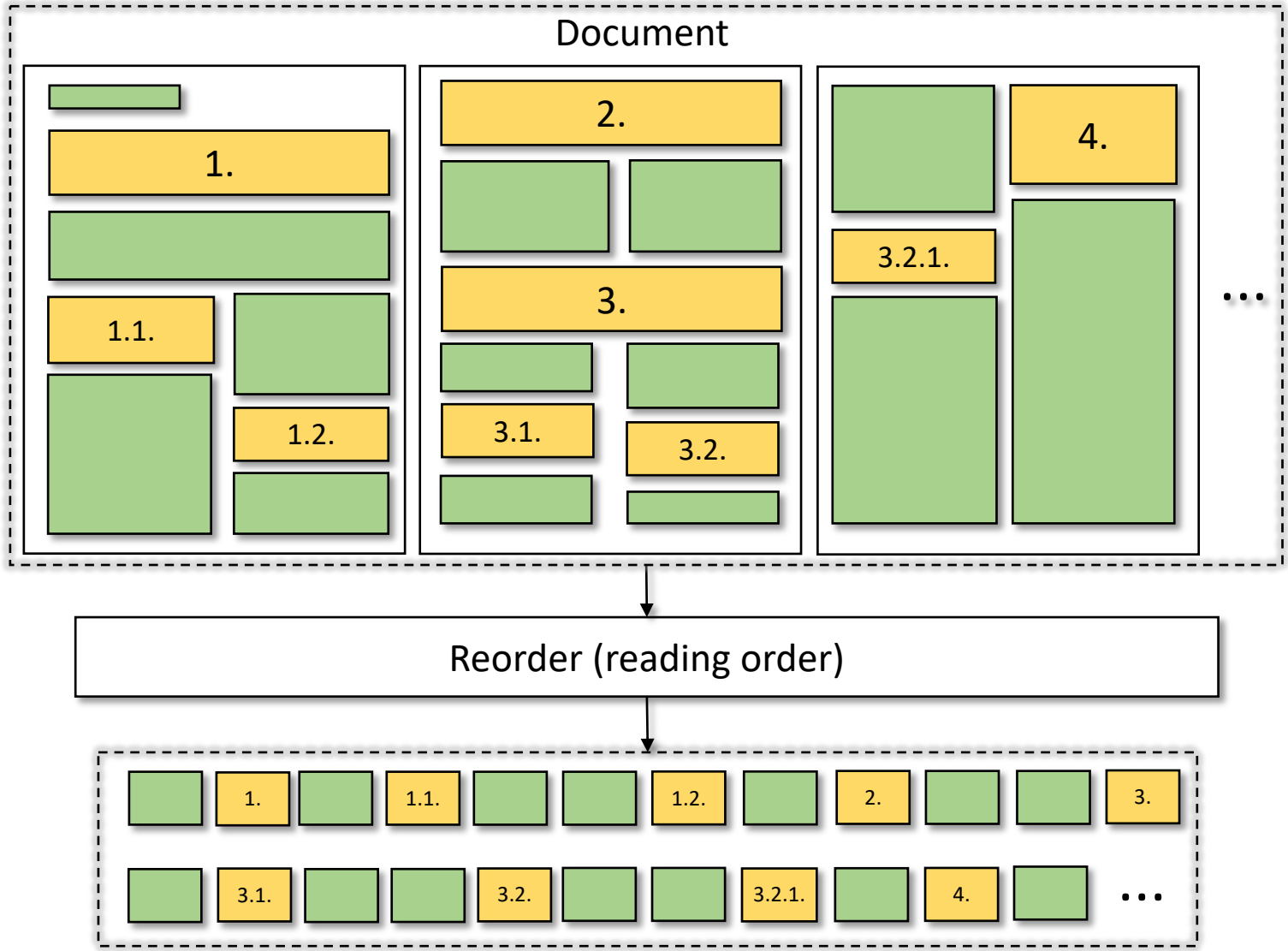
Our Method

Our method is based on the following assumptions:

- *Assumption 1*: Humans typically read documents in a left-to-right, top-to-bottom order, and a higher-level heading is read before its corresponding subheading and body text.
- *Assumption 2*: In a table of contents, the font size of a higher-level heading is no smaller than that of a lower-level heading or body text.
- *Assumption 3*: In a table of contents, headings of the same hierarchical level share the same font size.

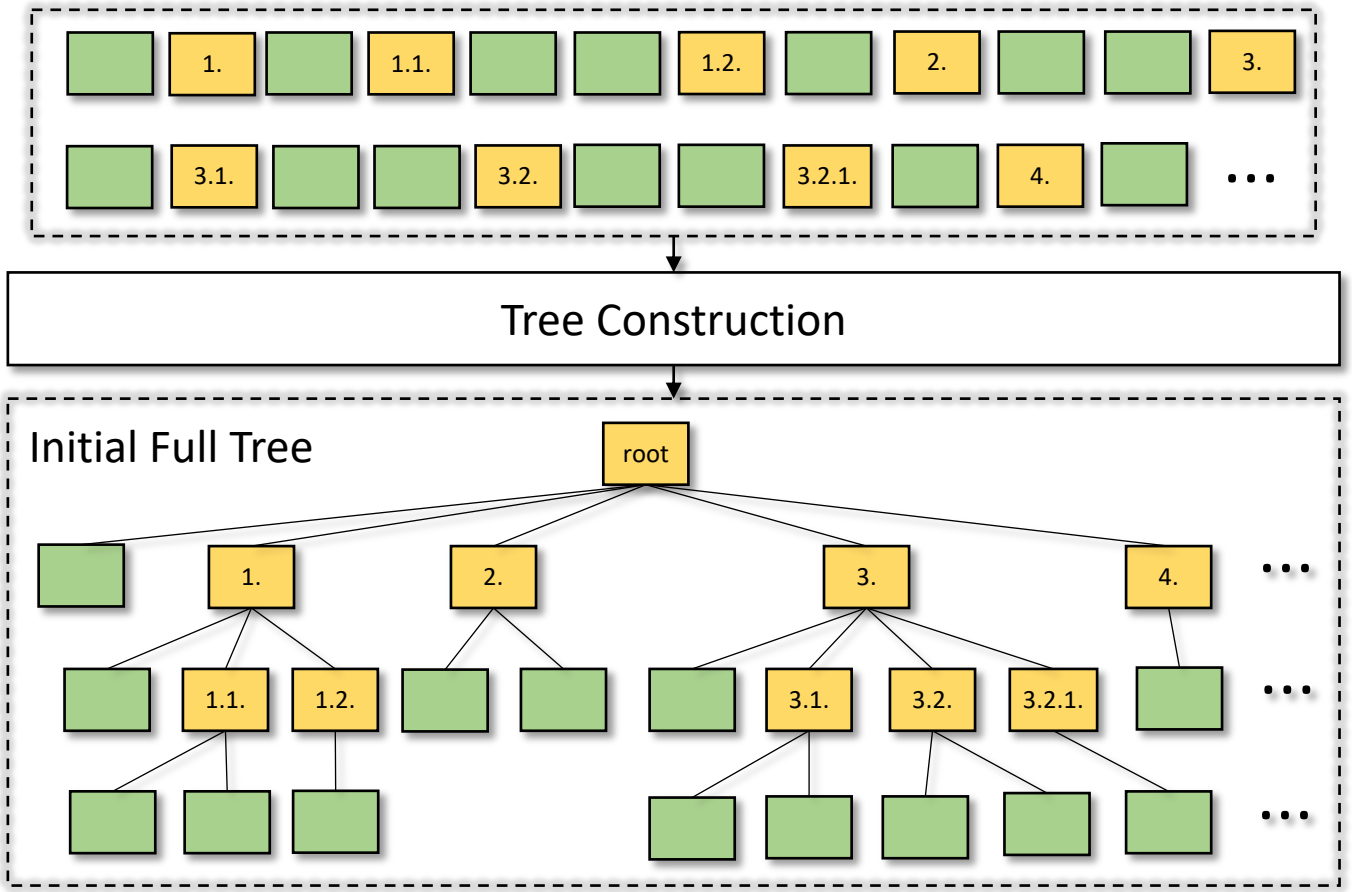
Our Method - Reorder

- Reorder the document based on reading order (from top-left to bottom-right) via xy-cut algorithm.



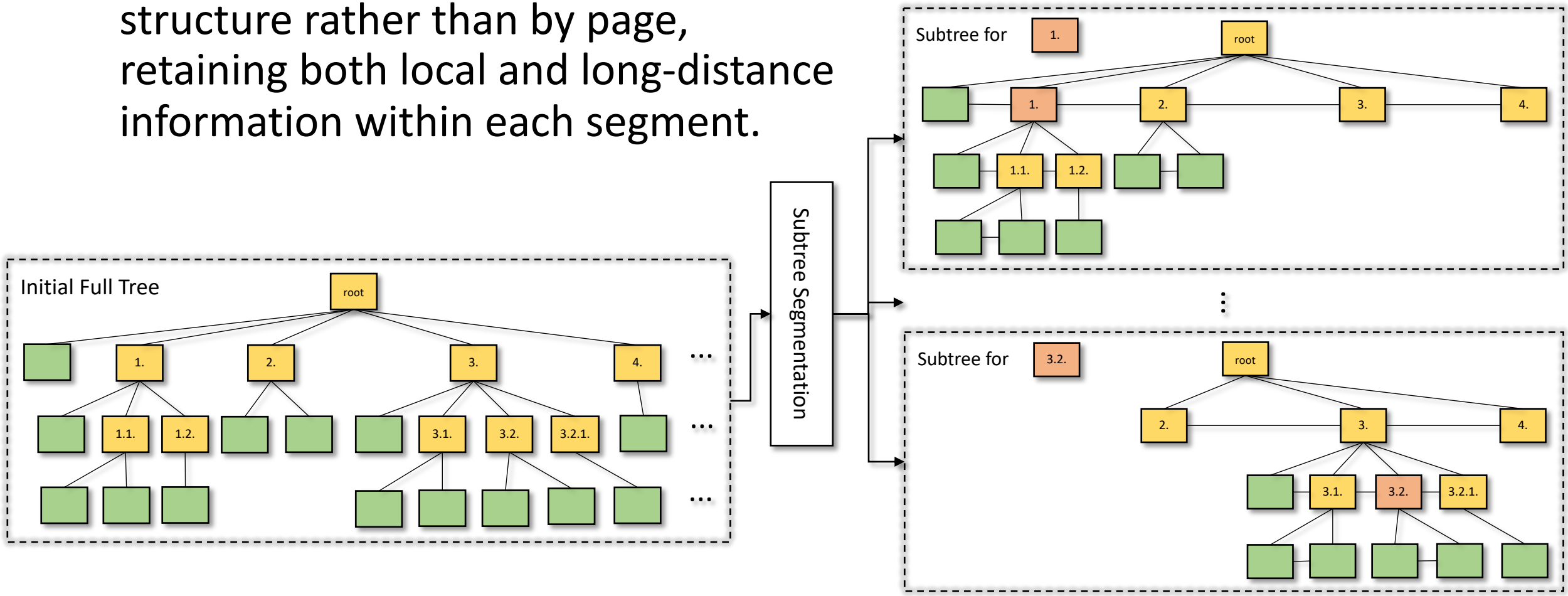
Our Method – Tree Construction

- Construct a tree based on the font size and reading order, wherein text with a larger font size is positioned at a higher level within the hierarchy.



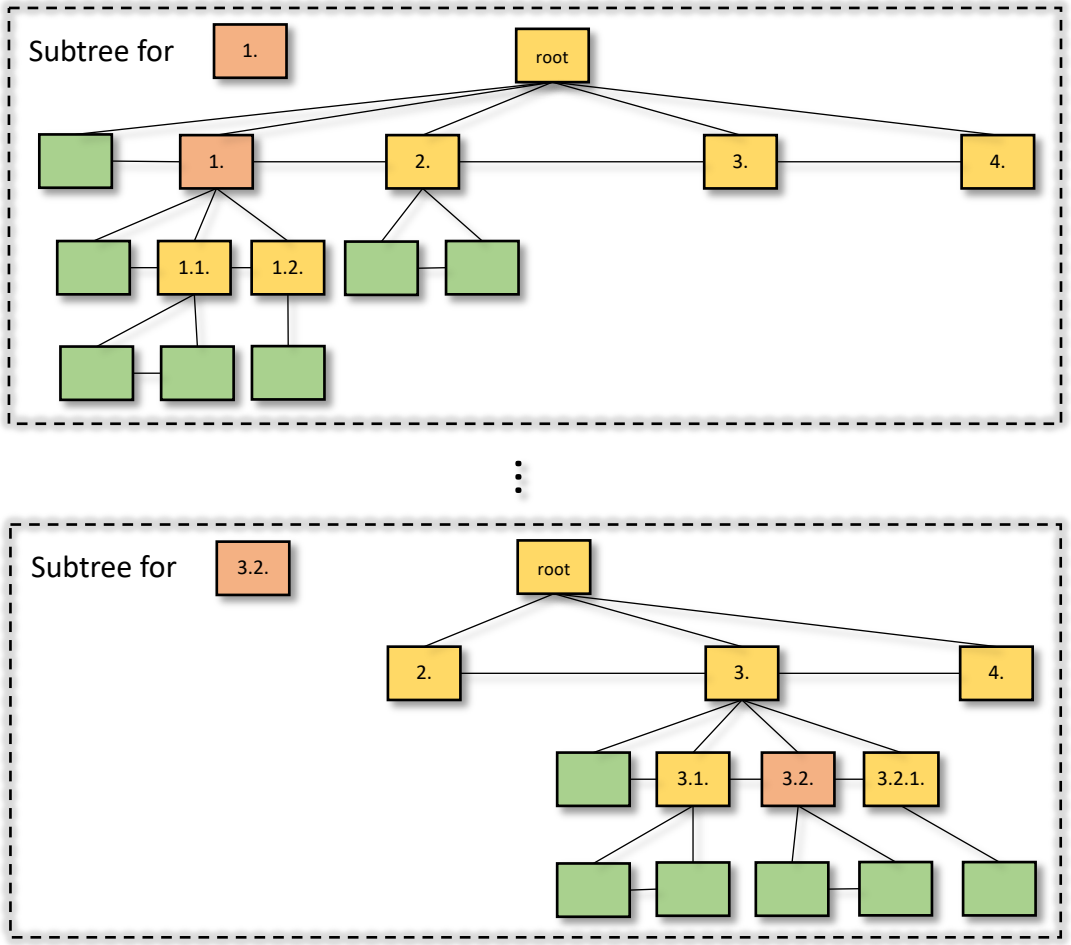
Our Method – Tree Segmentation

- Divide the document based on tree structure rather than by page, retaining both local and long-distance information within each segment.



Our Method - Modelling

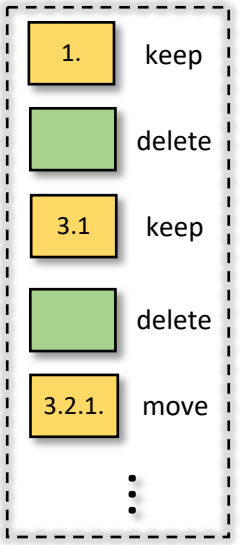
- Model each subtree separately via GNN, ensuring that GPU usage remains constant as the document lengthens.



Our Method – Modification Prediction

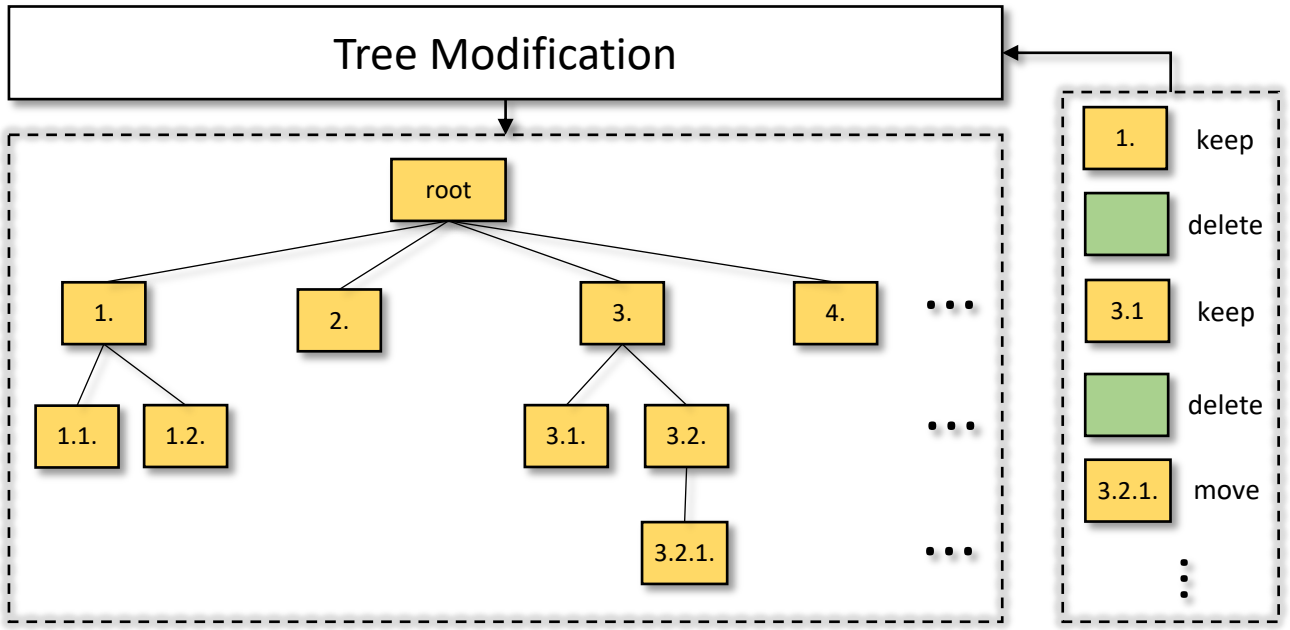
For each node, one of the following modifications is predicted:

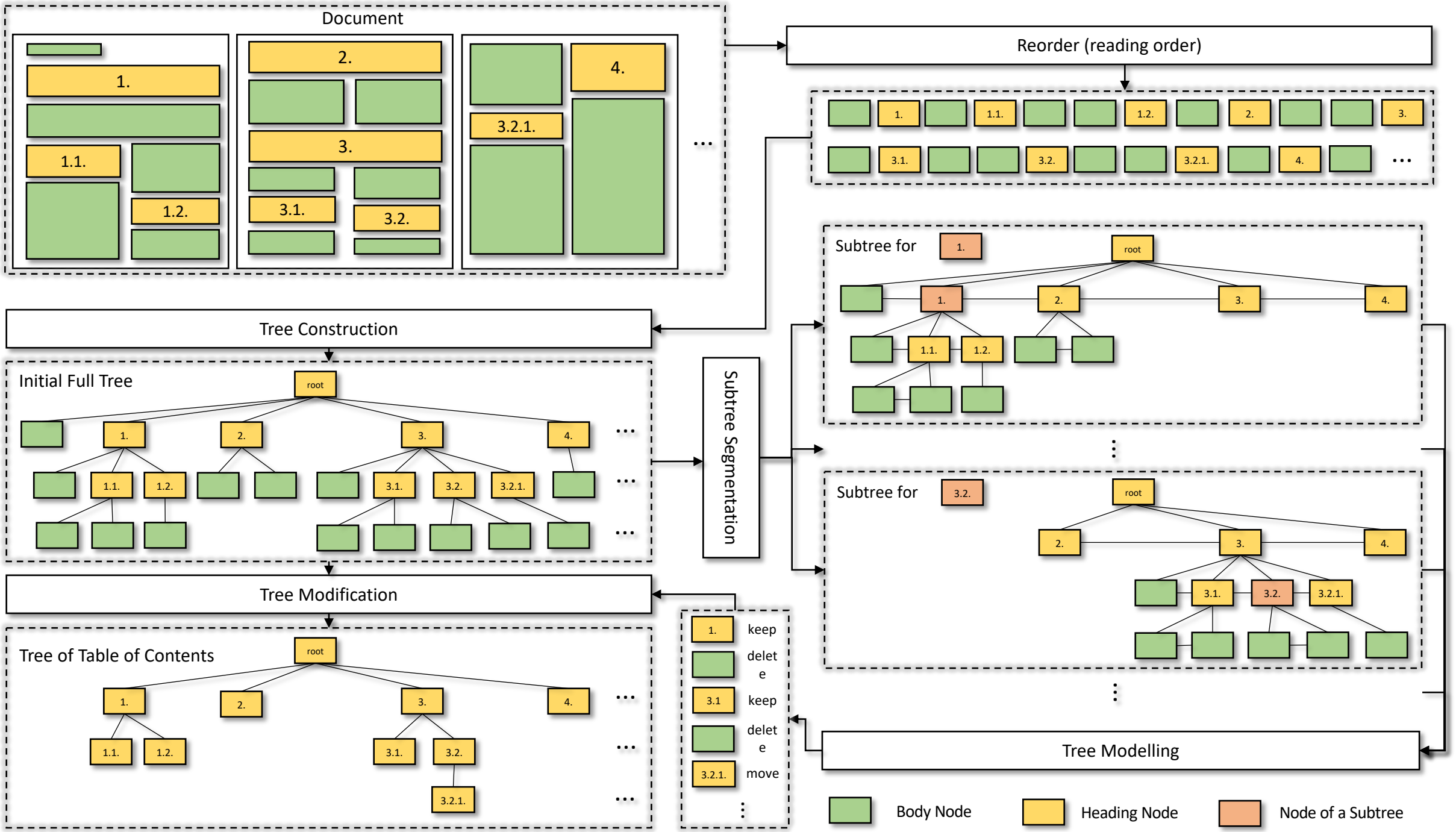
- Delete: This node is predicted as not a heading and will be deleted from the tree.
- Keep: This node is predicted as a heading and does not require any operations.
- Move: This node is predicted as a low-level heading that is a sibling of a high-level heading due to having the same font size in rare cases. This node will be relocated to be a child as its preceding sibling.



Our Method – Modification

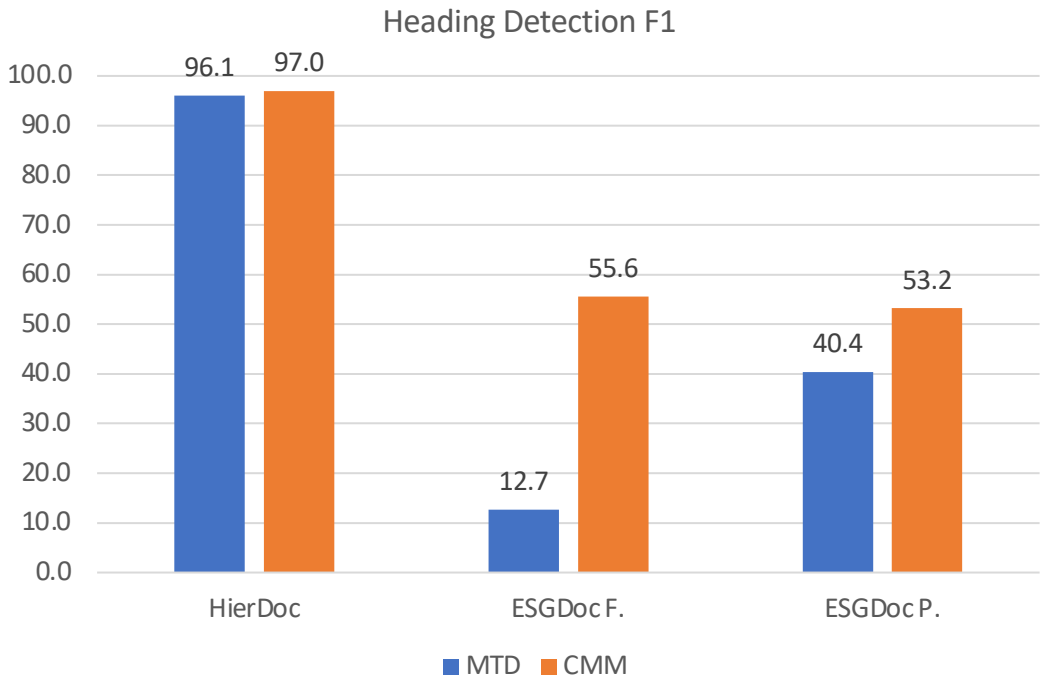
- Perform modification based on the prediction made for each node.

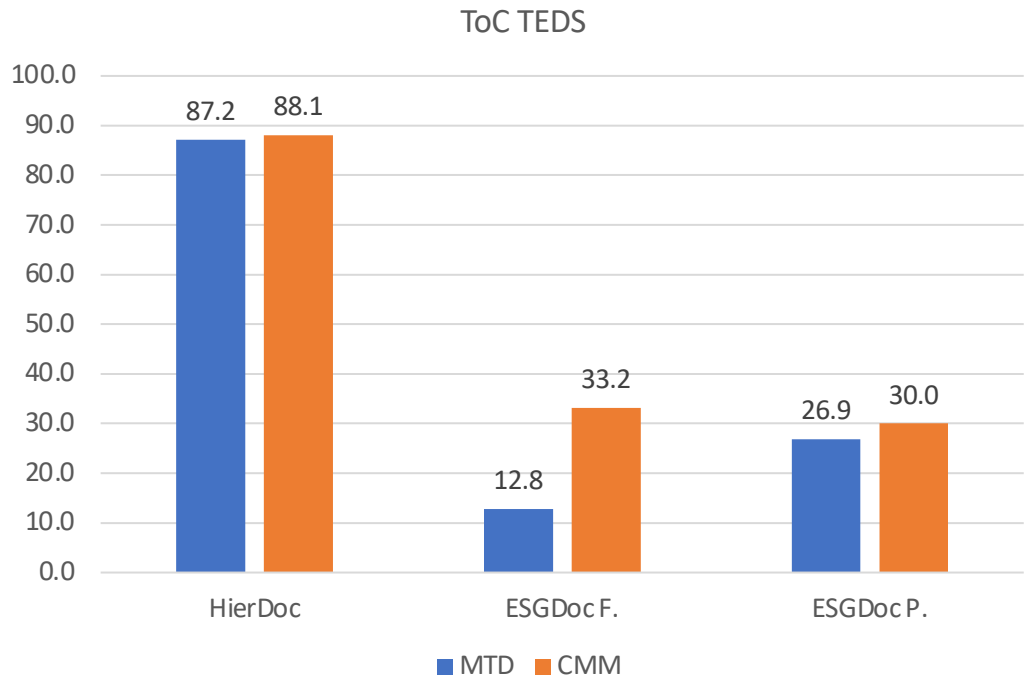



Document
1.
1.1.
1.2.
2.
3.
3.1.
3.2.
4.
3.2.1.
...
Reorder (reading order)
1. 1.1. 1.2. 2. 3.
3.1. 3.2. 3.2.1. 4. ...
Tree Construction
Initial Full Tree
root
Subtree Segmentation
Subtree for 1.
Subtree for 3.2.
Tree Modification
Tree of Table of Contents
1. keep
delet e
3.1 keep
delet e
3.2.1. move
Tree Modelling
Body Node
Heading Node
Node of a Subtree

Experiments

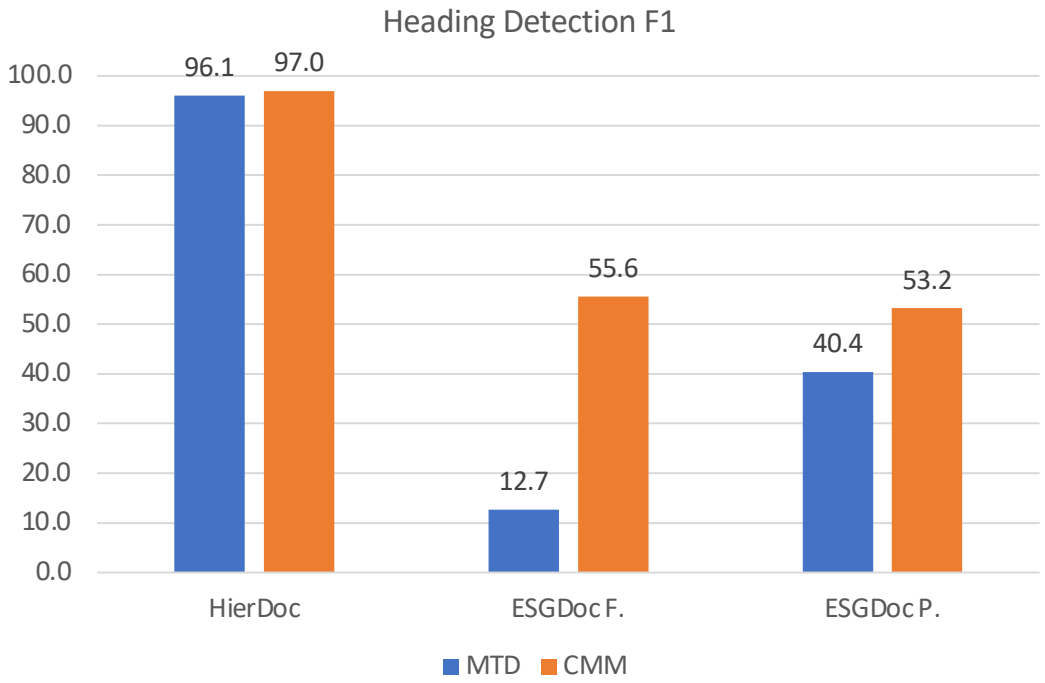
- MTD exhibits a low score in ESGDoc F. (Full) due to the out-of-memory issue when processing lengthy documents.
- ESGDoc P. (Partial) exclude documents longer than 50 pages.

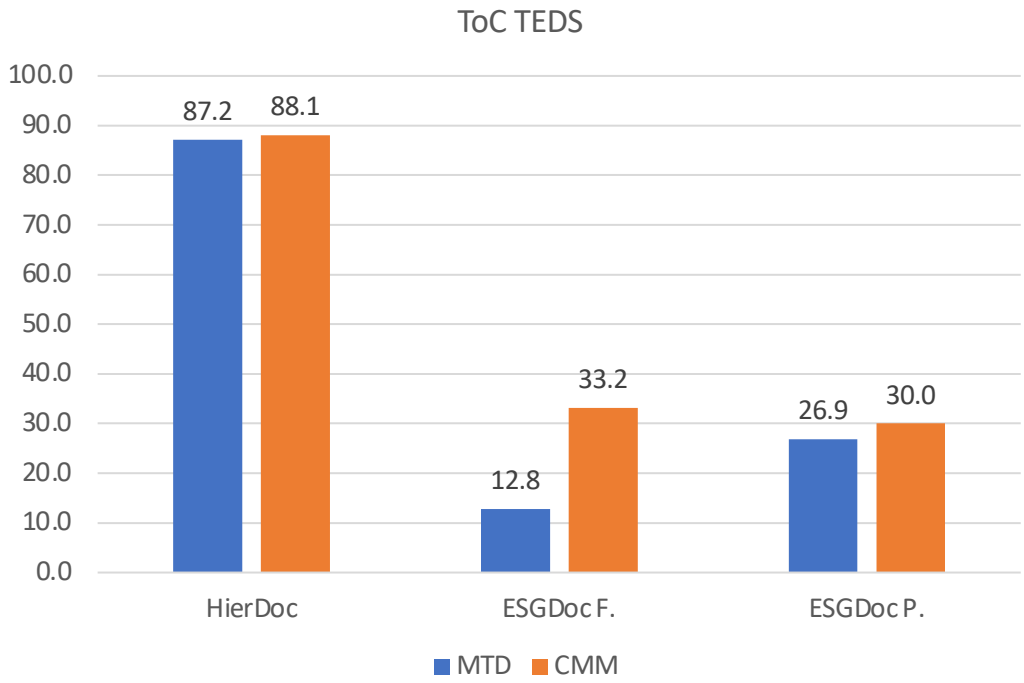




Experiments

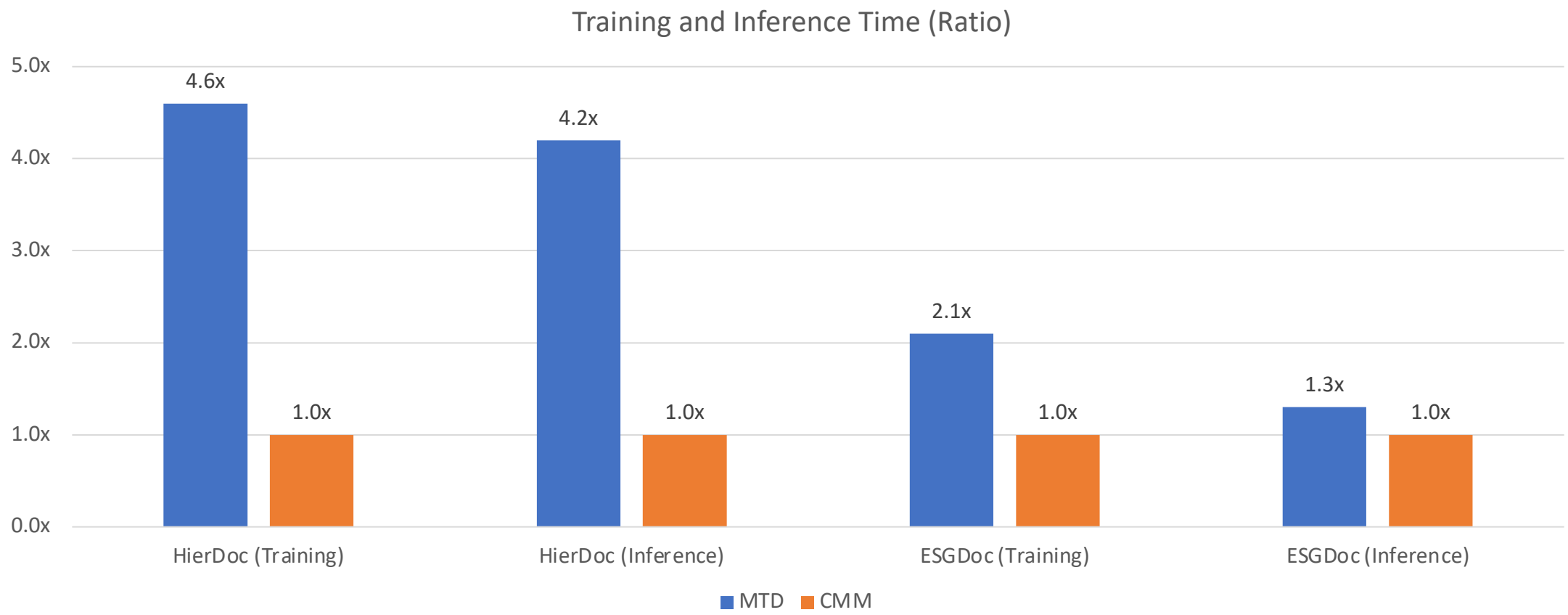
- ESGDoc is more challenging than HierDoc.
- CMM outperforms MTD and can handle documents in any length.





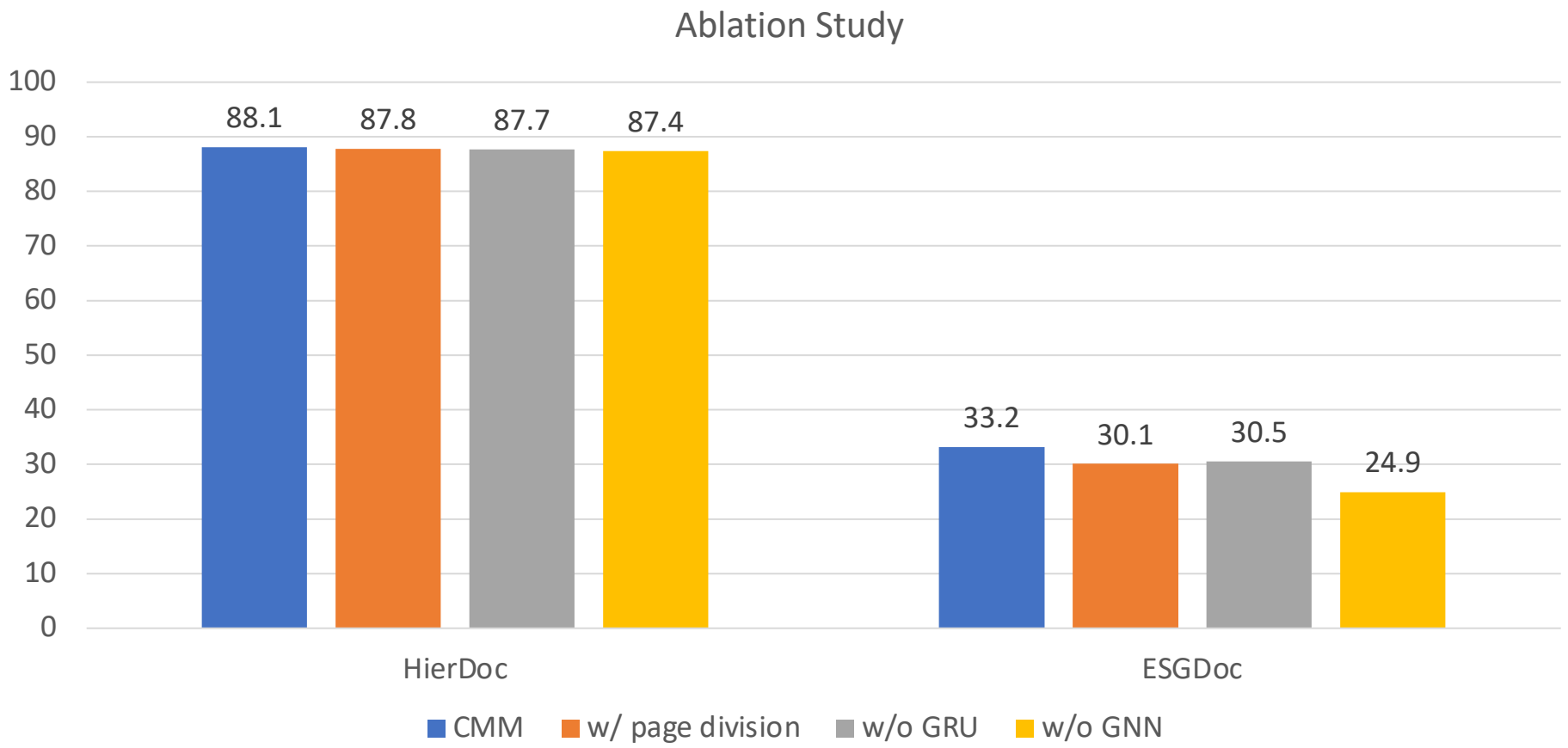
Experiments

- CMM is more computational efficient than MTD.



Experiments

- Ablation Study.



Scan to check the paper!

Thank You!