https://doi.org/10.2352/EI.2026.38.3.MOBMU-316

AROGD: A Benchmark Dataset for Automatic Reading-Order Generation of Document Collections

Md. Manzoor Murshed, Steven J. Simske
Colorado State University, Fort Collins, Colorado, United States

Abstract

Automatic reading-order generation aims to produce an effective sequence of related documents, such as textbook chapters, course lectures, and search results, using only their textual content. This capability is critical for applications such as automatic curriculum sequencing, adaptive reading-order construction for improved comprehension, and relevant search-result ranking. Despite its potential importance, the field is still relatively new, and no existing system matches the sequencing quality of expert human curation. There are no publicly available benchmark datasets for rigorous evaluation and research. To address this gap, we have proposed some algorithmic techniques based on surveying recent advances reported in leading journals and conferences and herein introduce the AROGD (Automatic Reading Order Generation Dataset), the first publicly available dataset designed for automatic reading-order generation for multiple related document collections. AROGD provides carefully curated document sets with expert-annotated orderings, enabling reproducible experimentation and standardized benchmarking of automatic sequence-generation models. This contribution establishes a foundation for future research on text-driven reading-order generation and supports the development of robust, human-competitive sequencing algorithms. The Dataset, benchmark, and leaderboard will be available at https://github.com/murshedm/AROGD-Dataset.

Introduction

Structuring a collection of related documents that facilitates better comprehension, and logical connection is a challenging task. Navigating a collection of related lecture notes on a subject or reading a collection of related documents needs logical sequence to retain knowledge and to understand the semantic linking and relationship among the documents. A well-structured and logically connected sequences of related documents can gradually build knowledge and coherent narratives on the subject matters. The reading order of the related documents significantly impacts semantic understanding, logical connection, and knowledge retention (Steven Simske, 2021). Automatic Reading-Order Generation (AROG) addresses this issue of finding such a suitable reading order automatically based on the contents of the documents in a scientific way. This is important in many Natural Language Processing (NLP) applications and tasks. Traditionally this type of task is done by human effort or by predetermined assumptions which can ignore some important relationship among them, like logical relationships, semantic similarity, and coherent sequence among them. AROG could automate the organization of documents and curriculum generation for a diverse collection of related documents in an efficient and effective manner using computational methods. For many practical NLP implementations, the automatic segmentation of text according to the topics covered is crucial (Alexander Clark, 2013). Potential applications of AROG include automatic curriculum sequencing, better search result ranking, adaptive reading order generation, among others.

Automatic Reading-Order Generation is a nascent discipline. In (Md. Manzoor Murshed, 2024) we tried to sequence documents automatically by calculating the information content using Entropy and by using different distance metrics like Cosine Distance, Jaccard Distance, Manhattan Distance, and Euclidean Distance. Our attempt at automatic sequencing was successful, however absence of having standardized evaluation datasets for this task makes the emergence of new algorithms and techniques challenging and difficult. To mitigate this gap, we introduce the dataset AROGD, the first benchmark dataset specifically curated for automatic reading-order generation. The introduction of new datasets and of new algorithms leads to new baselines. That raises some questions about how those baselines are established, about the fairness in the comparison between the newly introduced algorithm and the ones previously known (Ismaïla Seck, 2018). The dataset contains 2,128 documents, and 6,849,492 words, all available in different public domains in different formats. It consists of 7 educational textbooks, 41 novels, 7 online lecture materials on different subjects, and 6 dissertations. Each document is represented by its textual content only. All the metadata, like title, author, preface, index, pictures, graphs, index etc., are removed because that information is not related to the sequencing task. The goal of the AROG research is to find and implement an effective algorithm to automatically sequence a collection of related documents. To prove the effectiveness of the model we can apply the algorithm to a set of book chapters, dissertations, popular course materials, where we assume authors organize and sequence those chapters manually in a correct order, and our goal is to find that order mathematically, using only the text content of the documents. We assume that the author's generated sequences are the gold standard or correct sequence of the expected sequence. The opposite sequences of all the correct sequences are opposite sequences. The sequencing model will generate a sequence using an algorithm which we can termed as generated sequences, and in our dataset we will randomly generated sequences for all the data sets which we termed random sequences. ARGOD dataset contains all the text data of the documents, the correct sequences, random sequences, and opposite sequences of all the documents. To validate the performance of any automatically generated sequence, we can conduct comprehensive statistical and distance analyses across the four sequence types. Specifically, we can calculate KL Divergence, Jensen–Shannon Divergence, Pearson correlations, Normalized and weighted Hamming Distance (NHD), Manhattan Distance (NMD), Root Mean Square Error (RMSE), and Levenstein Distance (NLD) for the four sequence variants of all the test dataset. Using all four provides greater analytical insight than simply using a single metric for validation. Additionally, we can perform paired t-tests on the results to validate the model justification. These analyses can unequivocally establish whether generated sequences are more closely aligned with the correct sequences or not. And the

generated sequences should significantly outperform both random and opposite sequences for validity.

AROG is the task of taking an unordered set of related documents $D = \{d_1, d_2, \ldots, d_n\}$ and predict a permutation $S = \{d_{p1}, d_{p2}, \ldots, d_{pn}\}$ that are based on some utility function defined by prerequisite dependencies, logical flow, and chronological narrative among the documents text only. Currently we are considering the Entropy of the document, Cosine similarity, Jaccard similarity, Manhattan Distance, and Euclidean Distance as the utility function. Effective document sequencing requires understanding of macro dependencies among them, like information content of the documents, their semantic meaning, structural similarity, topics covered, logical relationship and overlaps among them (Sowmya Vajjala 2020). The dataset consists of the following contents.

Table 1: Dataset Description

Attribute	Description
# f dwmj r u|	#
# f r qwqw	1
	1
vht xhqf h	
# vht xhqf h	# #
# vht xhqf h	# #
vht xhqf h	# # 1

This paper makes the following contributions:

1. **Automatic Reading Order Generation Algorithm:** We presented different metrics suitable for automatic reading order generation and documents sequencing using the textual content of the document.
2. **The AROGD Dataset:** We release the first benchmark dataset containing 2,128 documents collections from public domain textbooks, novels, lecture materials, and dissertation, with ground-truth ordering annotated by authors of the text, as well as the random ordering and opposite ordering as a baseline. An effective reading order generation algorithm should generate a reading order better than the baseline and close to the ground-truth ordering.
3. **Standardized Documentation:** We provide a complete and comprehensive datasheets to ensure ethical and transparent usage.
4. **Algorithmic Benchmarks:** We used our dataset for a known sequencing algorithm, and the result indicates that our dataset can be used for Automatic Reading Order Generation and similar research.

Related work

Related work includes the research article, "Automatic Reading Order Sequencing: A Novel Reading Order Generator for Text-Based Documents" (Md. Manzoor Murshed, 2024): In this paper Murshed et al., proposed a novel framework for automatic reading order generation. The authors represent documents with term-frequency and TF–IDF vectors, extract topic distributions via k-means clustering, and compute document-level information scores with Shannon entropy; they then quantify pairwise relationships using cosine similarity, Jaccard overlap, and Euclidean distance. They developed a multi-metric system that identifies the optimal starting document via Entropy Relevance and determines the subsequent narrative flow using a priority of high overlap, similarity and low distance among the documents together with various tie-breaking rules. Using the metrics Euclidean, Manhattan distances and rank correlations (Spearman, Kendall) they compare their generated sequences with random, and author recommended gold orders. Experimental results showed that the algorithm produces sequences that are meaningfully closer to author order than random. However, they didn't publish their dataset, we applied their algorithm to our AROGD dataset and presented the benchmark result in this paper.

An earlier research article on reading order was "Generating Reading Orders Over Document Collections" (Georgia Koutrika, 2015): This paper presented a technique to generate a reading order for a set of documents using a tree structure where the nodes contain a set of equivalent documents and edges represent a reading order from general to more specific documents. They quantitatively define document generality and document overlap concept. By using these metrics, they presented an algorithm to read general documents to more focused documents by using only the document contents that require no external knowledge about the documents. The generated reading tree captures document specificity relationships and can generate multiple reading orders of the documents.

Earlier, in "On Statistical Sequencing of Document Collections" (Thinniyam, 2014), Thinniyam proposed two techniques for distance-based methods for sequencing. He found that Latent Semantic Analysis (LSA) based distance between documents is more effective than the Tableaux-based distancing method that uses similarity and distance measurements to find sequence elements that are similar.

Finally, in "A data mining approach to reading order detection" (Ceci, 2007), the reading order problem was investigated by implementing to data mining techniques that can generate the required knowledge from a set of training layout structures by using a naive Bayesian classification framework. This paper also provides solutions for reconstructing the chain over a block of layout.

Dataset Statistics

The dataset contains a collection of textbooks, novels, dissertations and lecture notes.

Table 2: Dataset Statistics of AROGD

Statistic	Value
v	
h	<
	6
	9
	:
h	
q	5
k	3
v	

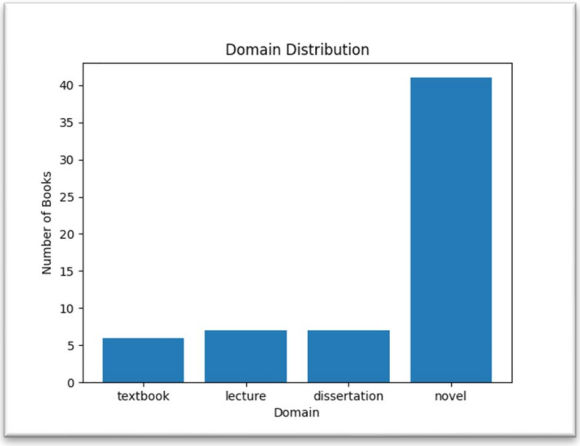


Figure 1. *Domain Distribution.*

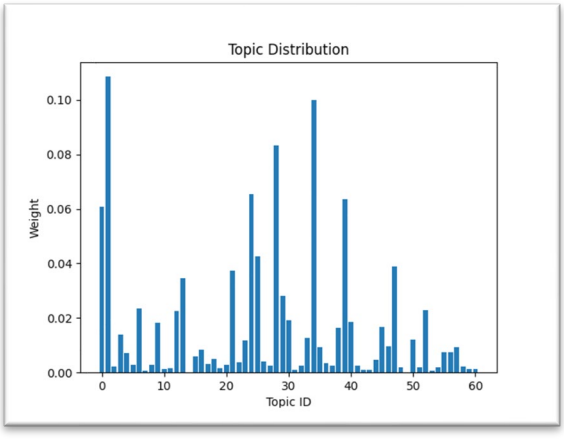


Figure 2. *Topic Distribution*

Datasheet for the AROGD Dataset

Following the framework proposed by Gebru et at. in "Datasheets for Datasets" (Timnit Gebru, 2021) we provide the datasheet for our dataset. This datasheet will promore transarent, responsible, upgradeble, and reproducible use of our dataset. This is an essential documentation artifacts explaning the motivation, composition, collection procedures, intended use, and potential risks associated with the AROGD. This standardized datasheet explains construction mechanisms, suitability for any downstream applications, known dangers, limitations and associate ethical concerns of the dataset for the researchers and developers prior to its use. As automatic sequencing of documents depends on many factors, a benchmark designed for evaluating different algorithms will be helpful. Dataset's structure, provenance, and intended use cases explained in the datasheet will promote fair and transparent evaluation of future algorithms on this topic.

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The AROGD was created to launch the first standardized, publicly available benchmark dataset for Automatic Reading-Order Generation (AROG) task. The goal of AROG is to find the most logical sequence of a collection of related documents automatically using only their textual content by using different techniques suitable for measuring their similarity, overlaps, topics, and presented information. There are no publicly available datasets for this type of task.

What (other) tasks could the dataset be used for?

The Dataset can be used for automatic curriculum generation, Curriculum sequencing and education content modeling, adaptive reading order creation, Search-result ranking based on semantic content, Evaluation of sequence-generation algorithms. It can also be used for training a neural network model.

Has the dataset been used for any tasks already? If so, where are the results so others can compare (e.g., links to published papers)?

The dataset was used for the benchmark algorithm of this paper and in the paper published at the symposium on electronic imaging (Md. Manzoor Murshed, 2024).

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The dataset was created by Md. Manzoor Murshed and Steven J. Simske. at Colorado State University, as part of Ph.D. research in automatic reading-order generation, document semantics analysis, and NLP sequence modeling.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

None.

Any other comments?

None.

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances are documents collected from textbooks, novels, dissertations, lecture materials. Only the text of the documents is collected. Other contents like images, indexes, prefaces, titles, author's information, table of contents are removed. An example instance is shown in figure 3 below.

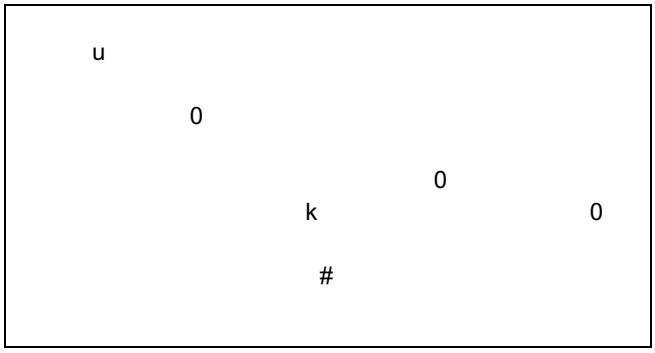


Figure 3. *An Example of text instance, taken from the file Introduction to Python Programming chapter2.txt.*

How many instances are there in total (of each type, if appropriate)?

There are 7 textbooks, 41 novels, 6 dissertations, and 7 lecture materials in the dataset. All together there are 2,128 documents consisting of total 6,849,492 words.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset is a sample of instances. It is intended to be a random sample of text documents. Due to the copywrite issue we did not use documents that has copywrites. We tried to use documents that are available free of charge in the public domain. In future to make the dataset representative, we will include documents from different domains, and subjects.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance consists of the text associated with the chapter of textbooks, dissertations, novels or one lecture material. Any graphics, images, diagrams, or tables associated with the document is removed. All the documents are a collection of words as a token only. Stop words are kept in the document, however during processing for algorithm those stop words and unnecessary punctuations are removed.

Is there a label or target associated with each instance? If so, please provide a description.

Documents associated with an instance are kept in a unique folder. All the documents inside the folders consist of unique names like chapter/lecture name or number, and the text contents are saved as a text file.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

All the required data for the algorithm is included in the dataset; there is no missing data.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

The author recommended chapters or lecture sequence are considered as gold standard and placed in the correct sequence folder, which consists of the text file named after the instance folder name. Similarly, the random and opposite folder contains the random sequence and opposite sequence of the documents. For simplicity random sequences are generated randomly and kept as static, because if we change this sequence during each run of the algorithm, the comparison statistics will not be consistent. Generated sequences from the algorithm are kept in a json file in the generated folder.

Are there recommended data splits (e.g. training, development, validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

There are no data splits, however, if anyone intends to use the dataset for training, validation and testing a Machine learning Model, they modify the dataset accordingly.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

No there are no known errors, sources of noise, or redundancies in the dataset.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is entirely self-contained, free to use and publicly available through GitHub repositories.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' nonpublic communications)? If so, please provide a description.

The whole datasets were collected electronically from the publicly available sources; we did not find any copywrite restriction to use the data for research during collection. Any special restrictions or confidentiality requirements are not known to the authors of the datasheet.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

Some novels might contain moderately inappropriate or offensive language, but we do not expect this to be the norm.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

No, the dataset does not identify any subpopulations by age or gender.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

We tried to remove all the personal information from the dataset. However, some personal information might be still available in the lecture notes materials.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

No, to the best of our knowledge, the dataset does not contain any sensitive information.

Any other comments?

No.

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

All data is provided as utf-8 encoded raw text. Each instance is stored in its own unique subfolder within a single main folder. All the text documents for each instance are kept as text files inside their respective subfolders. The correct, random, and opposite sequences are also stored as text files in separate folders that contain the name of all the text documents in their appropriate sequence. The text for the books was collected from Project Gutenberg (Hart, 2025) and Hugging Face (Hugging Face, 2025), the dissertations were collected from the Colorado State Library website, and the course materials were collected from various university professors' websites.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?

All the data is directly collected from the websites in electronic format and verified by the authors as much as possible during use. The data collection mechanisms of the websites are unknown to the authors of the datasheet.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

Unknown to the authors of the datasheet.

Who was involved in the data collection process (e.g., students, crowd workers, contractors) and how were they compensated (e.g., how much were crowd workers paid)?

Unknown to the authors of the datasheet.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please

describe the time frame in which the data associated with the instances was created.

Unknown to the authors of the datasheet.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

Unknown to the authors of the datasheet.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

As described above, the data was collected from different websites in electronic format.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

The data was downloaded from public web sources, and the owners of the websites were not explicitly informed that their data was to be used in this way.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

Not applicable for our dataset.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access points to the mechanism (if appropriate).

Not applicable for our dataset.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

Not applicable for our dataset.

Any other comments?

No.

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.

Yes, the chapters in the books and dissertations were separated into individual text files using a combination of software, programming, and manual methods as needed.

Was the "raw" data saved in addition to the preprocessed, cleaned, labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

The dataset contains all the raw text data,

Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.

Yes, all the utility programs are given with the dataset.

Any other comments?

No.

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

Yes, the dataset was used for the benchmark algorithm of this paper and in the paper published at the symposium on electronic imaging (Md. Manzoor Murshed, 2024).

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

Currently there is no such repository available.

What (other) tasks could the dataset be used for?

The Dataset can be used for automatic curriculum generation, Curriculum sequencing and education content modeling, adaptive reading order creation, Search-result ranking based on semantic content, Evaluation of sequence-generation algorithms. It can also be used for training a neural network model.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other risks or harm (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harm?

There is no known risk or harm to using the data as they were already publicly available.

Are there tasks for which the dataset should not be used? If so, please provide a description.

There are no known tasks for which the dataset should not be used.

Any other comments?

None.

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

Yes, the dataset is publicly available on the internet in GitHub repositories.

How will the dataset be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

The dataset will be distributed on GitHub repositories and does not have a DOI.

When will the dataset be distributed?

The dataset will be released with this paper.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The crawled data copyright belongs to the original authors of the data unless otherwise stated. There are no licenses and terms of use for the dataset, but there is a request to cite this paper and the corresponding paper if the dataset is used.

Have any third parties, imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

The authors are unaware of any such restrictions at the time of writing.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

The authors of this datasheet are unaware of that information.

Any other comments?

None.

Maintenance

Who will be supporting/hosting/maintaining the dataset?

Authors of this paper will maintain the dataset when necessary.

Is there an erratum? If so, please provide a link or other access point.

There are no such links or other access points at this time, if necessary, anyone can contact the corresponding authors of this paper.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (e.g., mailing list, GitHub)?

All future updates and modifications to this dataset will be published on its webpage.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

This is not applicable for this dataset.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.

If we update the dataset, the older versions will be kept around for consistency.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.

Modifications by others are allowed and encouraged when necessary, however, please contact the original authors to propose incorporating fixes and extensions.

Any other comments?

None.

Implementation of the benchmark algorithm using the dataset

In this section we used our dataset for the sequencing algorithm of the paper “Automatic Reading Order Sequencing: A Novel Reading Order Generator for Text-based Documents” (Md. Manzoor Murshed, 2024). We will comparatively analyze the model's generated output against our dataset’s correct sequences, as well as against the random sequences produced through random

permutation of the correct sequences, and opposite sequences by inverting the correct sequences or gold standard. The performance of the generated sequences compared to random and inverted baselines would provide strong evidence that the algorithm effectively approximates, or potentially improves the author recommended sequences. Our key metrics for sequence comparison are Normalized Manhattan Distance (NMD), Weighted Normalized Manhattan (WNMD), Normalized Hamming Distance (NHD), Weighted Normalized Hamming (WNHD), Normalized Levenstein Distance (NLD), Weighted Normalized Levenstein Distance (WNLD), Normalized Root Mean Square Error (RMSE), Weighted Normalized Root Mean Square (WNRD), and Pearson Correlation (PC). The following table summarizes the metric, showing that the generated sequences significantly outperform both random and opposite sequences.

Table 3: Summary of the Algorithm Output.

Metric Family	Metric Type	Generated	Random	Opposite
P dqkdwwdq	QP G	0.6435		4
	Z QP G	0.6564		4
Kdp p lqj	QKG	0.9014		<
	Z QKG	0.9029		<
Chyh0 qvwlq	QCG	0.1323		
	Z QCG	0.9029		<
# # Vt xduh	QUG	0.426		
	Z QUG	0.6905		4
Fr0 uhodwr q	Vshdu Ukr	0.0246	0	04

We performed t-test between the sequences, the results are given below

Table 4: t-test result between the sequences.

W0			
		0	3
d	qgr p =#w@	0	3
	@	0	3
	#		

Udqgr p @		0	3	
				0 #
Rssr vlwh@	3			
	@			0 #
Rssr vlwh@	3			

W0	#		
v0			
S0	0		
	1		
		#	
1			
W0	#		
W0 7			
S0	0		
	1		
		#	
1			

From the t-test results we can find that the generated sequences are statistically different from the correct sequences and not identical to the author curated sequences. However, they are better than the random and opposite sequences. The normalized results in Table 3 show that the generated sequence performs substantially closer to perfection than random guessing across all the metrics. Specifically, Manhattan-based measures (NMD and WNMD) indicate about **79.9%** progress toward the ideal sequence, while Hamming distances (NHD and WNHD) are slightly higher at approximately **81.5–81.6%**. The Levenstein distance demonstrates particularly strong performance, with NLD reaching **94.6%** and WNLD about **81.6%**, suggesting a high degree of structural similarity to the correct sequence. Similarly, Root Mean Square metrics show solid alignment, with NRD at **79.0%** and WNRD achieving **94.1%**, reinforcing the consistency of the generated output. In contrast, the Spearman Rho correlation is considerably lower at **30.5%**, indicating that while the absolute differences between sequences are small, the relative ordering of elements is less well preserved. Overall, these percentages suggest that the generated sequence is markedly closer to the perfect sequence than to random or opposite baselines, with especially strong performance in distance-based evaluations. In the future we need to find a better algorithm that will make the generated sequences closer to the ground truth. From the Kendall's Tau and Spearman's Rho data, we can find that the generated sequences have slightly positive correlations, which indicates a small but real structural similarity to the correct sequences. However, the negative values for random and opposite sequences indicate negative correlations and weak structural similarity to the correct sequences. From the pained t-test data between generated and random sequences we find that the p-value is extremely small, hence the null hypothesis (no difference between "Generated" and "Random") is rejected. The positive t-statistics value suggests that the generated sequences are significantly better than the random sequences. The paired t-test result between generated and opposite sequences suggests rejection

of null. The extremely high t-statistics and miniscule p-value show a massive, statistically significant difference, which indicated that the generated sequences are significantly better than the opposite sequences. From the overall average metrices chart of figure below we can find that the generated sequences are better than the opposite and random sequences for all metrics.

From the figure above we find that the mean error metric is not useful as it shows no significant differences and contains high variance. However, the correlation metrics (Kendall, Spearman) and divergence metrics (KL, JS) identified that the generated sequences consistently outperform Random and Opposite sequences. The positive values of Kendall's Tau and spearman's Rho for generated sequences indicate a small positive correlation with the correct ordering. Negative value for random sequences indicates that they are negatively correlated and contradict the correct order sequences. Opposite sequences are perfectly negatively correlated because they are exact reverse of the correct sequence.

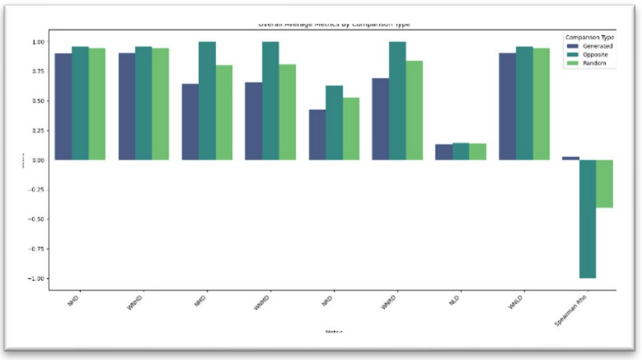


Figure 3. *Overall average metrics for all sequences.*

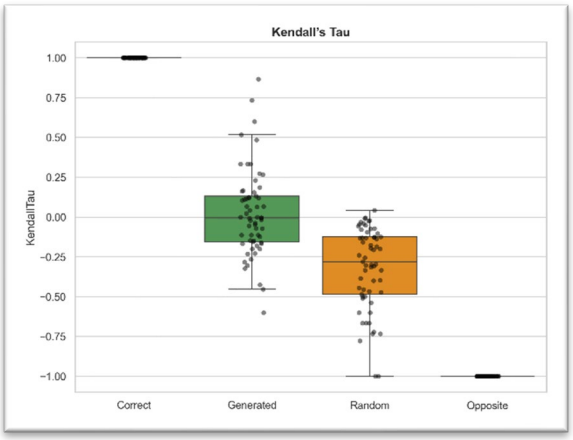


Figure 4. *Kendall's Tau output.*

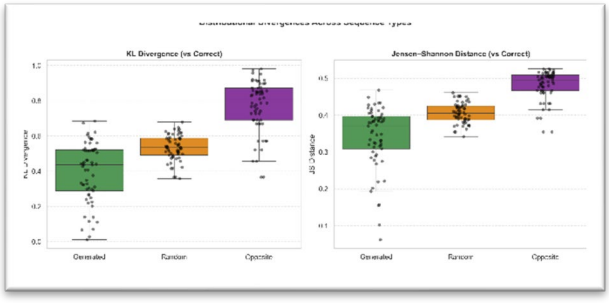


Figure 5. *Distributional Divergences Across Sequence Types.*

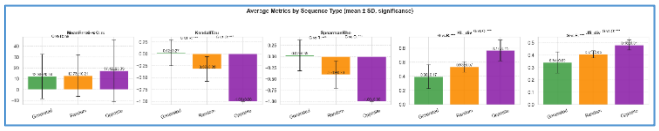


Figure 6. *Average Metrics by Sequence type.*

Random sequences are worse, and Opposite sequences are far from correct sequences and the worst. Generated sequences are significantly more similar to the correct sequences than both baselines.

Discussion and Limitations

In this research we presented a dataset suitable for Automatic Reading Order Generation or similar research. We used our dataset for a known Automatic Reading Order Generation algorithm and established a benchmark. Currently we are considering only the text data and assuming a single gold standard. In the future we can consider graphics and other relevant data inside the documents to better understand the content. For a set of related documents there might be more than one sequencing possible due to different needs and necessities. We can also incorporate large language model to include better semantic understanding of the documents. Because if we include semantic meanings, we can generate better similarity matrix suitable to getting better algorithm. Our dataset contains textbooks, novels, dissertations, and lecture materials and we applied the algorithm to the whole corpus. In future we can try to apply the algorithm separately to different domains because dissertations and lecture materials sequencing are more important than novels and textbooks. We can also extend our work for automatic curriculum generation by using lecture transcripts together with lecture materials.

Conclusion

In this research, we presented a dataset for automatic reading order generation using the text only features of a collection of related documents. We also implemented a benchmark algorithm and used our dataset and checked the quality of the generated sequences. Our work can predict document relationships like similarities, overlaps, and distances without using any external knowledge. Experimental

results established that our algorithm could sequence a collection of logically related text documents automatically. We used different metrics to judge the result of our algorithm. For all the metrics our generated sequences outperform random and opposite sequences and closer to the gold standard. We validated our finding by measuring the quality of the sequencing of textbooks, novels, dissertations, and course materials, where the original sequencing comes from the authors. As the proposed reading order is closer to the correct sequences than random and opposite sequences, we can conclude that the automated reading order sequencing algorithm is appropriate for any related text-based set of documents.

In future work we will use Large Language Model (LLM) to incorporate semantic relationships among the documents. In addition to the text, we can also use the graphics, diagrams, tables and other document contents to find relationships among documents. We want to extend our work for automatic curriculum generation by using the lecture transcripts in addition to the lecture presentation materials. We will also try to propose methods to personalize reading sequences for different levels and different users with background knowledge.

References

[1] Alexander Clark, Chris Fox, and Shalom Lappin. The Handbook of Computational Linguistics and Natural Language Processing. Vol. 53. Oxford, UK: Wiley-Blackwell, 2013.

[2] Ceci, M., Berardi, M., Porcelli, G., & Malerba, D. "A data mining approach to reading order detection." *International Conference on Document Analysis and Recognition*. IEEE, 2007.

[3] Georgia Koutrika, Lei Liu, Steve Simske. "Generating reading orders over document collections." *IEEE Xplore*. IEEE, 2015. https://ieeexplore.ieee.org/document/7113310.

[4] Hart, M. (2025, December 1). Project Gutenberg. Retrieved from https://www.gutenberg.org/

[5] huggingface. (2025, December 1). Retrieved from Hugging Face: https://huggingface.co/datasets

[6] Ismaïla Seck, Khouloud Dahmane, Pierre Duthon, Gaëlle Loosli. "Baselines and a datasheet for the Cerema AWP dataset." 2018. https://arxiv.org/abs/1806.04016.

[7] Md. Manzoor Murshed, Steven J. Simske, Arturo N. Villanueva. "Automatic Reading Order Sequencing: A Novel Reading Order Generator for Text-based Documents." *Electronic Imaging*. 2024. https://library.imaging.org/ei/articles/36/10/IPAS-252.

[8] Sowmya Vajjala, Bodhisattwa Majumder, Anuj Gupta, and Harshit Surana. *Practical Natural Language Processing, A Comprehensive Guide to Building Real-World NLP Systems*. Vol. 21. Sebastopol, CA: Oreilly, 2020.

[9] Steven Simske, Marie Vans. *Functional Applications of Text Analytics Systems*. Colorado State University, USA: River Publishers, 2021.

[10] Thinniyam, R. (2014). On Statistical Sequencing of Document Collections. University of Toronto. Ph.D. Thesis.

[11] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, Kate Crawford. "Datasheets for Datasets." *CACM* (2021), Retrieved from https://arxiv.org/abs/1803.09010

Author Biography

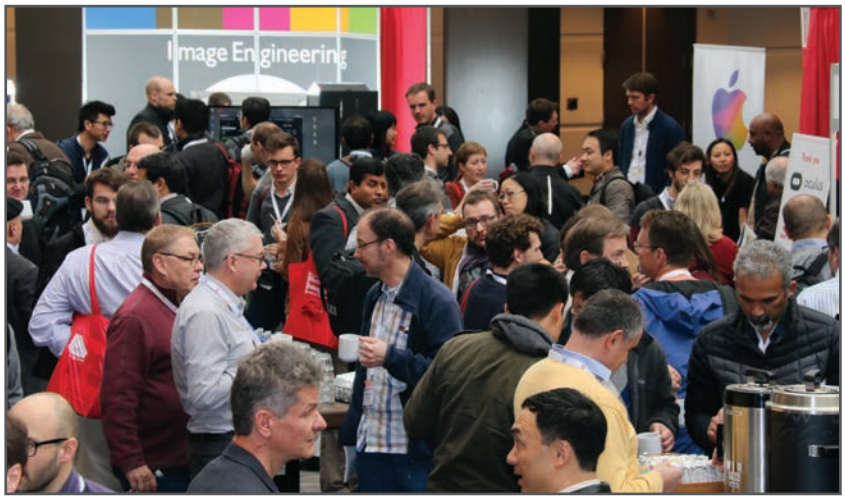
Md. Manzoor Murshed is a Ph.D. candidate (ABD) in Systems Engineering at Colorado State University, with an expected completion in Fall 2026. He currently serves as a full-time faculty member in the Computer Science and Security Department at Northeast Iowa Community College (NICC) and as an adjunct professor of Information Technology at Upper Iowa University (UIU). An enthusiastic and dedicated educator, Murshed previously served Upper Iowa University for over two decades. During this time, he was an Associate Professor in the School of Mathematics and Science, teaching courses in Information Systems, Information Technology, and Software Engineering. His extensive academic credentials include an M.S. in Computer Engineering and a Graduate Certificate in Information Assurance from Iowa State University, an MBA from Upper Iowa University, and an M.Sc. in Computer Science and a B.Sc. (Honors) in Applied Physics and Electronics from the University of Dhaka. His core research interests encompass machine learning, natural language processing, artificial intelligence, and computer privacy and security, with recent scholarly focus on automatic reading-order generation and text analytics.

Dr. Steve Simske is a Professor in Systems Engineering, and an affiliate of Biomedical and Mechanical Engineering, at Colorado State University (CSU). The author of 230 US patents and more than 450 publications, he is an IS&T, IEEE, and NAI Fellow. Steve is the Steering Committee Chair for the ACM DocEng Symposium, and a past member of the World Economic Forum Global Agenda Councils for Illicit Trade, Illicit Economy, and the Future of Electronics. At CSU, he is an FIIE (Faculty Institute for Inclusive Excellence) Fellow and a 2022 Best Teacher awardee. He has a cadre of on-campus students in his engineering departments, along with a larger contingent of online/remote graduate students researching in a wide variety of disciplines.

JOIN US AT THE NEXT EI!

electronic IMAGING

Imaging across applications . . . Where industry and academia meet!



• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

ADVANCING IMAGING SCIENCE 75 1947 IS&T 2022 imaging.org