Layout-Based Chunk Alignment: Utilizing Visual Information to Collect Parallel Texts From Image Documents

Masaki Kinouchi kinouchi.m.a7be@m.isct.ac.jp
Kayoko Nohara nohara.k.d485@m.isct.ac.jp
Xinru Zhu zhu.x.1137@m.isct.ac.jp
Yuma Miura tlfwiu078@gmail.com
School of Environment and Society, Institute of Science Tokyo, Tokyo, Japan

Abstract

This study proposes a layout-based chunk alignment method (Layout-CA) as an intermediate step between document- and sentence-level parallel text alignment for bilingual document images. Visually rich printed materials, such as institutional reports and magazines, often contain high-quality translations and are valuable sources of parallel data, yet their layout cues are underutilized. Layout-CA aligns semantically coherent text chunks across document pairs by integrating multi-modal cues from textual content and layout, and sentence alignment is then performed within the aligned chunk pairs. Experiments on English UNESCO reports and their Japanese translations show that introducing chunk alignment improves downstream sentence alignment for both Bleualign and Vecalign. When document order is disrupted, Layout-CA preserves alignment coverage by restricting sentence matching to corresponding chunks, enabling robust alignment in multilingual image documents.

1 Introduction

The construction of high-quality parallel corpora has long been a foundational topic in machine translation. Recent advances in corpus construction, driven by the vast availability of web documents, have enabled the collection of large-scale parallel data (Bañón et al., 2020; Morishita et al., 2020; Koehn, 2024), but ensuring their quality still remains a challenge. A key component in building parallel corpora is accurate sentence alignment, which is not only essential for the development of robust machine translation systems in general, but also plays a critical role in other language research fields, such as corpus linguistics and translation studies (Baker, 1993, 1996; Baorong, 2021). In these fields, it is often required to extract professional translations from printed documents to guarantee the quality of product in target. Although many alignment methods, especially at sentence level, have already been proposed, ranging from statistical (Gale and Church, 1993; Varga et al., 2007) or machine-translation-based approaches (Sennrich and Volk, 2011) to embedding-based approaches (Thompson and Koehn, 2019; Liu and Zhu, 2022), there is less attention to the complexities inherent in multilingual digitized documents or PDFs.

In practice, many professionally translated materials, such as institutional reports, magazines, etc., are published in visually rich layouts. These documents are often of high translation quality, making them valuable sources for parallel data. They contain rich visual information, such as complex layouts, font size, color and elements such as tables, figures, and captions that can indicate structural boundaries and correspondences. Conventional methods, which treat documents as simple sequences of text, leave room for further development to handle such structurally complex data when deal-

ing with documents as images.

To better capture the alignment structure of these documents, we draw inspiration from how humans approach translation alignment. When attempting to locate the translation of a particular sentence, one does not search the entire document linearly. Instead, one first identifies the corresponding article or section (based on headings, layout, or other visual cues), narrows down to a relevant paragraph, and finally locates the translated sentence. This top-down search reflects a hierarchical alignment process moving from document-level structure to semantically coherent chunks, down to sentences.

Motivated by this observation, we argue for an intermediate alignment step that operates at the level of chunk. Following the context of previous research (Zhao et al., 2025; Duarte et al., 2024), a *chunk* in this research is defined as a region for coherent texts in which the reader senses a persistent narrative theme while reading a document. These are semantically meaningful textual units such as articles, sections, or grouped paragraphs. These chunks can often be visually identified by layout features, such as headers indicated by font size or bounding boxes. In this paper, we propose a chunk-level text alignment framework for paired bilingual document images. By integrating layout-aware chunking into the pipeline, this work presents one of the first attempts to systematically incorporate multi-modal information into bilingual document alignment. Starting from document images, the framework applies OCR and document layout analysis to recover text content and structural information. It then uses them to align semantically corresponding chunks across document pairs before sentence-level alignment.

2 Related Work

2.1 Sentence Alignment

Sentence-level bilingual alignment research has evolved from simple statistical techniques to sophisticated neural methods, while leaving difficulties in extracting sentence pairs from raw image documents with text-based aligners. Early approaches like Gale-Church's algorithm (Gale and Church, 1993) align sentences by maximizing length-based probability models via dynamic programming. Later algorithms combined sentence-length models with bilingual word correspondences to improve alignment accuracy (Moore, 2002). Hybrid systems integrated multiple signals; for example, Hunalign (Varga et al., 2007) uses dictionaries plus length heuristics, and Bleualign (Sennrich and Volk, 2011) leverages rough machine translations with BLEU-based semantic similarity scoring to anchor matching sentences. In recent years, neural and embedding-based methods recast alignment as a bilingual retrieval problem: sentences are embedded into a shared multilingual vector space, and parallel pairs are identified by high semantic similarity (Thompson and Koehn, 2019; Liu and Zhu, 2022). These advances enable mining parallel sentences from documents for corpus construction. However, purely text-based aligners struggle with extracting from raw image documents as texts must be recovered from visually structured pages. Earlier work on parallel Web pages showed that HTML structure and text-chunk lengths can support translation detection and preliminary alignment (Resnik and Smith, 2003). This suggests that visual structure may likewise provide a valuable alignment signal in image documents, motivating chunk-aware approaches that exploit such information.

2.2 Document Layout Analysis

Document layout analysis (DLA) has been advanced by developments of CNN-based region detectors (Hao et al., 2016; Schreiber et al., 2017; Yang et al., 2017; Soto and Yoo, 2019). Recent work such as LayoutLM (Xu et al., 2020) introduces an architecture that integrates text, 2D layout coordinates, and visual features into multi-modal embedding, enabling strong performance on layout-aware understanding tasks. Building on this line, models such as LayoutReader (Wang et al., 2021) frame layout analysis as a reading-order recovery problem, explicitly modeling document structure through sequence prediction over detected regions.

Progress in DLA has been driven by large-scale annotated datasets. PubLayNet (Zhong et al., 2019) provides hundreds of thousands of document pages with automatically derived layout annotations, facilitating robust pretraining for region detection. DocLayNet (Pfitzmann et al., 2022) complements this with high-quality human annotations across diverse document types and finer-grained classes, improving generalization to real-world layouts. In addition, datasets such as ReadingBank (Wang et al.,

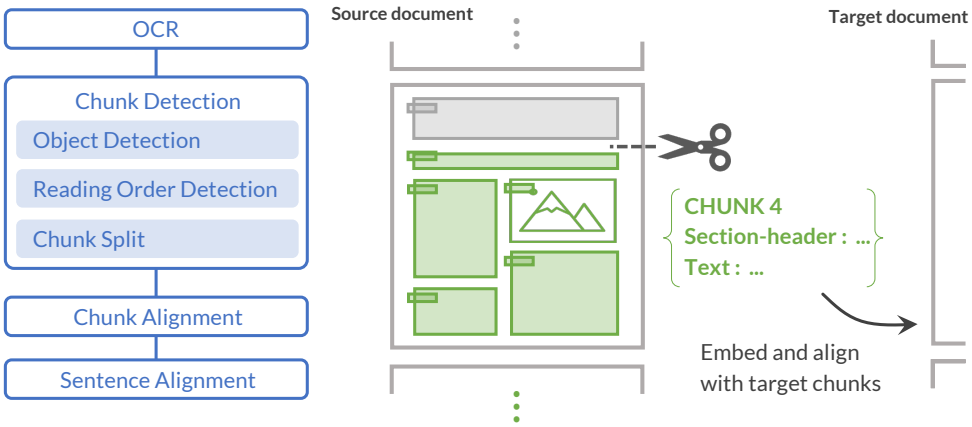

Figure 1: A diagram of layout-based chunk alignment method (Layout-CA).

2021) introduce explicit reading-order annotations, supporting the training and evaluation of models that reason over document flow rather than isolated regions.

3 Method

In this study, we propose layout-based chunk alignment method (Layout-CA), which consists of two principal stages (Figure 1). In the first stage, chunk detection, we transform raw document images into a linearized sequence of semantically coherent units, or chunks. This involves combining object detection with OCR, recovering reading order in complex multi-column layouts, and splitting the resulting sequence into segments that reflect the underlying discourse structure. In the second stage, chunk alignment, we establish bilingual correspondences between source and target chunks by computing similarity scores. The chunk layer functions as an intermediate alignment step that reduces the search space for sentence alignment while anchoring correspondences in document structure.

3.1 Chunk Detection

The task of chunk detection is to recover units of text that mirror the way readers and editors naturally perceive documents. Simply applying OCR to a page yields a collection of tokens or lines without structural coherence; on the other hand, treating detected layout regions as fixed units ignores the logical flow of reading. Our approach resolves this by integrating three sub-steps: OCR + object detection, reading order detection, and chunk splitting.

3.1.1 OCR + Object Detection

We first apply OCR and object detection page by page to capture the coordinates and labels of text boxes and textual information inside. We adopt labels defined in DocLayNet (Pfitzmann et al., 2022) and label each box using 11 categories: *Title, Section-header, Text, List-item, Formula, Picture, Table, Caption, Footnote, Page-header, and Page-footer*. These reflect common typographic and paratextual conventions of professionally edited publications. The OCR system outputs word-level tokens with bounding boxes, which are assigned to the detected layout regions based on maximum overlap. In this way, each region is enriched with its textual content, enabling joint reasoning over visual and linguistic signals. We also compute for each page a reference text height $\bar{h}_{Text}$ defined as the mean token height across *Text* segments. This statistic provides a robust scale for chunk split by identifying visually salient headers, since titles and section headers generally have larger font sizes than body text.

3.1.2 Reading Order Detection

We apply reading order detection to arrange text boxes linearly. Image documents often store multi-column layouts, sidebars, or figures, which break the

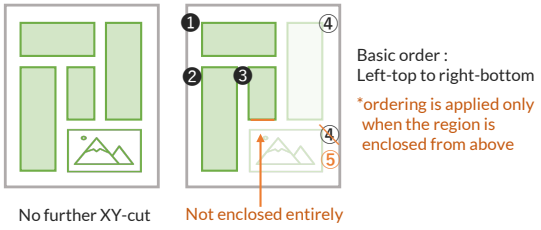

Figure 2: Reading order detection algorithm

simple top-to-bottom, left-to-right assumption. A key step, therefore, is to recover a plausible reading order that approximates how human readers traverse the page. We apply an XY-cut (George and Seth, 1984; Ha et al., 1995), which recursively identifies horizontal and vertical gaps in the distribution of bounding boxes and partitions the page into regions that correspond to columns or blocks.

However, as Liu et al. (2025) points out, XY-cut has a limitation in case like Figure 2. A further refinement prevents erroneous jumps across columns. Specifically, when linking a segment a to a next candidate segment b below it, we require that the horizontal coverage of above segments is greater than the width of a. This constraint enforces local column continuity. Finally, we normalize the position of paratextual segments: *Page-header* and *Page footer* are relocated to the beginning and end of the sequence in a page respectively.

3.1.3 Chunk Split

Having obtained a reading order, we next group segments into larger units that we call chunks. The central observation is that documents are typically organized around visually salient headers, such as article titles or section headers, which serve as anchors for subsequent blocks of text. To formalize this, we define salience r as a relative height ratio.

$$r = \bar{h}/\bar{h}_{Text}$$

$\bar{h}$ is the average height of tokens in the segment and $\bar{h}_{Text}$ is the reference average height of tokens labeled *Text* in the page. A new chunk is initiated when a segment is labeled as *Title* or *Section-header* and has salience r than or equal to a threshold r_{th}. Once a head is detected, all subsequent segments in reading order are appended to the same chunk until the next head is encountered. In this way, the document is decomposed into a sequence of semantically coherent, layout-aware chunks that reflect its discourse organization.

3.2 Chunk Alignment

The task of chunk alignment aims to establish correspondences between chunks across source and target documents. Let c_i^s and c_j^t denote the sets of source- and target-language chunks in reading order. We calculate a similarity score that integrates semantic and layout features to identify pairs of chunks between the languages.

Each chunk is embedded using a multilingual sentence encoder. In our implementation, we adopt multilingual Sentence-BERT[1] (Reimers and Gurevych, 2019), which provides language-agnostic embeddings in a shared multilingual semantic space. We embed the texts label by label in each chunk with mean pooling. Then we calculate cosine similarity s_l between the same-label texts in two chunks and normalize it with sigmoid function. Empty fields are mapped to the zero vector.

For each candidate chunk pair (i, j), we calculate a simple score function with s_l, label weight w_l and bias b:

$$S(i,j) = \sum_l w_l \cdot \sigma(s_l) + b$$

We trained a set of weights with a logistic regression model on annotated development dataset. The final alignment retains only pairs whose scores exceed the threshold. A source chunk is allowed to align with multiple target chunks (1:n), and the same target chunk can be shared across alignments with different source chunks.

3.3 From Chunks to Sentences

Layout-CA is integrated with existing sentence alignment methods by performing sentence alignment independently within each aligned chunk pair. The ultimate objective of our pipeline is to obtain reliable sentence-level bilingual alignments from raw image documents. The chunk layer contributes to this by constraining the alignment search space: rather than attempting to align sentences across entire documents, it performs within the narrower windows defined by aligned chunks.

[1] https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2

ID	Type	Pages	Chunks	Sentences	OCR
UNESCO (en)	5 documents	50	18	1425	Tesseract
UNESCO (ja)	5 documents, same order as (en)	51	18	1463	Yomitoku
UNESCO (shuffled, ja)	5 documents, randomly shuffled	51	18	1463	Yomitoku

Table 1: Evaluation dataset

4 Experiments

We conducted experiments on English UNESCO reports and their Japanese translations to evaluate chunk detection, chunk alignment, and their impact on improving sentence alignment accuracy. For OCR, we use Tesseract OCR[2] for English and Yomitoku[3] for Japanese. For object detection, we use YOLOv12 (Tian et al., 2025) pretrained on DocLayNet, while for Japanese, we further fine-tuned a model using the Japanese documents in the data. As a baseline for sentence alignment, we use the widely used Bleualign (Sennrich and Volk, 2011) with NLLB[4] (Team et al., 2022) as the MT system and Vecalign (Thompson and Koehn, 2019) with Sentence-BERT as a multilingual embedder. The threshold of the chunk similarity score was empirically set to 0.3 based on preliminary experiments.

4.1 Evaluation Dataset

We evaluate on documents collected from UNESCO Digital Library[5]. For each document, we obtained the corresponding source and target versions from publicly accessible materials (licensed under CC-BY-SA 3.0 IGO[6]). These documents are already aligned at document level using the metadata. We sampled five source documents and extracted ten pages per document, each beginning at a chapter header, resulting in a total of 50 pages in English. On this corpus, we created human gold annotations at both the chunk level and sentence level. Two annotators marked boundaries at the chapter or section granularity, and aligned corresponding source-target sentences. Source and target sentences can exhibit many-to-many correspondences. These annotations are used to evaluate chunk detection, chunk alignment, and downstream sentence alignment. Development dataset was also constructed following the same manner. (Appendix A)

To assess robustness to non-linear alignments, such as cases where the document structure in the target language differs from that of the source, we also conducted experiments on a shuffled dataset. Such non-linear correspondences can arise when editors deliberately restructure translated editions, especially in abridged translations, by omitting, combining, or reordering sections to fit a shorter format. They may also result from page-layout changes for right-to-left scripts or from explanatory notes added to the translation. This dataset was constructed by randomly sampling chunks from the original target Japanese corpus and permuting their order (Table 1).

All code and data sufficient to reproduce the experiments are released in our repository[7]. Due to licensing constraints on embedded images, however, we hide the pictures, logos, illustrations in the documents for this experiment.

4.2 Metrics

We evaluate our approach along three axes that mirror the pipeline: (i) whether chunks are correctly segmented within each monolingual document; (ii) whether source-language chunks are aligned to the expected target-language chunks; and (iii) whether these intermediate steps improve sentence-level alignment. We evaluate chunk detection from two perspectives: boundary accuracy and content consistency. Following metrics are used.

Boundary Accuracy. We compare predicted chunk boundary positions against human gold boundaries. For each monolingual document, we predict chunk boundaries by identifying the initial sentence of each chunk. A predicted boundary is

[2] https://github.com/tesseract-ocr/tesseract
[3] https://github.com/kotaro-kinoshita/yomitoku
[4] https://huggingface.co/facebook/nllb-200-distilled-600M
[5] https://unesdoc.unesco.org/
[6] https://creativecommons.org/licenses/by-sa/3.0/igo/deed.en
[7] https://github.com/mkino17/layout-ca

Model	EN: DocLayNet		JA: DocLayNet+	
	mAP50	50:90	mAP50	50:90
YOLOv11s	0.928	0.755	0.877	0.705
YOLOv12s	0.928	0.765	0.910	0.762
YOLOv12m	0.935	0.773	0.925	0.731
Detailed performance of YOLOv12s				
Title	0.895	0.785	0.995	0.839
Section-header	0.963	0.653	0.908	0.743
Text	0.967	0.838	0.943	0.860
List-item	0.946	0.818	0.955	0.853
Formula	0.836	0.661	0.954	0.797
Picture	0.939	0.867	0.851	0.746
Table	0.932	0.879	0.976	0.925
Caption	0.956	0.878	0.816	0.732
Footnote	0.853	0.647	0.745	0.449
Page-header	0.959	0.699	0.904	0.727
Page-footer	0.966	0.688	0.967	0.710

Table 2: Model comparison – object detection performance across train datasets.

counted as a true positive if it exactly matches a gold boundary position; otherwise it is counted as a false positive, and any gold boundary that is not predicted is counted as a false negative. Based on these counts, we compute precision, recall, and F1.

Average Chunk-level BLEU. Boundary-based metrics alone do not capture whether the content of a predicted chunk matches that of the gold chunk. We additionally evaluate content consistency using Average Chunk-level BLEU, adapted from BLEU-based evaluations (Papineni et al., 2002) used for page-level reading order (Wang et al., 2021). For each gold chunk c in chunks C of the document, we identify the predicted chunk $\hat{c}$ that contains the first sentence (boundary text) of c.

$$\textit{Avg. Chunk-level BLEU} = \frac{1}{|C|} \sum_{c \in C} \mathrm{BLEU}(c, \hat{c})$$

For chunk alignment, we anchor evaluation on the gold chunk structure and assess whether the system retrieves the correct target chunk(s) for each gold source chunk. Concretely, for each gold aligned pair $(c^{(s)}, c^{(t)})$, we map the gold source chunk $c^{(s)}$ to its corresponding predicted source chunk $\hat{c}^{(s)}$ with the same rule as chunk detection.

Label	w		
Title	-0.0002	Table	0.3809
Section-header	2.2551	Caption	-0.0002
Text	1.9326	Footnote	0.0657
List-item	0.5171	Page-header	0.5099
Formula	-0.0002	Page-footer	1.1570
Picture	-0.0002	b	-5.4418

Table 3: Weights for chunk similarity score function.

We then take the predicted matches for $\hat{c}^{(s)}$ and obtain the predicted target span $\hat{c}^{(t)}$. If the system produces a one-to-many match, we treat it as a target span composed of multiple consecutive predicted chunks and concatenate their texts. We evaluate whether $\hat{c}^{(t)}$ corresponds to the gold target chunk $c^{(t)}$ using the same two metrics as above: boundary accuracy and average chunk-level BLEU. As for boundary accuracy of chunk alignment, a one-to-many match contributes multiple boundary sentences.

Finally, for sentence alignment, we measure the gain attributable to the chunk layer by comparing a baseline sentence aligner with and without chunk-based preprocessing. We report sentence-level precision, recall, and F1 under both settings and highlight the relative improvement. All metrics are aggregated as document-level means, with sentence matches evaluated under a strict exact-match criterion.

5 Results and Discussions

5.1 Model Parameters

Table 2 and Table 3 present the object detection performance and the learned weights for chunk alignment, respectively. As shown in Table 3, some labels contribute little to alignment performance (largely because they do not appear in the development data), whereas labels such as *Section-header* and *Text* play a central role in aligning source and target chunks.

5.2 Chunk Detection and Chunk Alignment

Figure 3 illustrates how the number of split chunks in English and Japanese varies as a function of the salient ratio r. Although the optimal value of r depends on document type, the UNESCO dataset exhibits a clear tendency around $r \approx 1.3$. At this value, the chunk boundaries produced by our

Evaluation dataset	Number of chunks (pred/gold)	Boundary Accuracy			BLEU
		P	R	F1	
UNESCO (en)	23/18	0.739	0.944	0.829	0.755
UNESCO (ja)	18/18	0.889	0.889	0.889	0.201
UNESCO (shuffled, ja)	18/18	0.889	0.889	0.889	0.201

Table 4: Evaluation results for chunk detection (calculated with $r = 1.3$, chunk similarity score thresh. 0.3).

Evaluation dataset	Chunk pairs (pred/gold)	Boundary Accuracy			Avg. Chunk-level BLEU
		P	R	F1	
UNESCO	15/18	0.737	0.778	0.757	0.184
UNESCO (shuffled)	14/18	0.778	0.778	0.778	0.167

Table 5: Evaluation results for chunk alignment (calculated with $r = 1.3$, chunk similarity score threshold 0.3).

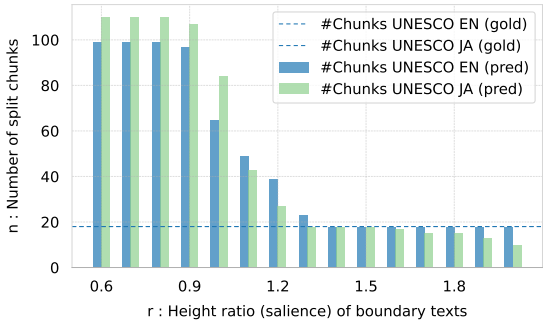

Figure 3: Relationship between salience r and chunk counts. Predicted chunks for English (blue), Japanese (green), and gold counts (horizontal lines).

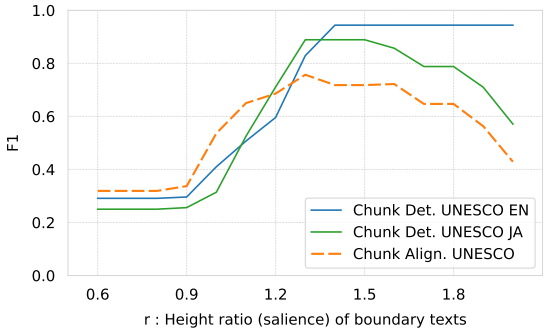

Figure 4: Relationship between salience r and F1 score. Chunk detection English (blue), Japanese (green), and chunk alignment (dashed orange).

method closely approximate the gold-standard segmentation based on chapters and sections.

Figure 4 shows the variation in F1 scores for chunk detection and chunk alignment across different values of r. The horizontal axes of Figures 3 and 4 are scaled identically to facilitate direct comparison. As these figures indicate, under-splitting conditions ($r < 1.0$) result in low precision, while excessive over-splitting ($r > 1.8$) also leads to a consistent decline in F1 scores. These results demonstrate that chunking at an appropriate granularity, corresponding to semantically coherent units, is crucial for achieving robust performance in both chunk detection and alignment.

Tables 4 and 5 present the results of chunk detection and chunk alignment, respectively, at $r = 1.3$, comparing performance across different datasets. Comparable chunk alignment scores and average chunk-level BLEU are observed between shuffled and non-shuffled documents, indicating that the proposed method maintains robust alignment performance even when the source and target documents differ substantially in document structure.

An important observation is that a simple layout-based salience measure can approximate discourse-level structure to a meaningful extent. In the UNESCO dataset, setting the salient ratio to $r \approx 1.3$ yields chunks that closely align with chapter- or section-based gold standards, with headers whose height is approximately 1.3 times that of the average body text treated as chunk boundaries. Under-splitting ($r > 1.8$) merges semantically heterogeneous content into a single unit, resulting in am-

Model	UNESCO				UNESCO (shuffled)			
	Sent. pairs (pred/gold)	P	R	F1	Sent. pairs	P	R	F1
VecAlign	976/1409	0.049	0.034	0.040	867/1409	0.030	0.019	0.023
VecAlign (w/ Layout-CA)	1197/1409	0.311	0.263	0.285	1197/1409	0.311	0.263	0.285
BleuAlign	913/1409	0.057	0.037	0.045	629/1409	0.079	0.036	0.049
BleuAlign (w/ Layout-CA)	1155/1409	0.355	0.290	0.319	1155/1409	0.355	0.290	0.319

Table 6: Sentence alignment performance with and without chunk alignment (calculated with $r = 1.3$, chunk similarity score threshold 0.3).

biguous alignment candidates and low precision. In contrast, excessive over-splitting ($r < 1.0$) fragments coherent discourse units and weakens contextual signals, which also leads to a degradation in F1 scores. Although this approach does not explicitly model discourse structure, it suggests that visual prominence in document layout can serve as a weak but effective proxy for semantic organization, particularly in heterogeneous or noisy document settings.

At the same time, the optimal value of r should not be considered universal. The sensitivity of both chunk detection and alignment performance to over- and under-splitting implies that the appropriate granularity is likely to vary across document types, layout conventions, and scripts. Treating the salient ratio as a tunable structural parameter, rather than a fixed heuristic, therefore offers a flexible design choice that can adapt to different datasets.

5.3 Improvement of Sentence Alignment

Table 6 shows sentence alignment performance with and without chunk alignment calculated with $r = 1.3$. The absolute scores are relatively small, reflecting the use of strict matching criteria and the effects of upstream processes, but sentence alignment performance consistently improves when Layout-CA is applied prior to alignment for both Bleualign and Vecalign. This indicates that structurally informed chunk alignment provides a more reliable context for downstream sentence-level alignment.

Without Layout-CA, shuffling of target document substantially reduces the number of aligned sentence pairs, indicating a breakdown of correspondence under disrupted document order. In contrast, when Layout-CA is applied, a similar number of sentence alignments is recovered in both shuffled and non-shuffled settings. This effect is observed for both Vecalign and Bleualign, which implicitly assume a linear, order-preserving correspondence between source and target texts and are therefore more susceptible to ordering mismatches. In such settings, directly aligning sentences without prior structural normalization tends to be less stable, whereas chunk-level alignment effectively constrains the search space and restores meaningful correspondence between semantically related chunks.

6 Conclusion

This paper proposed a layout-aware approach to document-level alignment that integrates chunk detection, chunk alignment, and sentence alignment. Overall, the results indicate that the proposed framework is effective across the alignment pipeline, from extracting coherent chunks to matching them across languages and enabling reliable sentence-level alignment. The results demonstrate that selecting an appropriate chunk granularity is essential for stable alignment performance, as both under- and over-splitting consistently degrade quality. Robust results across shuffled and non-shuffled settings indicate that the proposed method captures structural correspondences beyond surface-level ordering similarity. Sentence alignment further benefits from this structural normalization, with improvements observed for both Bleualign and Vecalign under shuffled conditions, highlighting the effectiveness of the approach in document image settings where ordering noise is common.

Limitations

In this study, we present an initial attempt to incorporate layout-aware chunk alignment into bilingual document alignment from document images, and the

evaluation is therefore limited in scope. Specifically, experiments are conducted on a restricted set of genres and a single language pair (English UNESCO reports and their Japanese translations), and the extent to which the approach generalizes to other document types or low-resource settings remains to be validated. Extending the range of languages and document genres is an important direction for future work.

The main purpose of this study was to test whether an intermediate chunk-level alignment stage improves sentence alignment. We therefore used a modular pipeline and did not optimize the full system for efficiency, measure its runtime or resource cost, or compare it with end-to-end vision-language models. Future work should examine whether the same structural insight can be incorporated into more efficient end-to-end or prompt-based multimodal methods.

Acknowledgements

This work was supported by JST SPRING, Japan Grant Number JPMJSP2180.

References

Baker, M. (1993). Corpus linguistics and translation studies: Implications and applications. In Baker, M., Francis, G., and Tognini-Bonelli, E., editors, *Text and Technology: In Honour of John Sinclair*, pages 233–250. John Benjamins, Amsterdam.

Baker, M. (1996). Corpus-based translation studies: The challenges that lie ahead. In Somers, H., editor, *Terminology, LSP and translation: Studies in language engineering in honour of Juan C. Sager*, pages 175–186. John Benjamins, Amsterdam.

Bañón, M., Chen, P., Haddow, B., Heafield, K., Hoang, H., Esplà-Gomis, M., Forcada, M. L., Kamran, A., Kirefu, F., Koehn, P., Ortiz Rojas, S., Pla Sempere, L., Ramírez-Sánchez, G., Sarrías, E., Strelec, M., Thompson, B., Waites, W., Wiggins, D., and Zaragoza, J. (2020). ParaCrawl: Web-scale acquisition of parallel corpora. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Baorong, H. (2021). Translation for professionals: corpus-based study of translation universals in computing. In Hu, K., Kim, J.-B., Zong, C., and Chersoni, E., editors, *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 80–88, Shanghai, China. Association for Computational Linguistics.

Duarte, A. V., Marques, J. D., Graça, M., Freire, M., Li, L., and Oliveira, A. L. (2024). LumberChunker: Long-form narrative document segmentation. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N., editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6473–6486, Miami, Florida, USA. Association for Computational Linguistics.

Gale, W. A. and Church, K. W. (1993). A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102.

George, N. and Seth, S. (1984). Hierarchical image representation with application to optically scanned documents. In *Proceedings of 7th International Conference on Pattern Recognition (ICPR)*, pages 347–349.

Ha, J., Haralick, R., and Phillips, I. (1995). Recursive x-y cut using bounding boxes of connected components. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 2, pages 952–955 vol.2.

Hao, L., Gao, L., Yi, X., and Tang, Z. (2016). A table detection method for pdf documents based on convolutional neural networks. In *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, pages 287–292.

Koehn, P. (2024). Neural methods for aligning large-scale parallel corpora from the web for south and East Asian languages. In Haddow, B., Kocmi, T., Koehn, P., and Monz, C., editors, *Proceedings of the Ninth Conference on Machine Translation*, pages 1454–1466, Miami, Florida, USA. Association for Computational Linguistics.

Liu, L. and Zhu, M. (2022). Bertalign: Improved word embedding-based sentence alignment for chinese–english parallel corpora of literary texts. *Digital Scholarship in the Humanities*, 38(2):621–634.

Liu, S., Li, Y., and Wei, J. (2025). Xy-cut++: Advanced layout ordering via hierarchical mask mechanism on a novel benchmark.

Moore, R. C. (2002). Fast and accurate sentence alignment of bilingual corpora. In Richardson, S. D., editor, *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 135–144, Tiburon, USA. Springer.

Morishita, M., Suzuki, J., and Nagata, M. (2020). JParaCrawl: A large scale web-based English-Japanese parallel corpus. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3603–3609, Marseille, France. European Language Resources Association.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In Isabelle, P., Charniak, E., and Lin, D., editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Pfitzmann, B., Auer, C., Dolfi, M., Nassar, A. S., and Staar, P. (2022). Doclaynet: A large human-annotated dataset for document-layout segmentation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 3743–3751, New York, NY, USA. Association for Computing Machinery.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Resnik, P. and Smith, N. A. (2003). The web as a parallel corpus. *American Journal of Computational Linguistics*, 29(3):349–380.

Schreiber, S., Agne, S., Wolf, I., Dengel, A., and Ahmed, S. (2017). Deepdesrt: Deep learning for detection and structure recognition of tables in document images. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 01, pages 1162–1167.

Sennrich, R. and Volk, M. (2011). Iterative, mt-based sentence alignment of parallel texts. In *Nordic Conference of Computational Linguistics*.

Soto, C. and Yoo, S. (2019). Visual detection with context for document layout analysis. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3464–3470, Hong Kong, China. Association for Computational Linguistics.

Team, N., Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., Sun, A., Wang, S., Wenzek, G., Youngblood, A., Akula, B., Barrault, L., Gonzalez, G. M., Hansanti, P., Hoffman, J., Jarrett, S., Sadagopan, K. R., Rowe, D., Spruit, S., Tran, C., Andrews, P., Ayan, N. F., Bhosale, S., Edunov, S., Fan, A., Gao, C., Goswami, V., Guzmán, F., Koehn, P., Mourachko, A., Ropers, C., Saleem, S., Schwenk, H., and Wang, J. (2022). No language left behind: Scaling human-centered machine translation.

Thompson, B. and Koehn, P. (2019). Vecalign: Improved sentence alignment in linear time and space. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1342–1348, Hong Kong, China. Association for Computational Linguistics.

Tian, Y., Ye, Q., and Doermann, D. (2025). Yolov12: Attention-centric real-time object detectors. *arXiv preprint arXiv:2502.12524*.

Varga, D., Halácsy, P., Kornai, A., Nagy, V., Németh, L., and Trón, V. (2007). *Parallel corpora for medium density languages*, pages 247–258. John Benjamins Publishing Company.

Wang, Z., Xu, Y., Cui, L., Shang, J., and Wei, F. (2021). LayoutReader: Pre-training of text and layout for reading order detection. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4735–4744, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., and Zhou, M. (2020). Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Confer-*

ence on Knowledge Discovery & Data Mining, KDD '20, page 1192–1200, New York, NY, USA. Association for Computing Machinery.

Yang, X., Yumer, E., Asente, P., Kraley, M., Kifer, D., and Giles, C. L. (2017). Learning to extract semantic structure from documents using multimodal fully convolutional neural networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4342–4351.

Zhao, J., Ji, Z., Feng, Y., Qi, P., Niu, S., Tang, B., Xiong, F., and Li, Z. (2025). Meta-chunking: Learning text segmentation and semantic completion via logical perception.

Zhong, X., Tang, J., and Yepes, A. J. (2019). Publaynet: largest dataset ever for document layout analysis. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1015–1022. IEEE.

A Dataset

Development

- UNESCO. 2018. Journalism, 'Fake News' & Disinformation Handbook for Journalism Education and Training. Paris, UNESCO.
- UNESCO. 2020. Education for Sustainable Development – A roadmap. Paris, UNESCO.

Evaluation

- UNESCO. 2018. Global Education Monitoring Report Summary 2019: Migration, Displacement and Education – Building Bridges, not Walls. Paris, UNESCO.
- UNESCO. 2020. Global Education Monitoring Report Summary 2020: Inclusion and education: All means all. Paris, UNESCO.
- UNESCO. 2021. Global Education Monitoring Report 2021/2: Non-state actors in education: Who chooses? Who loses? Paris, UNESCO.
- UNESCO. 2023. Global Education Monitoring Report Summary 2023: Technology in education: A tool on whose terms? Paris, UNESCO.
- UNESCO. 2024. Global Education Monitoring Report Summary 2024/5 – Leadership in education: Lead for learning. Paris, UNESCO.