Linking Representations with Multimodal Contrastive Learning

Abhishek Arora
Harvard University

Xinmei Yang
Renmin University

Shao Yu Jheng
Harvard University

Melissa Dell
Harvard University
NBER

Abstract

*Many applications require grouping instances contained in diverse document datasets into classes. Most widely used methods do not employ deep learning and do not exploit the inherently multimodal nature of documents. Notably, record linkage is typically conceptualized as a probabilistic or deterministic string matching problem. In record linkage problems, the number of classes is often extremely large, unknown ex ante, or constantly evolving. To address these challenges, this study develops `CLIPPINGS` (**C**ontrastively **LI**nking **P**ooled **P**re-trained Embedd**ings**). `CLIPPINGS` employs end-to-end training of symmetric vision and language bi-encoders, aligned through contrastive language-image pre-training, to learn a metric space where the pooled image-text representation for a given instance is close to representations in the same class and distant from representations in different classes. At inference time, instances can be linked by retrieving their nearest neighbor from an offline exemplar embedding index or by clustering their representations. The study examines two challenging applications: constructing comprehensive supply chains for mid-20th century Japan through linking firm level financial records - with each firm name represented by its crop in the document image and the corresponding OCR - and detecting which image-caption pairs in a massive corpus of historical U.S. newspapers came from the same underlying photo wire source. `CLIPPINGS` outperforms widely used string matching methods by a wide margin and also outperforms unimodal methods. Moreover, a purely self-supervised model trained on only image-OCR pairs - while below the supervised gold-standard - still outperforms popular string matching methods without requiring any labels.*

1. Introduction

Linking information across diverse sources is fundamental to many pipelines. To name a few examples, researchers and businesses frequently link individuals or firms across censuses and company records, governments de-duplicate benefit or voter rolls across locations, and analysts seek to identify how information from the same source spreads through media. In large swathes of the relevant literatures, deep neural methods have made few inroads. For example, a recent comprehensive review of the computer science, social science, and statistical record linkage literatures in *Science Advances* [8] concludes that other methods are preferred over deep neural models for record linkage in structured data. This contrasts to some seemingly similar problems, such as disambiguating entities in unstructured texts to a knowledgebase like Wikipedia, where transformer language models overwhelmingly predominate, *e.g.* [61, 15, 63].

This study examines whether deep learning can significantly enhance record linkage in contexts where it has not traditionally been applied, developing a novel multimodal framework - `CLIPPINGS` (**C**ontrastively **LI**nking **P**ooled **P**re-trained Embedd**ings**) - that can be applied to a wide variety of cross-document linking tasks. `CLIPPINGS` is motivated by the insight that many record linkage applications can be conceptualized as a multimodal classification problem, in which the document image crop of an entity's name and its corresponding text from optical character recognition (OCR) are the inputs, and each unique entity is a class. More generally, document texts and photographs that appear in the document could be used for cross-document linkage. Two example applications are illustrated in Figure 1. The first consists of the localized names of Japanese firms in historical document image scans and their OCR. The second consists of image-caption clippings from historical U.S. newspapers.

Standard powerful methods for image-text classification - used for applications such as hateful memes [32] or product descriptions [39, 31] - use a supervised classification objective that cannot be applied to problems like record linkage that have a very large number (potentially many millions) of classes, have classes

Figure 1. Data: This figure shows representative linked records from the Japanese firm and historical newspaper image-caption datasets.

that are unknown ex ante, or have classes that are constantly evolving as databases are updated. Much as in unimodal applications, such settings can be tackled by learning a metric space where classes can be assigned using kNN classification or clustering.

CLIPPINGS employs end-to-end training of symmetric vision and language bi-encoders to learn a metric space where the pooled image-text representation for an instance is close to representations in the same class and distant from representations in different classes. We first implement self-supervised contrastive language-image pre-training [1], starting with CLIP embeddings [1, 40], to align the image and text encoders. The self-supervised model can be used off-the-shelf, or contrastively tuned using the CLIPPINGS framework on labeled linked data. At inference time, instances can be linked by retrieving their nearest neighbor from an offline exemplar embedding index or by clustering their representations. The contrastive framework can flexibly incorporate blocking on database characteristics - common in record linkage - by using a type-specific loss [36].

CLIPPINGS significantly outperforms string matching methods - which predominate in the record linkage literature - as well as unimodal methods. The primary context examined is a traditional record linkage application, that constructs comprehensive supply chains for mid-20th century Japan through linking firm level financial records. The firm is represented by the image crop of its name in the document scan and the corresponding OCR (Figure 1). Supply chains are fundamental to the transmission of economic shocks [3, 4], agglomeration [17], and economic development [20, 42, 47, 7, 34]. Their role in long-run economic development has been difficult to study quantitatively, due largely to the difficulties of accurately linking large-scale historical records, as sampling in networks can significantly bias statistical analyses [13].

CLIPPINGS is trained on a modest number of linked firms using supervised contrastive learning, following self-supervised language-image pre-training on the image crop-OCR pairs. At inference time, customer-supplier records [25] are linked to their nearest neighbor in an exemplar offline embeddings index, computed using a comprehensive directory that contains rich firm-level data [56]. CLIPPINGS achieves a 94.5% record linkage accuracy, compared to a string matching maximum accuracy of 73.1%.

Particularly in academic research, fine-tuning on a modest amount of labeled data can be more of a feature than a bug, providing the fine-grained control that is typical of scientific research. Yet when we forgo supervised contrastive training - using only self-supervised language-image pre-training on the image-OCR pairs - this also outperforms traditional string matching, with an accuracy of 84.9%. Moreover, supervised contrastive training of the aligned vision encoder on only images outperforms the same training on an encoder with unimodal rather than multimodal pre-training, as the aligned vision encoder acquired some language understanding through multimodal pre-training. This underscores the power of multimodal methods for document analyses.

CLIPPINGS does not model cross-modal attention, because in record linkage applications where the inputs are an image of a text and the corresponding OCR, cross-modal attention is unlikely to lead to significantly richer representations. If desired, though, it would be straightforward to extend the framework to include cross-modal attention.

To test whether CLIPPINGS can apply to not only document images clips and their OCR but also to images and text more generally, we use it to measure which image-caption pairs in a large corpus of historical U.S. off-copyright newspapers came from the same underlying photo news wire source (Figure 1). Local papers obtained many of their images from news wire such as the Associated Press. Captions from the wire could be re-written or abridged. Images, on the other hand, can be very poor resolution (transmitted via telegraph), and there is also noise from cropping, scanning, etc. `CLIPPINGS` again significantly outperforms popular string matching methods.

`CLIPPINGS` can be trained and deployed with modest compute, which is central given its relevance to applications in academic research and government. This also makes it a feasible tool for removing noisy duplicates from massive image-caption training corpora, important for reducing the likelihood of training data regeneration [9].

The `CLIPPINGS` models and training data will be publicly released, in order to encourage further work on deep, multimodal record linkage.

The rest of this study is organized as follows: Section 2 discusses the literature, and Section 3 describes the `CLIPPINGS` architecture. Section 4 applies `CLIPPINGS` to the firm record linkage problem, and Section 5 applies it to detecting noisy image-caption duplicates in historical newspapers. Section 6 concludes.

2. Literature

Contrastive learning:

`CLIPPINGS` learns a metric space for pooled image-text representations that can be applied to linkage problems even when the number of classes is extremely large, unknown ex ante, or constantly changes as databases are updated. It is reminiscent of a variety of unimodal bi-encoder applications, such as semantic similarity [48], passage retrieval [28], and entity disambiguation [61] and co-reference resolution [21] in unstructured text. In order to create pooled image-text representations, it is necessary to have an aligned space. Contrastive Language-Image Pre-training (CLIP) [1] contrastively trained aligned text and image encoders using 400 million image-caption pairs. `CLIPPINGS` begins with pre-trained Japanese [40] and English [1] CLIP image and text encoders.

Record Linkage:

A variety of disciplines - including computer science, statistics, database management, economics, and political science - have made extensive methodological contributions to record linkage, alternatively referred to as entity resolution, fuzzy matching, approximate dictionary matching, and string matching.

Within this sprawling landscape, the literatures on entity resolution in structured databases [8] versus natural language text remain largely divorced. While deep learning has transformed methods for disambiguating entities in unstructured texts to a knowledgebase, *e.g.* [61, 15, 63], deep models have made few inroads in the large literature on record linkage using structured databases. This literature emphasizes linking noisy text fields that contain information such as individual names, firm names, organizations, or locations. Edit distance metrics are commonly used, *e.g.* [37, 23, 60].

Another widespread approach computes the cosine similarity between n-gram representations of strings, where n-grams are defined as all substrings of size n in a string [44]. There have also been some efforts to estimate machine learning models for record linkage. For example, [58] use a random forest classifier trained on labeled data to disambiguate authors of U.S. patents, applying clustering to the resulting dissimilarity scores to enforce transitivity.

Because labeled record linkage datasets are very small compared to the massive corpora used for training transformer models from scratch, a comprehensive 2022 review of the record linkage literature in *Science Advances* [8] concludes that deep neural models are unlikely to be applicable to entity resolution using structured data. Constructing training data for record linkage is indeed highly labor intensive, but much of the knowledge needed to improve record linkage is plausibly already encapsulated in pre-trained image and language encoders such as CLIP [1], or can be gleaned from the further self-supervised language-image pre-training pursued in this study. In their simplest form, approximate string matching methods simply count the required number of edits (insertions, deletions, substitutions) to transform one string into another. In practice, not all substitutions are equally probable, leading to efforts to construct rule-based lists that adjust the costs of substitutions. For example, the fuzzychinese [65] package uses strokes or radicals as the fundamental unit for n-grams substring representations of entities, where these strokes and radicals are drawn from an external database [29] covering a subset of the CJK script. Alternatively, the masala merge package [43] adjusts Levenshtein distance [37] to impose smaller penalties for common alternative spellings in Hindi. Soundex, first developed in 1918 - together with the updated 1970 New York State Identification and Intelligence System (NYSIIS) [51] - account for the fact that similar sounding substitutions are more likely since census enumerators misspelled names according to their sound. These remain a bedrock for record linkage in historical U.S. census data [2].

Such rule-based methods may perform well in the contexts to which they are tailored. However, they can be brittle and are labor-intensive to extend to new settings, due to the use of hand-crafted features. This heavily skews linked datasets towards a few high re-

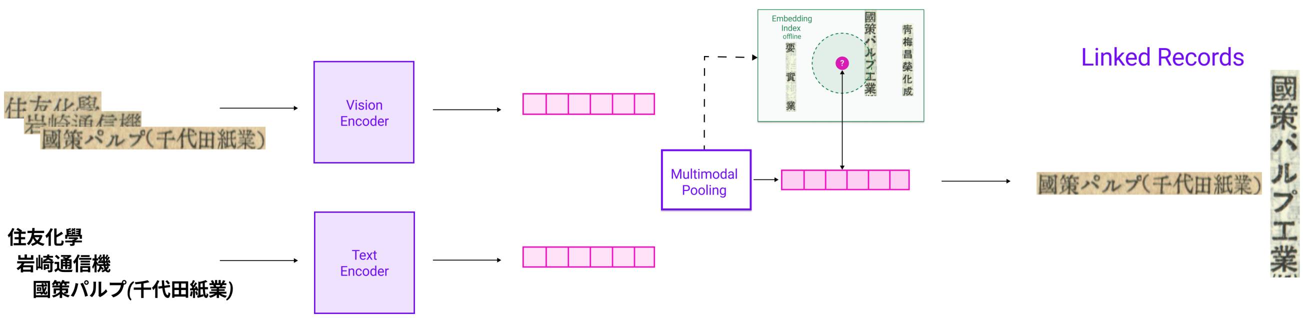

Figure 2. Model Architecture: This figure illustrates the `CLIPPINGS` model architecture.

source settings for which existing methods have been tailored, in turn skewing downstream knowledge. Even in high resource settings, low accuracy in some applications can require extensive human intervention in the matching process to achieve the desired accuracy [6], limiting the scale of problems.

An additional limitation of string matching is that it is unimodal, despite a growing emphasis on processing documents with multimodal models or OCR-free vision-only models [46, 18, 41, 49, 62, 33, 24, 5, 38, 22].

Noisy Duplicates: Identifying noisy duplicates in large corpora may be of inherent interest for downstream questions, for example, studying how the dissemination of iconic images through print media transformed public opinion. Second, duplicates significantly increase the regeneration of training data by large language models and Stable Diffusion [35, 27, 10, 9] raising copyright and plagiarism risks. Third, duplicates can contribute to test set leakage. [52] find that GPT-3 performs substantially better on samples from ReCoRD that have near duplicates in RealNews, a subset of Common Crawl included in the GPT-3 training data, compared to those that did not. While most de-duplication of noisy text focuses on n-grams or locally sensitive hashing, [52] show that neural de-duplication using a contrastively trained S-BERT bi-encoder significantly outperforms these methods and is highly scalable, consistent with `CLIPPINGS` being a potential tool for de-duplicating large image-caption datasets.

3. Model Architecture

Figure 2 shows the `CLIPPINGS` architecture. `CLIPPINGS` employs end-to-end training of symmetric vision and language bi-encoders to learn a metric space where the pooled image-text representation for a given instance is close to representations in the same class and distant from representations in different classes. The baseline uses a supervised contrastive loss [30] on the pooled representations:

$$\mathcal{L} = -\sum_{i\in\mathcal{B}} \frac{1}{|\mathcal{P}(i)|} \sum_{k\in\mathcal{P}(i)} \log \frac{\exp\left(\tau (z_i)^T (z_k)\right)}{\sum_{j\in\mathcal{B}} \exp\left(\tau (z_i)^T (z_j)\right)} \tag{1}$$

where $z_i = \dfrac{f(x_i) + g(t_i)}{2}$ is the mean of the image and text embeddings for instance i. $\mathcal{B}$ denotes the batch and τ is a temperature hyperparameter. This loss incentivizes alignment of image-image, text-text, image-text, and text-image representations across positive instances. It has the flavor of combining contrastive learning on text, contrastive learning on images, and UniCL [64], which has a bi-directional image-text and text-image similarity objective.

We begin with pre-trained, aligned image and text embeddings. Japanese language CLIP [40] was trained on image-caption pairs, using captions machine-translated into Japanese. For the newspapers, we use OpenAI's CLIP [1]. We continue self-supervised pre-training of the CLIP vision and text encoders on firm image crop-OCR pairs and newspaper image-caption pairs. `CLIPPINGS` then uses these encoders to initialize symmetric vision and language bi-encoders trained end-to-end on the pooled representations. It learns a metric space where the pooled image-text representation for an instance is close to representations in the same class and distant from representations in different classes. Hyperparameters, batching, and hard negative mining are detailed in the supplementary materials. All training was done on an A6000 40 GB GPU.

At inference time, instances can be linked by retrieving their nearest neighbor from an offline exemplar embedding index or by clustering their representations. The former is used for the Japanese record linkage and the latter - with highly scalable single linkage clustering - is used for detecting noisy duplicates in the newspaper clippings dataset. Facebook Artificial Intelligence Similarly Search (FAISS) [26], with `IndexFlatIP`, is used to calculate pairwise exact distances between the embeddings.[1]

Blocking - which incorporates type information from structured databases into record linkage - is important in many applications and has been the subject of extensive research (see [55, 45] for reviews). While not applicable to

[1]Because FAISS range search is not GPU-supported, we implement k nearest neighbor search, conservatively setting k to 900.

this study's applications, blocking is natural to incorporate into a contrastive setup, through using a type-specific loss function [36]. This will allow for matches to be made even when there is some noise in the type, a common scenario. For example, there might be noise in a firm's measured establishment date, used to link across censuses.

To examine how `CLIPPINGS` compares to an analogous framework trained using a state-of-the-art unimocal encoder, we train a symmetric DINO vision transformer [12] using the Japanese firm linked image crop data. Details on self-supervised pre-training and hyperparameters for the self-supervised and supervised training are detailed in the supplementary materials.

4. Record Linkage

4.1. Data

This study's first application is constructing historical Japanese supply chains. This entails matching suppliers and customers recorded in firm level records collected in 1956 for over 7,000 large Japanese firms [25] to a firm level directory that provides additional rich information about nearly 70,000 firms [56]. The former are written horizontally and the latter vertically, making a potentially challenging case for visual matching. Firm name crops were localized using a Mask R-CNN [19] model custom-trained with Layout Parser [50]. To create labeled data for training and evaluation, the customers and suppliers of randomly selected firms were hand merged with the firm index. Two highly skilled annotators completed this task and resolved all discrepancies. Many firms appear as customers and suppliers of multiple randomly selected firms, and the data were de-duplicated such that each customer or supplier appears only once, in order to avoid test set leakage. Sometimes a single firm is linked to multiple entries in the firm directory, as firms can appear there more than once if they were registered in multiple prefectures.

We used a 60-20-20 split to divide classes into a training set (772 examples), validation set (257 examples), and test set (257 examples). The test data links the customer-supplier list to all matching firms in the index (with multiple matches occurring when a firm is registered in multiple prefectures), whereas this costly labeling was not needed for the training data, where each firm has a single match. In the main text, we report results using a dataset that drops customers and suppliers like "the government" that do not appear in the firm index. In the supplementary materials, we report analyses including these instances, with similar patterns emerging. We also trained on 19,793 synthetically generated Japanese names, with an 80-20 train-val split. Each are rendered using different fonts and OCRed with two different OCR engines, that make different errors [11, 16]. Each epoch involved sampling 3 "views" of each image crop-ocr pair.

Additionally, we conducted self-supervised language-image pre-training of the Japanese CLIP encoders, using 111,750 firm image crops and their corresponding OCR, as well as the same 19,793 synthetically generated names and their OCR, with an 80-20 test-val split.

	Noisy OCR	Clean OCR
Panel A: String-Matching		
Levenshtein distance	0.630	0.731
Stroke n-gram similarity	0.689	0.731
Panel B: Language-Image Self-Supervised Training		
Visual Linking	0.769	0.769
Language Linking	0.740	0.790
Multimodal Linking	0.845	0.849
Panel C: Supervised Training on Linked Data with Vision Pre-training		
Visual Linking	0.878	0.878
Panel D: Supervised Training on Linked Data with Language-Image Pre-training		
Visual Linking	0.924	0.924
Language Linking	0.790	0.882
Multimodal Linking	0.937	**0.945**

Table 1. Baseline Matching Results: This table reports accuracy on the test set using a variety of different methods for linking Japanese firms from supply chain records to a large firm directory. *Noisy OCR* uses off-the-shelf Google Cloud Vision for OCR, and *Clean OCR* uses an accurate custom-trained OCR.

As record linkage accuracy with string matching is likely to relate to the quality of the OCR, we apply string matching comparisons using two different OCRs of the Japanese firm names: one created by using Google Cloud Vision off-the-shelf, as off-the-shelf OCR usage is the overwhelming norm (Google Cloud Vision does not support fine-tuning) and is typically noisier, as well as a custom-trained OCR that achieves a character error rate of 0.006 [11], a near best case scenario.

4.2. Results

`CLIPPINGS` achieves a record linkage accuracy of 94.5%, significantly outperforming string matching metrics on both noisy and clean OCR. Using the pooled representations also edges out tuning only the pre-trained image or pre-trained language encoders, which achieve an accuracy of 92.4% and 88.2%, respectively. When using multimodal representations, the accuracy of the OCR has only a very modest impact: 93.7% with noisy OCR versus 94.5% with clean OCR for the supervised multimodal `CLIPPINGS`. When supervised tuning is performed using only the linked OCR'ed texts, the performance deterioration from using a noisier OCR is larger (79.0% versus 88.2%), illustrating how working directly with images relieves the OCR information bottleneck.

Figure 3, Panel A shows representative errors made by the supervised vision only, language only, and multimodal encoders, that use language-image pre-training. In the vision-encoder error, the ground truth is Meiji Seika, a firm producing snacks and chocolates. Its prediction is Meiji

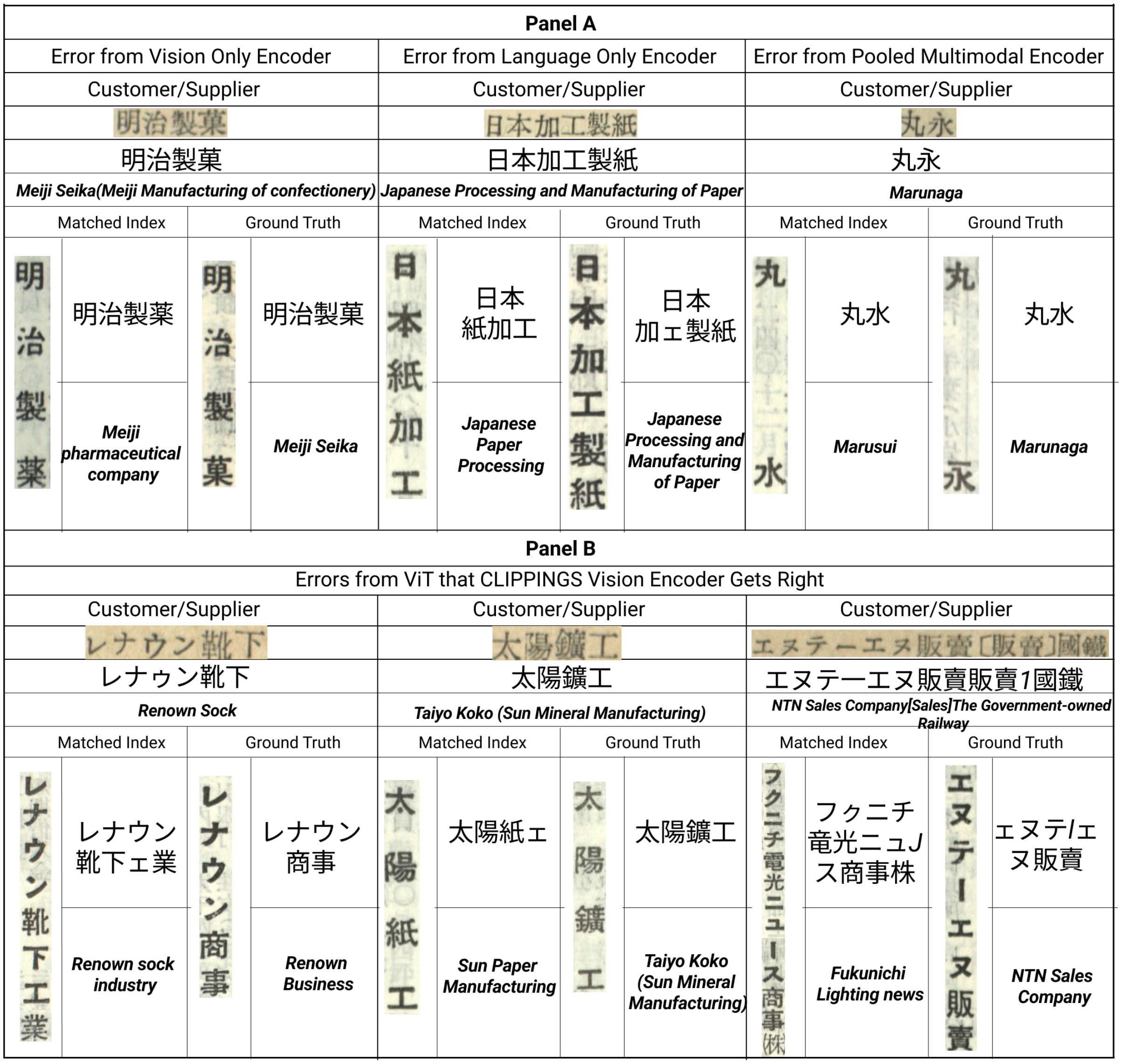

Figure 3. Errors: This figure shows errors in record linkage made by different models.

pharmaceutical company, as 菓 and 薬 look similar.

In the language-only encoder error, the ground truth is "Japanese Processing and Manufacturing of Paper"（日本加工製紙), and the prediction is "Japanese Paper Processing"（日本紙加工), very similar in meaning. The multimodal encoder predicts both cases correctly. Finally, the multimodal error is a company whose name is only two characters. The ground truth is 丸永, and the prediction is 丸水, with 永 and 水 very visually similar. These are names without language context, so the language encoder cannot distinguish them either.

Fascinatingly, the tuned vision only encoder with language-image pre-training gets some matches that require language understanding correct (accuracy 92.4%), that the tuned ViT model with vision-only pre-training gets wrong (accuracy 87.8%).[2] Figure 3, Panel B provides several examples. In the first example, レナウン靴下（Renown Sock) is matched by the ViT model to "レナウン商事(Renown Business), whereas the multimodally pre-trained encoder matches it correctly to レナウン靴下工業（Renown Sock Industry), despite the extra two characters. In the second example, the ground truth is 太陽鑛工 (Sun Mineral Manufacturing), whereas the ViT prediction is 太陽紙工（Sun

[2]Tuning a Japanese S-BERT language bi-encoder does even worse than the vision only encoder, but is less of an apples-to-apples comparisons to the CLIPPINGS text-only model due to the challenges of devising self-supervised language-only pre-training recipes in this context.

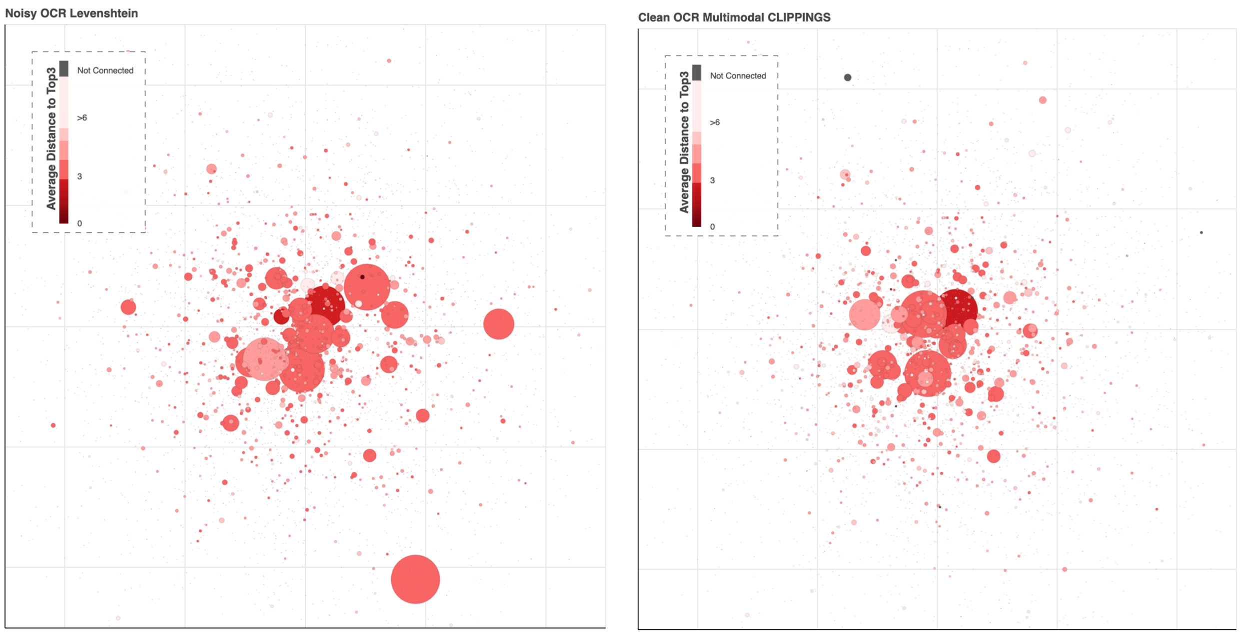

Figure 4. Input-Output Networks: This figure plots the average supply chain distance of Japanese firms to Mitsui, Mitsubishi, and Sumitomo, the three most prominent Japanese firms.

Paper Manufacturing). The third example results from an error in our custom trained layout detection model, that concatenates two customer-suppliers. The detected firm is "エヌティエヌ販賣會社[販賣]国鉄," which translates as "NTN Sales Company[Sales]The Government-owned Railway". The ViT simply predicts a company of similar length, whereas the encoder with multimodal pre-training matches it to NTN Sales Company. These examples show that the vision encoder gains language understanding through the contrastive language-image self-supervised pre-training.

The purely self-supervised multimodal model also outperforms the string matching methods, with an accuracy of 84.9%. n-gram similarity at the stroke level with fuzzychinese [65] achieves an accuracy of 73.1% on clean OCR, as coincidentally does Levenshtein distance. When only vision (76.9% accuracy) or language (79.0% accuracy) self-supervised embeddings are used at test time, the performance also exceeds that of standard string matching techniques. This shows that without any linked training data at all, neural methods can still outperform widely used rule-based methods.

`CLIPPINGS` plausibly outperforms string matching for several reasons. Since it leverages transfer learning from language-image pre-training, it can be precisely tailored to different settings with small, cheap-to-create training sets, or used as a purely self-supervised solution. Moreover, hard negatives in contrastive training encourage separation between different instance representations even if the instances are very similar. Our application does not incorporate blocking, as it is not necessary in this setting, but in the many settings where it is we conjecture that the expressiveness of a type-specific contrastive loss could be another benefit.

Deployment costs are important for extensibility, as record linkage problems often entail linking millions of entities on a limited compute budget. Experiments on the full dataset of 36,673 customer-suppliers, using an 18 core Intel(R) i9-9980XE CPU @ 3.00GHz and a single NVIDIA A6000 GPU, show that `CLIPPING`'s deployment costs are modest, making it a highly scalable solution. Mutlimodal `CLIPPINGS` takes 6 minutes, 21 seconds to run on the full data with a single GPU, all but one second of which are embedding the crops and OCR. This compares to 54 minutes to implement our optimized CPU Levenshtein distance calculations (which is an order of magnitude faster than R string matching, *e.g.* [57]) and 3 minutes and 7 seconds for stroke matching with fuzzychinese [65].

Figure 4 illustrates the supply chain networks created by applying `CLIPPINGS` to the full dataset, with the shading showing the average distance in the supply chain networks to the three largest Japanese conglomerates: Mitsui, Mitsubishi, and Sumitomo. Node position and scaling are fixed, with scaling showing average degree centrality. Using off-the-shelf OCR and Levenshtein distance - a prevalent approach in the literature - creates a visibly different network (left) than the multimodal method (right), which is much closer to the ground truth. 7,352 firm nodes (56%) are in the supply chains of these big-3 firms. A study of the Japanese economy based on the noisier network is likely to produce biased results [13].

5. Noisy Image-Caption Duplicates

5.1. Data

To examine whether the CLIPPINGS framework can be applied more broadly to cross-document vision-language linking, we examine the detection of noisy duplicate image-caption pairs in a massive corpora of off-copyright historical U.S. newspapers that we have digitized using a Mask R-CNN [19] model custom-trained with Layout Parser [50] and Tesseract OCR. Images and captions are associated with their coordinates using a rule-based method, given that captions almost always appear directly below their corresponding images. The classes are unknown ex ante, and so we use single linkage clustering to detect groups of near-duplicate image-caption pairs that came from the same underlying photo news wire source.

Images can be significantly cropped (Figure 1), flipped, or nearly illegible due to the low resolution and noise from scanning. (39 images that were entirely black due to poor scanner settings were removed from the evaluation data.) While the news wire included captions, papers frequently abridged or otherwise modified them to fit their space and appeal to their local audience. Frequently, different images also have the same generic caption (*e.g.*, "Main Street" or "John Smith").

The training dataset is drawn from 5 random days per odd year, between 1941 and 1989. The training set contains 910 classes, with all classes having at least two instances, and the validation set contains 255 classes. To create a test sample, as well as a validation sample for choosing the optimal threshold for single linkage clustering, we labeled a full day sample of every image-caption pair appearing on a randomly selected day, October 3rd, 1968. The test set has 606 classes (60% split), including singletons, and the validation set for choosing the single linkage clustering threshold has 405 classes (40% split), including singletons. Self-supervised pre-training is performed on 85,000 image-caption pairs, drawing from 1941 to 1989. 1968 is excluded to avoid leakage. Hyperparameters are provided in the supplementary materials.

5.2. Results

Multimodal CLIPPINGS achieves an adjusted rand index (ARI) - a widely used measure of clustering performance - of 61.5. This outperforms contrastively tuning just the multimodally pre-trained vision encoder (ARI of 59.7), whereas using just the language encoder performs substantially worse (ARI of 38.9).

For comparison, we also examine n-gram based methods, predominant in the literature measuring the spread of content in historical newspapers [59, 14, 53]. Texts are compared using Jaccard similarity , given by $\frac{|A \cap B|}{|A \cup B|}$. We compute overlaps between each pair of captions, drawing an edge between two captions if the overlap is above some minimum overlap threshold, selected on the validation sample.

To compute Jaccard similarity, we use 2-grams and a 10% overlap threshold. Both the n for n-grams and the similarity threshold were tuned on the same validation sample used to choose the single linkage clustering threshold. As is standard, text is uncased and stripped of punctuation. The text used for all experiments was spell-checked. Jaccard similarity achieves an ARI of 40.3. The purely self-supervised language-image model achieves an ARI of 43.0 (pooled embeddings), beating the ruled based method without requiring any annotation.

Ablations are reported in the supplementary materials.

	ARI
Panel A: String-Matching	
Jaccard similarity	40.3
Panel B: Language-Image Self-Supervised Training	
Visual Linking	40.7
Language Linking	31.0
Multimodal Linking	43.0
Panel C: Supervised Training on Linked Data with Language-Image Pre-training	
Visual Linking	59.7
Language Linking	38.9
Multimodal Linking	**61.5**

Table 2. Detecting Noisy Image-Caption Duplicates: This table reports adjusted rand index on the test set using a variety of different methods for detecting noisy duplicated image-caption pairs in a historical newspaper corpus.

6. Conclusion

This study demonstrates that multimodal methods can significantly improve record linkage, while remaining highly scalable. Compared to methods that predominate in the record linkage literature, deep multimodal approaches can leverage the power of transfer learning, facilitating their application to highly diverse settings with modest annotation requirements. Purely self-supervised multimodal methods can also outperform widely used string matching methods. While our focus here is on image-text linking, it would be straightforward to extend the CLIPPINGS framework to include audio, which is potentially central for datasets where noise enters through misspellings of entity names. Overall deep multimodal record linkage offers potentially substantial gains in a variety of settings where traditional string matching methods fall short of the required accuracy, including a variety of low resource settings, broadening the diversity and improving the quality of data for downstream applications.

Supplementary Materials

A. Methods

A.1. Japanese Multimodal Models

The Japanese multimodal models were initialized with a Japanese CLIP checkpoint [40]. Japanese CLIP was trained with the standard CLIP [1] loss but used a BERT-based text encoder and the vision transformer was initialized by weights from the AugReg ViT-B/16 model [54].

Synthetic Data

Both language-image pretraining and the supervised training of `CLIPPINGS` employed synthetic data. To create synthetic data, we rendered a list of common Japanese words as images using different fonts (a type of augmentation), applied image augmentations, and fed the resulting images to OCR. For the same word, this produced varying views of the word's image due to different augmentations, as well as different views of the word's text due to varying OCR errors induced by the augmentations.

We randomly sample one image-text pair per label (word) and use this subset for language-image pretraining. For training `CLIPPINGS`, we train on the full synthetic dataset and then fine-tune on our labelled data.

Other Training Details

Text crops are thin vertically or horizontally oriented rectangles, with highly diverse aspect ratios, whereas vision encoders are almost always trained on square images. Center cropping would discard important information and resizing often badly morphs the image, given that the underlying aspect ratios are far from a square. To preserve the aspect ratio, we pad the rectangular crop with the median value of the border pixel such that the text region is always centered.

For language-image pretraining, the standard CLIP loss was used to align the image and text encoders [1]. Supervised Contrastive loss [30] was used for all supervised training. We used the AdamW optimizer for all model training along with a Cosine Annealing with Warm Restarts learning rate (LR) scheduler where the maximum LR was specified for each run and the minimum LR was set to 0. 10 steps were chosen for the first restart with a restart factor of 2 that doubles the time to restart after every restart. An epoch involved sampling m views (image-text pair) of each label in the dataset and going through each of them once. It took a 24 hours to perform language-image pretraining, 10 hours to perform supervised training using synthetic data and 40 minutes to train on the labelled data - a total training time of 34.6 hours to train `CLIPPINGS` on a single A6000 GPU card. Hyperparameters and additional details about model training are listed in Table S-1.

At inference time, we used *IndexIPFlat* from FAISS [26] to find the nearest neighbor on L2-normalized embeddings.

Hard Negative Mining

`CLIPPINGS` was trained on a single A6000 GPU card, which could fit a batch size B of 153 image-text pairs. This compares to a batch size of 32,768 used to train the original CLIP [1] on 256 V100 GPUs. Offline hard negative mining was used to achieve sufficient in-batch negatives with the small batch sizes that can fit into compute setups that are realistic for diverse downstream users.

Define the data as a triple (x_n, t_n, y_n), where $x_n \in \mathcal{X}$ is the image, $t_n \in \mathcal{T}$ is the text, and $y_n \in \mathcal{Y}$ is an associated label (class). Let D be the set of all triples. For supervised pertaining using synthetic data, we randomly sampled one image-text pair per label in D to form $D' \subset D$. For each image-text pair (x_a, t_a) in D', we use the domain-adapted CLIP model to find its k nearest neighbor pairs (x_k, t_k) (including itself). This gives us the $k-1$ nearest neighbours for each label y_a in $D' \subset D$.

The anchor label y_a and its $k-1$ neighbors form a "hard-negative" set. In a batch, we sample $m = 3$ views of image-text pairs with replacement. A batch-size divisible by $k*m$ can fit $\frac{B}{k*m}$ unique classes, each with their own m views and the m views of all k neighbors. We shuffle the hard-negative sets and partition them into groups of $\frac{B}{k*m}$ such that each group can constitute a batch. Each constituent class has k neighbors and m views within the minibatch. For the next step - fine-tuning with labeled data - we follow the same approach but with the best model checkpoint from synthetic pretraining.

A.2. English Multimodal Models

Our English `CLIPPINGS` uses the official OpenAI CLIP model [1], which has a ViT-B/32 Transformer architecture for the image encoder and a masked self-attention Transformer as the text encoder. To train `CLIPPINGS` for this application, we used the standard image processor, which resized on the shortest edge, center cropped, and normalized. Language-image pretraining took 13 hours and the supervised training on labeled image-text pairs took 18 hours on a single NVIDIA GeForce RTX 3090. Thus, it took a total of 21 hours to train english multimodal clippings. Hyperparameters and other training details are listed in Table S-1.

To cluster the embeddings, we use Single Linkage Clustering. The distance threshold was tuned on the validation set jointly with the weight put on the image/text embedding to create the pooled embedding.

A.3. Vision Transformer

We initialized the weights of the Vision Transformer from the DINO-pretrained checkpoint for ViT/B16 [12]. Hyperparameters and other training details are listed in Table S-1.

As for `CLIPPINGS`, ViT training employed synthetic data. The same pipeline as above was used to generate synthetically noised images. For each word in a list of Japanese words, the text was rendered using different fonts and augmented to create synthetically noised views.

Offline hard-negative mining was used to train the ViT. The approach is similar to that described above. We used a pretrained checkpoint to find k nearest neighbors for each class. This was used to create a batch containing m views of that class, along with m views of $\frac{B}{m}-1$ other classes. When using hard negatives, we substituted "k-1" of these other classes with the nearest neighbor classes of the anchor.

B. Additional Results

Table S-2 examines the contribution of different elements of the CLIPPINGS training recipe. Panel A considers the performance of Japanese CLIP [40] off-the-shelf. Using the vision encoder alone, every instance is mispredicted. The off-the-shelf text encoder, while better than the vision encoder, is outperformed by traditional string matching methods.

Panel B examines the performance of CLIPPINGS when only supervised training is used, discarding the self-supervised training on image-OCR pairs. Performance declines significantly relative to when self-supervised language-image pre-training is employed - with accuracy similar to that of traditional string matching methods - illustrating the importance of first aligning the vision and text encoders for the document image-OCR domain.

Table S-3 examines the inclusion of instances without a match in the firm directory, for example, a range of government agencies. While performance declines somewhat relative to the results reported in the main text, the relative comparisons between models hold.

We also consider a similar set of ablations for detecting noisy duplicates in newspaper data. Panel A shows that off-the-shelf CLIP [1] (Panel A) does better in this case than in the record linkage case, though performance is still relatively poor and significantly worse than our self-supervised model. The model with supervised training only (Panel B) is again outperformed by the baseline that combines self-supervised language-image training with supervised training.

Model	lr	B	w_decay	temp	im_wt	m	k	epochs	nm_thresh
Panel A: Record Linkage Models									
Language-image Pretraining	5e-5	153	0.001	0.048	-	-	-	40	-
Supervised models									
ViT (synthetic)	5.8e-5	256	0.0398	0.048	-	8	8	5	-
ViT (labelled)	2e-6	252	0.1	0.09	-	3	8	10	0.88
Sup. Lang.-only (synthetic)	5e-6	153	0.001	0.1	0	3	3	30	0.85, 0.85
Sup. Image-only (synthetic)	5e-6	153	0.001	0.1	1	3	3	30	0.76
Sup. Mean-pool (synthetic)	5e-6	153	0.001	0.1	0.5	3	3	30	0.81, 0.80
Sup. Lang-only (labelled)	5e-6	153	0.001	0.1	0	3	3	30	0.84,0.82
Sup. Image-only (labelled)	5e-6	153	0.001	0.1	1	3	3	30	0.79
Sup. Mean-pooling (labelled)	5e-6	153	0.001	0.1	0.5	3	3	30	0.82,0.82
Panel B: Newspaper Image-Caption Models									
Language-image pretraining	5e-6	153	0.001	0.1	-	3	3	9	
Sup. mean pooling (labelled)	5e-9	153	0.001	0.1	0.5	3	3	228	0.227

Table S-1. Training Hyperparameters: *lr* is the maximum learning rate, *B* is the batch size, *w_decay* is the AdamW weight decay, *im_wt* is the weight of the image embedding in the pooled embedding, *m* is the number of views sampled in each epoch, *k* is the number of nearest neighbours in a hard-negative set, and epochs is the number of epochs. For Panel A, *nm_threshold* is a tuple with two tuned similarity thresholds (for noisy and clean OCR respectively) under which a retrieved neighbor is considered to not match with any of the target images. For Panel B, it corresponds to the *distance* threshold used in single linkage clustering. Note that for Panel B, *im_wt* at inference time does not have to be the same as the *im_wt* specified during multimodal training.

	Noisy OCR	Clean OCR
Panel A: Zero-shot Japanese-CLIP		
Visual Linking	0.000	0.000
Language Linking	0.639	0.626
Multimodal Linking	0.491	0.433
Panel B: Only Supervised Training		
Multimodal Linking	0.676	0.731

Table S-2. Record Linkage Ablations: This table reports accuracy for linking Japanese firms from supply chain records to a large firm directory. *Noisy OCR* uses off-the-shelf Google Cloud Vision for OCR, and *Clean OCR* uses an accurate custom-trained OCR. Panel A uses Japanese CLIP off-the-shelf, and Panel B uses only supervised training, without further self-supervised pre-training.

	Noisy OCR	Clean OCR
Panel A: String-Matching		
Levenshtein distance	0.605	0.625
Stroke n-gram similarity	0.625	0.650
Panel B: Language-Image Self-Supervised Training		
Visual Linking	0.693	0.693
Language Linking	0.680	0.741
Multimodal Linking	0.770	0.770
Panel C: Supervised Training on Linked Data with Vision Pre-training		
Visual Linking	0.819	0.819
Panel D: Supervised Training on Linked Data with Language-Image Pre-training		
Visual Linking	0.829	0.829
Language Linking	0.757	0.825
Multimodal Linking	0.845	**0.871**

Table S-3. Including instances without a match: This table reports accuracy on the test set using a variety of different methods for linking Japanese firms from supply chain records to a large firm directory. *Noisy OCR* uses off-the-shelf Google Cloud Vision for OCR, and *Clean OCR* uses an accurate custom-trained OCR.

	ARI
Panel A: Zero-shot CLIP	
Visual Linking	0.182
Language Linking	0.291
Multimodal Linking	0.314
Panel B: Only Supervised Training	
Visual Linking	0.200
Language Linking	0.303
Multimodal Linking	0.559

Table S-4. Noisy Duplicate Ablations: This table reports the adjusted rand index for detecting noisy duplicated image-caption pairs in a historical newspaper corpus, using CLIP off-the-shelf (Panel A) and using only supervised training, with no language-image pre-training (Panel B).

References

[1] Alec , Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 3, 4, 9, 10

[2] Ran Abramitzky, Leah Boustan, Katherine Eriksson, James Feigenbaum, and Santiago Pérez. Automated linking of historical data. *Journal of Economic Literature*, 59(3):865–918, 2021. 3

[3] Daron Acemoglu, Ufuk Akcigit, and William Kerr. Networks and the macroeconomy: An empirical exploration. *Nber macroeconomics annual*, 30(1):273–335, 2016. 2

[4] Daron Acemoglu, Vasco M Carvalho, Asuman Ozdaglar, and Alireza Tahbaz-Salehi. The network origins of aggregate fluctuations. *Econometrica*, 80(5):1977–2016, 2012. 2

[5] Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R Manmatha. Docformer: End-to-end transformer for document understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 993–1003, 2021. 4

[6] Martha J Bailey, Connor Cole, Morgan Henderson, and Catherine Massey. How well do automated linking methods perform? lessons from us historical data. *Journal of Economic Literature*, 58(4):997–1044, 2020. 4

[7] Dominick Bartelme and Yuriy Gorodnichenko. Linkages and economic development. Technical report, National Bureau of Economic Research, 2015. 2

[8] Olivier Binette and Rebecca C Steorts. (almost) all of entity resolution. *Science Advances*, 8(12):eabi8021, 2022. 1, 3

[9] Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. *arXiv preprint arXiv:2301.13188*, 2023. 3, 4

[10] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. *arXiv preprint arXiv:2202.07646*, 2022. 4

[11] Jacob Carlson, Tom Bryan, and Melissa Dell. Efficient ocr for building a diverse digital history. *arXiv e-prints arxiv:2304.02737*, 2023. 5

[12] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294*, 2021. 5, 9

[13] Arun Chandrasekhar and Randall Lewis. Econometrics of sampled networks. *Unpublished manuscript, MIT.[422]*, 2011. 2, 7

[14] Ryan Cordell. Reprinting, circulation, and the network author in antebellum newspapers. *American Literary History*, 27(3):417–445, 2015. 8

[15] Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. Autoregressive entity retrieval. *arXiv preprint arXiv:2010.00904*, 2020. 1, 3

[16] Yongkun Du, Zhineng Chen, Caiyan Jia, Xiaoting Yin, Tianlun Zheng, Chenxia Li, Yuning Du, and Yu-Gang Jiang. Svtr: Scene text recognition with a single visual model. *arXiv preprint arXiv:2205.00159*, 2022. 5

[17] Glenn Ellison, Edward L Glaeser, and William R Kerr. What causes industry agglomeration? evidence from coagglomeration patterns. *American Economic Review*, 100(3):1195–1213, 2010. 2

[18] He Guo, Xiameng Qin, Jiaming Liu, Junyu Han, Jingtuo Liu, and Errui Ding. Eaten: Entity-aware attention for single shot visual text extraction. *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 254–259, 2019. 4

[19] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 5, 8

[20] Albert O Hirschman. *The strategy of economic development.* Yale Univ. Press, New Haven, Conn, 1958. 2

[21] Benjamin Hsu and Graham Horwood. Contrastive representation learning for cross-document coreference resolution of events and entities. *arXiv preprint arXiv:2205.11438*, 2022. 3

[22] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*, 2023. 4

[23] Matthew A Jaro. Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the American Statistical Association*, 84(406):414–420, 1989. 3

[24] Abhinav Java, Shripad Deshmukh, Milan Aggarwal, Surgan Jandial, Mausoom Sarkar, and Balaji Krishnamurthy. One-shot doc snippet detection: Powering search in document beyond text. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5437–5446, 2023. 4

[25] Jinji Kō¯shinjo. *Nihon shokuinrokj.* Jinji Kō¯shinjo, 1954. 2, 5

[26] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547, 2019. 4, 9

[27] Nikhil Kandpal, Eric Wallace, and Colin Raffel. Deduplicating training data mitigates privacy risks in language models. *arXiv preprint arXiv:2202.06539*, 2022. 4

[28] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain

question answering. *arXiv preprint arXiv:2004.04906*, 2020. 3

[29] kfcd. chaizi. https://github.com/kfcd/chaizi, 2015. 3

[30] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*, 2020. 4, 9

[31] Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, and Davide Testuggine. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950*, 2019. 1

[32] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in Neural Information Processing Systems*, 33:2611–2624, 2020. 1

[33] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVIII*, pages 498–517. Springer, 2022. 4

[34] Nathan Lane. Manufacturing revolutions: Industrial policy and industrialization in south korea. *Available at SSRN 3890311*, 2022. 2

[35] Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. *arXiv preprint arXiv:2107.06499*, 2021. 4

[36] Megan Leszczynski, Daniel Y Fu, Mayee F Chen, and Christopher Ré. Tabi: Type-aware bi-encoders for open-domain entity retrieval. *arXiv preprint arXiv:2204.08173*, 2022. 2, 5

[37] Vladimir I Levenshtein et al. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union, 1966. 3

[38] Peizhao Li, Jiuxiang Gu, Jason Kuen, Vlad I Morariu, Handong Zhao, Rajiv Jain, Varun Manjunatha, and Hongfu Liu. Selfdoc: Self-supervised document representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5652–5660, 2021. 4

[39] Huidong Liu, Shaoyuan Xu, Jinmiao Fu, Yang Liu, Ning Xie, Chien-Chih Wang, Bryan Wang, and Yi Sun. Cma-clip: Cross-modality attention clip for image-text classification. *arXiv preprint arXiv:2112.03562*, 2021. 1

[40] Kei Sawada Makoto Sheen, Tenu Cho. nihongo niokeru gengo gazō jizen gakushū moderu no kōchiku to kōkai. In *The 25th Meeting on Image Recognition and Understanding*, 7 2022. 2, 3, 4, 9, 10

[41] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021. 4

[42] Gunnar Myrdal and Paul Sitohang. Economic theory and under-developed regions. *Regional Studies*, 1957. 2

[43] Paul Novosad. Masala merge. https://github.com/paulnov/masala-merge, 2018. 3

[44] Naoaki Okazaki and Jun'ichi Tsujii. Simple and efficient algorithm for approximate dictionary matching. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 851–859, 2010. 3

[45] George Papadakis, Dimitrios Skoutas, Emmanouil Thanos, and Themis Palpanas. A survey of blocking and filtering techniques for entity resolution. *arXiv preprint arXiv:1905.06167*, 2019. 4

[46] Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. Cord: a consolidated receipt dataset for post-ocr parsing. *Workshop on Document Intelligence at NeurIPS 2019*, 2019. 4

[47] Poul Nørregaard Rasmussen. *Studies in inter-sectoral relations*, volume 15. E. Harck, 1956. 2

[48] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019. 3

[49] Phillip Rust, Jonas F Lotz, Emanuele Bugliarello, Elizabeth Salesky, Miryam de Lhoneux, and Desmond Elliott. Language modelling with pixels. *arXiv preprint arXiv:2207.06991*, 2022. 4

[50] Zejiang Shen, Ruochen Zhang, Melissa Dell, Benjamin Charles Germain Lee, Jacob Carlson, and Weining Li. Layoutparser: A unified toolkit for deep learning based document image analysis. *International Conference on Document Analysis and Recognition*, 12821, 2021. 5, 8

[51] Jeffrey M Silbert. The world's first computerized criminal-justice information-sharing system-the new york state identification and intelligence system (nysiis). *Criminology*, 8:107, 1970. 3

[52] Emily Silcock, Luca D'Amico-Wong, Jinglin Yang, and Melissa Dell. Noise-robust de-duplication at scale. *arXiv preprint arXiv:2210.04261*, 2022. 4

[53] David A Smith, Ryan Cordell, and Abby Mullen. Computational methods for uncovering reprinted texts in antebellum newspapers. *American Literary History*, 27(3):E1–E15, 2015. 8

[54] Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *CoRR*, abs/2106.10270, 2021. 9

[55] Rebecca C Steorts, Samuel L Ventura, Mauricio Sadinle, and Stephen E Fienberg. A comparison of blocking methods for record linkage. In *Privacy in Statistical Databases: UNESCO Chair in Data Privacy,*

International Conference, PSD 2014, Ibiza, Spain, September 17-19, 2014. Proceedings, pages 253–268. Springer, 2014. 4

[56] Teikoku Kōshinjo. *Teikoku Ginkō Kaisha Yōroku*. Teikoku Kōshinjo, 1957. 2, 5

[57] M.P.J. van der Loo. The stringdist package for approximate string matching. *The R Journal*, 6:111–122, 2014. 7

[58] Samuel L Ventura, Rebecca Nugent, and Erica RH Fuchs. Seeing the non-stars:(some) sources of bias in past disambiguation approaches and a new public tool leveraging labeled records. *Research Policy*, 44(9):1672–1701, 2015. 3

[59] Aleksi Vesanto, Filip Ginter, Hannu Salmi, Asko Nivala, and Tapio Salakoski. A system for identifying and exploring text repetition in large historical document corpora. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 330–333, 2017. 8

[60] William E Winkler. String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage., 1990. 3

[61] Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. Scalable zero-shot entity linking with dense entity retrieval. *arXiv preprint arXiv:1911.03814*, 2019. 1, 3

[62] Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, et al. Layoutlmv2: Multimodal pre-training for visually-rich document understanding. *arXiv preprint arXiv:2012.14740*, 2020. 4

[63] Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. Luke: Deep contextualized entity representations with entity-aware self-attention. *arXiv preprint arXiv:2010.01057*, 2020. 1, 3

[64] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Ce Liu, Lu Yuan, and Jianfeng Gao. Unified contrastive learning in image-text-label space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19163–19173, 2022. 4

[65] znwang25. fuzzychinese. `https://github.com/znwang25/fuzzychinese`, 2020. 3, 7